Signals and Communication Technology

Frank Nielsen *Editor*

Geometric Theory of Information

Springer

Signals and Communication Technology

For further volumes:
http://www.springer.com/series/4748

Frank Nielsen
Editor

Geometric Theory of Information

Springer

Editor
Frank Nielsen
Sony Computer Science Laboratories Inc
Shinagawa-Ku, Tokyo
Japan

and

Laboratoire d'Informatique (LIX)
Ecole Polytechnique
Palaiseau Cedex
France

ISSN 1860-4862 ISSN 1860-4870 (electronic)
ISBN 978-3-319-05316-5 ISBN 978-3-319-05317-2 (eBook)
DOI 10.1007/978-3-319-05317-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939042

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Audrey Léna and Julien Léo,
To my whole family, heartily

Preface

The first conference of the *Geometric Sciences of Information* (GSI, website at http://www.gsi2013.org/ includes slides and photos) took place downtown Paris (France), in August 2013. The call for papers received an enthusiastic worldwide response, resulting in about a 100 accepted submissions of an average length of eight pages, and organized into the following broad topics:

- Computational Information Geometry,
- Hessian/Symplectic Information Geometry,
- Optimization on Matrix Manifolds,
- Probability on Manifolds,
- Optimal Transport Geometry,
- Divergence Geometry and Ancillarity,
- Machine/Manifold/Topology Learning,
- Tensor-Valued Mathematical Morphology,
- Differential Geometry in Signal Processing,
- Geometry of Audio Processing,
- Geometry for Inverse Problems,
- Shape Spaces: Geometry and Statistic,
- Geometry of Shape Variability,
- Relational Metric,
- Discrete Metric Spaces,
- Etc.

The GSI proceedings has been published as a (thick!) Springer LNCS (volume number 8085, XX, 879 pages, 157 illustrations). Since we could not accommodate long papers in GSI, we decided to solicit renown researchers to contribute to a full length chapter on the latest advances in information geometry and some of its applications (like computational anatomy, image morphology, statistics, and textures, to cite a few examples). Those selected applications emphasize on algorithmic aspects when programming the *Methods of Information Geometry* [1]. The 13 chapters of this book have been organized as follows:

- *Divergence Functions and Geometric Structures They Induce on a Manifold* by Jun Zhang,
- *Geometry on Positive Definite Matrices Deformed by V-Potentials and Its Submanifold Structure* by Atsumi Ohara and Shinto Eguchi,
- *Hessian Structures and Divergence Functions on Deformed Exponential Families* by Hiroshi Matsuzoe and Masayuki Henmi,
- *Harmonic Maps Relative to α-Connections* by Keiko Uohashi,
- *A Riemannian Geometry in the q-Exponential Banach Manifold Induced by q-Divergences* by Héctor R. Quiceno, Gabriel I. Loaiza and Juan C. Arango,
- *Computational Algebraic Methods in Efficient Estimation* by Kei Kobayashi and Henry P. Wynn,
- *Eidetic Reduction of Information Geometry Through Legendre Duality of Koszul Characteristic Function and Entropy: From Massieu–Duhem Potentials to Geometric Souriau Temperature and Balian Quantum Fisher Metric* by Frédéric Barbaresco,
- *Distances on Spaces of High-Dimensional Linear Stochastic Processes*: *A Survey*, by Bijan Afsari and René Vidal,
- *Discrete Ladders for Parallel Transport in Transformation Groups with an Affine Connection Structure* by Marco Lorenzi and Xavier Pennec,
- *A Diffeomorphic Iterative Centroid Method*, by Claire Cury, Joan A. Glaunès and Olivier Colliot,
- *Hartigan's Method for k-MLE: Mixture Modeling with Wishart Distributions and Its Application to Motion Retrieval* by Christophe Saint-Jean and Frank Nielsen,
- *Morphological Processing of Univariate Gaussian Distribution-Valued Images Based on Poincaré Upper-Half Plane Representation* by Jesús Angulo and Santiago Velasco-Forero,
- *Dimensionality Reduction for Classification of Stochastic Texture Images* by C. T. J. Dodson and W. W. Sampson.

There is an exciting time ahead for computational information geometry in studying the fundamental concepts and relationships of Information, Geometry, and Computation!

Acknowledgments

First of all, I would like to thank the chapter contributors for providing us with the latest advances in information geometry, its computational methods, and applications. I express my gratitude to the peer reviewers for their careful feedback that led to this polished, revised work. Each chapter received from two to eight review reports, with an average number of about three to five reviews per chapter.

I thank the following reviewers (in alphabetical order of their first name): Akimichi Takemura, Andrew Wood, Anoop Cherian, Arnaud Dessein, Atsumi Ohara, Bijan Afsari, Frank Critchley, Frank Nielsen, Giovanni Pistone, Hajime

Urakawa, Hirohiko Shima, Hiroshi Matsuzoe, Hitoshi Furuhata, Isabelle Bloch, Keiko Uohashi, Lipeng Ning, Manfred Deistler, Masatoshi Funabashi, Mauro Dalla-Mura, Olivier Alata, Richard Nock, Silvère Bonnabel, Stéphanie Allassonnière, Stefan Sommer, Stephen Marsland, Takashi Kurose, Tryphon T. Georgiou, Yoshihiro Ohnita, and Yu Fujimoto.

I would also like to reiterate my warmest thanks to our scientific, organizing, and financial sponsors: CNRS (GdRs MIA and Maths and Entreprises), Ecole des Mines de Paris, Supelec, Université Paris-Sud, Institut Mathématique de Bordeaux, Société de l'Electricité, de l'Électronique et des technologies de l'information et de la communication (SEE), Société Mathématique de France (SMF), Sony Computer Science Laboratories, and Thales.

Last but not least, I am personally indebted to Dr. Mario Tokoro and Dr. Hiroaki Kitano (Sony Computer Science Laboratories, Inc) for their many encouragements and continuing guidance over the years.

Tokyo, January 2014 Frank Nielsen

Reference

1. Amari, S., Nagaoka, H.: Method of information geometry, AMS Monograph. Oxford University Press, Oxford (2000)

Contents

Chapter 1
Divergence Functions and Geometric Structures They Induce on a Manifold

Jun Zhang

Abstract Divergence functions play a central role in information geometry. Given a manifold $\mathfrak{M}$, a divergence function $\mathcal{D}$ is a smooth, nonnegative function on the product manifold $\mathfrak{M} \times \mathfrak{M}$ that achieves its global minimum of zero (with semi-positive definite Hessian) at those points that form its diagonal submanifold $\Delta_{\mathfrak{M}} \subset \mathfrak{M} \times \mathfrak{M}$. In this chapter, we review how such divergence functions induce (i) a statistical structure (i.e., a Riemannian metric with a pair of conjugate affine connections) on $\mathfrak{M}$; (ii) a symplectic structure on $\mathfrak{M} \times \mathfrak{M}$ if they are "proper"; (iii) a Kähler structure on $\mathfrak{M} \times \mathfrak{M}$ if they further satisfy a certain condition. It is then shown that the class of $\mathcal{D}_{\Phi}$-divergence functions [23], as induced by a strictly convex function Φ on $\mathfrak{M}$, satisfies all these requirements and hence makes $\mathfrak{M} \times \mathfrak{M}$ a Kähler manifold (with Kähler potential given by Φ). This provides a larger context for the α-Hessian structure induced by the $\mathcal{D}_{\Phi}$-divergence on $\mathfrak{M}$, which is shown to be equiaffine admitting α-parallel volume forms and biorthogonal coordinates generated by Φ and its convex conjugate Φ^*. As the α-Hessian structure is dually flat for $\alpha = \pm 1$, the $\mathcal{D}_{\Phi}$-divergence provides richer geometric structures (compared to Bregman divergence) to the manifold $\mathfrak{M}$ on which it is defined.

1.1 Introduction

Divergence functions (also called "contrast functions", "york") are non-symmetric measurements of proximity. They play a central role in statistical inference, machine learning, optimization, and many other fields. The most familiar examples include Kullback-Leibler divergence, Bregman divergence [4], α-divergence [1], f-divergence [6], etc. Divergence functions are also a key construct of information

J. Zhang (✉)
Department of Psychology and Department of Mathematics, University of Michigan,
Ann Arbor, MI 48109, USA
e-mail: junz@umich.edu

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_1,

geometry. Just as L_2-distance is associated with Euclidean geometry, Bregman divergence and Kullback-Leibler divergence are associated with a pair of flat structures (where flatness means free of torsion and free of curvature) that are "dual" to each other; this is called Hessian geometry [18, 19] and it is the dualistic extension of the Euclidean geometry. So just as Riemannian geometry extends Euclidean geometry by allowing non-trivial metric structure, Hessian geometry extends Euclidean geometry by allowing non-trivial affine connections that come in pairs. The pairing of connections are with respect to a Riemannian metric g, which is uniquely specified in the case of Hessian geometry; yet the metric-induced Levi-Civita connection has non-zero curvature in general. The apparent inconvenience is offset by the existence of biorthogonal coordinates in any dually flat (i.e., Hessian) structure and a canonical divergence, along with the tools of convex analysis which is powerful in many practical applications.

In a quite general setting, any divergence function induces a Riemannian metric and a pair of torsion-free connections on the manifold where they are defined [8]. This so-called *statistical structure* is at the core of information geometry. Recently, other geometric structures induced by divergence functions are being investigated, including conformal structure [17], symplectic structure [3, 28], and complex structures [28].

The goal of this chapter is to review the relationship between divergence function and various information geometric structures. In Sect. 1.2, we provide background materials of various geometric structures on a manifold. In Sect. 1.3, we show how these structures can be induced from a divergence function. Starting from a general divergence function which always induces a statistical structure, we define the notion of "properness" for it to be a generating function of a symplectic structure. Imposing a further condition leads to complexification of the product manifold where divergence functions are defined. In Sect. 1.4, we show that a quite broad class of divergence functions, $\mathcal{D}_\Phi$-divergence functions [23] as induced by a strictly convex function, satisfies all these requirements and induces a Kähler structure (Riemannian and complex structures simultaneously) on the tangent bundle. Therefore, just as the full-fledged α-Hessian geometry extends the dually-flat Hessian manifold ($\alpha = \pm 1$), $\mathcal{D}_\Phi$-divergence generalizes Bregman divergence in the "nicest" way possible. Section 1.5 closes with a summary of this approach to information geometric structures through divergence functions.

1.2 Background: Structures on Smooth Manifolds

1.2.1 Differentiable Manifold: Metric and Connection Structures on $T\mathfrak{M}$

A differentiable manifold $\mathfrak{M}$ is a space which locally "looks like" a Euclidean space $\mathbb{R}^n$. By "looks like", we mean that for any base (reference) point $x \in \mathfrak{M}$, there

exists a bijective mapping ("coordinate functions") between the neighborhood of x (i.e., a patch of the manifold) and a subset V of $\mathbb{R}^n$. By locally, we mean that various such mappings must be smoothly related to one another (if they are centered at the same reference point) or consistently glued together (if they are centered at different reference points). Globally, they must cover the entire manifold. Below, we assume that a coordinate system is chosen such that each point is indexed by a vector in V, with the origin as the reference point.

A manifold is specified with certain structures. First, there is an inner-product structure associated with tangent spaces of the manifold. This is given by the metric 2-tensor field g which is, when evaluated at each location x, a symmetric bilinear form $g(\cdot, \cdot)$ of tangent vectors $X, Y \in T_x(\mathfrak{M}) \simeq \mathbb{R}^n$ such that $g(X, X)$ is always positive for all non-zero vector $X \in V$. In local "holonomic" coordinates[1] with bases $\partial_i \equiv \partial/\partial x^i$, $i = 1, \ldots, n$, (i.e., X, Y are expressed as $X = \sum_i X^i \partial_i$, $Y = \sum_i Y^i \partial_i$), the components of g are denoted as

$$g_{ij}(x) = g(\partial_i, \partial_j). \tag{1.1}$$

Metric tensor allows us to define distance on a manifold as shortest curve (called "geodesic") connecting two points, to measure angle and hence define orthogonality of vectors—projections of vectors to a lower dimensional submanifold become possible once a metric is given. Metric tensor also provides a linear isomorphism of tangent space with cotangent space at any point on the manifold.

Second, there is a structure implementing the notion of "parallelism" of vector fields and curviness of a manifold. This is given by the affine (linear) connection ∇, mapping two vector fields X and Y to a third one denoted by $\nabla_Y X$: $(X, Y) \mapsto \nabla_Y X$. Intuitively, it represents the "intrinsic" difference of a tangent vector $X(x)$ at point x and another tangent vector $X(x')$ at a nearby point x', which is connected to x in the direction given by the tangent vector $Y(x)$. Here "intrinsic" means that vector comparison across two neighboring points of the manifold is through a process called "parallel transport," whereby vector components are adjusted as the vector moves across points on the base manifold. Under the local coordinate system with bases $\partial_i \equiv \partial/\partial x^i$, components of ∇ can be written out in its "contravariant" form denoted $\Gamma^l_{ij}(x)$

$$\nabla_{\partial_i} \partial_j = \sum_l \Gamma^l_{ij} \partial_l. \tag{1.2}$$

Under coordinate transform $x \mapsto \tilde{x}$, the new coefficients $\widetilde{\Gamma}$ are related to old ones Γ via

$$\widetilde{\Gamma}^l_{mn}(\tilde{x}) = \sum_k \left(\sum_{i,j} \frac{\partial x^i}{\partial \tilde{x}^m} \frac{\partial x^j}{\partial \tilde{x}^n} \Gamma^k_{ij}(x) + \frac{\partial^2 x^k}{\partial \tilde{x}^m \partial \tilde{x}^n} \right) \frac{\partial \tilde{x}^l}{\partial x^k}; \tag{1.3}$$

[1] A holonomic coordinate system means that the coordinates have been properly "scaled" in unit-length with respect to each other such that the directional derivatives commute: their Lie bracket $[\partial_i, \partial_j] = \partial_i \partial_j - \partial_j \partial_i = 0$, i.e., the mixed partial derivatives are exchangeable in their order of application.

A curve whose tangent vectors are parallel along the curve is said to be "auto-parallel".

As a primitive on a manifold, affine connections can be characterized in terms of their (i) torsion and (ii) curvature. The torsion T of a connection Γ, which is a tensor itself, is given by the asymmetric part of the connection $T(\partial_i, \partial_j) = \nabla_{\partial_i}\partial_j - \nabla_{\partial_j}\partial_i = \sum_k T^k_{ij}\partial_k$, where T^k_{ij} is its local representation given as

$$T^k_{ij}(x) = \Gamma^k_{ij}(x) - \Gamma^k_{ji}(x).$$

The curviness/flatness of a connection Γ is described by the Riemann curvature tensor R, defined as

$$R(\partial_i, \partial_j)\partial_k = (\nabla_{\partial_i}\nabla_{\partial_j} - \nabla_{\partial_j}\nabla_{\partial_i})\partial_k.$$

Writing $R(\partial_i, \partial_j)\partial_k = \sum_l R^l_{kij}\partial_l$ and substituting (1.2), the components of the Riemann curvature tensor are[2]

$$R^l_{kij}(x) = \frac{\partial\Gamma^l_{jk}(x)}{\partial x^i} - \frac{\partial\Gamma^l_{ik}(x)}{\partial x^j} + \sum_m \Gamma^l_{im}(x)\Gamma^m_{jk}(x) - \sum_m \Gamma^l_{jm}(x)\Gamma^m_{ik}(x).$$

By definition, R^l_{kij} is anti-symmetric when $i \longleftrightarrow j$.

A connection is said to be flat when $R^l_{kij}(x) \equiv 0$ and $T^k_{ij} \equiv 0$. Note that this is a tensorial condition, so that the flatness of a connection ∇ is a coordinate-independent property even though the local expression of the connection (in terms of Γ) is coordinate-dependent. For any flat connection, there exists a local coordinate system under which $\Gamma^k_{ij}(x) \equiv 0$ in a neighborhood; this is the affine coordinate for the given flat connection.

In the above discussions, metric and connections are treated as separate structures on a manifold. When both are defined on the same manifold, then it is convenient to express affine connection Γ in its "covariant" form

$$\Gamma_{ij,k} = g(\nabla_{\partial_i}\partial_j, \partial_k) = \sum_l g_{lk}\Gamma^l_{ij}. \tag{1.4}$$

Though Γ^k_{ij} is the more primitive quantity that does not involve metric, $\Gamma_{ij,k}$ represents the projection of Γ onto the manifold spanned by the bases ∂_k. The covariant form of Riemann curvature is (c.f. footnote 2)

$$R_{lkij} = \sum_m g_{lm} R^m_{kij}.$$

[2] This component-wise notation of Riemann curvature tensor followed standard differential geometry textbook, such as [16]. On the other hand, information geometers, such as [2], adopt the notation that $R(\partial_i, \partial_j)\partial_k = \sum_l R^l_{ijk}\partial_l$, with $R_{ijkl} = \sum_l R^m_{ijk} g_{ml}$.

When the connection is torsion free, R_{lkij} is anti-symmetric when $i \longleftrightarrow j$ or when $k \longleftrightarrow l$, and symmetric when $(i, j) \longleftrightarrow (l, k)$. It is related to the Ricci tensor Ric via $\text{Ric}_{kj} = \sum_{i,l} R_{lkij} g^{il}$.

1.2.2 Coupling Between Metric and Connection: Statistical Structure

A fundamental theorem of Riemannian geometry states that given a metric, there is a unique connection (among the class of torsion-free connections) that "preserves" the metric, i.e., the following condition is satisfied

$$\partial_k g(\partial_i, \partial_j) = g(\widehat{\nabla}_{\partial_k} \partial_i, \partial_j) + g(\partial_i, \widehat{\nabla}_{\partial_k} \partial_j). \tag{1.5}$$

Such a connection, denoted as $\widehat{\nabla}$, is called the Levi-Civita connection. Its component forms, called Christoffel symbols, are specified by the components of the metric tensor as ("Christoffel symbols of the second kind")

$$\widehat{\Gamma}_{ij}^{k} = \sum_{l} \frac{g^{kl}}{2} \left(\frac{\partial g_{il}}{\partial x^j} + \frac{\partial g_{jl}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^l} \right).$$

and ("Christoffel symbols of the first kind")

$$\widehat{\Gamma}_{ij,k} = \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial x^j} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^k} \right).$$

The Levi-Civita connection $\widehat{\Gamma}$ is compatible with the metric g, in the sense that it treats tangent vectors of the shortest curves on a manifold as being parallel (equivalently speaking, it treats geodesics as auto-parallel curves).

It turns out that one can define a kind of "compatibility" relation more general than expressed by (1.5), by introducing the notion of "conjugacy" (denoted by $*$) between two connections. A connection ∇^* is said to be *conjugate* (or dual) to ∇ with respect to g if

$$\partial_k g(\partial_i, \partial_j) = g(\nabla_{\partial_k} \partial_i, \partial_j) + g(\partial_i, \nabla^*_{\partial_k} \partial j). \tag{1.6}$$

Clearly, $(\nabla^*)^* = \nabla$. Moreover, $\widehat{\nabla}$, which satisfies (1.5), is special in the sense that it is self-conjugate $(\widehat{\nabla})^* = \widehat{\nabla}$.

Because metric tensor g provides a one-to-one mapping between points in the tangent space (i.e., vectors) and points in the cotangent space (i.e., co-vectors), (1.6) can also be seen as characterizing how co-vector fields are to be parallel-transported in order to preserve their dual pairing $\langle \cdot, \cdot \rangle$ with vector fields.

Writing out (1.6):

$$\frac{\partial g_{ij}}{\partial x^k} = \Gamma_{ki,j} + \Gamma^*_{kj,i}, \tag{1.7}$$

where analogous to (1.2) and (1.4),

$$\nabla^*_{\partial_i} \partial_j = \sum_l \Gamma^{*l}_{ij}\, \partial_l$$

so that

$$\Gamma^*_{kj,i} = g(\nabla^*_{\partial_j} \partial_k, \partial_i) = \sum_l g_{il} \Gamma^{*l}_{kj}.$$

There is an alternative way of imposing "compatibility" condition between a metric g and a connection ∇, through investigating the behavior of how the metric tensor g behaves under ∇. We introduce a 3-tensor field, called "cubic form", as the covariant derivative of g: $C = \nabla g$, or in component forms

$$C(\partial_i, \partial_j, \partial_k) = (\nabla_{\partial_k} g)(\partial_i, \partial_j) = \partial_k g(\partial_i, \partial_j) - g(\nabla_{\partial_k} \partial_i, \partial_j) - g(\partial_i, \nabla_{\partial_k} \partial_j).$$

Writing out the above:

$$C_{ijk} = \frac{\partial g_{ij}}{\partial x_k} - \Gamma_{ki,j} - \Gamma_{kj,i} (= \Gamma^*_{kj,i} - \Gamma_{kj,i}).$$

From its definition, $C_{ijk} = C_{jik}$, that is, symmetric with respective to its first two indices. It can be further shown that:

$$C_{ijk} - C_{ikj} = \sum_l g_{il}\, (T^l_{jk} - T^{*l}_{jk})$$

where T, T^* are torsions of ∇ and ∇^*, respectively. Therefore, $C_{ijk} = C_{ikj}$, and hence C is totally symmetric in all (pairwise permutation of) indices, when $T^l_{jk} = T^{*l}_{jk}$. So conceptually, requiring C_{ijk} to be totally symmetric imposes a compatibility condition between g and ∇, making them the so-called "Codazzi pair" (see [20]). The Codazzi pairing generalizes the Levi-Civita coupling whose corresponding cubic form C_{ijk} is easily seen to be identically zero. Lauritzen [10] defined a "statistical manifold" $(\mathfrak{M}, g, \nabla)$ to be a manifold $\mathfrak{M}$ equipped with g and ∇ such that (i) ∇ is torsion free; (ii) $\nabla g \equiv C$ is totally symmetric. Equivalently, a manifold is said to have statistical structure when the conjugate connection ∇^* (with respect to g) of a torsion-free connection ∇ is also torsion-free. In this case, $\nabla^* g = -C$, and that the Levi-Civita connection $\hat{\nabla} = (\nabla + \nabla^*)/2$.

Two torsion-free connections Γ and Γ' are said to be *projectively equivalent* if there exists a function τ such that:

$$\Gamma'^k_{ij} = \Gamma^k_{ij} + \delta^k_i (\partial_j \tau) + \delta^k_j (\partial_i \tau),$$

where δ_i^k is the Kronecker delta. When two connections are projectively equivalent, their corresponding auto-parallel curves have identical shape (i.e., considered as unparameterized curves); these so-called "pre-geodesics" differ only by a change of parameterization τ.

Two torsion-free connections Γ and Γ' are said to be *dual-projectively equivalent* if there exists a function τ such that:

$$\Gamma'_{ij,k} = \Gamma_{ij,k} - g_{ij}(\partial_k \tau).$$

When two connections are dual-projectively equivalent, then their conjugate connections (with respect to g) have identical pre-geodesics (identical shape).

Recall that when the two Riemannian metric g, g' are *conformally equivalent*, i.e., there exists a function τ such that

$$g'_{ij} = e^{2\tau} g_{ij},$$

then their respective Levi-Civita connections $\widehat{\Gamma'}$ and $\widehat{\Gamma}$ are related via

$$\widehat{\Gamma'}_{ij,k} = \widehat{\Gamma}_{ij,k} - (\partial_k \tau) g_{ij} + (\partial_j \tau) g_{ik} + (\partial_i \tau) g_{jk}.$$

(This relation is obtained by directly substituting in the expressions of the corresponding Levi-Civita connections.) This motivates the definition of the more general notion of *conformally-projectively equivalent* of two statistical structures $(\mathfrak{M}, g, \Gamma)$ and $(\mathfrak{M}, g', \Gamma')$, through the existence of two functions ψ, ϕ such that:

$$g'_{ij} = e^{\psi+\phi} g_{ij} \tag{1.8}$$

$$\Gamma'_{ij,k} = \Gamma_{ij,k} - (\partial_k \psi) g_{ij} + (\partial_j \phi) g_{ik} + (\partial_i \phi) g_{jk}. \tag{1.9}$$

When $\phi = const$ (or $\psi = const$), then the corresponding connections are projectively (dual-projectively, resp) equivalent.

1.2.3 Equiaffine Structure and Parallel Volume Form

For a restrictive set of connections, called "equiaffine" connections, the manifold $\mathfrak{M}$ may admit, in a unique way, a volume form $\Omega(x)$ that is "parallel" under the given connection. Here, a volume form is a skew-symmetric multilinear map from n linearly independent vectors to a non-zero scalar at any point $x \in \mathfrak{M}$, and "parallel" is in the sense that $\nabla \Omega = 0$, or $(\partial_i \Omega)(\partial_1, \ldots, \partial_n) = 0$ where

$$(\partial_i \Omega)(\partial_1, \ldots, \partial_n) \equiv \partial_i(\Omega(\partial_1, \ldots, \partial_n)) - \sum_{l=1}^{n} \Omega(\ldots, \nabla_{\partial_i} \partial_l, \ldots).$$

Applying (1.2), the equiaffine condition becomes

$$\begin{aligned}\partial_i(\Omega(\partial_1,\ldots,\partial_n)) &= \sum_{l=1}^{n} \Omega\left(\ldots,\sum_{k=1}^{n}\Gamma_{il}^{k}\partial_k,\ldots\right)\\ &= \sum_{l=1}^{n}\sum_{k=1}^{n}\Gamma_{il}^{k}\,\delta_k^l\,\Omega(\partial_1,\ldots,\partial_n) = \Omega(\partial_1,\ldots,\partial_n)\sum_{l=1}^{n}\Gamma_{il}^{l}\end{aligned}$$

or

$$\sum_{l}\Gamma_{il}^{l}(x) = \frac{\partial \log \Omega(x)}{\partial x^i}. \tag{1.10}$$

Whether or not a connection is equiaffine is related to the so-called Ricci tensor Ric, defined as the contraction of the Riemann curvature tensor R

$$\mathrm{Ric}_{ij}(x) = \sum_{k} R_{ikj}^{k}(x). \tag{1.11}$$

For a torsion-free connection $\Gamma_{ij}^{k} = \Gamma_{ji}^{k}$, we can verify that

$$\begin{aligned}\mathrm{Ric}_{ij} - \mathrm{Ric}_{ji} &= \frac{\partial}{\partial x^i}\left(\sum_{l}\Gamma_{jl}^{l}(x)\right) - \frac{\partial}{\partial x^j}\left(\sum_{l}\Gamma_{il}^{l}(x)\right)\\ &= \sum_{k} R_{kij}^{k}.\end{aligned} \tag{1.12}$$

One immediately sees that the existence of a function Ω satisfying (1.10) is equivalent to the right side of (1.12) to be identically zero.

Making use of (1.10), it is easy to show that the parallel volume form of a Levi-Civita connection $\widehat{\Gamma}$ is given by

$$\widehat{\Omega}(x) = \sqrt{\det[g_{ij}(x)]}.$$

Making use of (1.7), the parallel volume forms Ω, Ω^* associated with Γ and Γ^* satisfy (apart from a multiplicative constant which must be positive)

$$\Omega(x)\,\Omega^*(x) = (\widehat{\Omega}(x))^2 = \det[g_{ij}(x)]. \tag{1.13}$$

The equiaffine condition can also be expressed using a quantity related to the cubic form C_{ijk}. We may introduce the *Tchebychev form* (also known as the *first Koszul form*), expressed in the local coordinates,

$$T_i = \sum_{j,k} C_{ijk}g^{jk}. \tag{1.14}$$

A tedious calculation shows that

$$\frac{\partial T_i}{\partial x^j} - \frac{\partial T_j}{\partial x^i} = \frac{\partial}{\partial x^j}\left(\sum_l \Gamma^l_{li}\right) - \frac{\partial}{\partial x^i}\left(\sum_l \Gamma^l_{lj}\right),$$

the righthand side of (1.12). Therefore, an equivalent requirement for equiaffine structure is that Tchebychev 1-form T is "closed":

$$\frac{\partial T_i}{\partial x^j} = \frac{\partial T_j}{\partial x^i}. \tag{1.15}$$

This expresses the integrability condition. When Eq. (1.15) is satisfied, there exits a function ϕ such that $T_i = \partial_i \tau$. Furthermore, it can be shown that

$$\tau = -2\log(\Omega/\widehat{\Omega}).$$

Proposition 1 ([13, 25]) *The necessary and sufficient condition for a torsion-free connection ∇ to be equiaffine is for any of the following to hold:*

1. *There exists a ∇-parallel volume element Ω : $\nabla\Omega = 0$.*
2. *Ricci tensor of ∇ is symmetric: $Ric_{ij} = Ric_{ji}$.*
3. *Curvature tensor $\sum_k R^k_{kij} = 0$.*
4. *The Tchebychev 1-form T is closed, $dT = 0$.*
5. *There exists a function τ, called Tchebychev potential, such that $T_i = \partial_i \tau$.*

It is known that the Ricci tensor of the Levi-Civita connection is always symmetric—this is why Riemannian volume form $\widehat{\Omega}$ always exists.

1.2.4 α-Structure and α-Hessian Structure

On a statistical manifold, one can define a one-parameter family of affine connections $\Gamma^{(\alpha)}$, called "α-connections" ($\alpha \in \mathbb{R}$):

$$\Gamma^{(\alpha)k}_{ij} = \frac{1+\alpha}{2}\Gamma^k_{ij} + \frac{1-\alpha}{2}\Gamma^{*k}_{ij}. \tag{1.16}$$

Obviously, $\Gamma^{(0)} = \widehat{\Gamma}$ is the Levi-Civita connection. Using cubic form, this amounts to $\nabla^{(\alpha)} g = \alpha C$. The α-parallel volume element is given by:

$$\Omega^{(\alpha)} = e^{-\frac{\alpha}{2}\tau}\widehat{\Omega}$$

where τ is the Tchebychev potential. The Riemannian volume element $\widehat{\Omega}$ is only parallel with respect to the Levi-Civita connection $\hat{\nabla}$ of g, that is, $\hat{\nabla}\widehat{\Omega} = 0$, but not other α-connections ($\alpha \neq 0$). Rather, $\nabla^{(\alpha)}\Omega^{(\alpha)} = 0$.

It can be further shown that the curvatures R_{lkij}, R^*_{lkij} for the pair of conjugate connections Γ, Γ^* satisfy

$$R_{lkij} = R^*_{lkij}.$$

So, Γ is flat if and only if Γ^* is flat. In this case, the manifold is said to be "dually flat". When Γ, Γ^* are dually flat, then $\Gamma^{(\alpha)}$ is called "α-transitively flat" [21]. In such case, $\{\mathfrak{M}, g, \Gamma^{(\alpha)}, \Gamma^{(-\alpha)}\}$ is called an α-Hessian structure [26]. They are all compatible with a metric g that is induced from a strictly convex (potential) function, see next subsection.

For an α-Hessian manifold, the Tchebychev form (1.14) is given by

$$T_i = \frac{\partial \log(\det[g_{kl}])}{\partial x^i}$$

and its derivative (known as *the second Koszul form*) is

$$\beta_{ij} = \frac{\partial T_i}{\partial x^j} = \frac{\partial^2 \log(\det[g_{kl}])}{\partial x^i \partial x^j}.$$

1.2.5 Biorthogonal Coordinates

A key feature for α-Hessian manifolds is biorthogonal coordinates, as we shall discuss now. They are the "best" coordinates one can have when the Riemannian metric is non-trivial.

Consider coordinate transform $x \mapsto u$,

$$\partial^i \equiv \frac{\partial}{\partial u_i} = \sum_l \frac{\partial x^l}{\partial u_i} \frac{\partial}{\partial x^l} = \sum_l F^{li} \partial_l$$

where the Jacobian matrix F is given by

$$F_{ij}(x) = \frac{\partial u_i}{\partial x^j}, \quad F^{ij}(u) = \frac{\partial x^i}{\partial u_j}, \quad \sum_l F_{il} F^{lj} = \delta^l_i \tag{1.17}$$

where δ^j_i is Kronecker delta (taking the value of 1 when $i = j$ and 0 otherwise). If the new coordinate system $u = [u_1, \ldots, u_n]$ (with components expressed by subscripts) is such that

$$F_{ij}(x) = g_{ij}(x), \tag{1.18}$$

then the x-coordinate system and the u-coordinate system are said to be "biorthogonal" to each other since, from the definition of metric tensor (1.1),

$$g(\partial_i, \partial^j) = g(\partial_i, \sum_l F^{lj}\partial_l) = \sum_l F^{lj} g(\partial_i, \partial_l) = \sum_l F^{lj} g_{il} = \delta_i^j.$$

In such case, denote

$$g^{ij}(u) = g(\partial^i, \partial^j), \tag{1.19}$$

which equals F^{ij}, the Jacobian of the inverse coordinate transform $u \mapsto x$. Also introduce the (contravariant version) of the affine connection Γ under u-coordinate and denote it by an unconventional notation Γ_t^{rs} defined by

$$\nabla_{\partial^r}\partial^s = \sum_t \Gamma_t^{rs}\partial^t;$$

similarly Γ_t^{*rs} is defined via

$$\nabla^*_{\partial^r}\partial^s = \sum_t \Gamma_t^{*rs}\partial^t.$$

The covariant version of the affine connections will be denoted by superscripted Γ and Γ^*

$$\Gamma^{ij,k}(u) = g(\nabla_{\partial^i}\partial^j, \partial^k), \quad \Gamma^{*ij,k}(u) = g(\nabla^*_{\partial^i}\partial^j, \partial^k). \tag{1.20}$$

The affine connections in u-coordinates (expressed in superscript) and in x-coordinates (expressed in subscript) are related via

$$\Gamma_t^{rs}(u) = \sum_k \left(\sum_{i,j} \frac{\partial x^r}{\partial u_i}\frac{\partial x^s}{\partial u_j}\Gamma_{ij}^k(x) + \frac{\partial^2 x^k}{\partial u_r \partial u_s} \right) \frac{\partial u_k}{\partial x^t} \tag{1.21}$$

and

$$\Gamma^{rs,t}(u) = \sum_{i,j,k} \frac{\partial x^r}{\partial u_i}\frac{\partial x^s}{\partial u_j}\frac{\partial x^t}{\partial u_k}\Gamma_{ij,k}(x) + \frac{\partial^2 x^t}{\partial u_r \partial u_s}. \tag{1.22}$$

Similarly relations hold between $\Gamma_t^{*rs}(u)$ and $\Gamma_{ij}^{*k}(x)$, and between $\Gamma^{*rs,t}(u)$ and $\Gamma^*_{ij,k}(x)$.

In analogous to (1.7), we have the following identity

$$\frac{\partial^2 x^t}{\partial u_s \partial u_r} = \frac{\partial g^{rt}(u)}{\partial u_s} = \Gamma^{rs,t}(u) + \Gamma^{*ts,r}(u).$$

Therefore, we have

Proposition 2 *Under biorthogonal coordinates, a pair of conjugate connections Γ, Γ^* satisfy*

$$\Gamma^{*ts,r}(u) = -\sum_{i,j,k} g^{ir}(u) g^{js}(u) g^{kt}(u) \Gamma_{ij,k}(x) \tag{1.23}$$

and

$$\Gamma_r^{*\,ts}(u) = -\sum_{j} g^{js}(u) \Gamma_{jr}^{t}(x). \tag{1.24}$$

Let us now express parallel volume forms $\Omega(x)$, $\Omega(u)$ under biorthogonal coordinates x or u. Contracting the indices t with r in (1.24), and invoking (1.10), we obtain

$$\frac{\partial \log \Omega^*(u)}{\partial u_s} + \sum_j \frac{\partial x^j}{\partial u_s} \frac{\partial \log \Omega(x)}{\partial x^j} = \frac{\partial \log \Omega^*(u)}{\partial u_s} + \frac{\partial \log \Omega(x)}{\partial u_s} = 0.$$

After integration,

$$\Omega^*(u)\, \Omega(x) = const. \tag{1.25}$$

From (1.13) and (1.25),

$$\Omega(u)\, \Omega^*(x) = const. \tag{1.26}$$

The relations (1.25) and (1.26) indicate that the volume forms of the pair of conjugate connections, when expressed in biorthogonal coordinates respectively, are inversely proportional to each other.

The $\Gamma^{(\alpha)}$-parallel volume element $\Omega^{(\alpha)}$ can be shown to be given by (in either x and u coordinates)

$$\Omega^{(\alpha)} = \Omega^{\frac{1+\alpha}{2}} (\Omega^*)^{\frac{1-\alpha}{2}}.$$

Clearly,

$$\Omega^{(\alpha)}(x)\Omega^{(-\alpha)}(x) = \det[g_{ij}(x)] \longleftrightarrow \Omega^{(\alpha)}(u)\Omega^{(-\alpha)}(u) = \det[g^{ij}(u)].$$

1.2.6 Existence of Biorthogonal Coordinates

From its definition (1.18), we can easily show that

Proposition 3 *A Riemannian manifold with metric g_{ij} admits biorthogonal coordinates if and only if $\frac{\partial g_{ij}}{\partial x^k}$ is totally symmetric*

$$\frac{\partial g_{ij}(x)}{\partial x^k} = \frac{\partial g_{ik}(x)}{\partial x^j}. \tag{1.27}$$

That (1.27) is satisfied for biorthogonal coordinates is evident by virtue of (1.17) and (1.18). Conversely, given (1.27), there must be n functions $u_i(x)$, $i = 1, 2, \ldots, n$

such that

$$\frac{\partial u_i(x)}{\partial x^j} = g_{ij}(x) = g_{ji}(x) = \frac{\partial u_j(x)}{\partial x^i}.$$

The above identity implies that there exist a function Φ such that $u_i = \partial_i \Phi$ and, by positive definiteness of g_{ij}, Φ would have to be a strictly convex function! In this case, the x- and u-variables satisfy (1.37), and the pair of convex functions, Φ and its conjugate $\widetilde{\Phi}$, are related to g_{ij} and g^{ij} by

$$g_{ij}(x) = \frac{\partial^2 \Phi(x)}{\partial x^i \partial x^j} \longleftrightarrow g^{ij}(u) = \frac{\partial^2 \widetilde{\Phi}(u)}{\partial u_i \partial u_j}.$$

It follows from the above Lemma that a necessary and sufficient condition for a Riemannian manifold to admit biorthogonal coordinates it that its Levi-Civita connection is given by

$$\widehat{\Gamma}_{ij,k}(x) \equiv \frac{1}{2}\left(\frac{\partial g_{ik}}{\partial x^j} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^k}\right) = \frac{1}{2}\frac{\partial g_{ij}}{\partial x^k}.$$

From this, the following can be shown:

Proposition 4 *A Riemannian manifold $\{\mathfrak{M}, g\}$ admits a pair of biorthogonal coordinates x and u if and only if there exists a pair of conjugate connections γ and γ^* such that $\gamma_{ij,k}(x) = 0$, $\gamma^{*rs,t}(u) = 0$.*

In other words, biorthogonal coordinates are affine coordinates for the dually-flat pair of connections. In fact, we can now define a pair of torsion-free connections by

$$\gamma_{ij,k}(x) = 0, \quad \gamma^*_{ij,k}(x) = \frac{\partial g_{ij}}{\partial x^k}$$

and show that they are conjugate with respect to g, that is, they satisfy (1.6). This is to say that we select an affine connection γ such that x is its affine coordinate. From (1.22), when γ^* is expressed in u-coordinates,

$$\begin{aligned}
\gamma^{*rs,t}(u) &= \sum_{i,j,k} g^{ir}(u) g^{js}(u) \frac{\partial x^k}{\partial u_t} \frac{\partial g_{ij}(x)}{\partial x^k} + \frac{\partial g^{ts}(u)}{\partial u_r} \\
&= \sum_{i,j} g^{ir}(u) \left(-\frac{\partial g^{js}(u)}{\partial u_t} g_{ij}(x)\right) + \frac{\partial g^{ts}(u)}{\partial u_r} \\
&= -\sum_j \delta^r_j \frac{\partial g^{js}(u)}{\partial u_t} + \frac{\partial g^{ts}(u)}{\partial u_r} = 0.
\end{aligned}$$

This implies that u is an affine coordinate system with respect to γ^*. Therefore, biorthogonal coordinates are affine coordinates for a pair of dually-flat connections.

1.2.7 Symplectic, Complex, and Kähler Structures

Symplectic structure on a manifold refers to the existence of a closed, non-degenerate 2-tensor, i.e., a skew-symmetric bilinear map ω: $W \times W \to \mathbb{R}$, with $\omega(X, Y) = -\omega(Y, X)$ for all $X, Y \in W \subseteq \mathbb{R}^{2n}$. For ω to be well-defined, the vector space W is required to be orientable and even-dimensional. In this case, there exists a base $\{e_1, \ldots, e_n, f_1, \ldots, f_n\}$ of W, $\dim(W) = 2n$ such that $\omega(e_i, e_j) = 0, \omega(f_i, f_j) = 0, \omega(e_i, f_j) = \delta_{ij}$ for all indices i, j taking values in $1, \ldots, n$.

Symplectic structure is closely related to inner-product structure (the existence of a positive-definite symmetric bilinear map G: $W \times W \to \mathbb{R}$) and complex structure (linear mapping J: $W \to W$ such that $J^2 = -Id$) on an even-dimensional vector space W. The complex structure J on W is said to be compatible with a symplectic structure ω if $\omega(JX, JY) = \omega(X, Y)$ (symplectomorphism condition) and $\omega(X, JY) > 0$ (taming condition) for any $X, Y \in W$. With ω, J given, $G(X, Y) \equiv \omega(X, JY)$ can be shown to be symmetric and positive-definite, and hence an inner-product on W.

The cotangent bundle $\mathcal{T}^*\mathfrak{M}$ of any manifold $\mathfrak{M}$ admits a canonical symplectic form written as

$$\omega = \sum_{i=1}^{n} dx^i \wedge dp_i,$$

where $(x^1, \ldots, x^n, p_1, \ldots, p_n)$ are coordinates of $\mathcal{T}^*\mathfrak{M}$. That ω is closed can be shown by the existence of the tautological (or Liouville) 1-form

$$\alpha = \sum_{i=1}^{n} p_i dx^i$$

(which can be checked to be coordinate-independent on $\mathcal{T}^*\mathfrak{M}$) and then verifying $\omega = -d\alpha$. Hence, ω is also coordinate-independent. Denote $\partial_i = \partial/\partial x^i$, $\widetilde{\partial}_i = \partial/\partial p_i$ as the base of the tangent bundle $\mathcal{T}\mathfrak{M}$, then

$$\omega(\partial_i, \partial_j) = \omega(\widetilde{\partial}_i, \widetilde{\partial}_j) = 0; \;\; \omega(\partial_i, \widetilde{\partial}_j) = -\omega(\widetilde{\partial}_j, \partial_i) = \omega_{ij}. \tag{1.28}$$

That is, when viewed as 2×2 blocks of $n \times n$ matrix, ω vanishes on diagonal blocks and has non-zero entries ω_{ij} and $-\omega_{ij}$ only on off-diagonal blocks.

The aforementioned linear map J of the tangent space $T_x\mathfrak{M} \simeq W$ at any point $x \in \mathfrak{M}$

$$J : J\partial_i = \widetilde{\partial}_j, J\widetilde{\partial}_j = -\partial_i,$$

gives rise to an "almost complex structure" on $T_x\mathfrak{M}$. For $\mathcal{T}\mathfrak{M}$ to be complex, that is, admitting complex coordinates, an integrable condition needs to be imposed for the J-maps at various base points x of $\mathfrak{M}$, and hence at various tangent spaces $T_x\mathfrak{M}$, to be "compatible" with one another. The condition is that the so-called Nijenhuis tensor N

$$N(X, Y) = [JX, JY] - J[X, JY] - J[JX, Y] - [X, Y]$$

must vanish for arbitrary tangent vector fields X, Y.

The Riemannian metric tensor G on $\mathcal{T}\mathfrak{M}$ compatible with ω has the form

$$\begin{aligned} G_{ij'} &\equiv G(\partial_i, \tilde{\partial}_j) = 0; \\ G_{i'j} &\equiv G(\tilde{\partial}_i, \partial_j) = 0; \\ G_{ij} &\equiv G(\partial_i, \partial_j) = g_{ij} \\ G_{i'j'} &\equiv G(\tilde{\partial}_i, \tilde{\partial}_j) = g_{ij}. \end{aligned}$$

where $i' = n + i$, $j' = n + j$ and i, j takes values in $1, \ldots, n$. When viewed as 2×2 blocks of $n \times n$ matrix, G vanishes on the off-diagonal blocks and has non-zero entries g_{ij} only on the two diagonal blocks. Such a metric on $\mathcal{T}\mathfrak{M}$ is in the form of Sasaki metric, which can also result from an appropriate "lift" of the Riemannian metric on $\mathfrak{M}$ into $\mathcal{T}\mathfrak{M}$, via an affine connection on $\mathcal{T}\mathfrak{M}$ which induces a splitting of $\mathcal{T}\mathcal{T}\mathfrak{M}$, the tangent bundle of $\mathcal{T}\mathfrak{M}$ as the base manifold. We omit the technical details here, but refer interested readers to Yano and Ishihara [22] and, in the context of statistical manifold, to Matsuzoe and Inoguchi [12].

It is a basic conclusion in symplectic geometry that for any symplectic form, there exists a compatible almost complex structure J. Along with the Riemannian metric, the three structures (G, ω, J) are said to form a *compatible triple* if any two gives rise to the third one. When a manifold has a compatible triple (G, ω, J) in which J is integrable, it is called a Kähler manifold. On a Kähler manifold, using complex coordinates, the metric $\tilde{G}$ associated with the complex line-element

$$ds^2 = \tilde{G}_{ij} dz^i d\bar{z}^j,$$

is given by

$$\tilde{G}_{ij}(z, \bar{z}) = \frac{\partial^2 \Phi}{\partial z^i \partial \bar{z}^j}.$$

Here the real-valued function Φ (of complex variables) is called the "Kähler potential".

It is known that the tangent bundle $\mathcal{T}\mathfrak{M}$ of a manifold $\mathfrak{M}$ with a flat connection on it admits a complex structure [7]. As [18] pointed out, Hessian manifold can be seen as the "real" Kähler manifold.

Proposition 5 ([7]) *$(\mathfrak{M}, g, \nabla)$ is a Hessian manifold if and only if $(T\mathfrak{M}, J, G)$ is a Kähler manifold, where G is the Sasaki lift of g.*

1.3 Divergence Functions and Induced Structures

1.3.1 Statistical Structure Induced on $T\mathfrak{M}$

A divergence function $\mathcal{D}\colon \mathfrak{M} \times \mathfrak{M} \to \mathbb{R}_{\geq 0}$ on a manifold $\mathfrak{M}$ under a local chart $V \subseteq \mathbb{R}^n$ is defined as a smooth function (differentiable up to third order) which satisfies

(i) $\mathcal{D}(x, y) \geq 0\ \forall x, y \in V$ with equality holding if and only if $x = y$;
(ii) $\mathcal{D}_i(x, x) = \mathcal{D}_{,j}(x, x) = 0, \forall i, j \in \{1, 2, \ldots, n\}$;
(iii) $-\mathcal{D}_{i,j}(x, x)$ is positive definite.

Here $\mathcal{D}_i(x, y) = \partial_{x^i}\mathcal{D}(x, y)$, $\mathcal{D}_{,i}(x, y) = \partial_{y^i}\mathcal{D}(x, y)$ denote partial derivatives with respect to the i-th component of point x and of point y, respectively, $\mathcal{D}_{i,j}(x, y) = \partial_{x^i}\partial_{y^j}\mathcal{D}(x, y)$ the second-order mixed derivative, etc.

On a manifold, divergence functions act as "pseudo-distance" functions that are non-negative but need not be symmetric. That dualistic Riemannian manifold structure (i.e., statistical structure) can be induced from a divergence function was first demonstrated by S. Eguchi.

Proposition 6 ([8, 9]) *A divergence function $\mathcal{D}$ induces a Riemannian metric g and a pair of torsion-free conjugate connections Γ, Γ^* given as*

$$
\begin{aligned}
g_{ij}(x) &= -\,\mathcal{D}_{i,j}(x, y)\big|_{x=y}\,; \\
\Gamma_{ij,k}(x) &= -\,\mathcal{D}_{ij,k}(x, y)\big|_{x=y}\,; \\
\Gamma^*_{ij,k}(x) &= -\,\mathcal{D}_{k,ij}(x, y)\big|_{x=y}\,.
\end{aligned}
$$

It is easily verifiable that $\Gamma_{ij,k}$, $\Gamma^*_{ij,k}$ as given above are torsion-free[3] and satisfy the conjugacy condition with respect to the induced metric g_{ij}. Hence $\{\mathfrak{M}, g, \Gamma, \Gamma^*\}$ as induced is a "statistical manifold' [10].

A natural question is whether/how the statistical structures induced from different divergence functions are related. The following is known:

Proposition 7 ([14]) *Let $\mathcal{D}$ be a divergence function and ψ, ϕ be two arbitrary functions. If $\mathcal{D}'(x, y) = e^{\psi(x)+\phi(y)}\mathcal{D}(x, y)$, then $\mathcal{D}'(x, y)$ is also a divergence function, and the induced $(\mathfrak{M}, g', \Gamma')$ and $(\mathfrak{M}, g, \Gamma)$ induced from $\mathcal{D}(x, y)$ are conformally-projectively equivalent. In particular, when $\phi(x) = const$, then Γ' and Γ are projectively equivalent; when $\psi(y) = const$, then Γ' and Γ are dual-projectively equivalent.*

[3] Conjugate connections which admit torsion has been recently studied by Calin et al. [5] and Matsuzoe [15].

1.3.2 Symplectic Structure Induced on $\mathfrak{M} \times \mathfrak{M}$

A divergence function $\mathcal{D}$ is given as a bi-variable function on $\mathfrak{M}$ (of dimension n). We now view it as a (single-variable) function on $\mathfrak{M} \times \mathfrak{M}$ (of dimension $2n$) that assumes zero value along the diagonal $\Delta_{\mathfrak{M}} \subset \mathfrak{M} \times \mathfrak{M}$. In this subsection, we investigate the condition under which a divergence function can serve as a "generating function" of a symplectic structure on $\mathfrak{M} \times \mathfrak{M}$. A compatible metric on $\mathfrak{M} \times \mathfrak{M}$ will also be derived.

First, we fix a particular y or a particular x in $\mathfrak{M} \times \mathfrak{M}$—this results in two n-dimensional submanifolds of $\mathfrak{M} \times \mathfrak{M}$ that will be denoted, respectively, $\mathfrak{M}_x \simeq \mathfrak{M}$ (with y point fixed) and $\mathfrak{M}_y \simeq \mathfrak{M}$ (with x point fixed). Let us write out the canonical symplectic form ω_x on the cotangent bundle $\mathcal{T}^*\mathfrak{M}_x$ given by

$$\omega_x = dx^i \wedge d\xi^i.$$

Given $\mathcal{D}$, we define a map $L_{\mathcal{D}}$ from $\mathfrak{M} \times \mathfrak{M} \to \mathcal{T}^*\mathfrak{M}_x$, $(x, y) \mapsto (x, \xi)$ given by

$$L_{\mathcal{D}}\colon\ (x, y) \mapsto (x, \mathcal{D}_i(x, y)dx^i).$$

(Recall that the comma separates the variable being in the first slot versus the second slot for differentiation.) It is easy to check that in a neighborhood of the diagonal $\Delta_{\mathfrak{M}} \subset \mathfrak{M} \times \mathfrak{M}$, the map $L_{\mathcal{D}}$ is a diffeomorphism since the Jacobian matrix of the map

$$\begin{pmatrix} \delta_{ij} & \mathcal{D}_{ij} \\ 0 & \mathcal{D}_{i,j} \end{pmatrix}$$

is non-degenerate in such a neighborhood of the diagonal $\Delta_{\mathfrak{M}}$.

We calculate the pullback of this symplectic form (defined on $\mathcal{T}^*\mathfrak{M}_x$) to $\mathfrak{M} \times \mathfrak{M}$:

$$\begin{aligned} L_{\mathcal{D}}^* \omega_x &= L_{\mathcal{D}}^* (dx^i \wedge d\xi^i) = dx^i \wedge d\mathcal{D}_i(x, y) \\ &= dx^i \wedge (\mathcal{D}_{ij}(x, y)dx^j + \mathcal{D}_{i,j}dy^j) = \mathcal{D}_{i,j}(x, y)dx^i \wedge dy^j. \end{aligned}$$

(Here $\mathcal{D}_{ij}dx^i \wedge dx^j = 0$ since $\mathcal{D}_{ij}(x, y) = \mathcal{D}_{ji}(x, y)$ always holds.)

Similarly, we consider the canonical symplectic form $\omega_y = dy^i \wedge d\eta^i$ on $\mathfrak{M}_y$ and define a map $R_{\mathcal{D}}$ from $\mathfrak{M} \times \mathfrak{M} \to \mathcal{T}^*\mathfrak{M}_y$, $(x, y) \mapsto (y, \eta)$ given by

$$R_{\mathcal{D}}\colon\ (x, y) \mapsto (y, \mathcal{D}_{,i}(x, y)dy^i).$$

Using $R_{\mathcal{D}}$ to pullback ω_y to $\mathfrak{M} \times \mathfrak{M}$ yields an analogous formula:

$$R_{\mathcal{D}}^* \omega_y = -\mathcal{D}_{i,j}(x, y)dx^i \wedge dy^j.$$

Therefore, based on canonical symplectic forms on $\mathcal{T}^*\mathfrak{M}_x$ and $\mathcal{T}^*\mathfrak{M}_y$, we obtained the same symplectic form on $\mathfrak{M} \times \mathfrak{M}$

$$\omega_{\mathcal{D}}(x, y) = -\mathcal{D}_{i,j}(x, y)dx^i \wedge dy^j. \tag{1.29}$$

Proposition 8 *A divergence function $\mathcal{D}$ induces a symplectic form $\omega_{\mathcal{D}}$ (1.29) on $\mathfrak{M} \times \mathfrak{M}$ which is the pullback of the canonical symplectic forms ω_x and ω_y by the maps $L_{\mathcal{D}}$ and $R_{\mathcal{D}}$*

$$L_{\mathcal{D}}^* \omega_y = \mathcal{D}_{i,j}(x, y)dx^i \wedge dy^j = -R_{\mathcal{D}}^* \omega_x \tag{1.30}$$

It was Barndorff-Nielsen and Jupp [3] who first proposed (1.29) as an induced symplectic form on $\mathfrak{M} \times \mathfrak{M}$, apart from a minus sign; the divergence function $\mathcal{D}$ was called a "york". As an example, Bregman divergence $\mathcal{B}_{\Phi}$ (given by (1.33) below) induces the symplectic form $\sum \Phi_{ij}dx^i \wedge dy^j$.

1.3.3 Almost Complex Structure and Hermite Metric on $\mathfrak{M} \times \mathfrak{M}$

An almost complex structure J on $\mathfrak{M} \times \mathfrak{M}$ is defined by a vector bundle isomorphism (from $\mathcal{T}(\mathfrak{M} \times \mathfrak{M})$ to itself), with the property that $J^2 = -Id$. Requiring J to be compatible with $\omega_{\mathcal{D}}$, that is,

$$\omega_{\mathcal{D}}(JX, JY) = \omega_{\mathcal{D}}(X, Y), \ \ \forall X, Y \in \mathcal{T}_{(x,y)}(\mathfrak{M} \times \mathfrak{M}),$$

we may obtain a constraint on the divergence function $\mathcal{D}$. From

$$\omega_{\mathcal{D}}\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial y^j}\right) = \omega_{\mathcal{D}}\left(J\frac{\partial}{\partial x^i}, J\frac{\partial}{\partial y^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial y^i}, -\frac{\partial}{\partial x^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial y^i}\right),$$

we require

$$\mathcal{D}_{i,j} = \mathcal{D}_{j,i}, \tag{1.31}$$

or explicitly

$$\frac{\partial^2 \mathcal{D}}{\partial x^i \partial y^j} = \frac{\partial^2 \mathcal{D}}{\partial x^j \partial y^i}.$$

Note that this condition is always satisfied on $\Delta_{\mathfrak{M}}$, by the definition of a divergence function $\mathcal{D}$, which has allowed us to define a Riemannian structure on $\Delta_{\mathfrak{M}}$ (Proposition 6). We now require it to be satisfied on $\mathfrak{M} \times \mathfrak{M}$ (at least a neighborhood of $\Delta_{\mathfrak{M}}$).

For divergence functions satisfying (1.31), we can consider inducing a metric $G_{\mathcal{D}}$ on $\mathfrak{M} \times \mathfrak{M}$—the induced Riemannian (Hermit) metric $G_{\mathcal{D}}$ is defined by

$$G_{\mathcal{D}}(X, Y) = \omega_{\mathcal{D}}(X, JY).$$

It is easy to verify $G_{\mathcal{D}}$ is invariant under the almost complex structure J. The metric components are given by:

$$G_{ij} = G_{\mathcal{D}}\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial x^i}, J\frac{\partial}{\partial x^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial y^j}\right) = -\mathcal{D}_{i,j},$$

$$G_{i'j'} = g_{\mathcal{D}}\left(\frac{\partial}{\partial y^i}, \frac{\partial}{\partial y^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial y^i}, J\frac{\partial}{\partial y^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial y^i}, -\frac{\partial}{\partial x^j}\right) = -\mathcal{D}_{j,i},$$

$$G_{ij'} = g_{\mathcal{D}}\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial y^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial x^i}, J\frac{\partial}{\partial y^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial x^i}, -\frac{\partial}{\partial x^j}\right) = 0.$$

$$G_{i'j} = g_{\mathcal{D}}\left(\frac{\partial}{\partial y^i}, \frac{\partial}{\partial x^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial y^i}, J\frac{\partial}{\partial x^j}\right) = \omega_{\mathcal{D}}\left(\frac{\partial}{\partial y^i}, -\frac{\partial}{\partial y^j}\right) = 0.$$

So the desired Riemannian metric on $\mathfrak{M} \times \mathfrak{M}$ is

$$G_{\mathcal{D}} = -\mathcal{D}_{i,j}\Big(dx^i dx^j + dy^i dy^j\Big).$$

So for $G_{\mathcal{D}}$ to be a Riemannian metric, we require $-\mathcal{D}_{i,j}$ to be positive-definite.

We call a divergence function $\mathcal{D}$ *proper* if and only if $-\mathcal{D}_{i,j}$ is symmetric and positive-definite on $\mathfrak{M} \times \mathfrak{M}$. Just as any divergence function induces a Riemannian structure on the diagonal manifold $\Delta_{\mathfrak{M}}$ of $\mathfrak{M} \times \mathfrak{M}$, any proper divergence function induces a Riemannian structure on $\mathfrak{M} \times \mathfrak{M}$ that is compatible with the symplectic structure $\omega_{\mathcal{D}}$ on it.

1.3.4 Complexification and Kähler Structure on $\mathfrak{M} \times \mathfrak{M}$

We now discuss possible existence of a Kähler structure on the product manifold $\mathfrak{M} \times \mathfrak{M}$. By definition,

$$\begin{aligned} ds^2 &= G_{\mathcal{D}} - \sqrt{-1}\omega_{\mathcal{D}} \\ &= -\mathcal{D}_{i,j}\Big(dx^i \otimes dx^j + dy^i \otimes dy^j\Big) + \sqrt{-1}\mathcal{D}_{i,j}\Big(dx^i \otimes dy^j - dy^i \otimes dx^j\Big) \\ &= -\mathcal{D}_{i,j}\Big(dx^i + \sqrt{-1}dy^i\Big) \otimes \Big(dx^j - \sqrt{-1}dy^j\Big) = -\mathcal{D}_{i,j}dz^i \otimes d\bar{z}^j. \end{aligned}$$

Now introduce complex coordinates $z = x + \sqrt{-1}y$,

$$\mathcal{D}(x, y) = \mathcal{D}\left(\frac{z + \bar{z}}{2}, \frac{z - \bar{z}}{2\sqrt{-1}}\right) \equiv \widehat{D}(z, \bar{z}),$$

so

$$\frac{\partial^2 \mathcal{D}}{\partial z^i \partial \bar{z}^j} = \frac{1}{4}(\mathcal{D}_{ij} + \mathcal{D}_{,ij}) = \frac{1}{2}\frac{\partial^2 \widehat{\mathcal{D}}}{\partial z^i \partial \bar{z}^j}.$$

If $\mathcal{D}$ satisfies

$$\mathcal{D}_{ij} + \mathcal{D}_{,ij} = \kappa \mathcal{D}_{i,j} \tag{1.32}$$

where κ is a constant, then $\mathfrak{M} \times \mathfrak{M}$ admits a Kähler potential (and hence $\widehat{\mathcal{D}}$ is a Kähler manifold)

$$ds^2 = \frac{\kappa}{2}\frac{\partial^2 \widehat{\mathcal{D}}}{\partial z^i \partial \bar{z}^j} dz^i \otimes d\bar{z}^j.$$

1.3.5 Canonical Divergence for Hessian Manifold

On dually flat (i.e., Hessian) manifold, there is a canonical divergence as shown below. Recall that the Hessian metric

$$g_{ij}(x) = \frac{\partial^2 \Phi(x)}{\partial x^i \partial x^j}$$

and the dual connections

$$\Gamma_{ij,k}(x) = 0, \quad \Gamma^*_{ij,k}(x) = \frac{\partial^3 \Phi(x)}{\partial x^i \partial x^j \partial x^k}$$

are induced from a convex potential function Φ. In the (biorthogonal) u-coordinates, these geometric quantities can be expressed as

$$g^{ij}(u) = \frac{\partial^2 \widetilde{\Phi}(u)}{\partial u_i \partial u_j}, \quad \Gamma^{*ij,k}(u) = 0, \quad \Gamma^{ij,k}(u) = \frac{\partial^3 \widetilde{\Phi}(u)}{\partial u_i \partial u_j \partial u_k},$$

where $\widetilde{\Phi}$ is the convex conjugate of Φ.

Integrating the Hessian structure reveals the so-called Bregman divergence $\mathcal{B}_\Phi(x, y)$ [4] as the generating function:

$$\mathcal{B}_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle x - y, \partial\Phi(y)\rangle \tag{1.33}$$

where $\partial\Phi = [\partial_1\Phi, \ldots, \partial_n\Phi]$ with $\partial_i \equiv \partial/\partial x^i$ denotes the gradient valued in the co-vector space $\widetilde{\mathbb{R}}^n$, and $\langle \cdot, \cdot \rangle_n$ denotes the canonical pairing of a point/vector $x = [x^1, \ldots, x^n] \in \mathbb{R}^n$ and a point/co-vector $u = [u_1, \ldots, u_n] \in \widetilde{\mathbb{R}}_n$ (dual to $\mathbb{R}^n$):

$$\langle x, u\rangle_n = \sum_{i=1}^{n} x^i u_i. \tag{1.34}$$

(Where there is no danger of confusion, the subscript n in $\langle \cdot, \cdot \rangle_n$ is often omitted.) A basic fact in convex analysis is that the necessary and sufficient condition for a smooth function Φ to be strictly convex is

$$\mathcal{B}_\Phi(x, y) > 0 \tag{1.35}$$

for $x \neq y$.

Recall that, when Φ is convex, its convex conjugate $\widetilde{\Phi}\colon \widetilde{V} \subseteq \widetilde{\mathbb{R}}_n \to \mathbb{R}$ is defined through the Legendre transform:

$$\widetilde{\Phi}(u) = \langle (\partial\Phi)^{-1}(u), u \rangle - \Phi((\partial\Phi)^{-1}(u)), \tag{1.36}$$

with $\widetilde{\widetilde{\Phi}} = \Phi$ and $(\partial\Phi) = (\partial\widetilde{\Phi})^{-1}$. The function $\widetilde{\Phi}$ is also convex, and through which (1.35) precisely expresses the Fenchel inequality

$$\Phi(x) + \widetilde{\Phi}(u) - \langle x, u \rangle \geq 0$$

for any $x \in V$, $u \in \widetilde{V}$, with equality holding if and only if

$$u = (\partial\Phi)(x) = (\partial\widetilde{\Phi})^{-1}(x) \longleftrightarrow x = (\partial\widetilde{\Phi})(u) = (\partial\Phi)^{-1}(u), \tag{1.37}$$

or, in component form,

$$u_i = \frac{\partial \Phi}{\partial x^i} \longleftrightarrow x^i = \frac{\partial \widetilde{\Phi}}{\partial u_i}. \tag{1.38}$$

With the aid of conjugate variables, we can introduce the "canonical divergence" $\mathcal{A}_\Phi\colon V \times \widetilde{V} \to \mathbb{R}_+$ (and $\mathcal{A}_{\widetilde{\Phi}}\colon \widetilde{V} \times V \to \mathbb{R}_+$)

$$\mathcal{A}_\Phi(x, v) = \Phi(x) + \widetilde{\Phi}(v) - \langle x, v \rangle = \mathcal{A}_{\widetilde{\Phi}}(v, x).$$

They are related to the Bregman divergence (1.33) via

$$\mathcal{B}_\Phi(x, (\partial\Phi)^{-1}(v)) = \mathcal{A}_\Phi(x, v) = \mathcal{B}_{\widetilde{\Phi}}((\partial\widetilde{\Phi})(x), v).$$

1.4 $\mathcal{D}_\Phi$-Divergence and Its Induced Structures

In this section, we study a particular parametric family of divergence functions, called $\mathcal{D}_\Phi$, induced by a strictly convex function Φ, with α as the parameter. This family was first introduced by Zhang [23], who showed that it included many familiar families (see also [27]). The resulting geometric structures will be studied below.

1.4.1 $\mathcal{D}_\Phi$-Divergence Functions

Recall that, by definition, a strictly convex function $\Phi: V \subseteq \mathbb{R}^n \to \mathbb{R}, x \mapsto \Phi(x)$ satisfies

$$\frac{1-\alpha}{2}\Phi(x) + \frac{1+\alpha}{2}\Phi(y) - \Phi\left(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right) > 0 \tag{1.39}$$

for all $x \neq y$ for any $|\alpha| < 1$ (the inequality sign is reversed when $|\alpha| > 1$). Assume Φ to be sufficiently smooth (differentiable up to fourth order).

Zhang [23] introduced the following family of function on $V \times V$ as indexed by $\alpha \in \mathbb{R}$

$$\mathcal{D}_\Phi^{(\alpha)}(x,y) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}\Phi(x) + \frac{1+\alpha}{2}\Phi(y) - \Phi\left(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right)\right). \tag{1.40}$$

From its construction, $\mathcal{D}_\Phi^{(\alpha)}(x,y)$ is non-negative for $|\alpha| < 1$ due to Eq. (1.39), and for $|\alpha| = 1$ due to Eq. (1.35). For $|\alpha| > 1$, assuming $(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y) \in V$, the non-negativity of $\mathcal{D}_\Phi^{(\alpha)}(x,y)$ can also be proven due to the inequality (1.39) reversing its sign. Furthermore, $\mathcal{D}_\Phi^{(\pm 1)}(x,y)$ is defined by taking $\lim_{\alpha\to\pm 1}$:

$$\begin{aligned}\mathcal{D}_\Phi^{(1)}(x,y) &= \mathcal{D}_\Phi^{(-1)}(y,x) = B_\Phi(x,y),\\ \mathcal{D}_\Phi^{(-1)}(x,y) &= \mathcal{D}_\Phi^{(1)}(y,x) = B_\Phi(y,x).\end{aligned}$$

Note that $\mathcal{D}_\Phi^{(\alpha)}(x,y)$ satisfies the relation (called "referential duality" in [24])

$$\mathcal{D}_\Phi^{(\alpha)}(x,y) = \mathcal{D}_\Phi^{(-\alpha)}(y,x),$$

that is, exchanging the asymmetric status of the two points (in the directed distance) amounts to $\alpha \leftrightarrow -\alpha$.

1.4.2 Induced α-Hessian Structure on $T\mathfrak{M}$

We start by reviewing a main result from [23] linking the divergence function $\mathcal{D}_\Phi^{(\alpha)}(x,y)$ defined in (1.40) and the α-Hessian structure.

Proposition 9 ([23]) *The manifold* $\{\mathcal{M}, g(x), \Gamma^{(\alpha)}(x), \Gamma^{(-\alpha)}(x)\}$[4] *associated with* $\mathcal{D}_\Phi^{(\alpha)}(x,y)$ *is given by*

$$g_{ij}(x) = \Phi_{ij} \tag{1.41}$$

[4] The functional argument of x (or u-below) indicates that x-coordinate system (or u-coordinate system) is being used. Recall from Sect. 1.2.5 that under x (u, resp) local coordinates, g and Γ, in component forms, are expressed by lower (upper, resp) indices.

and

$$\Gamma^{(\alpha)}_{ij,k}(x) = \frac{1-\alpha}{2}\,\Phi_{ijk}, \quad \Gamma^{*(\alpha)}_{ij,k}(x) = \frac{1+\alpha}{2}\,\Phi_{ijk}. \tag{1.42}$$

Here, Φ_{ij}, Φ_{ijk} denote, respectively, second and third partial derivatives of $\Phi(x)$

$$\Phi_{ij} = \frac{\partial^2 \Phi(x)}{\partial x^i \partial x^j}, \quad \Phi_{ijk} = \frac{\partial^3 \Phi(x)}{\partial x^i \partial x^j \partial x^k}.$$

Recall that an α-Hessian manifold is equipped with an α-independent metric and a family of α-transitively flat connections $\Gamma^{(\alpha)}$ (i.e., $\Gamma^{(\alpha)}$ satisfying (1.16) and $\Gamma^{(\pm 1)}$ are dually flat). From (1.42),

$$\Gamma^{*(\alpha)}_{ij,k} = \Gamma^{(-\alpha)}_{ij,k},$$

with the Levi-Civita connection given as:

$$\widehat{\Gamma}_{ij,k}(x) = \frac{1}{2}\Phi_{ijk}.$$

Straightforward calculation shows that:

Proposition 10 ([26]) *For α-Hessian manifold $\{\mathfrak{M}, g(x), \Gamma^{(\alpha)}(x), \Gamma^{(-\alpha)}(x)\}$,*

(i) *the Riemann curvature tensor of the α-connection is given by:*

$$R^{(\alpha)}_{\mu\nu ij}(x) = \frac{1-\alpha^2}{4}\sum_{l,k}(\Phi_{il\nu}\Phi_{jk\mu} - \Phi_{il\mu}\Phi_{jk\nu})\Psi^{lk} \;=\; R^{*(\alpha)}_{ij\mu\nu}(x),$$

with Ψ^{ij} being the matrix inverse of Φ_{ij};

(ii) *all α-connections are equiaffine, with the α-parallel volume forms (i.e., the volume forms that are parallel under α-connections) given by*

$$\omega^{(\alpha)}(x) = \det[\Phi_{ij}(x)]^{\frac{1-\alpha}{2}}.$$

It is worth pointing out that while $\mathcal{D}_\Phi$-divergence induces the α-Hessian structure, it is not unique, as the same structure can arise from the following divergence function, which is a mixture of Bregman divergences in conjugate forms:

$$\frac{1-\alpha}{2}\mathcal{B}_\Phi(x,y) + \frac{1+\alpha}{2}\mathcal{B}_\Phi(y,x).$$

1.4.3 The Family of α-Geodesics

The family of auto-parallel curves on α-Hessian manifold have analytic expression. From

$$\frac{d^2x^i}{ds^2} + \sum_{j,k} \Gamma_{jk}^{i(\alpha)} \frac{dx^j}{ds}\frac{dx^k}{ds} = 0$$

and substituting (1.42), we obtain

$$\sum_i \Phi_{ki}\frac{d^2x^i}{ds^2} + \frac{1-\alpha}{2}\sum_{i,l}\Phi_{kij}\frac{dx^j}{ds}\frac{dx^k}{ds} = 0 \longleftrightarrow \frac{d^2}{ds^2}\Phi_k\left(\frac{1-\alpha}{2}x\right) = 0.$$

So the auto-parallel curves of an α-Hessian manifold all have the form

$$\Phi_k\left(\frac{1-\alpha}{2}x\right) = a^k s^{(\alpha)} + b^k$$

where the scalar s is the arc length and $a^k, b^k, k = 1, 2\ldots, n$ are constant vectors (determined by a point and the direction along which the auto-parallel curve flows through). For $\alpha = -1$, the auto-parallel curves are given by $u^k = \Phi_k(x) = a^k s + b^k$ are affine coordinates as previously noted.

1.4.3.1 Related Divergences and Geometries

Note that the metric and conjugated connections in the forms (1.41) and (1.42) are induced from (1.40). Using the convex conjugate $\widetilde{\Phi}\colon \widetilde{V} \to \mathbb{R}$ given by (1.36), we introduce the following family of divergence functions $\widetilde{\mathcal{D}}_{\widetilde{\Phi}}^{(\alpha)}(x, y)$ defined by

$$\widetilde{\mathcal{D}}_{\widetilde{\Phi}}^{(\alpha)}(x, y) \equiv \mathcal{D}_{\widetilde{\Phi}}^{(\alpha)}((\partial\Phi)(x), (\partial\Phi)(y)).$$

Explicitly written, this new family of divergence functions is

$$\widetilde{\mathcal{D}}_{\widetilde{\Phi}}^{(\alpha)}(x, y) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}\widetilde{\Phi}(\partial\Phi(x)) + \frac{1+\alpha}{2}\widetilde{\Phi}(\partial\Phi(y)) - \widetilde{\Phi}\left(\frac{1-\alpha}{2}\partial\Phi(x) + \frac{1+\alpha}{2}\partial\Phi(y)\right)\right).$$

Straightforward calculation shows that $\widetilde{\mathcal{D}}_{\widetilde{\Phi}}^{(\alpha)}(x, y)$ induces the α-Hessian structure $\{\mathcal{M}, g, \Gamma^{(-\alpha)}, \Gamma^{(\alpha)}\}$ where $\Gamma^{(\mp\alpha)}$ are given by (1.42); that is, the pair of α-connections are themselves "conjugate" (in the sense of $\alpha \leftrightarrow -\alpha$) to those induced by $\mathcal{D}_{\Phi}^{(\alpha)}(x, y)$.

If, instead of choosing $x = [x^1, \ldots, x^n]$ as the local coordinates for the manifold $\mathfrak{M}$, we use its biorthogonal counterpart $u = [u_1, \ldots, u_n]$ related to x via (1.38) to index points on $\mathcal{M}$. Under this u-coordinate system, the divergence function $\mathcal{D}_{\Phi}^{(\alpha)}$ between the same two points on $\mathcal{M}$ becomes

$$\widetilde{\mathcal{D}}_{\Phi}^{(\alpha)}(u, v) \equiv \mathcal{D}_{\Phi}^{(\alpha)}((\partial\widetilde{\Phi})(u), (\partial\widetilde{\Phi})(v)).$$

Explicitly written,

$$\widetilde{\mathcal{D}}_{\Phi}^{(\alpha)}(u, v) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}\Phi((\partial\Phi)^{-1}(u)) + \frac{1+\alpha}{2}\Phi((\partial\Phi)^{-1}(v)) - \Phi\left(\frac{1-\alpha}{2}(\partial\Phi)^{-1}(u) + \frac{1+\alpha}{2}(\partial\Phi)^{-1}(v)\right)\right).$$

Proposition 11 ([23]) *The α-Hessian manifold $\{\mathfrak{M}, g(u), \Gamma^{(\alpha)}(u), \Gamma^{(-\alpha)}(u)\}$ associated with $\widetilde{\mathcal{D}}_{\Phi}^{(\alpha)}(u, v)$ is given by*

$$g^{ij}(u) = \widetilde{\Phi}^{ij}(u), \tag{1.43}$$

$$\Gamma^{(\alpha)ij,k}(u) = \frac{1+\alpha}{2}\widetilde{\Phi}^{ijk}, \quad \Gamma^{*(\alpha)ij,k}(u) = \frac{1-\alpha}{2}\widetilde{\Phi}_{ijk}, \tag{1.44}$$

Here, $\widetilde{\Phi}^{ij}$, $\widetilde{\Phi}^{ijk}$ denote, respectively, second and third partial derivatives of $\widetilde{\Phi}(u)$

$$\widetilde{\Phi}^{ij}(u) = \frac{\partial^2\widetilde{\Phi}(u)}{\partial u_i \partial u_j}, \quad \widetilde{\Phi}^{ijk}(u) = \frac{\partial^3\widetilde{\Phi}(u)}{\partial u_i \partial u_j \partial u_k}.$$

We remark that the same metric (1.43) and the same α-connections (1.44) are induced by $\mathcal{D}_{\widetilde{\Phi}}^{(-\alpha)}(u, v) \equiv \mathcal{D}_{\widetilde{\Phi}}^{(\alpha)}(v, u)$—this follows as a simple application of Eguchi relation.

An application of (1.23) gives rise to the following relations:

$$\Gamma^{(\alpha)mn,l}(u) = -\sum_{i,j,k} g^{im}(u)g^{jn}(u)g^{kl}(u)\Gamma_{ij,k}^{(-\alpha)}(x),$$

$$\Gamma^{*(\alpha)mn,l}(u) = -\sum_{i,j,k} g^{im}(u)g^{jn}(u)g^{kl}(u)\Gamma_{ij,k}^{(\alpha)}(x),$$

$$R^{(\alpha)klmn}(u) = \sum_{i,j,\mu,\nu} g^{ik}(u)g^{jl}(u)g^{\mu m}(u)g^{\nu n}(u)R_{ij\mu\nu}^{(\alpha)}(x).$$

The volume form associated with $\Gamma^{(\alpha)}$ is

$$\omega^{(\alpha)}(u) = \det[\widetilde{\Phi}^{ij}(u)]^{\frac{1+\alpha}{2}}.$$

Table 1.1 Various divergence functions under biorthogonal coordinates x or u and their induced geometries

Divergence function	Induced geometry
$\mathcal{D}_{\Phi}^{(\alpha)}(x, y)$	$\{\Phi_{ij}(x), \Gamma(x)^{(\alpha)}, \Gamma(x)^{(-\alpha)}\}$
$\mathcal{D}_{\widetilde{\Phi}}^{(\alpha)}((\partial\Phi)(x), (\partial\Phi)(y))$	$\{\Phi_{ij}(x), \Gamma(x)^{(-\alpha)}, \Gamma(x)^{(\alpha)}\}$
$\mathcal{D}_{\widetilde{\Phi}}^{(\alpha)}(u, v)$	$\{\widetilde{\Phi}^{ij}(u), \Gamma(u)^{(-\alpha)}, \Gamma(u)^{(\alpha)}\}$
$\mathcal{D}_{\Phi}^{(\alpha)}((\partial\widetilde{\Phi})(x), (\partial\widetilde{\Phi})(y))$	$\{\widetilde{\Phi}^{ij}(u), \Gamma(u)^{(\alpha)}, \Gamma(u)^{(-\alpha)}\}$

When $\alpha = \pm 1$, $\widetilde{\mathcal{D}}_{\Phi}^{(\alpha)}(u, v)$, as well as $\widetilde{\mathcal{D}}_{\widetilde{\Phi}}^{(\alpha)}(x, y)$ introduced earlier, take the form of Bregman divergence (1.33). In this case, the manifold is dually flat, with Riemann curvature tensor $R_{ij\mu\nu}^{(\pm 1)}(u) = R^{(\pm 1)klmn}(x) = 0$.

We summarize the relations between the convex-based divergence functions and the geometry they generate in Table 1.1. The duality associated with $\alpha \leftrightarrow -\alpha$ is called "referential duality" whereas the duality associated with $x \leftrightarrow u$ is called representational duality [23, 24, 27].

1.4.4 Induced Symplectic and Kähler Structures on $\mathfrak{M} \times \mathfrak{M}$

With respect to the $\mathcal{D}_{\Phi}$-divergence (1.40), observe that

$$\Phi\left(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right) = \Phi\Big((\frac{1-\alpha}{4} + \frac{1+\alpha}{4\sqrt{-1}})z + (\frac{1-\alpha}{4} - \frac{1+\alpha}{4\sqrt{-1}})\bar{z}\Big) \equiv \widehat{\Phi}^{(\alpha)}(z, \bar{z}), \tag{1.45}$$

we have

$$\frac{\partial^2\widehat{\Phi}^{(\alpha)}}{\partial z^i \partial\bar{z}^j} = \frac{1+\alpha^2}{8}\Phi_{ij}\Big(\left(\frac{1-\alpha}{4} + \frac{1+\alpha}{4\sqrt{-1}}\right)z + \left(\frac{1-\alpha}{4} - \frac{1+\alpha}{4\sqrt{-1}}\right)\bar{z}\Big)$$

which is symmetric in i, j. Both (1.31) and (1.32) are satisfied. The symplectic form, under the complex coordinates, is given by

$$\omega^{(\alpha)} = \Phi_{ij}\left(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}\right)dx^i \wedge dy^j = \frac{4\sqrt{-1}}{1+\alpha^2}\frac{\partial^2\widehat{\Phi}^{(\alpha)}}{\partial z^i \partial\bar{z}^j}dz^i \wedge d\bar{z}^j$$

and the line-element is given by

$$ds^2 = \frac{8}{1+\alpha^2}\frac{\partial^2\widehat{\Phi}^{(\alpha)}}{\partial z^i \partial\bar{z}^j}dz^i \otimes d\bar{z}^j.$$

Proposition 12 ([28]) *A smooth, strictly convex function Φ: $dom(\Phi) \subset \mathfrak{M} \to \mathbb{R}$ induces a a family of Kähler structure $(\mathfrak{M}, \omega^{(\alpha)}, G^{(\alpha)})$ defined on $dom(\Phi) \times dom(\Phi) \subset \mathfrak{M} \times \mathfrak{M}$ with*

1. *the symplectic form* $\omega^{(\alpha)}$ *is given by*

$$\omega^{(\alpha)} = \Phi_{ij}^{(\alpha)} dx^i \wedge dy^j$$

which is compatible with the canonical almost complex structure J

$$\omega^{(\alpha)}(JX, JY) = \omega^{(\alpha)}(X, Y),$$

where X, Y are vector fields on dom$\Phi \times dom(\Phi)$*;*

2. *the Riemannian metric* $G^{(\alpha)}$*, compatible with J and* $\omega^{(\alpha)}$ *above, is given by* $\Phi_{ij}^{(\alpha)}$

$$G^{(\alpha)} = \Phi_{ij}^{(\alpha)}(dx^i dx^j + dy^i dy^j);$$

3. *the Kähler structure*

$$ds^{2(\alpha)} = \Phi_{ij}^{(\alpha)} dz^i \otimes d\bar{z}^j = \frac{8}{1+\alpha^2} \frac{\partial^2 \widehat{\Phi}^{(\alpha)}}{\partial z^i \partial \bar{z}^j},$$

with the Kähler potential given by

$$\frac{2}{1+\alpha^2} \widehat{\Phi}^{(\alpha)}(z, \bar{z}).$$

Here, $\Phi_{ij}^{(\alpha)} = \Phi_{ij}\left(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right)$.

For the diagonal manifold $\Delta_{\mathfrak{M}} = \{(x, x) : x \in \mathfrak{M}\}$, a basis of its tangent space $\mathcal{T}_{(x,x)}\Delta_{\mathfrak{M}}$ can be selected as

$$e_i = \frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial x^i} + \frac{\partial}{\partial y^i}\right).$$

The Riemannian metric on the diagonal, induced from $G^{(\alpha)}$ is

$$\begin{aligned} G^{(\alpha)}(e_i, e_j)|_{x=y} &= \langle G^{(\alpha)}, e_i \otimes e_j \rangle \\ &= \langle \Phi_{kl}^{(\alpha)}(dx^k \otimes dx^l + dy^k \otimes dy^l), \frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial x^i} + \frac{\partial}{\partial y^i}\right) \otimes \frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial x^j} + \frac{\partial}{\partial y^j}\right) \rangle \\ &= \Phi_{ij}^{(\alpha)}(x, x) = \Phi_{ij}(x), \end{aligned}$$

where $\langle \alpha, a \rangle$ denotes a form α operating on a tensor field a. Therefore, restricting to the diagonal $\Delta_{\mathfrak{M}}$, $g^{(\alpha)}$ reduces to the Riemannian metric induced by the divergence $\mathcal{D}_{\Phi}^{(\alpha)}$ through the Eguchi method.

We next calculate the Levi-Civita connection $\tilde{\Gamma}$ associated with $G^{(\alpha)}$. Denote $x^{i'} = y^i$, and that

$$\tilde{\Gamma}_{i'jk'} = G^{(\alpha)}\left(\nabla_{\frac{\partial}{\partial x^{i'}}}\frac{\partial}{\partial x^{j}}, \frac{\partial}{\partial x^{k'}}\right) = G^{(\alpha)}\left(\nabla_{\frac{\partial}{\partial y^{i}}}\frac{\partial}{\partial x^{j}}, \frac{\partial}{\partial y^{k}}\right),$$

and so on. The Levi-Civita connection on $\mathfrak{M} \times \mathfrak{M}$ is

$$\tilde{\Gamma}_{ijk} = \frac{1}{2}\Big(\frac{\partial G^{(\alpha)}_{ik}}{\partial x^{j}} + \frac{\partial G^{(\alpha)}_{jk}}{\partial x^{i}} - \frac{\partial G^{(\alpha)}_{ij}}{\partial x^{k}}\Big) = \frac{1-\alpha}{4}\Phi^{(\alpha)}_{ijk}.$$

$$\tilde{\Gamma}_{ijk'} = \frac{1}{2}\Big(\frac{\partial G^{(\alpha)}_{ik'}}{\partial x^{j}} + \frac{\partial G^{(\alpha)}_{jk'}}{\partial x^{i}} - \frac{\partial G^{(\alpha)}_{ij}}{\partial x^{k'}}\Big) = -\frac{1+\alpha}{4}\Phi^{(\alpha)}_{ijk}.$$

$$\tilde{\Gamma}_{i'jk'} = \tilde{\Gamma}_{ij'k'} = \frac{1}{2}\Big(\frac{\partial G^{(\alpha)}_{ik'}}{\partial x^{j'}} + \frac{\partial G^{(\alpha)}_{j'k'}}{\partial x^{i}} - \frac{\partial G^{(\alpha)}_{ij'}}{\partial x^{k'}}\Big) = \frac{1-\alpha}{4}\Phi^{(\alpha)}_{ijk}.$$

$$\tilde{\Gamma}_{i'jk} = \tilde{\Gamma}_{ij'k} = \frac{1}{2}\Big(\frac{\partial G^{(\alpha)}_{ik}}{\partial x^{j'}} + \frac{\partial G^{(\alpha)}_{j'k}}{\partial x^{i}} - \frac{\partial G^{(\alpha)}_{ij'}}{\partial x^{k}}\Big) = \frac{1+\alpha}{4}\Phi^{(\alpha)}_{ijk}.$$

$$\tilde{\Gamma}_{i'j'k} = \frac{1}{2}\Big(\frac{\partial G^{(\alpha)}_{i'k}}{\partial x^{j'}} + \frac{\partial G^{(\alpha)}_{j'k}}{\partial x^{i'}} - \frac{\partial G^{(\alpha)}_{i'j'}}{\partial x^{k}}\Big) = -\frac{1-\alpha}{4}\Phi^{(\alpha)}_{ijk}.$$

$$\tilde{\Gamma}_{i'j'k'} = \frac{1}{2}\Big(\frac{\partial G^{(\alpha)}_{i'k'}}{\partial x^{j'}} + \frac{\partial G^{(\alpha)}_{j'k'}}{\partial x^{i'}} - \frac{\partial G^{(\alpha)}_{i'j'}}{\partial x^{k'}}\Big) = \frac{1+\alpha}{4}\Phi^{(\alpha)}_{ijk}.$$

1.5 Summary

In order to construct divergence functions in a principled way, this chapter considered the various geometric structures on the underlying manifold $\mathfrak{M}$ induced from a divergence function. Among the geometric structures considered are: statistical structure (Riemannian metric with a pair of torsion-free dual connections, or by simple construction, a family of α-connections), equiaffine structure (those connections that admit parallel volume forms), and Hessian structure (those connections that are dually flat)—they are progressively more restrictive: while any divergence function will induce a statistical manifold, only canonical divergence (i.e., Bregman divergence) will induce a Hessian manifold. Lying in-between these extremes is the equiaffine α-Hessian geometry induced from, say, the class of $\mathcal{D}_\Phi$-divergence. The α-Hessian structure has the advantage of the existence of biorthogonal coordinates, induced from the convex function Φ and its conjugate; these coordinates are convenient for computation. It should be noted that the above geometric structures, from statistical to Hessian, are all induced on the tangent bundle $\mathcal{T}\mathfrak{M}$ of the manifold $\mathfrak{M}$ on which the divergence function is defined.

On the cotangent bundle $\mathcal{T}^*\mathfrak{M}$ side, a divergence function can be viewed as a generating function for a symplectic structure on $\mathfrak{M} \times \mathfrak{M}$ that can be constructed in a "canonical" way. This imposes a "properness" condition on divergence function, stating that the mixed second derivatives of $\mathcal{D}(x, y)$ with respect to x and y must commute. For such divergence functions, a Riemannian structure on $\mathfrak{M} \times \mathfrak{M}$ can be constructed, which can be seen as an extension of the Riemannian structure on $\Delta_{\mathfrak{M}} \subset \mathfrak{M} \times \mathfrak{M}$. If a further condition on $\mathcal{D}$ is imposed, then $\mathfrak{M} \times \mathfrak{M}$ may be complexified, so it becomes a Kähler manifold. It was shown that $\mathcal{D}_\Phi$-divergence [23] satisfies this Kählerian condition, in addition to itself being proper—the Kähler potential is simply given by the real-valued convex function Φ. These properties, along with the α-Hessian structure it induces on the tangent bundle, makes $\mathcal{D}_\Phi$ a class of divergence functions that enjoy a special role with "nicest" geometric properties, extending the canonical (Bregman) divergence for dually flat manifolds. This will have implications for machine learning, convex optimization, geometric mechanics, etc.

References

1. Amari, S.: Differential Geometric Methods in Statistics. Lecture Notes in Statistics, vol. 28. Springer, New York (1985) (Reprinted in 1990)
2. Amari, S., Nagaoka, H.: Method of Information Geometry. AMS Monograph. Oxford University Press, Oxford (2000)
3. Barndorff-Nielsen, O.E., Jupp, P.E.: Yorks and symplectic structures. J. Stat. Plan. Inference **63**, 133–146 (1997)
4. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Phys. **7**, 200–217 (1967)
5. Calin, O., Matsuzoe, H., Zhang. J.: Generalizations of conjugate connections. In: Sekigawa, K., Gerdjikov, V., Dimiev, S. (eds.) Trends in Differential Geometry, Complex Analysis and Mathematical Physics: Proceedings of the 9th International Workshop on Complex Structures and Vector Fields, pp. 24–34. World Scientific Publishing, Singapore (2009)
6. Csiszár, I.: On topical properties of f-divergence. Studia Mathematicarum Hungarica **2**, 329–339 (1967)
7. Dombrowski, P.: On the geometry of the tangent bundle. Journal fr der reine und angewandte Mathematik **210**, 73–88 (1962)
8. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**, 793–803 (1983)
9. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**, 631–647 (1992)
10. Lauritzen, S.: Statistical manifolds. In: Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., Rao, C.R. (eds.) Differential Geometry in Statistical Inference. IMS Lecture Notes, vol. 10, pp. 163–216. Institute of Mathematical Statistics, Hayward (1987)
11. Matsuzoe, H.: On realization of conformally-projectively flat statistical manifolds and the divergences. Hokkaido Math. J. **27**, 409–421 (1998)
12. Matsuzoe, H., Inoguchi, J.: Statistical structures on tangent bundles. Appl. Sci. **5**, 55–65 (2003)
13. Matsuzoe, H., Takeuchi, J., Amari, S.: Equiaffine structures on statistical manifolds and Bayesian statistics. Differ. Geom. Appl. **24**, 567–578 (2006)
14. Matsuzoe, H.: Computational geometry from the viewpoint of affine differential geometry. In: Nielsen, F. (ed.) Emerging Trends in Visual Computing, pp. 103–123. Springer, Berlin, Heidelberg (2009)

15. Matsuzoe, M.: Statistical manifolds and affine differential geometry. Adv. Stud. Pure Math. **57**, 303–321 (2010)
16. Nomizu, K., Sasaki, T.: Affine Differential Geometry—Geometry of Affine Immersions. Cambridge University Press, Cambridge (1994)
17. Ohara, A., Matsuzoe, H., Amari, S.: Conformal geometry of escort probability and its applications. Mod. Phys. Lett. B **26**, 1250063 (2012)
18. Shima, H.: Hessian Geometry. Shokabo, Tokyo (2001) (in Japanese)
19. Shima, H., Yagi, K.: Geometry of Hessian manifolds. Differ. Geom. Appl. **7**, 277–290 (1997)
20. Simon, U.: Affine differential geometry. In: Dillen, F., Verstraelen, L. (eds.) Handbook of Differential Geometry, vol. I, pp. 905–961. Elsevier Science, Amsterdam (2000)
21. Uohashi, K.: On α-conformal equivalence of statistical manifolds. J. Geom. **75**, 179–184 (2002)
22. Yano, K., Ishihara, S.: Tangent and Cotangent Bundles: Differential Geometry, vol. 16. Dekker, New York (1973)
23. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**, 159–195 (2004)
24. Zhang, J.: Referential duality and representational duality on statistical manifolds. Proceedings of the 2nd International Symposium on Information Geometry and Its Applications, Tokyo, pp. 58–67 (2006)
25. Zhang, J.: A note on curvature of α-connections on a statistical manifold. Ann. Inst. Stat. Math. **59**, 161–170 (2007)
26. Zhang, J., Matsuzoe, H.: Dualistic differential geometry associated with a convex function. In: Gao, D.Y., Sherali, H.D. (eds.) Advances in Applied Mathematics and Global Optimization (Dedicated to Gilbert Strang on the occasion of his 70th birthday), Advances in Mechanics and Mathematics, vol. III, Chap. 13, pp. 439–466. Springer, New York (2009)
27. Zhang, J.: Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. Entropy **15**, 5384–5418 (2013)
28. Zhang, J., Li, F.: Symplectic and Kähler structures on statistical manifolds induced from divergence functions. In: Nielson, F., Barbaresco, F. (eds.) Proceedings of the 1st International Conference on Geometric Science of Information (GSI2013), pp. 595–603 (2013)

Chapter 2
Geometry on Positive Definite Matrices Deformed by V-Potentials and Its Submanifold Structure

Atsumi Ohara and Shinto Eguchi

Abstract In this paper we investigate dually flat structure of the space of positive definite matrices induced by a class of convex functions called *V-potentials*, from a viewpoint of information geometry. It is proved that the geometry is invariant under special linear group actions and naturally introduces a foliated structure. Each leaf is proved to be a homogeneous statistical manifold with a negative constant curvature and enjoy a special decomposition property of canonically defined divergence. As an application to statistics, we finally give the correspondence between the obtained geometry on the space and the one on elliptical distributions induced from a certain Bregman divergence.

Keywords Information geometry · Divergence · Elliptical distribution · Negative constant curvature · Affine differential geometry

2.1 Introduction

The space of positive definite matrices has been studied both geometrically and algebraically, e.g., as a Riemannian symmetric space and a cone of squares in Jordan algebra ([9, 12, 17, 37, 45] and many others), and the results have broad applications to analysis, statistics, physics and convex optimization and so on.

In the development of Riemannian geometry on the space of positive definite matrices P, the function $\varphi(P) = -k \log\det P$ with a positive constant k plays a cen-

A. Ohara (✉)
Department of Electrical and Electronics Engineering,
University of Fukui, Fukui 910-8507, Japan
e-mail: ohara@fuee.u-fukui.ac.jp

S. Eguchi
The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-0014, Japan
e-mail: eguchi@ism.ac.jp

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_2,

tral role. Important facts are that the function φ is strictly convex, and that the Riemmanian metric defined as the Hessian of φ and the associated Levi-Civita connection are invariant under automorphism group actions on the space, i.e., congruent transformations by general linear group [37, 45]. The group invariance, i.e, homogeneity of the space is also crucial in several applications such as multivariate statistical analysis [23] and mathematical program called the *semidefinite program* (*SDP*) [28, 46].

On the other hand, the space of positive definite matrices admits dualistic geometry (information geometry [1, 2], Hessian geometry [40]), which involves a certain pair of two affine connections instead of the single Levi-Civita connection. Then the geometry on the space not only maintains the group invariance but also reveals its abundant dualistic structure, e.g., Legendre transform, dual flatness, divergences and Pythagorean type relation and so on [34]. Such concepts and results have been proved very natural also in study of applications, e.g., structures of stable matrices [31], properties of means defined as midpoints of dual geodesics [30], or iteration complexity of interior point methods for SDP [15]. In this case, the function φ is still an essential source of the structure and induces the dualistic geometry as a potential function.

The first purpose of this paper is to investigate the geometry of the space of positive definite matrices induced by a special class of convex functions, including φ, from the viewpoints of the dualistic geometry. We here consider the class of functions of the form $\varphi^{(V)} = V(\det P)$ with smooth functions V on positive real numbers $\mathbf{R}_{++}$, and call $\varphi^{(V)}$ a *V-potential* [32].[1] While the dualistic geometry induced from the log-potential $\varphi = \varphi^{(-\log)}$ is standard in the sense that it is invariant for the action of the every automorphism group, the geometries induced from general V-potentials lose the prominent property. We, however, show that they still preserve the invariance for the special linear group actions. On the basis of this fact and by means of the affine hypersurface theory [29], we exploit a foliated structure and discuss the common properties of leaves as homogeneous submanifolds.

Another purpose is to show that the dualistic geometry on the space of positive definite matrices induced from V-potentials naturally corresponds to the geometry of fairly general class of multivariate statistical models:

For a given $n \times n$ positive definite matrix P the n-dimensional Gaussian distribution with zero mean vector is defined by the density function

$$f(x, P) = \exp\{-\frac{1}{2}x^T Px - \varphi(P)\}, \quad \varphi(P) = -\frac{1}{2}\log\det P + \frac{n}{2}\log 2\pi,$$

which has a wide variety of applications in the field of probability theory, statistics, physics and so forth. Information geometry successfully provides elegant geometric insight, such as the Fisher information metric, e-connection and m-connection, with the theory of exponential family of distributions including the Gaussian distribution family [1, 2, 16, 25]. Note that in this geometry, the structure is derived from the

[1] The reference [32] is a conference paper of this chapter omitting proofs and the whole Sect. 2.5.

so-called *Kullback-Leibler divergence (relative entropy)* [18] and the log-potential φ again plays a key role through the conjugate convexity.

The Gaussian distribution is naturally extended to a general non-exponential but symmetric distribution called an *elliptical distribution* [8, 23, 36] with a density function of the form

$$f(x, P) = u\Big(-\frac{1}{2}x^T Px - c_U(\det P)\Big), \quad u(s) = \frac{dU(s)}{ds},$$

where U is a convex function on an interval in $\mathbf{R}$ with the positive derivative u and $c_U(\det P)$ is the normalizing constant. On the family of such distributions we can introduce a new Bregman divergence called a *U-divergence*, yielding geometric counterparts that define natural dually flat geometry of the family. Refer to [4, 24, 33] for the detail of the U-divergence and its applications. Further, recent progress as for theory and applications of *generalized Gaussian distributions* in statistics or statistical physics can be found in, e.g., [3, 10, 26, 27, 42].

We prove that the geometry of elliptical distributions derived by the U-divergence coincides with the dualistic geometry on positive definite matrices induced from the corresponding V-potential. This implies that the various geometric properties of the family can be translated to those of dualistic geometry induced from the corresponding V-potential.

In addition to statistical applications it should be noted that the geometry induced from V-potentials can be applied to derive a class of quasi-Newton update formulas in the field of numerical optimization [14].

The paper is organized as follows:

Section 2.2 provides basic results on information geometry including the idea of statistical manifold.

In Sect. 2.3, we introduce the V-potential function derive the corresponding Riemannian metric. The conditions for the V-potential to have the positive definite Hessian are given.

Section 2.4 focuses on the dual affine connections induced from the V-potential function. It is observed the particular class of potentials called the *power potentials* induce $GL(n, \mathbf{R})$-invariant dual connections while the others generally do not.

In Sect. 2.5, we introduce a foliated structure, which is naturally defined by the special linear group action, and discuss the geometric properties of its associated leaves of constant determinants. We show that each leaf is a Riemmanian symmetric space with respect to the induced Riemannian metric, and that the induced dual connection is independent of the choice of V-potentials. Further, on the basis of affine hypersurface theory, each leaf is proved to be a statistical submanifold with constant curvature. As a consequence, we demonstrate that a new decomposition theorem of the divergence holds for the leaf and a point outside it.

Section 2.6 gives the correspondence between two geometries induced by a V-potential and U-divergence, as an application to statistics.

2.2 Preliminary: Information Geometry and Statistical Manifolds

We recall the basic notions and results in information geometry [2], i.e., dualistic differential geometry of Hessian domain [40] and statistical manifolds [19–21].

For a torsion-free affine connection ∇ and a pseudo-Riemannian metric g on a manifold $\mathcal{M}$, the triple $(\mathcal{M}, \nabla, g)$ is called a *statistical (Codazzi) manifold* if it admits another torsion-free connection ∇^* satisfying

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z) \tag{2.1}$$

for arbitrary X, Y and Z in $\mathcal{X}(\mathcal{M})$, where $\mathcal{X}(\mathcal{M})$ is the set of all tangent vector fields on $\mathcal{M}$. We call ∇ and ∇^* *duals* of each other with respect to g and (g, ∇, ∇^*) is said *dualistic structure* on $\mathcal{M}$.

A statistical manifold $(\mathcal{M}, \nabla, g)$ is said to be of *constant curvature k* if the curvature tensor R of ∇ satisfies

$$R(X, Y)Z = k\{g(Y, Z)X - g(X, Z)Y\}.$$

When the constant k is zero, the statistical manifold is said to be *flat*, or *dually flat*, because the curvature tensor R^* of ∇^* also vanishes automatically [2].

For $\alpha \in \mathbf{R}$, statistical manifolds $(\mathcal{M}, \nabla, g)$ and $(\mathcal{M}, \nabla', g')$ are said to be α-*conformally equivalent* if there exists a function ϕ on $\mathcal{M}$ such that

$$\begin{aligned} g'(X, Y) &= e^{\phi} g(X, Y) \\ g(\nabla'_X Y, Z) &= g(\nabla_X Y, Z) - \frac{1+\alpha}{2} d\phi(Z) g(X, Y) \\ &\quad + \frac{1-\alpha}{2}\{d\phi(X) g(Y, Z) + d\phi(Y) g(X, Z)\}. \end{aligned}$$

Statistical manifolds $(\mathcal{M}, \nabla, g)$ and $(\mathcal{M}, \nabla', g')$ are α-conformally equivalent if and only if $(\mathcal{M}, \nabla^*, g)$ and $(\mathcal{M}, \nabla'^*, g')$ are $-\alpha$-conformally equivalent [20]. A statistical manifold $(\mathcal{M}, \nabla, g)$ is called α-*conformally flat* if it is locally α-conformally equivalent to a flat statistical manifold.

Let $\{x^1, \ldots, x^n\}$ be an affine coordinate system of $\mathbf{R}^n$ with respect to certain basis vectors and ∇ be the canonical flat affine connection, i.e., $\nabla_{\partial/\partial x^i} \partial/\partial x^j = 0$. If the Hessian $g^{(\varphi)} = \sum(\partial^2\varphi/\partial x^i \partial x^j) dx^i dx^j$ of a function φ on a domain Ω in $\mathbf{R}^n$ is non-degenerate, we call $(\Omega, \nabla, g^{(\varphi)})$ a *Hessian domain*, which is a manifold with a pseudo-Riemannian metric $g^{(\varphi)}$ and an affine connection ∇. A Hessian domain $(\Omega, \nabla, g^{(\varphi)})$ is a flat statistical manifold. Conversely, a flat statistical manifold is locally a Hessian domain [2, 40].

We denote by $\langle y^*, x\rangle$ the pairing of x in $\mathbf{R}^n$ and y^* in the dual space denoted by $\mathbf{R}^*_n$, and define the *gradient mapping* $\mathrm{grad}\varphi$ at a point $p \in \Omega$ by $\langle \mathrm{grad}\varphi, X\rangle = d\varphi(X)$, identifying $X \in T_p\Omega$ with $X \in \mathbf{R}^n$. Since a Hessian domain $(\Omega, \nabla, g^{(\varphi)})$ possesses

non-degenerate Hessian $g^{(\varphi)}$, the gradient mapping at $p = (x^1, \cdots, x^n)$, expressed in the dual coordinate system $\{x_1^*, \cdots, x_n^*\}$

$$\mathrm{grad}\varphi \colon x \mapsto x^* = (x_1^*, \cdots, x_n^*), \quad x_i^* = \partial\varphi/\partial x^i$$

on local neighborhood, is invertible on its image. The *conjugate function* φ^* of x^* is locally defined by

$$\varphi^*(x^*) = \langle x^*, (\mathrm{grad'})^{-1}(x^*)\rangle - \varphi((\mathrm{grad'})^{-1}(x^*)).$$

We call the mapping from (x, φ) to (x^*, φ^*) the *Legendre transformation*. Both functions φ and φ^* are said *potentials*. We regard $\{x_1^*, \cdots, x_n^*\}$ is another coordinate system on Ω. For a Hessian domain $(\Omega, \nabla, g^{(\varphi)})$, the dual connection for ∇ denoted by ${}^*\nabla^{(\varphi)}$ is characterized by the flat connection with the affine coordinate system $\{x_1^*, \cdots, x_n^*\}$. Further, the locally defined Hessian $g^{(\varphi^*)} = \sum(\partial^2\varphi^*/\partial x_i^* \partial x_j^*) dx_i^* dx_j^*$ is equal to $g^{(\varphi)}$, i.e., $(\Omega, {}^*\nabla^{(\varphi)}, g^{(\varphi)} = g^{(\varphi^*)})$ is locally a Hessian domain [2].

When the conjugate function φ^* of φ can be globally defined on Ω, we define the *canonical divergence* $D^{(\varphi)}$ of $(\Omega, \nabla, g^{(\varphi)})$ by

$$D^{(\varphi)}(p, q) = \varphi(x(p)) + \varphi^*(x^*(q)) - \langle x^*(q), x(p)\rangle$$

for two points p and q on Ω, where $x(\cdot)$ and $x^*(\cdot)$ respectively represent x- and x^*-coordinates of a point $\cdot$. If $g^{(\varphi)}$ is positive definite and Ω is a convex domain, then

$$\text{(i) } D^{(\varphi)}(p, q) \geq 0 \quad \text{and} \quad \text{(ii) } D^{(\varphi)}(p, q) = 0 \Leftrightarrow p = q \tag{2.2}$$

hold because the unique maximum $\varphi^*(x^*(q))$ of $\sup_{p\in\Omega}\{\langle x^*(q), x(p)\rangle - \varphi(x(p))\}$ is attained when $x(p) = (\mathrm{grad}\varphi)^{-1}(x^*(q))$ holds.

Conversely, for a function $D \colon \mathcal{M} \times \mathcal{M} \to \mathbf{R}$ satisfying two conditions (2.2) for any points p and q in $\mathcal{M}$, define the positive semidefinite form $g^{(D)}$, and two affine connections $\nabla^{(D)}$ and ${}^*\nabla^{(D)}$ by

$$\begin{aligned} g^{(D)}(X, Y) &= -D(X|Y), \\ g^{(D)}(\nabla_X^{(D)} Y, Z) = -D(XY|Z), &\quad g^{(D)}({}^*\nabla_X^{(D)} Y, Z) = -D(Z|XY), \end{aligned} \tag{2.3}$$

for X, Y and $Z \in \mathcal{X}(\mathcal{M})$, where the notation in the right-hand sides stands for

$$D(X_1 \cdots X_n|Y_1 \cdots Y_m)(p) = (X_1)_p \cdots (X_n)_p (Y_1)_q \cdots (Y_m)_q D(p, q)|_{p=q}$$

for $X_i, Y_j \in \mathcal{X}(\mathcal{M})$. If $g^{(D)}$ is positive definite, then we call D a *divergence* or a *contrast function* on $\mathcal{M}$. The following result is used in the Sect. 2.6.

Proposition 1 ([2, 4]) *If D is a divergence on $\mathcal{M}$, then $(\mathcal{M}, \nabla^{(D)}, g^{(D)})$ is a statistical manifold with the dual connection ${}^*\nabla^{(D)}$.*

2.3 V-Potential Function and Riemannian Metric

Consider the vector space of $n \times n$ real symmetric matrices, denoted by $Sym(n, \mathbf{R})$, with an inner product $(X|Y) = \mathrm{tr}(XY)$ for $X, Y \in Sym(n, \mathbf{R})$. We identify an element Y^* in the dual space $Sym(n, \mathbf{R})^*$ with Y^* (denoted by the same symbol) in $Sym(n, \mathbf{R})$ via $\langle Y^*, X\rangle = (Y^*|X)$. Let $\{E_1, \ldots, E_{n(n+1)/2}\}$ be basis matrices of $Sym(n, \mathbf{R})$, and consider the affine coordinate system $\{x^1, \ldots, x^{n(n+1)/2}\}$ with respect to them and the canonical flat affine connection ∇. Denote by $PD(n, \mathbf{R})$ the convex cone of positive definite matrices in $Sym(n, \mathbf{R})$.

Let $\mathcal{X}(PD(n, \mathbf{R}))$ and $T_P PD(n, \mathbf{R})$ be the set of all tangent vector fields on $PD(n, \mathbf{R})$ and the tangent vector space at $P \in PD(n, \mathbf{R})$, respectively. By identifying E_i and the natural basis $(\partial/\partial x^i)_P$, we represent $X_P \in T_P PD(n, \mathbf{R})$ by $X \in Sym(n, \mathbf{R})$. Similarly we regard a $Sym(n, \mathbf{R})$-valued smooth function X on $PD(n, \mathbf{R})$ as $X \in \mathcal{X}(PD(n, \mathbf{R}))$ via the identification of constant function E_i with $\partial/\partial x^i$.

Let τ_G denote the congruent transformation by matrices G, i.e., $\tau_G X = GXG^T$. The differential of τ_G is denoted by τ_{G*}. If G is nonsingular, the transformation τ_G is an element of automorphism group that acts transitively on $PD(n, \mathbf{R})$.

In the Sects. 2.3 and 2.4, we consider the dually flat structure on $PD(n, \mathbf{R})$ as a Hessian domain induced from a certain class of potential functions.

Definition 1 Let $V(s)$ be a smooth function on positive real numbers $s \in \mathbf{R}_{++}$. The function defined by

$$\varphi^{(V)}(P) = V(\det P) \tag{2.4}$$

is called a *V-potential* on $PD(n, \mathbf{R})$.

Let $\nu_i(s)$, $i = 1, 2, 3$ be functions defined by

$$\nu_i(s) = \frac{d\nu_{i-1}(s)}{ds} s, \quad i = 1, 2, 3, \quad \text{where } \nu_0(s) = V(s).$$

We assume that $V(s)$ satisfies the following two conditions:

$$\text{(i) } \nu_1(s) < 0 \quad (s > 0), \quad \text{(ii) } \beta^{(V)}(s) = \frac{\nu_2(s)}{\nu_1(s)} < \frac{1}{n} \quad (s > 0), \tag{2.5}$$

which are later shown to ensure the convexity of $\varphi^{(V)}(P)$ on $PD(n, \mathbf{R})$. Note that the first condition $\nu_1(s) < 0$ for all $s > 0$ implies the function $V(s)$ is strictly decreasing on $s > 0$. Important examples of $V(s)$ satisfying (2.5) are $-\log s$ or $c_1 + c_2 s^\beta$ for real parameters c_1, c_2, β with $c_2\beta < 0$ and $\beta < 1/n$. Another example $V(s) = c\log(cs + 1) - \log s$ with $0 \le c < 1$ is proved useful for the quasi-Newton updates [14].

Using the formula $\mathrm{grad}\det P = (\det P)P^{-1}$, we have the gradient mapping $\mathrm{grad}\varphi^{(V)}$ and the differential form $d\varphi^{(V)}$, respectively, as

$$\mathrm{grad}\varphi^{(V)} : P \mapsto P^* = \nu_1(\det P)P^{-1}, \tag{2.6}$$

$$d\varphi^{(V)} : X \mapsto d\varphi^{(V)}(X) = \nu_1(\det P)\mathrm{tr}(P^{-1}X). \tag{2.7}$$

The Hessian of $\varphi^{(V)}$ at $P \in PD(n, \mathbf{R})$, which we denote by $g_P^{(V)}$, is calculated as

$$\begin{aligned} g_P^{(V)}(X, Y) &= d(d\varphi^{(V)}(X))(Y) \\ &= -\nu_1(\det P)\mathrm{tr}(P^{-1}XP^{-1}Y) + \nu_2(\det P)\mathrm{tr}(P^{-1}X)\mathrm{tr}(P^{-1}Y). \end{aligned}$$

Theorem 1 *The Hessian $g^{(V)}$ is positive definite on $PD(n, \mathbf{R})$ if and only if the conditions* (2.5) *hold.*

The proof can be found in the Appendix A.

Remark 1 The Hessian $g^{(V)}$ is $SL(n, \mathbf{R})$-invariant, i.e., $g_{P'}^{(V)}(X', Y') = g_P^{(V)}(X, Y)$ for any $G \in SL(n, \mathbf{R})$, where $P' = \tau_G P$, $X' = \tau_{G*}X$ and $Y' = \tau_{G*}Y$.

The conjugate function of $\varphi^{(V)}$ denoted by $\varphi^{(V)*}$ is

$$\varphi^{(V)*}(P^*) = \sup_P \{\langle P^*, P\rangle - \varphi^{(V)}(P)\}. \tag{2.8}$$

Since the extremal condition is

$$P^* = \mathrm{grad}\varphi^{(V)}(P) = \nu_1(\det P)P^{-1}$$

and $\mathrm{grad}\varphi^{(V)}$ is invertible by the positive definiteness of $g^{(V)}$, we have the following expression for $\varphi^{(V)*}$ with respect to P:

$$\varphi^{(V)*}(P^*) = n\nu_1(\det P) - \varphi^{(V)}(P). \tag{2.9}$$

Hence the canonical divergence $D^{(V)}$ of $(PD(n, \mathbf{R}), \nabla, g^{(V)})$ is obtained as

$$\begin{aligned} D^{(V)}(P, Q) &= \varphi^{(V)}(P) + \varphi^{(V)*}(Q^*) - \langle Q^*, P\rangle \\ &= V(\det P) - V(\det Q) + \langle Q^*, Q - P\rangle. \end{aligned} \tag{2.10}$$

2.4 Dual Affine Connection Induced from V-Potential

Let ∇ be the canonical flat affine connection on $Sym(n, \mathbf{R})$. To discuss dually flat structure on $PD(n, \mathbf{R})$ by regarding $g^{(V)}$ as positive definite Riemannian metric, we derive the dual connection ${}^*\nabla^{(V)}$ with respect to $g^{(V)}$ introduced in the Sect. 2.2.

Consider a smooth curve $\gamma = \{P_t \mid -\epsilon < t < \epsilon\}$ in $PD(n, \mathbf{R})$ satisfying

$$(P_t)_{t=0} = P \in PD(n, \mathbf{R}), \qquad \left(\frac{dP_t}{dt}\right)_{t=0} = X \in T_P PD(n, \mathbf{R}).$$

Lemma 1 *The differential of the gradient mapping* $\mathrm{grad}\varphi^{(V)}$ *is*

$$(\mathrm{grad}\varphi^{(V)})_*: X \mapsto \nu_2(\det P)\mathrm{tr}(P^{-1}X)P^{-1} - \nu_1(\det P)P^{-1}XP^{-1}.$$

Proof Differentiate the Legendre transform $P_t^* = \mathrm{grad}\varphi^{(V)}(P_t)$ along the curve γ, then

$$\begin{aligned}\frac{dP_t^*}{dt} &= \frac{d\nu_1(\det P_t)}{dt}P_t^{-1} - \nu_1(\det P_t)P_t^{-1}\frac{dP_t}{dt}P_t^{-1}\\ &= \nu_1'(\det P_t)\frac{d\det P_t}{dt}P_t^{-1} - \nu_1(\det P_t)P_t^{-1}\frac{dP_t}{dt}P_t^{-1}\\ &= \nu_1'(\det P_t)\det P_t\mathrm{tr}\left(P_t^{-1}\frac{dP_t}{dt}\right)P_t^{-1} - \nu_1(\det P_t)P_t^{-1}\frac{dP_t}{dt}P_t^{-1},\end{aligned}$$

where $\nu_1' = d\nu_1/ds$. By evaluating both sides at $t = 0$, we see the statement holds. □

Theorem 2 *Let* π_t *denote the parallel shift operator of the connection* ${}^*\nabla^{(V)}$. *Then the parallel shift* $\pi_t(Y)$ *of the tangent vector* $Y(= \pi_0(Y)) \in T_PPD(n, \mathbf{R})$ *along the curve* γ *satisfies*

$$\left(\frac{d\pi_t(Y)}{dt}\right)_{t=0} = XP^{-1}Y + YP^{-1}X + \Phi(X, Y, P) + \Phi^{\perp}(X, Y, P),$$

where

$$\Phi(X, Y, P) = \frac{\nu_2(s)tr(P^{-1}X)}{\nu_1(s)}Y + \frac{\nu_2(s)tr(P^{-1}Y)}{\nu_1(s)}X, \tag{2.11}$$

$$\Phi^{\perp}(X, Y, P) = \rho P, \tag{2.12}$$

$$\rho = \frac{\{\nu_3(s)\nu_1(s) - 2\nu_2^2(s)\}tr(P^{-1}X)tr(P^{-1}Y) + \nu_2(s)\nu_1(s)\mathrm{tr}(P^{-1}XP^{-1}Y)}{\nu_1(s)\{\nu_1(s) - n\nu_2(s)\}}$$

and $s = \det P$.

The proof can be found in the Appendix B.

Remark 2 The denominator $\nu_1(s)\{\nu_1(s) - n\nu_2(s)\}$ in ρ is always positive because (2.5) is assumed.

Corollary 1 *Let* E_i *be the matrix representation of the natural basis vector fields* $\partial/\partial x^i$ *described in the beginning of the* Sect. 2.3. *Then, their covariant derivatives at* P *defined by the dual connection* ${}^*\nabla^{(V)}$ *have the following matrix representations:*

$$\left({}^*\nabla^{(V)}_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}\right)_P = -E_iP^{-1}E_j - E_jP^{-1}E_i - \Phi(E_i, E_j, P) - \Phi^{\perp}(E_i, E_j, P),$$

Proof Recall the identification $(\partial/\partial x^i)_P = E_i$ for any $P \in PD(n, \mathbf{R})$. Then the statement follows from Theorem 2 and the definition of the covariant derivatives, i.e.,

$$\left({}^*\nabla^{(V)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}\right)_P = \left(\frac{d\pi_{-t}((\partial/\partial x^j)_{P_t})}{dt}\right)_{t=0}, \quad P_t \in PD(n, \mathbf{R}),\ P = P_0.$$

This completes the proof. □

Remark 3 From Corollary 1 we observe that both of the connections ∇ and ${}^*\nabla^{(V)}$ are generally $SL(n, \mathbf{R})$-invariant, i.e.,

$$\tau_{G*}(\nabla_X Y)_P = \left(\nabla_{X'} Y'\right)_{P'}, \quad \tau_{G*}\left({}^*\nabla^{(V)}_X Y\right)_P = \left({}^*\nabla^{(V)}_{X'} Y'\right)_{P'}$$

holds for any $G \in SL(n, \mathbf{R})$, where $P' = \tau_G P$, $X' = \tau_{G*} X$ and $Y' = \tau_{G*} Y$. Particularly, both of the connections induced from the *power potential* $\varphi^{(V)}$, defined via $V(s) = c_1 + c_2 s^\beta$ with real constants c_1, c_2 and β, are $GL(n, \mathbf{R})$-invariant. In addition, so is the orthogonality with respect to $g^{(V)}$. Hence, we conclude that both ∇-*and* ${}^*\nabla^{(V)}$-*projections* [1, 2] are $GL(n, \mathbf{R})$-invariant for the power potentials, while so is not $g^{(V)}$.

The power potential function $\varphi^{(V)}$ with normalizing conditions $V(1) = 0$ and $\nu_1(1) = -1$, i.e., $V(s) = (1 - s^\beta)/\beta$ is called the *beta potential*. In this case, $\nu_1(s) = -s^\beta$, $\nu_2(s) = -\beta s^\beta$, $\nu_3(s) = -\beta^2 s^\beta$ and $\beta^{(V)}(s) = \beta$. Note that setting β to zero leads to $V(s) = -\log s$, which recovers the standard dualistic geometry induced by the logarithmic characteristic function $\varphi^{(-\log)}(P) = -\log\det P$ on $PD(n, \mathbf{R})$ [34]. See [7, 33] for detailed discussion related with the power potential function.

2.5 Foliated Structures and Their Geometry

When $V(s) = -k \log s$ for a positive constant k, we know the geometry induced by the corresponding V-potential function is standard in several senses [2, 34]. In particular it is invariant under the automorphism group action of τ_G where $G \in GL(n, \mathbf{R})$. On the other hand, for general $V(s)$ satisfying (2.5), we have observed in the previous sections that the dualistic structure $(g^{(V)}, \nabla, {}^*\nabla^{(V)})$ on $PD(n, \mathbf{R})$ is invariant under the action of τ_G only when $G \in SL(n, \mathbf{R})$.

In this section we are devoted to investigate the $SL(n, \mathbf{R})$-invariant geometry on $PD(n, \mathbf{R})$ induced by V-potentials.

2.5.1 Foliated Structures

Define foliated structure of $PD(n, \mathbf{R})$ specified by a real s as follows (Fig. 2.1):

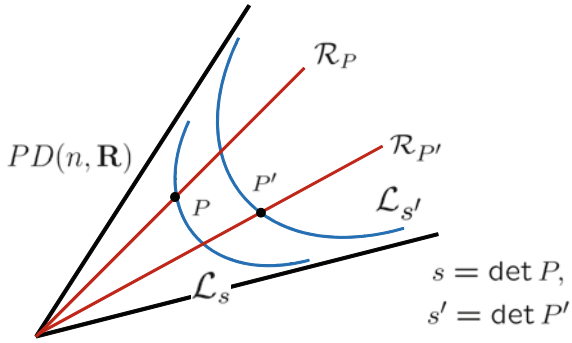


Fig. 2.1 Foliated structures of $PD(n, \mathbf{R})$

$$PD(n,\mathbf{R}) = \bigcup_{s>0} \mathcal{L}_s, \quad \mathcal{L}_s = \{P | P > 0, \det P = s\}.$$

The tangent vector space of $\mathcal{L}_s$ at P is characterized as

$$T_P\mathcal{L}_s = \{X | \mathrm{tr}(P^{-1}X) = 0,\ X = X^T\} = \{P^{1/2}YP^{1/2} | \mathrm{tr}Y = 0,\ Y = Y^T\}.$$

Denote by $\mathcal{R}_P$ the ray through P in the convex cone $PD(n, \mathbf{R})$, i.e, $\mathcal{R}_P = \{Q | Q = \lambda P, 0 < \lambda \in \mathbf{R}\}$. Another foliated structure we consider is

$$PD(n,\mathbf{R}) = \bigcup_{P\in\mathcal{L}_s} \mathcal{R}_P.$$

Proposition 2 *Let P be in $\mathcal{L}_s$. For any functions V satisfying (2.5), each leaf $\mathcal{L}_s$ is orthogonal to the ray $\mathcal{R}_P$ at P with respect to $g^{(V)}$.*

Proof The tangent vectors of $\mathcal{R}_P$ are just P multiplied by scalars. Hence for $X \in T_P\mathcal{L}_s$ we have

$$g_P^{(V)}(X, P) = -\nu_1(s)\mathrm{tr}(P^{-1}X) + n\nu_2(s)\mathrm{tr}(P^{-1}X) = 0.$$

This completes the proof. □

Proposition 3 *Every ray $\mathcal{R}_P$ is simultaneously a ∇- and ${}^*\nabla^{(V)}$-geodesic for an arbitrary function V satisfying (2.5).*

Proof The statement follows from that $\mathcal{R}_P$ and its Legendre transform $\{Q^* | Q^* = \lambda^* P^*, 0 < \lambda^* \in \mathbf{R}\}$ are, respectively, straight lines with respect to the affine coordinate systems of ∇ and ${}^*\nabla^{(V)}$. □

We know the mean of dual connections is metric and torsion-free [2], i.e.,

$$\hat{\nabla}^{(V)} = \frac{1}{2}\left(\nabla + {}^*\nabla^{(V)}\right) \tag{2.13}$$

is the Levi-Civita connection for $g^{(V)}$. Proposition 3 implies that every $\mathcal{R}_P$ is always $\hat{\nabla}^{(V)}$-geodesic for any V satisfying (2.5). On the other hand, we have the following:

Proposition 4 *Each leaf $\mathcal{L}_s$ is autoparallel with respect to the Levi-Civita connection $\hat{\nabla}^{(V)}$ if and only if $\nu_2(s) = 0$, i.e, $V(s) = -k \log s$ for a positive constant k.*

The proof can be found in the Appendix C.

2.5.2 Geometry of Submanifolds

Next, we further investigate the geometry of the submanifolds $\mathcal{L}_s$. Let $(\tilde{g}^{(V)}, \tilde{\nabla}, {}^*\tilde{\nabla}^{(V)})$ be the dualistic structure on $\mathcal{L}_s$ induced from $(g^{(V)}, \nabla, {}^*\nabla^{(V)})$ on $PD(n, \mathbf{R})$.

Since each transformation τ_G where $G \in SL(n, \mathbf{R})$ is an isometry of $\tilde{g}^{(V)}$ and an element of automorphism group acting on $\mathcal{L}_s$ transitively, the Riemannian manifold $(\mathcal{L}_s, \tilde{g}^{(V)})$ is a homogeneous space. The isotropy group at λI consists of τ_G where $G \in SO(n)$. Hence $\mathcal{L}_s$ can be identified with $SL(n, \mathbf{R})/SO(n)$. We have a slightly extended result of Proposition 6 in [38], which treated the case of the standard potential $V(s) = -\log s$:

Proposition 5 *Each manifold $(\mathcal{L}_s, \tilde{g}^{(V)})$ is a globally Riemannian symmetric space for any V satisfying (2.5).*

Proof Define the mapping $\iota_s : \mathcal{L}_s \to \mathcal{L}_s$ by

$$\iota_s P = -\left(\frac{s}{s^*}\right)^{\frac{1}{n}} \mathrm{grad}\varphi^{(V)}(P) = s^{\frac{2}{n}} P^{-1}, \quad s^* = \frac{(-\nu_1(s))^n}{s}.$$

It is seen that ι_s is the involutive isometry of $(\mathcal{L}_s, \tilde{g}^{(V)})$ with the isolated fixed point $\lambda I \in \mathcal{L}_s$, where $\lambda = s^{1/n}$. In particular, $-\mathrm{grad}\varphi^{(V)}(P)$ is the involutive isometry of $(\mathcal{L}_s, \tilde{g}^{(V)})$ satisfying $s = s^*$. Since isometries τ_G where $G \in SL(n, \mathbf{R})$ act transitively on $\mathcal{L}_s$, each manifold $(\mathcal{L}_s, \tilde{g}^{(V)})$ is globally Riemannian symmetric [12]. □

Proposition 6 *For any $V(s)$ satisfying (2.5), the following statements hold:*

(i) The metric $\tilde{g}^{(V)}$ on $\mathcal{L}_s$ is given by $\tilde{g}^{(V)} = -\nu_1(s)\tilde{g}^{(-log)}$,
(ii) For $P, Q \in \mathcal{L}_s$, it holds that $D^{(V)}(P, Q) = -\nu_1(s)D^{(-log)}(P, Q)$,
(iii) The induced connections ${}^*\tilde{\nabla}^{(V)}$ on $\mathcal{L}_s$ are actually independent of $V(s)$, i.e., ${}^*\tilde{\nabla}^{(V)} = {}^*\tilde{\nabla}^{(-log)}$ for all $V(s)$ (See the remark below).

The proof can be found in the Appendix D.

Remark 4 Since $\tilde{\nabla}$ is common, Proposition 6 demonstrates that the V-potential does not deform the submanifold geometry of $\mathcal{L}_s$ except a multiplicative constant for $\tilde{g}^{(V)}$. Hence, we simply write $\tilde{\nabla}^*$ instead of ${}^*\tilde{\nabla}^{(-\log)}$ from now on.

2.5.3 Curvature of Each Leaf and Decomposition of Divergence

In [11, 40, 43, 44] level surfaces of the potential function in a Hessian domain were studied via the affine hypersurface theory [29]. Since the statistical submanifold $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ is a level surface of $\varphi^{(V)}$ in a Hessian domain $(PD(n, \mathbf{R}), \nabla, g^{(V)})$, the general results in the literature can be applied. The immediate consequences along our context are summarized as follows:

Consider the gradient vector field E of $\varphi^{(V)}$ and a scaled gradient vector field N respectively defined by

$$g^{(V)}(X, E) = d\varphi^{(V)}(X), \quad \forall X \in \mathcal{X}(PD(n, \mathbf{R}))$$

and

$$N = -\frac{1}{d\varphi^{(V)}(E)} E.$$

Straightforward calculation shows they are respectively represented at each $P \in PD(n, \mathbf{R})$ by

$$E = \frac{\nu_1(\det P)}{n\nu_2(\det P) - \nu_1(\det P)} P, \quad N = -\frac{1}{n\nu_1(\det P)} P. \tag{2.14}$$

Using N as the transversal vector field, the results in [11, 43] demonstrate that the geometry $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ coincides with the one realized by the *affine immersion* [29] of $\mathcal{L}_s$ into $Sym(n, \mathbf{R})$ with the canonical flat connection ∇:

$$\nabla_X Y = \tilde{\nabla}_X Y + h(X, Y)N, \tag{2.15}$$

$$\nabla_X N = -A(X) + \tau(X)N. \tag{2.16}$$

Here h and τ are, respectively, the *affine fundamental form* and the *transversal connection form* satisfying the relation:

$$h = \tilde{g}^{(V)}, \quad \tau = 0, \tag{2.17}$$

for all $X, Y \in \mathcal{X}(\mathcal{L}_s)$ at each $P \in \mathcal{L}_s$. The *affine shape operator* A is determined in the proof of Theorem 3. The condition (2.17) shows this affine immersion is *non-degenerate* and *equiaffine*. In addition the form of N in (2.14) demonstrates that the immersion is *centro-affine* [29], which implies a $\tilde{\nabla}$-geodesic on $\mathcal{L}_s$ is the intersection of $\mathcal{L}_s$ and a two-dimensional subspace in $Sym(n, \mathbf{R})$.

Further, via Proposition 2 and 3, the following results are proved for the level surface $\mathcal{L}_s$ of $\varphi^{(V)}$ (Proposition 8 can be also verified by direct calculations):

Proposition 7 ([43]) *The statistical submanifold* $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ *is 1-conformally flat.*

Proposition 8 ([44]) *Let points P and Q be in* $\mathcal{L}_s$*, and R be in* $\mathcal{R}_Q$*, s.t.* $R = \lambda Q, 0 < \lambda \in \mathbf{R}$. *Then*

$$D^{(V)}(P,R) = \mu D^{(V)}(P,Q) + D^{(V)}(Q,R), \tag{2.18}$$

holds for $\mu \in \mathbf{R}$ *satisfying* $R^* = \mu Q^*$, *i.e.,* $\mu = \lambda^{-1}\nu_1(\det R)/\nu_1(\det Q) > 0$.

Using these preliminaries, we have the main result in this section as follows:

Theorem 3 *The following two statements hold:*

(i) The statistical submanifold $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ *is* ± 1*-conformally flat.*
(ii) The statistical submanifold $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ *has negative constant curvature* $k_s = 1/(\nu_1(s)n)$.

Proof (i) Note that $\mathcal{L}_s$ is a level surface of not only $\varphi^{(V)}(P)$ but also $\varphi^{(V)*}(P^*)$ in (2.9), which implies both $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ and $(\mathcal{L}_s, \tilde{\nabla}^*, \tilde{g}^{(V)})$ are 1-conformally flat from Proposition 7. By the duality of α-conformal equivalence, they are also -1-conformally flat.

(ii) For arbitrary $X \in \mathcal{X}(\mathcal{L}_s)$, the equalities

$$\nabla_X N = -\frac{1}{\nu_1(s)n}\nabla_X P = -\frac{1}{\nu_1(s)n}X$$

hold at each $P \in \mathcal{L}_s$. The second equality follows from that ∇ is a canonical flat connection of the vector space $Sym(n, \mathbf{R})$. Comparing this equation with (2.16) and (2.17), we have $A = k_s I$. By substituting into the Gauss equation for the affine immersion [29]:

$$R(X,Y)Z = h(Y,Z)A(X) - h(X,Z)A(Y),$$

we see that $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ is of constant curvature k_s. □

Remark 5 For a general statistical manifold $\mathcal{M}$ of $\dim\mathcal{M} \geq 3$, the equivalence of the statements (i) and (ii) in the above theorem is known [2, 19]. The theorem claims that the equivalence holds even when $n = 2$, i.e, $\dim\mathcal{L}_s = 2$

Kurose proved that a modified form of the Pythagorean relation holds for canonical divergences on statistical manifolds with constant curvature [20]. As a consequence of (ii) in Theorem 3, his result holds on each $\mathcal{L}_s$.

Proposition 9 ([20]) *Suppose that points* P, Q *and* R *be in a statistical submanifold* $\mathcal{L}_s$ *with constant curvature* k_s. *Let* $\tilde{\gamma}$ *be the* $\tilde{\nabla}$*-geodesic joining* P *and* Q, *and* $\tilde{\gamma}^*$ *the* $\tilde{\nabla}^*$*-geodesic joining* Q *and* R. *If* $\tilde{\gamma}$ *and* $\tilde{\gamma}^*$ *are mutually orthogonal at* Q *with respect to* $\tilde{g}^{(V)}$, *the divergence* $D^{(V)}$ *satisfies*

$$D^{(V)}(P,R) = D^{(V)}(P,Q) + D^{(V)}(Q,R) - k_s D^{(V)}(P,Q)D^{(V)}(Q,R). \tag{2.19}$$

The -1-conformal flatness of $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ in Theorem 3 ensures another relation that is dual to Proposition 8:

Lemma 2 *Suppose that three points P, Q and R and the parameter λ meet the same assumptions in Proposition 8. Then, the following relation holds:*

$$D^{(V)}(R,P) = D^{(V)}(R,Q) + \lambda D^{(V)}(Q,P). \tag{2.20}$$

Proof By direct calculation the right-hand side of (2.20) can be modified as

$$\begin{aligned}
&V(\det R) - V(\det Q) + \langle Q^*, Q - R\rangle + \lambda\langle P^*, P - Q\rangle \\
&= V(\det R) - V(\det P) + \nu_1(s)n - \langle Q^*, \lambda Q\rangle + \lambda\nu_1(s)n - \lambda\langle P^*, Q\rangle \\
&= V(\det R) - V(\det P) + \nu_1(s)n - \langle P^*, R\rangle = D^{(V)}(R,P).
\end{aligned}$$

□

Remark 6 By duality the -1-conformal flatness of $(\mathcal{L}_s, \tilde{\nabla}, \tilde{g}^{(V)})$ implies 1-conformal flatness of $(\mathcal{L}_s, \tilde{\nabla}^*, \tilde{g}^{(V)})$. Hence, introducing *dual divergence* [2] defined by

$${}^*D^{(V)}(P,Q) = D^{(V)}(Q,P) = \varphi^{(V)*}(P^*) + \varphi^{(V)}(Q) - \langle P^*, Q\rangle,$$

we see that the following dual result to Proposition 8 with respect to $(\mathcal{L}_s, \tilde{\nabla}^*, \tilde{g}^{(V)})$:

$${}^*D^{(V)}(P,R) = {}^*D^{(V)}(Q,R) + \lambda\, {}^*D^{(V)}(P,Q)$$

alternatively proves the above lemma.

Now we have the following new decomposing relation of the divergence for $PD(n,\mathbf{R})$ and its submanifold $\mathcal{L}_s$ (Fig. 2.2), which is an immediate consequence of Proposition 9 and Lemma 2:

Proposition 10 *Suppose P be in $PD(n,\mathbf{R})$, and R and S be on $\mathcal{L}_s$. Let Q be the uniquely defined point by P as $Q \in \mathcal{L}_s \cap \mathcal{R}_P$. If the $\tilde{\nabla}$-geodesic $\tilde{\gamma}$ joining Q and R and the $\tilde{\nabla}^*$-geodesic $\tilde{\gamma}^*$ joining R and S are mutually orthogonal at R with respect to $\tilde{g}^{(V)}$, then the divergence $D^{(V)}$ satisfies*

$$D^{(V)}(P,S) = D^{(V)}(P,R) + \kappa D^{(V)}(R,S), \quad \kappa = \lambda\{1 - k_s D^{(V)}(Q,R)\}, \tag{2.21}$$

where the constant $\lambda > 0$ is defined by $Q = \lambda P$.

Proof From Lemma 2 and Proposition 9, we have the following relations:

$$\begin{aligned}
D^{(V)}(P,S) &= D^{(V)}(P,Q) + \lambda D^{(V)}(Q,S), \\
D^{(V)}(P,R) &= D^{(V)}(P,Q) + \lambda D^{(V)}(Q,R), \\
D^{(V)}(Q,S) &= D^{(V)}(Q,R) + D^{(V)}(R,S) - k_s D^{(V)}(Q,R) D^{(V)}(R,S).
\end{aligned}$$

Using these equations, we see the statement holds. □

Owing to the negativity of the curvature k_s, we have the following application of the Proposition 10 on the minimality condition of $D^{(V)}(\bullet,\bullet)$ measured from a point $P \in PD(n,\mathbf{R})$ to a certain submanifold in $\mathcal{L}_s$.

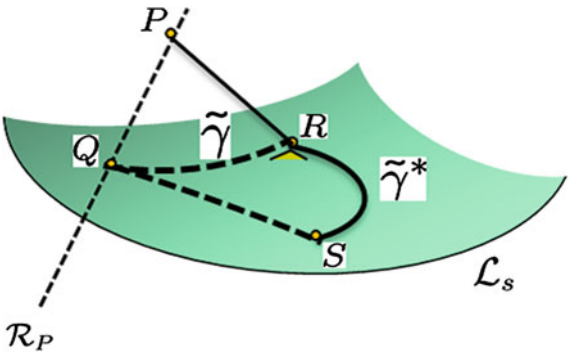


Fig. 2.2 Decomposition of divergences when R and S are on a leaf $\mathcal{L}_s$ and $\tilde{\gamma}$ and $\tilde{\gamma}^*$ are orthogonal at R

Corollary 2 *Let P be a point in $PD(n, \mathbf{R})$ and let $\mathcal{N}$ be a $\tilde{\nabla}^*$-autoparallel submanifold in $\mathcal{L}_s$. A point R in $\mathcal{N}$ satisfies $D^{(V)}(P, R) = \min_{S\in\mathcal{N}} D^{(V)}(P, S)$ if and only if the $\tilde{\nabla}$-geodesic connecting $Q = \lambda P \in \mathcal{R}_P \cap \mathcal{L}_s$ and R is orthogonal to $\mathcal{N}$ at R.*

Proof For arbitrary point $S \in \mathcal{N}$ it follows from Proposition 10

$$D^{(V)}(P, S) = D^{(V)}(P, R) + \lambda\{1 - k_s D^{(V)}(Q, R)\}D^{(V)}(R, S), \tag{2.22}$$

where Q and λ are uniquely defined by $Q \in \mathcal{R}_P \cap \mathcal{L}_s$ and $Q = \lambda P$.

Since $1-k_s D^{(V)}(Q, R)$ and λ are positive constants respectively, $D^{(V)}(P, S)$ takes the minimum value $D^{(V)}(P, R)$ if and only if $D^{(V)}(R, S) = 0$, i.e., $S = R$. □

Dual results for ${}^*D^{(V)}$ to the above ones also hold.

2.6 Dualistic Geometries on U-Model and Positive Definite Matrices

We explore a close relation between the dualistic geometries induced from U-divergence and V-potential. In the field of statistical inference, the well-established method is the maximum likelihood method, which is based on the Kullback-Leibler divergence.

To improve robustness performance of the method maintaining its theoretical advantages, such as efficiency, the methods of minimizing general divergences have been recently proposed as alternatives to the maximum likelihood method, in robust statistical analysis for pattern recognition, machine learning, principal component analysis and so on [4, 6, 13, 22, 24, 41].

For example the *beta-divergence*

$$D_\beta(f, g) = \int \frac{g(x)^{\beta+1} - f(x)^{\beta+1}}{\beta + 1} - \frac{f(x)\{g(x)^\beta - f(x)^\beta\}}{\beta} dx \tag{2.23}$$

is utilized in the literature. As β goes to 0, it reduces to the Kullback-Leibler divergence; On the other hand, as β goes to 1, it reduces to the squared L^2-distance. Thus the efficiency increases as β goes to 0, while the robustness increases as β goes to 1 [5, 39]. In this sense we could find an appropriate β between 0 and 1 as a trade-off between efficiency and robustness. The beta-divergence is strongly connected to the Tsallis entropy [42].

Let us make more general discussion on divergence functionals.

Definition 2 [4] Let $U(s)$ be a smooth convex function with the positive derivatives $u(s) = U'(s)$ and $u'(s)$ on $\mathbf{R}$ or its (semi-infinite) interval and ξ be the inverse function of u there. If the following functional for two functions $f(x)$ and $g(x)$ on $\mathbf{R}^n$

$$D_U(f, g) = \int U(\xi(g)) - U(\xi(f)) - \big[\xi(g) - \xi(f)\big] f dx$$

exists and converges, we call it the *U-divergence*.

It follows that $D_U(f, g) \geq 0$ and $D_U(f, g) = 0$ if and only if $f = g$ because the integrand $U(\xi_g) - [U(\xi_f) + u(\xi_f)(\xi_g - \xi_f)]$, where $\xi_f = \xi(f)$ and $\xi_g = \xi(g)$, is interpreted as the difference of the convex function U and its supporting function. While our U-divergence can be regarded as a dual expression [35] of the ordinary Bregman divergence, the expression is proved convenient in statistical inference from empirical data rather than the ordinary one [4, 24]. If we set $U(s) = \frac{1}{\beta+1}(1 + \beta s)^{(\beta+1)/\beta}, s > -1/\beta$, then the corresponding U-divergence is the beta-divergence defined in (2.23).

When we consider the family of functions parametrized by elements in a manifold $\mathcal{M}$, the U-divergence induces the dualistic structure on $\mathcal{M}$ in such a way as Proposition 1. Here, we confine our attention to the family of multivariate probability density functions specified by P in $PD(n, \mathbf{R})$. The family is natural in the sense that it is a dually flat statistical manifold with respect to the dualistic geometry induced by the U-divergence.

Definition 3 [4] Let U and u be the functions given in Definition 2. The family of elliptical distributions with the following density functions

$$\mathcal{M}_U = \left\{ f(x, P) = u\left(-\frac{1}{2} x^T P x - c_U(\det P)\right) \middle| P \in PD(n, \mathbf{R}) \right\},$$

is called the (zero-mean) *U-model* associated with the U-divergence. Here, we set $f(x, P) = 0$ if the right-hand side of f is undefined, and $c_U(\det P)$ is a normalizing constant satisfying

$$\int f(x, P) dx = (\det P)^{-\frac{1}{2}} \int u\left(-\frac{1}{2} y^T y - c_U(\det P)\right) dy = 1,$$

where we assume that the integral converges for the specified P.

For the case when the support of f is bounded, a straightforward computation shows that

$$\int u\left(-\frac{1}{2}y^T y - c_U(\det P)\right) dy = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \int_0^{\rho} u\left(-\frac{r}{2} - c_U(\det P)\right) r^{\frac{n}{2}-1} dr$$

where ρ satisfies $u(-\rho/2 - c_U(\det P)) = 0$. Hence, the normalizing constant is given by

$$c_U(\det P) = \Gamma_{\frac{n}{2},u}^{-1}\left(\frac{\Gamma(\frac{n}{2})(\det P)^{\frac{1}{2}}}{\pi^{\frac{n}{2}}}\right),$$

with the inverse function of $\Gamma_{a,u}$, which is defined by

$$\Gamma_{a,u}(c) = \int_0^{-2c-2\xi(0)} u\left(-\frac{r}{2} - c\right) r^{a-1} dr.$$

A similar argument is also valid for the case of unbounded supports. See for examples of the calculation in [33].

Note that if the function $U(s)$ satisfies a sort of self-similarity [33], the density function f in the U-model can be expressed in the usual form of an elliptical distribution [8, 23], i.e.,

$$f(x, P) = c_f (\det P)^{\frac{1}{2}} u\left(-\frac{1}{2}x^T P x\right)$$

with a constant c_f.

Now we consider the correspondence between the dualistic geometry induced by D_U on the U-model and that on $PD(n, \mathbf{R})$ induced by the V-potential function discussed in the Sects. 2.3 and 2.4.

Theorem 4 *Define the V-potential function $\varphi^{(V)}$ via*

$$V(s) = s^{-\frac{1}{2}} \int U\left(-\frac{1}{2}x^T x - c_U(s)\right) dx + c_U(s), \quad s > 0. \tag{2.24}$$

Assume that V satisfies the conditions (2.5), then the dualistic structure $(g^{(V)}, \nabla, {}^*\nabla^{(V)})$ on $PD(n, \mathbf{R})$ coincides with that on the U-model induced by the U-divergence in such a way as Proposition 1.

The proof can be found in the Appendix E.

From the above theorem it follows that the geometry of U-model derived from U-divergence is completely characterized by the geometry associated with V-potential function defined via (2.24).

2.7 Conclusion

We have studied dualistic geometry on positive definite matrices induced from the V-potential instead of the standard characteristic function.

First, we have derived the associated Riemannian metric, mutually dual affine connections and the canonical divergence. The induced geometry is, in general, proved $SL(n, \mathbf{R})$-invariant while it is $GL(n, \mathbf{R})$-invariant in the case of the characteristic function ($V(s) = -\log s$). However, when V is of the power form, it is shown that orthogonality and a pair of mutually dual connections are $GL(n, \mathbf{R})$-invariant.

Next, we have investigated a foliated structure via the induced geometry. Each leaf (the set of positive definite matrices with a constant determinant) is proved to be a homogeneous statistical manifold with a constant curvature depending on V. As a consequence, we have given a new decomposition relation for the canonical divergence that would be useful to solve the nearest point on a specified leaf.

Finally, to apply the induced geometry to robust statistical inferences we have established a relation with geometry of U-model (or symmetric elliptical densities).

Further applications of such structures and investigation of the other form of potential functions are left in the future work.

Acknowledgments We thank the anonymous referees for their constructive comments and careful checks of the original manuscript.

Appendices

A Proof of Theorem 1

It is observed that $-\nu_1(\det P) \neq 0$ on $PD(n, \mathbf{R})$ is necessary because the second term is not positive definite. Hence, the Hessian can be represented as

$$\begin{aligned} g_P^{(V)}(X, Y) &= -\nu_1(\det P)\{\mathrm{tr}(P^{-1}XP^{-1}Y) - \beta^{(V)}(\det P)\mathrm{tr}(P^{-1}X)\mathrm{tr}(P^{-1}Y)\} \\ &= -\nu_1(\det P)\mathrm{vec}^T(\tilde{X})\left(I_{n^2} - \beta^{(V)}(\det P)\mathrm{vec}(I_n)\mathrm{vec}^T(I_n)\right)\mathrm{vec}(\tilde{Y}). \end{aligned}$$

Here $\tilde{X} = P^{-1/2}XP^{-1/2}$, $\tilde{Y} = P^{-1/2}YP^{-1/2}$, $\mathrm{vec}(\bullet)$ is the operator that maps $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ to $[a_{11}, \cdots, a_{n1}, a_{12}, \cdots, a_{n2}, \cdots, a_{1n}, \cdots, a_{nn}]^T \in \mathbf{R}^{n^2}$, and I_n and I_{n^2} denote the unit matrices of order n and n^2, respectively. By congruently transforming the matrix $I_{n^2} - \beta^{(V)}(\det P)\mathrm{vec}(I_n)\mathrm{vec}^T(I_n)$ with a proper permutation matrix, we see the positive definiteness of $g^{(V)}$ is equivalent with $-\nu_1(\det P) > 0$ and

$$I_n - \beta^{(V)}(\det P)\mathbf{1}\mathbf{1}^T > 0, \quad \text{where } \mathbf{1} = [1, 1, \cdots, 1]^T \in \mathbf{R}^n.$$

Let W be an orthogonal matrix that has $\mathbf{1}/\sqrt{n}$ as the first column vector. Since the following eigen-decomposition

$$I_n - \beta^{(V)}(\det P)\mathbf{1}\mathbf{1}^T = W \begin{pmatrix} 1 - n\beta^{(V)}(\det P) & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} W^T$$

holds, the conditions (2.5) are necessary and sufficient for positive definiteness of $g^{(V)}$. Thus, the statement follows. □

B Proof of Theorem 2

Since the components of P^* is an affine coordinate for the connection ${}^*\nabla^{(V)}$, the parallel shift $\pi_t(Y)$ along the curve γ satisfies

$$(\mathrm{grad}\varphi^{(V)})_*(Y) = (\mathrm{grad}\varphi^{(V)})_*(\pi_t(Y))$$

for any t.

From Lemma 1, this implies

$$\frac{d}{dt}\left[\nu_2(\det P_t)\mathrm{tr}\{P_t^{-1}\pi_t(Y)\}P_t^{-1} - \nu_1(\det P_t)P_t^{-1}\pi_t(Y)P_t^{-1}\right] = 0$$

for any t $(-\epsilon < t < \epsilon)$.

By calculating the left-hand side, we get

$$\begin{aligned}
&\nu_3(s_t)\mathrm{tr}\left(P_t^{-1}\frac{dP_t}{dt}\right)\mathrm{tr}(P_t^{-1}\pi_t(Y))P_t^{-1} - \nu_2(s_t)\mathrm{tr}\left(P_t^{-1}\frac{dP_t}{dt}P_t^{-1}\pi_t(Y)\right)P_t^{-1} \\
&+ \nu_2(s_t)\mathrm{tr}\left(P_t^{-1}\frac{d\pi_t(Y)}{dt}\right)P_t^{-1} - \nu_2(s_t)\mathrm{tr}(P_t^{-1}\pi_t(Y))P_t^{-1}\frac{dP_t}{dt}P_t^{-1} \\
&- \nu_2(s_t)\mathrm{tr}\left(P_t^{-1}\frac{dP_t}{dt}\right)P_t^{-1}\pi_t(Y)P_t^{-1} + \nu_1(s_t)P_t^{-1}\frac{dP_t}{dt}P_t^{-1}\pi_t(Y)P_t^{-1} \\
&- \nu_1(s_t)P_t^{-1}\frac{d\pi_t(Y)}{dt}P_t^{-1} + \nu_1(s_t)P_t^{-1}\pi_t(Y)P_t^{-1}\frac{dP_t}{dt}P_t^{-1} = 0,
\end{aligned}$$

where $s_t = \det P_t$. If $t = 0$, then this equation implies that

$$\begin{aligned}
&\nu_3(s)\mathrm{tr}(P^{-1}X)\mathrm{tr}(P^{-1}Y)P^{-1} - \nu_2(s)\mathrm{tr}(P^{-1}XP^{-1}Y)P^{-1} \\
&+ \nu_2(s)\mathrm{tr}\left\{P^{-1}\left(\frac{d\pi_t(Y)}{dt}\right)_{t=0}\right\}P^{-1} - \nu_2(s)\mathrm{tr}(P^{-1}Y)P^{-1}XP^{-1}
\end{aligned}$$

$$
\begin{aligned}
&- \nu_2(s)\mathrm{tr}\left(P^{-1}X\right)P^{-1}YP^{-1} + \nu_1(s)P^{-1}XP^{-1}YP^{-1} \\
&- \nu_1(s)P^{-1}\left(\frac{d\pi_t(Y)}{dt}\right)_{t=0}P^{-1} + \nu_1(s)P^{-1}YP^{-1}XP^{-1} = 0,
\end{aligned}
$$

where $s = \det P$. Hence we observe that

$$
\begin{aligned}
&\nu_1(s)P^{-1}\left(\frac{d\pi_t(Y)}{dt}\right)_{t=0} \\
&= \nu_2(s)\mathrm{tr}\left\{P^{-1}\left(\frac{d\pi_t(Y)}{dt}\right)_{t=0}\right\} I \\
&\quad + \nu_1(s)(P^{-1}XP^{-1}Y + P^{-1}YP^{-1}X) \\
&\quad - \nu_2(s)\left\{\mathrm{tr}\left(P^{-1}Y\right)P^{-1}X + \mathrm{tr}\left(P^{-1}X\right)P^{-1}Y\right\} \\
&\quad + \nu_3(s)\mathrm{tr}\left(P^{-1}X\right)\mathrm{tr}\left(P^{-1}Y\right)I - \nu_2(s)\mathrm{tr}\left(P^{-1}YP^{-1}X\right)I.
\end{aligned} \tag{2.25}
$$

Taking the trace for both sides of (2.25), we get

$$
\begin{aligned}
&(\nu_1(s) - n\nu_2(s))\mathrm{tr}\left\{P^{-1}\left(\frac{d\pi_t(Y)}{dt}\right)_{t=0}\right\} \\
=&(2\nu_1(s) - n\nu_2(s))\mathrm{tr}\left(P^{-1}XP^{-1}Y\right) + (n\nu_3(s) - 2\nu_2(s))\mathrm{tr}\left(P^{-1}X\right)\mathrm{tr}\left(P^{-1}Y\right).
\end{aligned} \tag{2.26}
$$

From (2.25) and (2.26) it follows that

$$
\begin{aligned}
&P^{-1}\left(\frac{d\pi_t(Y)}{dt}\right)_{t=0} \\
&= P^{-1}XP^{-1}Y + P^{-1}YP^{-1}X \\
&\quad - \frac{\nu_2(s)}{\nu_1(s)}\left\{\mathrm{tr}\left(P^{-1}X\right)P^{-1}Y + \mathrm{tr}\left(P^{-1}Y\right)P^{-1}X\right\} \\
&\quad + \frac{(\nu_3(s)\nu_1(s) - 2\nu_2(s)^2)\mathrm{tr}\left(P^{-1}X\right)\mathrm{tr}\left(P^{-1}Y\right) + \nu_2(s)\nu_1(s)\mathrm{tr}\left(P^{-1}XP^{-1}Y\right)}{\nu_1(s)(\nu_1(s) - n\nu_2(s))}I.
\end{aligned}
$$

This completes the proof. □

C Proof of Proposition 4

Since geometric structure $(\mathcal{L}_s, g^{(V)})$ is also invariant under the transformation τ_G where $G \in SL(n, \mathbf{R})$, it suffices to consider at $\lambda I \in \mathcal{L}_s$, where $\lambda = s^{1/n}$.

Let $\tilde{X} \in \mathcal{X}(\mathcal{L}_s)$ be a vector field defined at each $P \in \mathcal{L}_s$ by

$$\tilde{X} = \sum_i \tilde{X}^i(P)\frac{\partial}{\partial x^i} = P^{1/2}XP^{1/2}, \quad X \in T_I\mathcal{L}_1 = \{X|\mathrm{tr}(X) = 0,\ X = X^T\},$$

where $\tilde{X}^i$ are certain smooth functions on $\mathcal{L}_s$. Consider the curve $P_t = \lambda \exp Xt \in \mathcal{L}_s$ starting at $t = 0$ and a vector field $\tilde{Y}$ along P_t defined by

$$\tilde{Y}_{P_t} = P_t^{1/2}Y_tP_t^{1/2} = \sum_i \tilde{Y}^i(P_t)\frac{\partial}{\partial x^i},$$

where Y_t is an arbitrary smooth curve in $T_I\mathcal{L}_1$ with $Y_0 = Y$ and $\tilde{Y}^i$ are smooth functions on P_t. We show that the $(T_{\lambda I}\mathcal{L}_s)^\perp$-component of $\left(\hat{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I}$, i.e., the covariant derivative at λI orthogonal to $T_{\lambda I}\mathcal{L}_s$, vanishes for any X and $Y \in T_I\mathcal{L}_1$ if and only if $\nu_2(s) = 0$.

We see

$$(P_t)_{t=0} = \lambda I, \quad \left(\frac{dP_t}{dt}\right)_{t=0} = \lambda X, \quad \tilde{Y}_{\lambda I} = \lambda Y$$

hold. Note that

$$\frac{d}{dt}\tilde{Y}_{P_t} = \frac{1}{2}(X\tilde{Y}_{P_t} + \tilde{Y}_{P_t}X) + P_t^{1/2}\frac{dY_t}{dt}P_t^{1/2},$$

then using (2.13) and corollary 1, we obtain

$$\begin{aligned}\left(\hat{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I} &= \left(\frac{d}{dt}\tilde{Y}_{P_t}\right)_{t=0} + \left(\sum_{i,j}\tilde{X}^i\tilde{Y}^j\hat{\nabla}^{(V)}_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}\right)_{\lambda I} \\ &= \frac{\lambda}{2}(XY + YX) + \lambda\left(\frac{d}{dt}Y_t\right)_{t=0} \\ &\quad -\frac{1}{2}\left\{\lambda(XY + YX) + \Phi(\lambda X, \lambda Y, \lambda I) + \Phi^\perp(\lambda X, \lambda Y, \lambda I)\right\} \\ &= \lambda\left(\frac{d}{dt}Y_t\right)_{t=0} - \frac{1}{2}\Phi^\perp(\lambda X, \lambda Y, \lambda I).\end{aligned}$$

For the third equality we have used that $\Phi(\lambda X, \lambda Y, \lambda I) = 0$ for any X and $Y \in T_I\mathcal{L}_1$.

Since it holds that

$$g^{(V)}_{\lambda I}\left(\left(\hat{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I}, I\right) = \lambda^{-1}(-\nu_1(s) + \nu_2(s)n)\mathrm{tr}\left(\hat{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I}$$

and $-\nu_1(s) + \nu_2(s)n \neq 0$ by (2.5), the $(T_{\lambda I}\mathcal{L}_s)^\perp$-component of $\left(\hat{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I}$ vanishes for any X and $Y \in T_I\mathcal{L}_1$ if and only if

$$\mathrm{tr}\left(\hat{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I} = -\frac{1}{2}\mathrm{tr}\Phi^\perp(\lambda X, \lambda Y, \lambda I) = 0.$$

Here, we have used $\mathrm{tr}((dY_t/dt)_{t=0}) = 0$. The above equality is equivalent to $\nu_2(s) = 0$. Hence, we conclude that the statement holds. □

D Proof of Proposition 6

The statements (i) and (ii) follow from direct calculations. Since $\left({}^*\tilde{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_P$ is the orthogonal projection of $\left({}^*\nabla^{(V)}_{\tilde{X}}\tilde{Y}\right)_P$ to $T_P\mathcal{L}_s$ with respect to $g^{(V)}_P$, it can be represented by

$$\left({}^*\tilde{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_P = \left({}^*\nabla^{(V)}_{\tilde{X}}\tilde{Y}\right)_P - \delta P, \quad \delta \in \mathbf{R},$$

where δ is determined from the orthogonality condition

$$g^{(V)}_P\left(\left({}^*\tilde{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_P, P\right) = 0.$$

Similarly to the proof of Proposition 4 where $\lambda = s^{1/n}$, we see that

$$\begin{aligned}
\left({}^*\nabla^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I} &= \left(\frac{d}{dt}\tilde{Y}_{P_t}\right)_{t=0} + \left(\sum_{i,j}\tilde{X}^i\tilde{Y}^j {}^*\nabla^{(V)}_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}\right)_{\lambda I} \\
&= \frac{\lambda}{2}(XY + YX) + \lambda\left(\frac{d}{dt}Y_t\right)_{t=0} \\
&\quad - \left\{\lambda(XY + YX) + \Phi(\lambda X, \lambda Y, \lambda I) + \Phi^{\perp}(\lambda X, \lambda Y, \lambda I)\right\} \\
&= \lambda\left(\frac{d}{dt}Y_t\right)_{t=0} - \frac{\lambda}{2}(XY + YX) - \Phi^{\perp}(\lambda X, \lambda Y, \lambda I).
\end{aligned}$$

Since $\Phi^{\perp}(\lambda X, \lambda Y, \lambda I) \in (T_{\lambda I}\mathcal{L}_s)^{\perp}$ and $(dY_t/dt)_{t=0} \in T_{\lambda I}\mathcal{L}_s$, the orthogonal projection of $\left({}^*\nabla^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I}$ to $T_{\lambda I}\mathcal{L}_s$ is that of $\lambda(dY_t/dt)_{t=0} - \lambda(YX + XY)/2$. Thus, from the orthogonality condition we have

$$\left({}^*\tilde{\nabla}^{(V)}_{\tilde{X}}\tilde{Y}\right)_{\lambda I} = \lambda\left(\frac{d}{dt}Y_t\right)_{t=0} - \frac{\lambda}{2}(XY + YX) + \frac{\lambda}{n}\mathrm{tr}(XY)I,$$

which is independent of $V(s)$. □

E Proof of Theorem 4

For P and Q in $PD(n, \mathbf{R})$, we shortly write two density functions in a U-model as $f_P(x) = f(x, P)$ and $f_Q(x) = f(x, Q)$.

It suffices to show the dual canonical divergence ${}^*D^{(V)}(P,Q) = D^{(V)}(Q,P)$ of $(PD(n,\mathbf{R}), \nabla, g^{(V)})$ given by (2.10) coincides with $D_U(f_P, f_Q)$. Note that an exchange of the order for two arguments in a divergence only causes that of the definitions of primal and dual affine connections in (2.3) but does not affect whole dualistic structure of the induced geometry.

Recalling (2.6), we have

$$\mathrm{grad}\varphi^{(V)}(P) = \Big(V'(\det P)\det P\Big)P^{-1} = \nu_1(\det P)P^{-1},$$

where V' denotes the derivative of V by s. On the other hand, we can directly differentiate $\varphi^{(V)}(P)$ defined via (2.24)

$$\begin{aligned}
&\mathrm{grad}\varphi^{(V)}(P)\\
&= \mathrm{grad}\left\{\int U\left(-\frac{1}{2}x^TPx - c_U(\det P)\right)dx + c_U(\det P)\right\}\\
&= \int f_P(x)\left\{-\frac{1}{2}xx^T - \Big(c_U'(\det P)\det P\Big)P^{-1}\right\}dx + \Big(c_U'(\det P)\det P\Big)P^{-1}\\
&= -\frac{1}{2}\int f_P(x)xx^Tdx = -\frac{1}{2}\mathrm{E}_P(xx^T),
\end{aligned}$$

where E_P is the expectation operator with respect to $f_P(x)$. Thus, we have

$$\nu_1(\det P)P^{-1} = -\frac{1}{2}\mathrm{E}_P(xx^T). \tag{2.27}$$

Note that

$$\xi(f_P) = -\frac{1}{2}x^TPx - c_U(\det P), \quad \xi(f_Q) = -\frac{1}{2}x^TQx - c_U(\det Q)$$

because $\xi(u)$ is the identity. From the definition, U-divergence is

$$\begin{aligned}
D_U(f_P,f_Q) = &\int U\left(-\frac{1}{2}x^TQx - c_U(\det Q)\right) - U\left(-\frac{1}{2}x^TPx - c_U(\det P)\right)\\
&-f_P(x)\left\{-\frac{1}{2}x^TQx - c_U(\det Q) + \frac{1}{2}x^TPx + c_U(\det P)\right\}dx\\
= &\,\varphi^{(V)}(\det Q) - \varphi^{(V)}(\det P) + \frac{1}{2}\mathrm{E}_P\left(x^TQx - x^TPx\right).
\end{aligned}$$

Using (2.27), the third term is expressed by

$$\begin{aligned}
\frac{1}{2}\mathrm{E}_P\left(x^TQx - x^TPx\right) &= \frac{1}{2}\mathrm{tr}\{\mathrm{E}_P(xx^T)(Q-P)\}\\
&= \nu_1(\det P)\mathrm{tr}(P^{-1}(P-Q)).
\end{aligned}$$

Hence, $D_U(f_P, f_Q) = {}^*D^{(V)}(P, Q) = D^{(V)}(Q, P)$. □

References

1. Amari, S.: Differential-geometrical methods in statistics, Lecture notes in Statistics. vol. 28, Springer, New York (1985)
2. Amari, S., Nagaoka, H.: Methods of information geometry, AMS & OUP, Oxford (2000)
3. David, A.P.: The geometry of proper scoring rules. Ann. Inst. Stat. **59**, 77–93 (2007)
4. Eguchi, S.: Information geometry and statistical pattern recognition. Sugaku Expositions Amer. Math. Soc. **19**, 197–216 (2006) (originally Sūgaku, **56**, 380–399 (2004) in Japanese)
5. Eguchi, S.: Information divergence geometry and the application to statistical machine learning. In: Emmert-Streib, F., Dehmer, M. (eds.) Information Theory and Statistical Learning, pp. 309–332. Springer, New York (2008)
6. Eguchi, S., Copas, J.: A class of logistic-type discriminant functions. Biometrika **89**(1), 1–22 (2002)
7. Eguchi, S., Komori, O., Kato, S.: Projective power entropy and maximum tsallis entropy distributions. Entropy **13**, 1746–1764 (2011)
8. Fang, K.T., Kotz, S., Ng, K.W.: Symmetric Multivariate and Related Distributions. Chapman and Hall, London (1990)
9. Faraut, J., Korányi, A.: Analysis on Symmetric Cones. Oxford University Press, New York (1994)
10. Grunwald, P.D., David, A.P.: Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. Ann. Stat. **32**, 1367–1433 (2004)
11. Hao, J.H., Shima, H.: Level surfaces of nondegenerate functions in r^{n+1}. Geom. Dedicata **50**(2), 193–204 (1994)
12. Helgason, S.: Differential Geometry and Symmetric Spaces. Academic Press, New York (1962)
13. Higuchi, I., Eguchi, S.: Robust principal component analysis with adaptive selection for tuning parameters. J. Mach. Learn. Res. **5**, 453–471 (2004)
14. Kanamori, T., Ohara, A.: A bregman extension of quasi-newton updates I: an information geometrical framework. Optim. Methods Softw. **28**(1), 96–123 (2013)
15. Kakihara, S., Ohara, A., Tsuchiya, T.: Information geometry and interior-point algorithms in semidefinite programs and symmetric cone programs. J. Optim. Theory Appl. **157**(3), 749–780 (2013)
16. Kass, R.E., Vos, P.W.: Geometrical Foundations of Asymptotic Inference. Wiley, New York (1997)
17. Koecher, M.: The Minnesota Notes on Jordan Algebras and their Applications. Springer, Berlin (1999)
18. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
19. Kurose, T.: Dual connections and affine geometry. Math. Z. **203**(1), 115–121 (1990)
20. Kurose, T.: On the divergences of 1-conformally flat statistical manifolds. Tohoku Math. J. **46**(3), 427–433 (1994)
21. Lauritzen, S.: Statistical manifolds. In: Amari, S.-I., et al. (eds.) Differential Geometry in Statistical Inference, Institute of Mathematical Statistics, Hayward (1987)
22. Minami, M., Eguchi, S.: Robust blind source separation by beta-divergence. Neural Comput. **14**, 1859–1886 (2002)
23. Muirhead, R.J.: Aspects of Multivariate Statistical Theory. Wiley, New York (1982)
24. Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S.: Information geometry of u-boost and bregman divergence. Neural Comput. **16**, 1437–1481 (2004)
25. Murray, M.K., Rice, J.W.: Differential Geometry and Statistics. Chapman & Hall, London (1993)

26. Naudts, J.: Continuity of a class of entropies and relative entropies. Rev. Math. Phys. **16**, 809–822 (2004)
27. Naudts, J.: Estimators, escort probabilities, and ϕ-exponential families in statistical physics. J. Ineq. Pure Appl. Math. **5**, 102 (2004)
28. Nesterov, Y.E., Todd, M.J.: Primal-dual interior-point methods for self-scaled cones. SIAM J. Optim. **8**, 324–364 (1998)
29. Nomizu, K., Sasaki, T.: Affine differential geometry. Cambridge University Press, Cambridge (1994)
30. Ohara, A.: Geodesics for dual connections and means on symmetric cones. Integr. Eqn. Oper. Theory **50**, 537–548 (2004)
31. Ohara, A., Amari, S.: Differential geometric structures of stable state feedback systems with dual connections. Kybernetika **30**(4), 369–386 (1994)
32. Ohara, A., Eguchi, S.: Geometry on positive definite matrices induced from V-potential function. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information; Lecture Notes in Computer Science 8085, pp. 621–629. Springer, Berlin (2013)
33. Ohara, A., Eguchi, S.: Group invariance of information geometry on q-gaussian distributions induced by beta-divergence. Entropy **15**, 4732–4747 (2013)
34. Ohara, A., Suda, N., Amari, S.: Dualistic differential geometry of positive definite matrices and its applications to related problems. Linear Algebra Appl. **247**, 31–53 (1996)
35. Ohara, A., Wada, T.: Information geometry of q-Gaussian densities and behaviors of solutions to related diffusion equations. J. Phys. A: Math. Theor. **43**, 035002 (18pp.) (2010)
36. Ollila, E., Tyler, D., Koivunen, V., Poor, V.: Complex elliptically symmetric distributions : survey, new results and applications. IEEE Trans. signal process. **60**(11), 5597–5623 (2012)
37. Rothaus, O.S.: Domains of positivity. Abh. Math. Sem. Univ. Hamburg **24**, 189–235 (1960)
38. Sasaki, T.: Hyperbolic affine hyperspheres. Nagoya Math. J. **77**, 107–123 (1980)
39. Scott, D.W.: Parametric statistical modeling by minimum integrated square error. Technometrics **43**, 274–285 (2001)
40. Shima, H.: The geometry of Hessian structures. World Scientific, Singapore (2007)
41. Takenouchi, T., Eguchi, S.: Robustifying adaboost by adding the naive error rate. Neural Comput. **16**(4), 767–787 (2004)
42. Tsallis, C.: Introduction to Nonextensive Statistical Mechanics. Springer, New York (2009)
43. Uohashi, K., Ohara, A., Fujii, T.: 1-conformally flat statistical submanifolds. Osaka J. Math. **37**(2), 501–507 (2000)
44. Uohashi, K., Ohara, A., Fujii, T.: Foliations and divergences of flat statistical manifolds. Hiroshima Math. J. **30**(3), 403–414 (2000)
45. Vinberg, E.B.: The theory of convex homogeneous cones. Trans. Moscow Math. Soc. **12**, 340–430 (1963)
46. Wolkowicz, H., et al. (eds.): Handbook of Semidefinite Programming. Kluwer Academic Publishers, Boston (2000)

Chapter 3
Hessian Structures and Divergence Functions on Deformed Exponential Families

Hiroshi Matsuzoe and Masayuki Henmi

Abstract A Hessian structure (∇, h) on a manifold is a pair of a flat affine connection ∇ and a semi-Riemannian metric h which is given by a Hessian of some function. In information geometry, it is known that an exponential family naturally has dualistic Hessian structures and their canonical divergences coincide with the Kullback-Leibler divergences, which are also called the relative entropies. A deformed exponential family is a generalization of exponential families. A deformed exponential family naturally has two kinds of dualistic Hessian structures and conformal structures of Hessian metrics. In this paper, geometry of such Hessian structures and conformal structures are summarized. In addition, divergence functions on these Hessian manifolds are constructed from the viewpoint of estimating functions. As an application of such Hessian structures to statistics, a generalization of independence and geometry of generalized maximum likelihood method are studied.

Keywords Hessian manifold · Statistical manifold · Deformed exponential family · Divergence · Information geometry · Tsallis statistics

3.1 Introduction

In information geometry, an exponential family is a useful statistical model and it is applied to various fields of statistical sciences (cf. [1]). For example, the set of Gaussian distributions is an exponential family. It is known that an exponential family can be naturally regarded as a Hessian manifold [28], which is also called a dually flat

H. Matsuzoe (✉)
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
e-mail: matsuzoe@nitech.ac.jp

M. Henmi
The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
e-mail: henmi@ism.ac.jp

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_3,

space [1] or a flat statistical manifold [12]. A pair of dually flat affine connections has essential roles in geometric theory of statistical inferences. In addition, a Hessian manifold has an asymmetric squared-distance like function, called the canonical divergence. On an exponential family, the canonical divergence coincides with the Kullback-Leibler divergence or the relative entropy. (See Sect. 3.3.)

A deformed exponential family is a generalization of exponential families, which was introduced in anomalous statistical physics [22]. (See also [23, 32] and [33].) A deformed exponential family naturally has two kinds of dualistic Hessian structures, and such geometric structures are independently studied in machine learning theory [21] and statistical physics [3, 26], etc. For example, a q-exponential family is a typical example of deformed exponential families. One of Hessian structures on a q-exponential family is related to geometry of β-divergences (or density power divergences [5]). The other Hessian structure is related to geometry of α-divergences. (In the q-exponential case, these geometry are studied in [18].) In addition, conformal structures of statistical manifolds play important roles in geometry of deformed exponential families.

In this paper, we summarize such Hessian structures and conformal structures on deformed exponential families. Then we construct a generalized relative entropy from the viewpoint of estimating functions. As an application, we consider generalization of independence of random variables, then elucidate geometry of the maximum q-likelihood estimator. This paper is written based on the proceeding [19].

3.2 Preliminaries

In this paper, we assume that all objects are smooth, and a manifold M is an open domain in $\boldsymbol{R}^n$.

Let (M, h) be a semi-Riemannian manifold, that is, h is assumed to be nondegenerate, which is not necessary to be positive definite (e.g. the Lorentzian metric in relativity). Let ∇ be an affine connection on M. We define the *dual connection* ∇^* of ∇ with respect to h by

$$Xh(Y, Z) = h(\nabla_X Y, Z) + h(Y, \nabla^*_X Z),$$

where X, Y and Z are arbitrary vector fields on M. It is easy to check that $(\nabla^*)^* = \nabla$.

For an affine connection ∇, we define the *curvature tensor field* R and the *torsion tensor field* T by

$$\begin{aligned} R(X, Y)Z &:= \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, \\ T(X, Y) &:= \nabla_X Y - \nabla_Y X - [X, Y], \end{aligned}$$

where $[X, Y] := XY - YX$. We say that ∇ is *curvature-free* if R vanishes everywhere on M, and the one is *torsion-free* if T vanishes everywhere.

For pair of dual affine connections, the following proposition holds (cf. [16]).

Proposition 1 *Consider the conditions below:*

1. ∇ *is torsion-free.*
2. ∇^* *is torsion-free.*
3. $\nabla^{(0)} = (\nabla + \nabla^*)/2$ *is the Levi-Civita connection with respect to* h.
4. ∇h *is totally symmetric, where* ∇h *is the* $(0, 3)$*-tensor field defined by*

$$(\nabla_X h)(Y, Z) := Xh(Y, Z) - h(\nabla_X Y, Z) - h(Y, \nabla_X Z).$$

Assume any two of the above conditions, then the others hold.

From now on, we assume that an affine connection ∇ is torsion-free.

We say that an affine connection ∇ is *flat* if ∇ is curvature-free. For a flat affine connection ∇, there exists a coordinate system $\{\theta^i\}$ on M locally such that the connection coefficients $\{\Gamma_{ij}^{\nabla\ k}\}$ $(i, j, k = 1, \ldots, n)$ of ∇ vanish on its coordinate neighbourhood. We call such a coordinate system $\{\theta^i\}$ an *affine coordinate system*.

Let (M, h) be a semi-Riemannian manifold, and let ∇ be a flat affine connection on M. We say that the pair (∇, h) is a *Hessian structure* on M if there exists a function ψ, at least locally, such that $h = \nabla d\psi$ [28]. In the coordinate form, the following formula holds:

$$h_{ij}(p(\theta)) = \frac{\partial^2}{\partial\theta^i \partial\theta^j}\psi(p(\theta)),$$

where p is an arbitrary point in M and $\{\theta^i\}$ is a ∇-affine coordinate system around p. Under the same assumption, we call the triplet (M, ∇, h) a *Hessian manifold*. For a Hessian manifold (M, ∇, h), we define a totally symmetric $(0, 3)$-tensor field C by $C := \nabla h$. We call C the *cubic form* for (M, ∇, h).

For a semi-Riemannian manifold (M, h) with a torsion-free affine connection ∇, the triplet (M, ∇, h) is said to be a *statistical manifold* if ∇h is totally symmetric [12]. Originally, the triplet (M, g, C) is called a statistical manifold [14], where (M, g) is a Riemannian manifold and C is a totally symmetric $(0, 3)$-tensor field on M. From Proposition 1, these definitions are essentially equivalent. In fact, for a semi-Riemannian manifold (M, h) with a totally symmetric $(0, 3)$-tensor field C, we can define mutually dual torsion-free affine connections ∇ and ∇^* by

$$h(\nabla_X Y, Z) := h(\nabla_X^{(0)} Y, Z) - \frac{1}{2}C(X, Y, Z), \tag{3.1}$$

$$h(\nabla_X^* Y, Z) := h(\nabla_X^{(0)} Y, Z) + \frac{1}{2}C(X, Y, Z), \tag{3.2}$$

where $\nabla^{(0)}$ is the Levi-Civita connection with respect to h. In this case, ∇h and $\nabla^* h$ are totally symmetric. Hence (M, ∇, h) and (M, ∇^*, h) are statistical manifolds.

A triplet (M, ∇, h) is a flat statistical manifold if and only if it is a Hessian manifold (cf. [28]). Suppose that R and R^* are curvature tensors of ∇ and ∇^*, respectively. Then we have

$$h(R(X, Y)Z, V) = -h(Z, R^*(X, Y)V).$$

Hence the condition that the triplet (M, ∇, h) is a Hessian manifold is equivalent to that the quadruplet (M, h, ∇, ∇^*) is a *dually flat space* [1].

For a Hessian manifold (M, ∇, h), we suppose that $\{\theta^i\}$ is a ∇-affine coordinate system on M. Then there exists a ∇^*-affine coordinate system $\{\eta_i\}$ such that

$$h\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\eta_j}\right) = \delta^i_j.$$

We call $\{\eta_i\}$ the *dual coordinate system* of $\{\theta^i\}$ with respect to h.

Proposition 2 *Let (M, ∇, h) be a Hessian manifold. Suppose that $\{\theta^i\}$ is a ∇-affine coordinate system, and $\{\eta_i\}$ is the dual coordinate system of $\{\theta^i\}$. Then there exist functions ψ and ϕ on M such that*

$$\frac{\partial\psi}{\partial\theta^i} = \eta_i, \quad \frac{\partial\phi}{\partial\eta_i} = \theta^i, \quad \psi(p) + \phi(p) - \sum_{i=1}^{n} \theta^i(p)\eta_i(p) = 0, \quad (p \in M), \tag{3.3}$$

$$h_{ij} = \frac{\partial^2\psi}{\partial\theta^i\partial\theta^j}, \quad h^{ij} = \frac{\partial^2\phi}{\partial\eta_i\partial\eta_j},$$

where (h_{ij}) is the component matrix of a semi-Riemannian metric h with respect to $\{\theta^i\}$, and (h^{ij}) is the inverse matrix of (h_{ij}). Moreover,

$$C_{ijk} = \frac{\partial^3\psi}{\partial\theta^i\partial\theta^j\partial\theta^k} \tag{3.4}$$

is the cubic form of (M, ∇, h).

For proof, see [1] and [28]. The functions ψ and ϕ are called the *θ-potential* and the *η-potential*, respectively. From the above proposition, the Hessians of θ-potential and η-potential coincide with the semi-Riemannian metric h:

$$\frac{\partial\eta_i}{\partial\theta^j} = \frac{\partial^2\psi}{\partial\theta^i\partial\theta^j} = h_{ij}, \qquad \frac{\partial\theta^i}{\partial\eta_j} = \frac{\partial^2\phi}{\partial\eta_i\partial\eta_j} = h^{ij}. \tag{3.5}$$

In addition, we obtain the original flat connection ∇ and its dual ∇^* from the potential function ψ. From Eq. (3.4), we have the cubic form of Hessian manifold (M, ∇, h). Then we obtain two affine connections ∇ and ∇^* by Eqs. (3.1), (3.2) and (3.4).

Under the same assumptions as in Proposition 2, we define a function D on $M \times M$ by

$$D(p, r) := \psi(p) + \phi(r) - \sum_{i=1}^{n} \theta^i(p)\eta_i(r), \quad (p, r \in M).$$

We call D the *canonical divergence* of (M, ∇, h). The definition is independent of choice of an affine coordinate system. The canonical divergence is an asymmetric

squared distance like function on M. In particular, the canonical divergence D is non-negative if the metric h is positive definite. However, we assumed that h is a semi-Riemannian metric, hence D can take negative values. (cf. [12] and [15].)

We remark that the canonical divergence induces the original Hessian manifold (M, ∇, h) by Eguchi's relation [7]. Suppose that D is a function on $M \times M$. We define a function on M by the following formula:

$$D[X_1, \ldots, X_i | Y_1, \ldots, Y_j](p) := (X_1)_p \cdots (X_i)_p (Y_1)_r \cdots (Y_j)_r D(p, r)|_{p=r},$$

where $X_1, \ldots, X_i$ and $Y_1, \cdots, Y_j$ are vector fields on M. We say that D is a *contrast function* on $M \times M$ if

$$\begin{aligned} &1.\ D[\ |\ ](p) = D(p, p) = 0, \\ &2.\ D[X|\ ](p) = D[\ |X](p) = 0, \\ &3.\ h(X, Y) := -D[X|Y] \\ &\qquad \text{is a semi-Riemannian metric on } M. \end{aligned} \tag{3.6}$$

For a contrast function D on $M \times M$, we define a pair of affine connections by

$$\begin{aligned} h(\nabla_X Y, Z) &= -D[XY|Z], \\ h(Y, \nabla^*_X Z) &= -D[Y|XZ]. \end{aligned}$$

By differentiating Eq. (3.6), two affine connections ∇ and ∇^* are mutually dual with respect to h. We can check that ∇ and ∇^* are torsion-free, and ∇h and $\nabla^* h$ are totally symmetric. Hence triplets (M, ∇, h) and (M, ∇^*, h) are statistical manifolds. We call (M, ∇, h) the *induced statistical manifold* from a contrast function D. If (M, ∇, h) is a Hessian manifold, we say that (M, ∇, h) is the *induced Hessian manifold* from D.

Proposition 3 *Suppose that D is the canonical divergence on a Hessian manifold (M, ∇, h). Then D is a contrast function on $M \times M$ which induces the original Hessian manifold (M, ∇, h).*

Proof From the definition and Eq. (3.3), we have $D[\ |\ ] = 0$ and $D[X|\ ] = D[\ |X] = 0$. Let $\{\theta^i\}$ be a ∇-affine coordinate and $\{\eta_j\}$ the dual affine coordinate of $\{\theta^j\}$. Set $\partial_i = \partial/\partial\theta^i$. From Eqs. (3.3) and (3.5), we have

$$\begin{aligned} D[\partial_i|\partial_j](p) &= (\partial_i)_p (\partial_j)_r D(p, q)|_{p=r} \ = \ (\partial_j)_r \{\eta_i(p) - \eta_i(r)\}|_{p=r} \\ &= -(\partial_j)_r \eta_i(r)|_{p=r} \ = \ -h_{ij}(p). \end{aligned}$$

This implies that the canonical divergence D is a contrast function on $M \times M$. Induced affine connections are given by

$$\begin{aligned} \Gamma_{ij,k} &= -D[\partial_i \partial_j | \partial_k] \ = \ (\partial_i)_p (\partial_k)_r \left\{\eta_j(p) - \eta_j(r)\right\}|_{p=r} \\ &= -(\partial_i)_p (\partial_k)_r \eta_j(r)|_{p=r} \ = \ 0, \end{aligned}$$

$$\begin{aligned}\Gamma^*_{ik,j} &= -D[\partial_j|\partial_i\partial_k] \;=\; (\partial_i)_r(\partial_k)_r\left\{\eta_j(p)-\eta_j(r)\right\}|_{p=r}\\ &= -(\partial_i)_r(\partial_k)_r\eta_j(r)|_{p=r} \;=\; -(\partial_i)_r(\partial_k)_r(\partial_j)_r\psi(r)|_{p=r}\\ &= C_{ikj},\end{aligned}$$

where $\Gamma_{ij,k}$ and $\Gamma^*_{ik,j}$ are Christoffel symbols of the first kind of ∇ and ∇^*, respectively. From Eqs. (3.1) and (3.2), since h is nondegenerate, the affine connection ∇ coincides with the original one of (M, ∇, h). □

At the end of this section, we review generalized conformal equivalence for statistical manifolds. Fix a number $\alpha \in \boldsymbol{R}$. We say that two statistical manifolds (M, ∇, h) and $(M, \bar{\nabla}, \bar{h})$ are α-*conformally equivalent* if there exists a function φ on M such that

$$\begin{aligned}\bar{h}(X, Y) &= e^{\varphi}h(X, Y),\\ \bar{\nabla}_X Y &= \nabla_X Y - \frac{1+\alpha}{2}h(X, Y)\mathrm{grad}_h\varphi + \frac{1-\alpha}{2}\left\{d\varphi(Y)\,X + d\varphi(X)\,Y\right\},\end{aligned}$$

where $\mathrm{grad}_h\varphi$ is the gradient vector field of φ with respect to h, that is,

$$h(\mathrm{grad}_h\varphi, X) := X\varphi.$$

(The vector field $\mathrm{grad}_h\varphi$ is often called the *natural gradient* of φ in neurosciences, etc.) We say that a statistical manifold (M, ∇, h) is α-*conformally flat* if it is locally α-conformally equivalent to some Hessian manifold [12].

Suppose that D and $\bar{D}$ are contrast functions on $M \times M$. We say that D and $\bar{D}$ are α-*conformally equivalent* if there exists a function φ on M such that

$$\bar{D}(p, r) = \exp\left[\frac{1+\alpha}{2}\varphi(p)\right]\exp\left[\frac{1-\alpha}{2}\varphi(r)\right]D(p, r).$$

In this case, induced statistical manifolds (M, ∇, h) and $(M, \bar{\nabla}, \bar{h})$ from D and $\bar{D}$, respectively, are α-conformally equivalent.

Historically, conformal equivalence of statistical manifolds was introduced in asymptotic theory of sequential estimation [27]. (See also [11].) Then it is generalized in affine differential geometry (e.g. [10, 12, 13] and [17]). As we will see in Sects. 3.5 and 3.6, conformal structures on a deformed exponential family play important roles. (See also [2, 20, 24] and [25].)

3.3 Statistical Models

Let $(\Omega, \mathcal{F}, P)$ be a probability space, that is, Ω is a sample space, $\mathcal{F}$ is a completely additive class on Ω, and P is a probability measure on Ω. Let Ξ be an open subset in $\boldsymbol{R}^n$. We say that S is a *statistical model* if S is a set of probability density functions on Ω with parameter $\xi = {}^t(\xi^1, \ldots, \xi^n) \in \Xi$ such that

$$S := \left\{ p(x;\xi) \;\middle|\; \int_{\Omega} p(x;\xi)dx = 1,\ p(x;\xi) > 0,\ \xi \in \Xi \subset \boldsymbol{R}^n \right\}.$$

Under suitable conditions, S can be regarded as a manifold with local coordinate system $\{\xi^i\}$ [1]. In particular, we assume that we can interchange differentials and integrals. Hence, the equation below holds

$$\int_{\Omega} \left(\frac{\partial}{\partial \xi^i} p(x;\xi) \right) dx = \frac{\partial}{\partial \xi^i} \int_{\Omega} p(x;\xi)dx = \frac{\partial}{\partial \xi^i} 1 = 0.$$

For a statistical model S, we define the Fisher information matrix $g^F(\xi) = (g^F_{ij}(\xi))$ by

$$\begin{aligned} g^F_{ij}(\xi) &:= \int_{\Omega} \left(\frac{\partial}{\partial \xi^i} \log p(x;\xi) \right) \left(\frac{\partial}{\partial \xi^j} \log p(x;\xi) \right) p(x;\xi)\, dx \\ &= E_p[\partial_i l_\xi \partial_j l_\xi], \end{aligned} \tag{3.7}$$

where $\partial_i = \partial/\partial \xi^i$, $l_\xi = l(x;\xi) = \log p(x;\xi)$, and $E_p[f]$ is the expectation of $f(x)$ with respect to $p(x;\xi)$. The Fisher information matrix g^F is semi-positive definite in general. Assuming that g^F is positive definite and all components are finite, then g^F can be regarded as a Riemannian metric on S. We call g^F the *Fisher metric* on S. The Fisher metric g^F has the following representations:

$$g^F_{ij}(\xi) = \int_{\Omega} \left(\frac{\partial}{\partial \xi^i} p(x;\xi) \right) \left(\frac{\partial}{\partial \xi^j} \log p(x;\xi) \right) dx \tag{3.8}$$

$$= \int_{\Omega} \frac{1}{p(x;\xi)} \left(\frac{\partial}{\partial \xi^i} p(x;\xi) \right) \left(\frac{\partial}{\partial \xi^j} p(x;\xi) \right) dx. \tag{3.9}$$

Next, let us define an affine connection on S. For a fixed $\alpha \in \boldsymbol{R}$, an α-*connection* $\nabla^{(\alpha)}$ on S is defined by

$$\Gamma^{(\alpha)}_{ij,k}(\xi) := E_p \left[\left(\partial_i \partial_j l_\xi + \frac{1-\alpha}{2} \partial_i l_\xi \partial_j l_\xi \right) (\partial_k l_\xi) \right],$$

where $\Gamma^{(\alpha)}_{ij,k}$ is the Christoffel symbol of the first kind of $\nabla^{(\alpha)}$.

We remark that $\nabla^{(0)}$ is the Levi-Civita connection with respect to the Fisher metric g^F. The connection $\nabla^{(e)} := \nabla^{(1)}$ is called the the *exponential connection* and $\nabla^{(m)} := \nabla^{(-1)}$ is called the *mixture connection*. Two connections $\nabla^{(e)}$ and $\nabla^{(m)}$ are expressed as follows:

$$\Gamma^{(e)}_{ij,k} = E_p[(\partial_i\partial_j l_\xi)(\partial_k l_\xi)] = \int_\Omega \partial_i\partial_j \log p(x;\xi)\partial_k p(x;\xi)dx, \tag{3.10}$$

$$\Gamma^{(m)}_{ij,k} = E_p[((\partial_i\partial_j l_\xi + \partial_i l_\xi \partial_j l_\xi)(\partial_k l_\xi)] = \int_\Omega \partial_i\partial_j p(x;\xi)\partial_k \log p(x;\xi)dx. \tag{3.11}$$

We can check that the α-connection $\nabla^{(\alpha)}$ is torsion-free and $\nabla^{(\alpha)}g^F$ is totally symmetric. These imply that $(S, \nabla^{(\alpha)}, g^F)$ forms a statistical manifold. In addition, it is known that the Fisher metric g^F and the α-connection $\nabla^{(\alpha)}$ are independent of choice of dominating measures on Ω. Hence we call the triplet $(S, \nabla^{(\alpha)}, g^F)$ an *invariant statistical manifold*. The cubic form C^F of the invariant statistical manifold $(S, \nabla^{(e)}, g^F)$ is given by

$$C^F_{ijk} = \Gamma^{(m)}_{ij,k} - \Gamma^{(e)}_{ij,k}.$$

A statistical model S_e is said to be an *exponential family* if

$$S_e := \left\{ p(x;\theta) \;\middle|\; p(x;\theta) = \exp\left[\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right],\ \theta \in \Theta \subset \boldsymbol{R}^n \right\},$$

under a choice of suitable dominating measure, where $F_1(x), \ldots, F_n(x)$ are functions on the sample space Ω, $\theta = (\theta^1, \ldots, \theta^n)$ is a parameter, and $\psi(\theta)$ is a function of θ for normalization. The following proposition is well-known in information geometry [1].

Theorem 1 (cf. [1]) *For an exponential family S_e, the following hold:*

1. *$(S_e, \nabla^{(e)}, g^F)$ and $(S_e, \nabla^{(m)}, g^F)$ are mutually dual Hessian manifolds, that is, $(S_e, g^F, \nabla^{(e)}, \nabla^{(m)})$ is a dually flat space.*
2. *$\{\theta^i\}$ is a $\nabla^{(e)}$-affine coordinate system on S_e.*
3. *For the Hessian structure $(\nabla^{(e)}, g^F)$ on S_e, $\psi(\theta)$ is the potential of g^F and C^F with respect to $\{\theta^i\}$:*

$$g^F_{ij}(\theta) = \partial_i\partial_j\psi(\theta), \quad (\partial_i = \partial/\partial\theta^i),$$
$$C^F_{ijk}(\theta) = \partial_i\partial_j\partial_k\psi(\theta).$$

4. *Set the expectation of $F_i(x)$ by $\eta_i := E_p[F_i(x)]$. Then $\{\eta_i\}$ is the dual affine coordinate system of $\{\theta^i\}$ with respect to g^F.*
5. *Set $\phi(\eta) := E_p[\log p(x;\theta)]$. Then $\phi(\eta)$ is the potential of g^F with respect to $\{\eta_i\}$.*

Since $(S_e, \nabla^{(e)}, g^F)$ is a Hessian manifold, the formulas in Proposition 2 hold.

For a statistical model S, we define a *Kullback-Leibler divergence* (or a *relative entropy*) by

$$D_{KL}(p, r) := \int_{\Omega} p(x) \log \frac{p(x)}{r(x)} dx$$
$$= E_p[\log p(x) - \log r(x)], \quad (p(x), r(x) \in S).$$

The Kullback-Leibler divergence D_{KL} on an exponential family S_e coincides with the canonical divergence D on $(S_e, \nabla^{(m)}, g^F)$.

We define an $\boldsymbol{R}^n$ valued function $s(x; \xi) = (s^1(x; \xi), \ldots, s^n(x; \xi))^T$ by

$$s^i(x; \xi) := \frac{\partial}{\partial \xi^i} \log p(x; \xi).$$

We call $s(x; \xi)$ the *score function* of $p(x; \xi)$ with respect to ξ. In information geometry, $s^i(x; \xi)$ is called the *e-(exponential) representation* of $\partial/\partial\xi^i$, and $\partial/\partial\xi^i p(x; \xi)$ is called the *m-(mixture) representation*. The duality of *e*- and *m*-representations is important. In fact, Eq. (3.8) implies that the Fisher metric g^F is nothing but an L^2 inner product of *e*- and *m*-representations.

Construction of the Kullback-Leibler divergence is as follows. We define a *cross entropy* $d_{KL}(p, r)$ by

$$d_{KL}(p, r) := -E_p[\log r(x)].$$

A cross entropy $d_{KL}(p, r)$ gives a bias of information $-\log r(x)$ with respect to $p(x)$. A cross entropy is also called a *yoke* on S [4]. Intuitively, a yoke measures a dissimilarity of two probability density functions on S. We should also note that the cross entropy is obtained by taking the expectation with respect to $p(x)$ of the integrated score function at $r(x)$. Then we have the Kullback-Leibler divergence by

$$D_{KL}(p, r) = -d_{KL}(p, p) + d_{KL(}p, r)$$
$$= E_p[\log p(x) - \log r(x)].$$

The Kullback-Leibler divergence D_{KL} is a normalized yoke on S, which satisfies $D_{KL}(p, p) = 0$. This argument suggests how to construct divergence functions. Once a function like the cross entropy is defined, we can construct divergence functions in the same way.

3.4 The Deformed Exponential Family

In this section, we review the deformed exponential family. For more details, see [3, 22, 23] and [26]. Geometry of deformed exponential families relates to so-called U-geometry [21].

Let χ be a strictly increasing function from $(0, \infty)$ to $(0, \infty)$. We define a *deformed logarithm function* (or a *χ-logarithm function*) by

$$\log_\chi(s) := \int_1^s \frac{1}{\chi(t)}\, dt.$$

We remark that $\log_\chi(s)$ is strictly increasing and satisfies $\log_\chi(1) = 0$. The domain and the target of $\log_\chi(s)$ depend on the function $\chi(t)$. Set $U = \{s \in (0,\infty) \,|\, |\log_\chi(s)| < \infty\}$ and $V = \{\log_\chi(s) \,|\, s \in U\}$. Then $\log_\chi(s)$ is a function from U to V. We also remark that the deformed logarithm is usually called the ϕ-logarithm [23]. However, we use ϕ as the dual potential on a Hessian manifold.

A *deformed exponential function* (or a χ-*exponential function*) is defined by the inverse of the deformed logarithm function $\log_\chi(s)$:

$$\exp_\chi(t) := 1 + \int_0^t \lambda(s) ds,$$

where $\lambda(s)$ is defined by the relation $\lambda(\log_\chi(s)) := \chi(s)$.

When $\chi(s)$ is a power function $\chi(s) = s^q$, $(q > 0, q \neq 1)$, the deformed logarithm and the deformed exponential are given by

$$\log_q(s) := \frac{s^{1-q} - 1}{1-q}, \qquad (s > 0),$$

$$\exp_q(t) := (1 + (1-q)t)^{\frac{1}{1-q}}, \qquad (1 + (1-q)t > 0).$$

The function $\log_q(s)$ is called the *q-logarithm* and $\exp_q(t)$ the *q-exponential*. Taking the limit $q \to 1$, the standard logarithm and the standard exponential are recovered, respectively.

A statistical model S_χ is said to be a *deformed exponential family* (or a χ-*exponential family*) if

$$S_\chi := \left\{ p(x;\theta) \,\middle|\, p(x;\theta) = \exp_\chi\left[\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right],\ \theta \in \Theta \subset \boldsymbol{R}^n \right\},$$

under a choice of suitable dominating measure, where $F_1(x), \ldots, F_n(x)$ are functions on the sample space Ω, $\theta = \{\theta^1, \ldots, \theta^n\}$ is a parameter, and $\psi(\theta)$ is the function of θ for normalization. We assume that S_χ is a statistical model in the sense of [1]. That is, $p(x;\theta)$ has support entirely on Ω, there exits a one-to-one correspondence between the parameter θ and the probability distribution $p(x;\theta)$, and differentiation and integration are interchangeable. In addition, functions $\{F_i(x)\}$, $\psi(\theta)$ and parameters $\{\theta^i\}$ must satisfy the anti-exponential condition. For example, in the q-exponential case, these functions satisfy

$$\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta) < \frac{1}{q-1}.$$

Then we can regard that S_χ is a manifold with local coordinate system $\{\theta^i\}$. We also assume that the function ψ is strictly convex since we consider Hessian metrics on S_χ later. A deformed exponential family has several different definitions. See [30] and [34], for example.

For a deformed exponential probability density $p(x;\theta) \in S_\chi$, we define the *escort distribution* $P_\chi(x;\theta)$ of $p(x;\theta)$ by

$$P_\chi(x;\theta) := \frac{1}{Z_\chi(\theta)}\chi\{p(x;\theta)\},$$

where $Z_\chi(\theta)$ is the normalization defined by

$$Z_\chi(\theta) := \int_\Omega \chi\{p(x;\theta)\}dx.$$

The *χ-expectation* $E_{\chi,p}[f]$ of $f(x)$ with respect to $P_\chi(x;\theta)$ is defined by

$$E_{\chi,p}[f] := \int_\Omega f(x)P_\chi(x;\theta)\,dx \;=\; \frac{1}{Z_\chi(\theta)}\int_\Omega f(x)\chi\{p(x;\theta)\}dx.$$

When χ is a power function $\chi(s) = s^q$, $(q > 0, q \neq 0)$, we denote the escort distribution of $p(x;\theta)$ by $P_q(x;\theta)$, and the χ-expectation with respect to $p(x;\theta)$ by $E_{q,p}[*]$.

Example 1 (*discrete distributions* [3]) The set of discrete distributions S_n is a deformed exponential family for an arbitrary χ. Suppose that Ω is a finite set: $\Omega = \{x_0, x_1, \ldots, x_n\}$. Then the statistical model S_n is given by

$$S_n := \left\{ p(x;\eta) \;\middle|\; \eta_i > 0,\ p(x;\eta) = \sum_{i=0}^{n} \eta_i \delta_i(x),\ \sum_{i=0}^{n} \eta_i = 1 \right\},$$

where $\eta_0 := 1 - \sum_{i=1}^{n} \eta_i$ and

$$\delta_i(x) := \begin{cases} 1 \ (x = x_i), \\ 0 \ (x \neq x_i). \end{cases}$$

Set $\theta^i = \log_\chi p(x_i) - \log_\chi p(x_0) = \log_\chi \eta_i - \log_\chi \eta_0$, $F_i(x) = \delta_i(x)$ and $\psi(\theta) = -\log_\chi \eta_0$. Then the χ-logarithm of $p(x) \in S_n$ is written by

$$\log_\chi p(x) = \sum_{i=1}^{n} \left(\log_\chi \eta_i - \log_\chi \eta_0\right) \delta_i(x) + \log_\chi(\eta_0)$$
$$= \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta).$$

This implies that S_n is a deformed exponential family.

Example 2 (*q-normal distributions* [20]) A q-normal distribution is the probability distribution defined by the following formula:

$$p_q(x;\mu,\sigma) := \frac{1}{Z_q(\sigma)} \left[1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right]_+^{\frac{1}{1-q}},$$

where $[*]_+ := \max\{0, *\}$, $\{\mu, \sigma\}$ are parameters $-\infty < \mu < \infty, 0 < \sigma < \infty$, and $Z_q(\sigma)$ is the normalization defined by

$$Z_q(\sigma) := \begin{cases} \dfrac{\sqrt{3-q}}{\sqrt{1-q}} \mathrm{B}\left(\dfrac{2-q}{1-q}, \dfrac{1}{2}\right)\sigma, & (-\infty < q < 1), \\ \dfrac{\sqrt{3-q}}{\sqrt{q-1}} \mathrm{B}\left(\dfrac{3-q}{2(q-1)}, \dfrac{1}{2}\right)\sigma, & (1 \le q < 3). \end{cases}$$

Here, $\mathrm{B}(*,*)$ is the beta function. We restrict ourselves to consider the case $q \ge 1$. Then the probability distribution $p_q(x;\mu,\sigma)$ has its support entirely on $\boldsymbol{R}$ and the set of q-normal distributions S_q is a statistical model. Set

$$\theta^1 := \frac{2}{3-q}\{Z_q(\sigma)\}^{q-1}\frac{\mu}{\sigma^2}, \quad \theta^2 := -\frac{1}{3-q}\{Z_q(\sigma)\}^{q-1}\frac{1}{\sigma^2},$$
$$\psi(\theta) := -\frac{(\theta^1)^2}{4\theta^2} - \frac{\{Z_q(\sigma)\}^{q-1} - 1}{1-q},$$

then we have

$$\log_q p_q(x;\theta) = \frac{1}{1-q}(\{p_q(x;\theta)\}^{1-q} - 1)$$
$$= \frac{1}{1-q}\left\{\frac{1}{\{Z_q(\sigma)\}^{1-q}}\left(1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right) - 1\right\}$$
$$= \frac{2\mu\{Z_q(\sigma)\}^{q-1}}{(3-q)\sigma^2}x - \frac{\{Z_q(\sigma)\}^{q-1}}{(3-q)\sigma^2}x^2$$
$$- \frac{\{Z_q(\sigma)\}^{q-1}}{3-q}\cdot\frac{\mu^2}{\sigma^2} + \frac{\{Z_q(\sigma)\}^{q-1} - 1}{1-q}$$
$$= \theta^1 x + \theta^2 x^2 - \psi(\theta).$$

This implies that the set of q-normal distributions S_q is a q-exponential family. For a q-normal distribution $p_q(x;\mu,\sigma)$, the q-expectation μ_q and a q-variance σ_q^2 are given by

$$\begin{aligned}\mu_q &= E_{q,p}[x] \;=\; \mu,\\ \sigma_q^2 &= E_{q,p}\left[(x-\mu)^2\right] \;=\; \sigma^2.\end{aligned}$$

We remark that a q-normal distribution is nothing but a three-parameter version of Student's t-distribution when $q \geq 1$. In fact, if $q = 1$, then the q-normal distribution is the normal distribution. If $q = 2$, then the distribution is the Cauchy distribution. We also remark that mathematical properties of q-normal distributions have been obtained by several authors. See [29, 31], for example.

3.5 Geometry of Deformed Exponential Families Derived from the Standard Expectation

In this section, we consider geometry of deformed exponential families by generalizing the e-representation with the deformed logarithm function. For more details, see [21, 26].

Let S_χ be a deformed exponential family. We define an $\boldsymbol{R}^n$ valued function $s^\chi(x;\theta) = \left((s^\chi)^1(x;\theta),\ldots,(s^\chi)^n(x;\theta)\right)^T$ by

$$(s^\chi)^i(x;\theta) := \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta), \quad (i=1,\ldots,n). \tag{3.12}$$

We call $s^\chi(x;\theta)$ the *χ-score function* of $p(x;\theta)$. Using the χ-score function, we define a $(0,2)$-tensor field g^M on S_χ by

$$g^M_{ij}(\theta) := \int_\Omega \partial_i p(x;\theta)\partial_j \log_\chi p(x;\theta)\,dx, \quad \left(\partial_i = \frac{\partial}{\partial\theta^i}\right). \tag{3.13}$$

Lemma 1 *The tensor field g^M on S_χ is semi-positive definite.*

Proof From the definitions of g^M and $\log_\chi$, the tensor field g^M is written as

$$g^M_{ij}(\theta) = \int_\Omega \chi(p(x;\theta))\,(F_i(x)-\partial_i\psi(\theta))\left(F_j(x)-\partial_j\psi(\theta)\right)dx. \tag{3.14}$$

Since χ is strictly increasing, g^M is semi-positive definite. □

From now on, we assume that g^M is positive definite. Hence g^M is a Riemannian metric on S_χ. This assumption is same as in the case of Fisher metric. The Riemannian metric g^M is a generalization of the Fisher metric in terms of the representation (3.8).

We can consider other types of generalizations of the Fisher metric as follows.

$$\begin{aligned} g^E_{ij}(\theta) &:= \int_\Omega \left(\partial_i \log_\chi p(x;\theta)\right)\left(\partial_j \log_\chi p(x;\theta)\right) P_\chi(x;\theta)dx \\ &= E_{\chi,p}[\partial_i l_\chi(\theta)\partial_j l_\chi(\theta)], \\ g^N_{ij}(\theta) &:= \int_\Omega \frac{1}{P_\chi(x;\theta)} \left(\partial_i p(x;\theta)\right)\left(\partial_j p(x;\theta)\right) dx, \end{aligned}$$

where $l_\chi(\theta) = \log_\chi p(x;\theta)$. Obviously, g^E and g^N are generalizations of the Fisher metic with respect to the representations (3.7) and (3.9), respectively.

Proposition 4 *Let S_χ be a deformed exponential family. Then Riemannian metrics g^E, g^M and g^N are mutually conformally equivalent. In particular, the following formulas hold:*

$$Z_\chi(\theta) g^E(\theta) = g^M(\theta) = \frac{1}{Z_\chi(\theta)} g^N(\theta),$$

where $Z_\chi(\theta)$ is the normalization of the escort distribution $P_\chi(x;\theta)$.

Proof For a deformed exponential family S_χ, the differentials of probability density functions are given as follows:

$$\begin{aligned} \frac{\partial}{\partial\theta^i} p(x;\theta) &= \chi(p(x;\theta))\left(F_i(x) - \frac{\partial}{\partial\theta^i}\psi(\theta)\right), \\ \frac{\partial}{\partial\theta^i} \log_\chi p(x;\theta) &= F_i(x) - \frac{\partial}{\partial\theta^i}\psi(\theta). \end{aligned}$$

From the above formula and the definitions of Riemannian metrics g^E and g^N, we have

$$\begin{aligned} g^E_{ij}(\theta) &= \frac{1}{Z_\chi(\theta)} \int_\Omega \chi(p(x;\theta))\left(F_i(x) - \partial_i\psi(\theta)\right)\left(F_j(x) - \partial_j\psi(\theta)\right) dx, \\ g^N_{ij}(\theta) &= Z_\chi(\theta) \int_\Omega \chi(p(x;\theta))\left(F_i(x) - \partial_i\psi(\theta)\right)\left(F_j(x) - \partial_j\psi(\theta)\right) dx. \end{aligned}$$

These equations and Eq. (3.14) imply that Riemannian metrics g^E, g^M and g^N are mutually conformally equivalent. □

Among the three possibilities of generalizations of the Fisher metric, g^M is especially associated with a Hessian structure on S_χ, as we will see below. Although the

meaning of g^E is unknown, g^N gives a kind of Cramér-Rao lower bound in statistical inferences. (See [22, 23].)

By differentiating Eq. (3.13), we can define mutually dual affine connections $\nabla^{M(e)}$ and $\nabla^{M(m)}$ on S_χ by

$$\Gamma_{ij,k}^{M(e)}(\theta) := \int_\Omega \partial_k p(x;\theta)\partial_i\partial_j \log_\chi p(x;\theta)dx,$$
$$\Gamma_{ij,k}^{M(m)}(\theta) := \int_\Omega \partial_i\partial_j p(x;\theta)\partial_k \log_\chi p(x;\theta)dx.$$

From the definitions of the deformed exponential family and the deformed logarithm function, $\Gamma_{ij,k}^{M(e)}$ vanishes identically. Hence the connection $\nabla^{M(e)}$ is flat, and $(\nabla^{M(e)}, g^M)$ is a Hessian structure on S_χ. Denote by C^M the cubic form of $(S_\chi, \nabla^{M(e)}, g^M)$, that is,

$$C_{ijk}^M = \Gamma_{ij,k}^{M(m)} - \Gamma_{ij,k}^{M(e)} = \Gamma_{ij,k}^{M(m)}.$$

For $t > 0$, set a function $V_\chi(t)$ by

$$V_\chi(t) := \int_1^t \log_\chi(s)\, ds.$$

We assume that $V_\chi(0) = \lim_{t\to+0} V_\chi(t)$ is finite. Then the *generalized entropy functional* I_χ and the *generalized Massieu potential* Ψ are defined by

$$I_\chi(p_\theta) := -\int_\Omega \left\{V_\chi(p(x;\theta)) + (p(x;\theta) - 1)V_\chi(0)\right\} dx,$$
$$\Psi(\theta) := \int_\Omega p(x;\theta)\log_\chi p(x;\theta)dx + I_\chi(p_\theta) + \psi(\theta),$$

respectively, where ψ is the normalization of the deformed exponential family.

Theorem 2 (cf. [21, 26]) *For a deformed exponential family S_χ, the following hold:*

1. *$(S_\chi, \nabla^{M(e)}, g^M)$ and $(S_\chi, \nabla^{M(m)}, g^M)$ are mutually dual Hessian manifolds, that is, $(S_\chi, g^M, \nabla^{M(e)}, \nabla^{M(m)})$ is a dually flat space.*
2. *$\{\theta^i\}$ is a $\nabla^{M(e)}$-affine coordinate system on S_χ.*
3. *$\Psi(\theta)$ is the potential of g^M and C^M with respect to $\{\theta^i\}$, that is,*

$$g_{ij}^M(\theta) = \partial_i\partial_j\Psi(\theta),$$
$$C_{ijk}^M(\theta) = \partial_i\partial_j\partial_k\Psi(\theta).$$

4. *Set the expectation of $F_i(x)$ by $\eta_i := E_p[F_i(x)]$. Then $\{\eta_i\}$ is a $\nabla^{M(m)}$-affine coordinate system on S_χ and the dual of $\{\theta^i\}$ with respect to g^M.*
5. *Set $\Phi(\eta) := -I_\chi(p_\theta)$. Then $\Phi(\eta)$ is the potential of g^M with respect to $\{\eta_i\}$.*

Let us construct a divergence function which induces the Hessian manifold $(S_\chi, \nabla^{M(e)}, g^M)$. We define the *bias corrected χ-score function* $u_p^\chi(x;\theta)$ of $p(x;\theta)$ by

$$(u_p^\chi)^i(x;\theta) := \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta) - E_p\left[\frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta)\right].$$

Set a function $U_\chi(t)$ by

$$U_\chi(s) := \int_0^s \exp_\chi(t)\,dt.$$

Then we have

$$\begin{aligned} V_\chi(s) &= s\log_\chi(s) - \int_1^s t\left(\frac{d}{dt}\log_\chi(t)\right)dt \\ &= s\log_\chi(s) - \int_0^{\log_\chi(s)} \exp_\chi(u)du \\ &= s\log_\chi(s) - U_\chi(\log_\chi(s)). \end{aligned}$$

Since $\partial/\partial\theta^i V_\chi(p(x;\theta)) = (\partial/\partial\theta^i p(x;\theta))\log_\chi p(x;\theta)$, we have

$$p(x;\theta)\left(\frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta)\right) = \frac{\partial}{\partial\theta^i}U_\chi(\log_\chi p(x;\theta)).$$

Hence, by integrating the bias corrected χ-score function at $r(x;\theta) \in S_\chi$ with respect to θ, and by taking the standard expectation with respect to $p(x;\theta)$, we define a *χ-cross entropy of Bregman type* by

$$d_\chi^M(p,r) = -\int_\Omega p(x)\log_\chi r(x)dx + \int_\Omega U_\chi(\log_\chi r(x))dx.$$

Then we obtain the *χ-divergence* (or *U-divergence*) by

$$\begin{aligned} D_\chi(p,r) &= -d_\chi^M(p,p) + d_\chi^M(p,r) \\ &= \int_\Omega \big\{U_\chi(\log_\chi r(x)) - U_\chi(\log_\chi p(x)) \\ &\qquad -p(x)(\log_\chi r(x) - \log_\chi p(x))\big\}\,dx. \end{aligned}$$

In the q-exponential case, the *bias corrected q-score function* is given by

$$\begin{aligned} u_q^i(x;\theta) &= \frac{\partial}{\partial\theta^i}\log_q p(x;\theta) - E_p\left[\frac{\partial}{\partial\theta^i}\log_q p(x;\theta)\right] \\ &= \frac{\partial}{\partial\theta^i}\left\{\frac{1}{1-q}p(x;\theta)^{1-q} - \frac{1}{2-q}\int_\Omega p(x;\theta)^{2-q}dx\right\} \\ &= p(x;\theta)^{1-q}s^i(x;\theta) - E_p[p(x;\theta)^{1-q}s^i(x;\theta)]. \end{aligned}$$

This score function is nothing but a weighted score function in robust statistics. The χ-divergence constructed from the bias corrected q-score function coincides with the β-*divergence* ($\beta = 1-q$):

$$\begin{aligned} D_{1-q}(p,r) &= -d_{1-q}(p,p) + d_{1-q}(p,r) \\ &= \frac{1}{(1-q)(2-q)}\int_\Omega p(x)^{2-q}dx \\ &\quad -\frac{1}{1-q}\int_\Omega p(x)r(x)^{1-q}dx + \frac{1}{2-q}\int_\Omega r(x)^{2-q}dx. \end{aligned}$$

3.6 Geometry of Deformed Exponential Families Derived from the χ-Expectation

Since S_χ is linearizable by the deformed logarithm function, we can naturally define geometric structures from the potential function ψ.

A χ-*Fisher metric* g^χ and a χ-*cubic form* C^χ are defined by

$$\begin{aligned} g^\chi_{ij}(\theta) &:= \partial_i\partial_j\psi(\theta), \\ C^\chi_{ijk}(\theta) &:= \partial_i\partial_j\partial_k\psi(\theta), \end{aligned}$$

respectively [3]. In the q-exponential case, we denote the χ-Fisher metric by g^q, and the χ-cubic form by C^q. We call g^q and C^q a q-*Fisher metric* and a q-*cubic form*, respectively.

Let $\nabla^{\chi(0)}$ be the Levi-Civita connection with respect to the χ-Fisher metric g^χ. Then a χ-*exponential connection* $\nabla^{\chi(e)}$ and a χ-*mixture connection* $\nabla^{\chi(m)}$ are defined by

$$\begin{aligned} g^\chi(\nabla^{\chi(e)}_X Y, Z) &:= g^\chi(\nabla^{\chi(0)}_X Y, Z) - \frac{1}{2}C^\chi(X,Y,Z), \\ g^\chi(\nabla^{\chi(m)}_X Y, Z) &:= g^\chi(\nabla^{\chi(0)}_X Y, Z) + \frac{1}{2}C^\chi(X,Y,Z), \end{aligned}$$

respectively. The following theorem is known in [3].

Theorem 3 (cf. [3]) *For a deformed exponential family S_χ, the following hold:*

1. *$(S_\chi, \nabla^{\chi(e)}, g^\chi)$ and $(S_\chi, \nabla^{\chi(m)}, g^\chi)$ are mutually dual Hessian manifolds, that is, $(S_\chi, g^\chi, \nabla^{\chi(e)}, \nabla^{\chi(m)})$ is a dually flat space.*
2. *$\{\theta^i\}$ is a $\nabla^{\chi(e)}$-affine coordinate system on S_χ.*
3. *$\psi(\theta)$ is the potential of g^χ and C^χ with respect to $\{\theta^i\}$.*
4. *Set the χ-expectation of $F_i(x)$ by $\eta_i := E_{\chi,p}[F_i(x)]$. Then $\{\eta_i\}$ is a $\nabla^{\chi(m)}$-affine coordinate system on S_χ and the dual of $\{\theta^i\}$ with respect to g^χ.*
5. *Set $\phi(\eta) := E_{\chi,p}[\log_\chi p(x;\theta)]$. Then $\phi(\eta)$ is the potential of g^χ with respect to $\{\eta_i\}$.*

Proof Statements 1, 2 and 3 are easily obtained from the definitions of χ-Fisher metric and χ-cubic form. From Eq. (3.3) and $\eta_i = E_{\chi,p}[F_i(x)]$, Statements 4 and 5 follow from the fact that

$$E_{\chi,p}[\log_\chi p(x;\theta)] = E_{\chi,p}\left[\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right] = \sum_{i=1}^n \theta^i \eta_i - \psi(\theta). \quad \square$$

Suppose that $s^\chi(x;\theta)$ is the χ-score function defined by (3.12). The χ-score is unbiased with respect to χ-expectation, that is, $E_{\chi,p}[(s^\chi)^i(x;\theta)] = 0$. Hence we regard that $s^\chi(x;\theta)$ is a generalization of unbiased estimating functions.

By integrating a χ-score function, we define the *χ-cross entropy* by

$$\begin{aligned} d^\chi(p,r) &:= -E_{\chi,p}[\log_\chi r(x)] \\ &= -\int_\Omega P(x)\log_\chi r(x)dx. \end{aligned}$$

Then we obtain the *generalized relative entropy* $D^\chi(p,r)$ by

$$\begin{aligned} D^\chi(p,r) &:= -d^\chi(p,p) + d^\chi(p,r) \\ &= E_{\chi,p}[\log_\chi p(x) - \log_\chi r(x)]. \end{aligned} \tag{3.15}$$

The generalized relative entropy $D^\chi(p,r)$ coincides with the canonical divergence $D(r,p)$ for $(S_\chi, \nabla^{\chi(e)}, g^\chi)$. In fact, from (3.15), we can check that

$$\begin{aligned} D^\chi(p(\theta), p(\theta')) &= E_{\chi,p}\left[\left(\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right) - \left(\sum_{i=1}^n (\theta')^i F_i(x) - \psi(\theta')\right)\right] \\ &= \psi(\theta') + \sum_{i=1}^n \theta^i \eta_i - \psi(\theta) - \sum_{i=1}^n (\theta')^i \eta_i \;=\; D(p(\theta'), p(\theta)). \end{aligned}$$

Let us consider the q-exponential case. We assume that a q-exponential family S_q admits an invariant statistical manifold structure $(S_q, \nabla^{(\alpha)}, g^F)$.

Theorem 4 ([20]) *For a q-exponential family S_q, the invariant statistical manifold $(S_q, \nabla^{(2q-1)}, g^F)$ and the Hessian manifold $(S_q, \nabla^{q(e)}, g^q)$ are 1-conformally equivalent. In this case, the invariant statistical manifold $(S_q, \nabla^{(2q-1)}, g^F)$ is 1-conformally flat.*

Divergence functions for $(S_q, \nabla^{q(e)}, g^q)$ and $(S_q, \nabla^{(2q-1)}, g^F)$ are given as follows. The *α-divergence* $D^{(\alpha)}(p, r)$ with $\alpha = 1 - 2q$ is defined by

$$D^{(1-2q)}(p, r) := \frac{1}{q(1-q)} \left\{ 1 - \int_{\Omega} p(x)^q r(x)^{1-q} dx \right\}.$$

On the other hand, the *normalized Tsallis relative entropy* $D_q^T(p, r)$ is defined by

$$\begin{aligned} D_q^T(p, r) &:= \int_{\Omega} P_q(x) \left(\log_q p(x) - \log_q r(x) \right) dx \\ &= E_{q,p}[\log_q p(x) - \log_q r(x)]. \end{aligned}$$

We remark that the invariant statistical manifold $(S_q, \nabla^{(1-2q)}, g^F)$ is induced from the α-divergence with $\alpha = 1 - 2q$, and that the Hessian manifold $(S_q, \nabla^{q(e)}, g^q)$ is induced from the dual of the normalized Tsallis relative entropy. In fact, for a q-exponential family S_q, divergence functions have the following relations:

$$\begin{aligned} D(r, p) &= D_q^T(p, r) \\ &= \int_{\Omega} \frac{p(x)^q}{Z_q(p)} \left(\log_q p(x) - \log_q r(x) \right) dx \\ &= \frac{1}{Z_q(p)} \int_{\Omega} \left(\frac{p(x) - p(x)^q}{1-q} - \frac{p(x)^q r(x)^{1-q} - p(x)^q}{1-q} \right) dx \\ &= \frac{1}{(1-q) Z_q(p)} \left\{ 1 - \int_{\Omega} p(x)^q r(x)^{1-q} dx \right\} \\ &= \frac{q}{Z_q(r)} D^{(1-2q)}(p, r), \end{aligned}$$

where D is the canonical divergence of the Hessian manifold $(S_q, \nabla^{q(e)}, g^q)$.

3.7 Maximum q-Likelihood Estimators

In this section, we generalize the maximum likelihood method from the viewpoint of generalized independence. To avoid complicated arguments, we restrict ourselves to consider the q-exponential case. However, we can generalize it to the χ-exponential case (cf. [8, 9]).

Let X and Y be random variables which follow probability distributions $p_1(x)$ and $p_2(y)$, respectively. We say that two random variables X and Y are *independent* if the joint probability $p(x, y)$ is decomposed by a product of marginal distributions $p_1(x)$ and $p_2(Y)$:

$$p(x, y) = p_1(x)p_2(y).$$

When $p_1(x) > 0$ and $p_2(y) > 0$, the independence can be written with an exponential function and a logarithm function by

$$p(x, y) = \exp\left[\log p_1(x) + \log p_2(x)\right].$$

We generalize the notion of independence using the q-exponential and q-logarithm. Suppose that $x > 0$, $y > 0$ and $x^{1-q} + y^{1-q} - 1 > 0$ ($q > 0$). We say that $x \otimes_q y$ is a *q-product* [6] of x and y if

$$\begin{aligned} x \otimes_q y &:= \left[x^{1-q} + y^{1-q} - 1\right]^{\frac{1}{1-q}} \\ &= \exp_q\left[\log_q x + \log_q y\right]. \end{aligned}$$

In this case, the following low of exponents holds:

$$\exp_q x \otimes_q \exp_q y = \exp_q(x + y),$$

in other words,

$$\log_q(x \otimes_q y) = \log_q x + \log_q y.$$

Let X_i be a random variable on $\mathcal{X}_i$ which follows $p_i(x)$ $(i = 1, 2, \ldots, N)$. We say that $X_1, X_2, \ldots, X_N$ are *q-independent with m-normalization* (mixture normalization) if

$$p(x_1, x_2, \ldots, x_N) = \frac{p_1(x_1) \otimes_q p_2(x_2) \otimes_q \cdots \otimes_q p_N(x_N)}{Z_{p_1, p_2, \cdots, p_N}}$$

where $p(x_1, x_2, \ldots, x_N)$ is the joint probability density of $X_1, X_2, \ldots, X_N$ and $Z_{p_1,p_2,\cdots,p_N}$ is the normalization of $p_1(x_1) \otimes_q p_2(x_2) \otimes_q \cdots \otimes_q p_N(x_N)$ defined by

$$Z_{p_1,p_2,\cdots,p_N} := \int \cdots \int_{\mathcal{X}_1 \cdots \mathcal{X}_N} p_1(x_1) \otimes_q p_2(x_2) \otimes_q \cdots \otimes_q p_N(x_N) dx_1 \cdots dx_N.$$

Let $S_q = \{p(x;\xi)|\xi \in \Xi\}$ be a q-exponential family, and let $\{x_1, \ldots, x_N\}$ be N-observations from $p(x;\xi) \in S_q$. We define a *q-likelihood function* $L_q(\xi)$ by

$$L_q(\xi) = p(x_1;\xi) \otimes_q p(x_2;\xi) \otimes_q \cdots \otimes_q p(x_N;\xi).$$

Equivalently, a *q-log-likelihood function* is given by

$$\log_q L_q(\xi) = \sum_{i=1}^{N} \log_q p(x_i;\xi).$$

In the case $q \to 1$, L_q is the standard likelihood function on Ξ.

The *maximum q-likelihood estimator* $\hat{\xi}$ is the maximizer of the q-likelihood functions, which is defined by

$$\hat{\xi} := \operatorname*{argmax}_{\xi\in\Xi} L_q(\xi) \quad \left(= \operatorname*{argmax}_{\xi\in\Xi} \log_q L_q(\xi)\right).$$

Let us consider geometry of maximum q-likelihood estimators. Let S_q be a q-exponential family. Suppose that $\{x_1, \ldots, x_N\}$ are N-observations generated from $p(x;\theta) \in S_q$.

The q-log-likelihood function is calculated as

$$\begin{aligned}\log_q L_q(\theta) = \sum_{j=1}^{N} \log_q p(x_j;\theta) &= \sum_{j=1}^{N} \left\{ \sum_{i=1}^{n} \theta^i F_i(x_j) - \psi(\theta) \right\} \\ &= \sum_{i=1}^{n} \theta^i \sum_{j=1}^{N} F_i(x_j) - N\psi(\theta).\end{aligned}$$

The q-log-likelihood equation is

$$\partial_i \log_q L_q(\theta) = \sum_{j=1}^{N} F_i(x_j) - N\partial_i\psi(\theta) = 0.$$

Thus, the maximum q-likelihood estimator for η is given by

$$\hat{\eta}_i = \frac{1}{N} \sum_{j=1}^{N} F_i(x_j).$$

On the other hand, the canonical divergence for $(S_q, \nabla^{q(e)}, g^q)$ can be calculated as

$$\begin{aligned} D_q^T(p(\hat{\eta}), p(\theta)) &= D(p(\theta), p(\hat{\eta})) \\ &= \psi(\theta) + \phi(\hat{\eta}) - \sum_{i=1}^{n} \theta^i \hat{\eta}_i \\ &= \phi(\hat{\eta}) - \frac{1}{N} \log_q L_q(\theta). \end{aligned}$$

This implies that the q-likelihood attains the maximum if and only if the normalized Tsallis relative entropy attains the minimum.

Let M be a *curved q-exponential family* in S_q, that is, M is a submanifold in S_q and is a statistical model itself. Suppose that $\{x_1, \ldots, x_N\}$ are N-observations generated from $p(x; u) = p(x; \theta(u)) \in M$. The above arguments implies that the maximum q-likelihood estimator for M is given by the orthogonal projection of data with respect to the normalized Tsallis relative entropy.

We remark that the maximum q-likelihood estimator can be generalized by U-geometry. (See [8, 9] by Fujimoto and Murata.) However, their approach and ours are slightly different. They applied the χ-divergence (U-divergence) projection for a parameter estimation, whereas we applied the generalized relative entropy. As we discussed in this paper, the induced Hessian structures from those divergences are different.

3.8 Conclusion

In this paper, we considered two Hessian structures from the viewpoints of the standard expectation and the χ-expectation. Though the former and the later are known as U-geometry ([21, 26]) and χ-geometry ([3]), respectively, they turn out to be different Hessian structures in the same deformed exponential family through a comparison of each other.

We note that, from the viewpoint of estimating functions, the former is geometry of bias-corrected χ-score functions with the standard expectation, whereas the later is geometry of unbiased χ-score functions with the χ-expectation.

As an application to statistics, we considered generalization of maximum likelihood method for q-exponential family. We used the normalized Tsallis relative entropy for orthogonal projection, whereas the previous results used χ-divergences of Bregman type.

Acknowledgments The authors would like to express their sincere gratitude to the anonymous reviewers for constructive comments for preparation of this paper. The first named author is partially supported by JSPS KAKENHI Grant Number 23740047.

References

1. Amari, S., Nagaoka, H.: Method of Information Geometry. American Mathematical Society, Providence, Oxford University Press, Oxford (2000)
2. Amari, S., Ohara, A.: Geometry of q-exponential family of probability distributions. Entropy **13**, 1170–1185 (2011)
3. Amari, S., Ohara, A., Matsuzoe, H.: Geometry of deformed exponential families: invariant, dually-flat and conformal geometry. Phys. A. **391**, 4308–4319 (2012)
4. Barondorff-Nielsen, O.E., Jupp, P.E.: Statistics, yokes and symplectic geometry. Ann. Facul. Sci. Toulouse **6**, 389–427 (1997)
5. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**, 549–559 (1998)
6. Borgesa, E.P.: A possible deformed algebra and calculus inspired in nonextensive thermostatistics. Phys. A **340**, 95–101 (2004)
7. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**, 631–647 (1992)
8. Fujimoto, Y., Murata, N.: A generalization of independence in naive bayes model. Lect. Notes Comp. Sci. **6283**, 153–161 (2010)
9. Fujimoto Y., Murata N.: A generalisation of independence in statistical models for categorical distribution. Int. J. Data Min. Model. Manage. **2**(4), 172–187 (2012)
10. Ivanov, S.: On dual-projectively flat affine connections. J. Geom. **53**, 89–99 (1995)
11. Kumon, M., Takemura, A., Takeuchi, K.: Conformal geometry of statistical manifold with application to sequential estimation. Sequential Anal. **30**, 308–337 (2011)
12. Kurose, T.: On the divergences of 1-conformally flat statistical manifolds. Tôhoku Math. J. **46**, 427–433 (1994)
13. Kurose, T.: Conformal-projective geometry of statistical manifolds. Interdiscip. Inform. Sci. **8**, 89–100 (2002)
14. Lauritzen, S. L.: Statistical Manifolds, Differential Geometry in Statistical Inferences, IMS Lecture Notes Monograph Series, vol. 10, pp. 96–163. Hayward, California (1987)
15. Matsuzoe, H.: Geometry of contrast functions and conformal geometry. Hiroshima Math. J. **29**, 175–191 (1999)
16. Matsuzoe, H.: Geometry of statistical manifolds and its generalization. In: Proceedings of the 8th International Workshop on Complex Structures and Vector Fields, pp. 244–251. World Scientific, Singapore (2007)
17. Matsuzoe, H.: Computational geometry from the viewpoint of affine differential geometry. Lect. Notes Comp. Sci. **5416**, 103–113 (2009)
18. Matsuzoe, H.: Statistical manifolds and geometry of estimating functions, pp. 187–202. Recent Progress in Differential Geometry and Its Related Fields World Scientific, Singapore (2013)
19. Matsuzoe, H., Henmi, M.: Hessian structures on deformed exponential families. Lect. Notes Comp. Sci. **8085**, 275–282 (2013)
20. Matsuzoe, H., Ohara, A.: Geometry for q-exponential families. In: Recent progress in differential geometry and its related fields, pp. 55–71. World Scientific, Singapore (2011)
21. Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S.: Information geometry of u-boost and bregman divergence. Neural Comput. **16**, 1437–1481 (2004)
22. Naudts, J.: Estimators, escort probabilities, and ϕ-exponential families in statistical physics. J. Ineq. Pure Appl. Math. **5**, 102 (2004)
23. Naudts, J.: Generalised Thermostatistics, Springer, New York (2011)
24. Ohara, A.: Geometric study for the legendre duality of generalized entropies and its application to the porous medium equation. Euro. Phys. J. B. **70**, 15–28 (2009)
25. Ohara, A., Matsuzoe H., Amari S.: Conformal geometry of escort probability and its applications. Mod. Phys. Lett. B. **10**, 26:1250063 (2012)
26. Ohara A., Wada, T.: Information geometry of q-Gaussian densities and behaviors of solutions to related diffusion equations. J. Phys. A: Math. Theor. **43**, 035002 (2010)
27. Okamoto, I., Amari, S., Takeuchi, K.: Asymptotic theory of sequential estimation procedures for curved exponential families. Ann. Stat. **19**, 961–961 (1991)

28. Shima, H.: The Geometry of Hessian Structures, World Scientific, Singapore (2007)
29. Suyari, H., Tsukada, M.: Law of error in tsallis statistics. IEEE Trans. Inform. Theory **51**, 753–757 (2005)
30. Takatsu, A.: Behaviors of φ-exponential distributions in wasserstein geometry and an evolution equation. SIAM J. Math. Anal. **45**, 2546–2546 (2013)
31. Tanaka, M.: Meaning of an escort distribution and τ-transformation. J. Phys.: Conf. Ser. **201**, 012007 (2010)
32. Tsallis, C.: Possible generalization of boltzmann—gibbs statistics. J. Stat. Phys. **52**, 479–487 (1988)
33. Tsallis, C.: Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World. Springer, New York (2009)
34. Vigelis, R.F., Cavalcante, C.C.: On ϕ-families of probability distributions. J. Theor. Probab. **21**, 1–25 (2011)

Chapter 4
Harmonic Maps Relative to α-Connections

Keiko Uohashi

Abstract In this paper, we study harmonic maps relative to α-connections, but not necessarily relative to Levi-Civita connections, on Hessian domains. For the purpose, we review the standard harmonic map and affine harmonic maps, and describe the conditions for harmonicity of maps between level surfaces of a Hessian domain in terms of the parameter α and the dimension n. To illustrate the theory, we describe harmonic maps between the level surfaces of convex cones.

4.1 Introduction

Harmonic maps are important objects in certain branches of geometry and physics. Geodesics on Riemannian manifolds and holomorphic maps between Kähler manifolds are typical examples of harmonic maps. In addition a harmonic map has a variational characterization by the energy of smooth maps between Riemannian manifolds and several existence theorems for harmonic maps are already known. On the other hand the notion of a Hermitian harmonic map from a Hermitian manifold to a Riemannian manifold was introduced and investigated by [4, 8, 10]. It is not necessary a harmonic map if the domain Hermitian manifold is non-Kähler. The similar results are pointed out for affine harmonic maps, which is analogy to Hermitian harmonic maps [7].

Statistical manifolds have mainly been studied in terms of their affine geometry, information geometry, and statistical mechanics [1]. For example, Shima established conditions for harmonicity of gradient mappings of level surfaces on a Hessian domain, which is a typical example of a dually flat statistical manifold [14]. Level surfaces on a Hessian domain are known as 1- and (-1)-conformally flat statistical

K. Uohashi (✉)
Department of Mechanical Engineering and Intelligent Systems, Faculty of Engineering, Tohoku Gakuin University, 1-13-1 Chuo, Tagajo, Miyagi 985-8537, Japan
e-mail: uohashi@mail.tohoku-gakuin.ac.jp

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_4,

manifolds for primal and dual connections, respectively [17, 19]. The gradient mappings are then considered to be harmonic maps relative to the dual connection, i.e., the (-1)-connection [13].

In this paper, we review the notions of harmonic maps, affine harmonic maps and α-affine harmonic maps, and investigate different kinds of harmonic maps relative to α-connections. In Sect. 4.2, we give definitions of an affine harmonic map, a harmonic map and the standard Laplacian. In Sect. 4.3, we explain the generalized Laplacian which defines a harmonic map relative to an affine connection. In Sect. 4.4, we present the Laplacian of a gradient mapping on a Hessian domain, as an example of the generalized Laplacian. Moreover, we compare the harmonic map defined by Shima with an affine harmonic map defined in Sect. 4.2. In Sect. 4.5, α-connections of statistical manifolds are explained. In Sect. 4.6, we define α-affine harmonic maps which are generalization of affine harmonic maps and also a generalization of harmonic maps defined by Shima. In Sect. 4.7, we describe the α-conformal equivalence of statistical manifolds and a harmonic map relative to two α-connections. In Sect. 4.8, we review α-conformal equivalence of level surfaces of a Hessian domain. In Sect. 4.9, we study harmonic maps of level surfaces relative to two α-connections, for examples of a harmonic map in Sect. 4.7, and provide examples on level surfaces of regular convex cones.

Shima [13] investigated harmonic maps of n-dimensional level surfaces into an $(n+1)$-dimensional dual affine space, rather than onto other level surfaces. Although Nomizu and Sasaki calculated the Laplacian of centro-affine immersions into an affine space, which generate projectively flat statistical manifolds (i.e. (-1)-conformally flat statistical manifolds), they did not discuss any harmonic maps between two centro-affine hypersurfaces [12]. Then, we study harmonic maps between hypersurfaces with the same dimension relative to general α-connections that may not satisfy $\alpha = -1$ or 0 (where the 0-connection implies the Levi-Civita connection). In particular, we demonstrate the existence of non-trivial harmonic maps between level surfaces of a Hessian domain with α-parameters and the dimension n.

4.2 Affine Harmonic Maps and Harmonic Maps

First, we recall definitions of an affine harmonic map and a harmonic map.

Let M an m-dimensional affine manifold and $\{x^1, \ldots, x^m\}$ a local affine coordinate system of M. If there exist a symmetric tensor field of degree 2

$$g = g_{ij} dx^i dx^j$$

on M satisfying locally

$$g_{ij} = \frac{\partial^2 \varphi}{\partial x^i \partial x^j} \tag{4.1}$$

for a convex function φ, M is said to be a *Kähler affine manifold* [2, 7]. A matrix $[g_{ij}]$ is positive definite and defines a Riemannian metric. Then for the Kähler affine manifold M, (M, D, g) is a Hessian manifold, where D is a canonical flat affine connection for $\{x^1, \ldots, x^m\}$. We will mention details of Hessian manifolds and Hessian domains in later sections of this paper.

The *Kähler affine structure* (4.1) defines an affinely invariant operator L by

$$L = \sum_{i,j=1}^{m} g_{ij} \frac{\partial^2}{\partial x^i \partial x^j}. \tag{4.2}$$

A smooth function $f : M \to \mathbf{R}$ is said to be *affine harmonic* if

$$Lf = 0.$$

For a Kähler affine manifold (M, g) and a Riemannian manifold (N, h), a smooth map $\phi : M \to N$ is said to be *affine harmonic* if

$$\sum_{i,j=1}^{m} g^{ij} \left(\frac{\partial^2 \phi^\gamma}{\partial x^i \partial x^j} + \sum_{\delta,\beta=1}^{n} \hat{\Gamma}^\gamma_{\delta\beta} \frac{\partial \phi^\delta}{\partial x^i} \frac{\partial \phi^\beta}{\partial x^j} \right) = 0, \quad \gamma = 1, \ldots, n, \tag{4.3}$$

where $\hat{\Gamma}$ is the Christoffel symbol of the Levi-Civita connection for a Riemannian metric h, and $n = \dim N$.

Let us compare an affine harmonic map with a harmonic map. For this purpose, we give a definition of a harmonic function at first. For a Riemannian manifold (M, g), a smooth function $f : M \to \mathbf{R}$ is said to be a *harmonic function* if

$$\Delta f = 0,$$

where Δ is the standard *Laplacian*, i.e.,

$$\Delta f = \operatorname{div} \operatorname{grad} f = \frac{1}{\sqrt{g}} \sum_{i,j=1}^{m} \frac{\partial}{\partial x^i} \left(\sqrt{g} g^{ij} \frac{\partial f}{\partial x^j} \right) \tag{4.4}$$

$$= \sum_{i,j=1}^{m} g^{ij} \left(\frac{\partial^2 f}{\partial x^i \partial x^j} - \sum_{k=1}^{m} \Gamma^k_{ij} \frac{\partial f}{\partial x^k} \right) \tag{4.5}$$

$$= \sum_{i=1}^{m} \{ e_i(e_i f) - (\nabla^{LC}_{e_i} e_i) f \},$$

$$g = \det[g_{ij}],$$

$\{e_1, \ldots, e_m\}$ is a local orthogonal frame on a neighborhood of $x \in M$, and ∇^{LC}, Γ are the Levi-Civita connection, the Christoffel symbol of ∇^{LC}, respectively. Remark that the sign of definition (4.4) is inverse to the sign of the Laplacian in [3, 21].

For Riemannian manifolds (M, g), (N, h), a smooth map $\phi : M \to N$ is said to be a *harmonic map* if

$$\tau(\phi) \equiv 0; \quad \text{the } \textit{Euler-Lagrange equation},$$

where $\tau(\phi) \in \Gamma(\phi^{-1}TN)$ is the standard tension field of ϕ defined by

$$\begin{aligned} \tau(\phi)(x) &= \sum_{i=1}^{m} (\tilde{\nabla}^{LC}_{e_i} \phi_* e_i - \phi_* \nabla^{LC}_{e_i} e_i)(x), \quad x \in M, \\ \tilde{\nabla}^{LC}_{e_i} \phi_* e_i &= \hat{\nabla}^{LC}_{\phi_* e_i} \phi_* e_i; \quad \text{the pull-back connection}, \end{aligned} \tag{4.6}$$

and ∇^{LC}, $\hat{\nabla}^{LC}$ are the Levi-Civita connections for g, h, respectively. For local coordinate systems $\{x^1, \ldots, x^m\}$ and $\{y^1, \ldots, y^n\}$ on M and N, the γ-th component of $\tau(\phi)$ at $x \in M$ is described by

$$\begin{aligned} \tau(\phi)^\gamma(x) &= \sum_{i,j=1}^{m} g^{ij} \left\{ \frac{\partial^2 \phi^\gamma}{\partial x^i \partial x^j} - \sum_{k=1}^{m} \Gamma^k_{ij}(x) \frac{\partial \phi^\gamma}{\partial x^k} + \sum_{\delta,\beta=1}^{n} \hat{\Gamma}^\gamma_{\delta\beta}(\phi(x)) \frac{\partial \phi^\delta}{\partial x^i} \frac{\partial \phi^\beta}{\partial x^j} \right\} \\ &= \Delta \phi^\gamma + \sum_{i,j=1}^{m} \sum_{\delta,\beta=1}^{n} g^{ij} \hat{\Gamma}^\gamma_{\delta\beta}(\phi(x)) \frac{\partial \phi^\delta}{\partial x^i} \frac{\partial \phi^\beta}{\partial x^j}, \\ \phi^\delta &= y^\delta \circ \phi, \quad \gamma = 1, \ldots, n, \end{aligned} \tag{4.7}$$

where

$$\tau(\phi)(x) = \sum_{\gamma=1}^{n} \tau(\phi)^\gamma(x) \frac{\partial}{\partial y^\gamma},$$

and Γ^k_{ij}, $\hat{\Gamma}^\gamma_{\delta\beta}$ are the Christoffel symbols of ∇^{LC}, $\hat{\nabla}^{LC}$, respectively. The original definition of a harmonic map is described in [3, 21], and so on.

Remark 1 Term (4.5) is not equal to the definition (4.2). Hence an affine harmonic function is not necessary a harmonic function.

Remark 2 Term (4.7) is not equal to the definition (4.3). Hence an affine harmonic map is not necessary a harmonic map.

4.3 Affine Harmonic Maps and Generalized Laplacians

In Sect. 4.2, the Laplacian is defined for a function on a Riemannian manifold. In this section, we treat Laplacians for maps between Riemannian manifolds.

For Riemannian manifolds (M, g) and (N, h), a tension field of a smooth map $\phi : M \to N$ is defined by

$$\begin{aligned}\tau(\phi) &= \sum_{i=1}^{m}(\hat{\nabla}_{e_i}(\phi_* e_i) - \phi_*(\nabla^{LC}_{e_i} e_i)) \quad \in \Gamma(\phi^{-1}TN) \\ &= \sum_{i,j=1}^{m} g^{ij}\left\{\hat{\nabla}_{\frac{\partial}{\partial x^i}}\left(\phi_* \frac{\partial}{\partial x^j}\right) - \phi_*\left(\nabla^{LC}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}\right)\right\},\end{aligned} \tag{4.8}$$

where $\{e_1, \ldots, e_m\}$ is a local orthonormal frame for g, $\{x^1, \ldots, x^m\}$ is a local coordinate system on M, ∇^{LC} is the Levi-Civita connection of g, and $\hat{\nabla}$ is a torsion free affine connection on N [12]. The affine connection $\hat{\nabla}$ does not need to be the Levi-Civita connection. We also denote by $\hat{\nabla}$ the pull-back connection of $\hat{\nabla}$ to M. Then ϕ is said to be a *harmonic map relative to* $(g, \hat{\nabla})$ if

$$\tau(\phi) = \sum_{i=1}^{m}(\hat{\nabla}_{e_i}(\phi_* e_i) - \phi_*(\nabla^{LC}_{e_i} e_i)) \equiv 0.$$

If a Riemannian manifold N is an finite dimensional real vector space V, the tension field $\tau(\phi)$ is said to be a *Laplacian of a map* $\phi : M \to V$. Then a notation Δ for the standard Laplacian is often used for the Laplacian of a map as the following;

$$\Delta\phi = \Delta_{(g,\hat{\nabla})}\phi = \tau(\phi) : M \to V. \tag{4.9}$$

For $V = \mathbf{R}$, $\Delta\phi$ defined by Eqs. (4.8) and (4.9) coincides with the standard Laplacian for a function defined by (4.4).

See in [12] for an affine immersion and the Laplacian of a map, and see in [13, 14] for the gradient mapping and the Laplacian on a Hessian domain.

4.4 Gradient Mappings and Affine Harmonic Maps

In this section, we investigate the Laplacian of a gradient mapping in view of geometry of affine harmonic maps.

Let D be the canonical flat affine connection on an $(n+1)$-dimensional real affine space $\mathbf{A}^{n+1}$ and let $\{x^1, \ldots, x^{n+1}\}$ be the canonical affine coordinate system on $\mathbf{A}^{n+1}$, i.e., $Ddx^i = 0$. If the Hessian $Dd\varphi = \sum_{i,j=1}^{n+1}(\partial^2\varphi/\partial x^i \partial x^j)dx^i dx^j$ of a function φ is

non-degenerate on a domain Ω in $\mathbf{A}^{n+1}$, then $(\Omega, D, g = Dd\varphi)$ is a *Hessian domain* [14].

For the dual affine space $\mathbf{A}^*_{n+1}$ and the dual affine coordinate system $\{x^*_1, \ldots, x^*_{n+1}\}$ of $\mathbf{A}^{n+1}$, the *gradient mapping* ι from a Hessian domain $(\Omega, D, g = Dd\varphi)$ into $(\mathbf{A}^*_{n+1}, D^*)$ is defined by

$$x^*_i \circ \iota = -\frac{\partial \varphi}{\partial x^i}.$$

The dually flat affine connection D' on Ω is given by

$$\iota_*(D'_X Y) = D^*_X \iota_*(Y) \quad \text{for } X, Y \in \Gamma(T\Omega), \tag{4.10}$$

where $D^*_X \iota_*(Y)$ denotes the covariant derivative along ι induced by the canonical flat affine connection D^* on $\mathbf{A}^*_{n+1}$.

The Laplacian of ι with respect to (g, D^*) is given by

$$\begin{aligned}
\Delta_{(g,D^*)}\iota &= \sum_{i,j} g^{ij} \left\{ D^*_{\frac{\partial}{\partial x^i}} \left(\iota_* \frac{\partial}{\partial x^j} - \iota_* \left(\nabla^{LC}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) \right) \right\} \\
&= \iota_* \left\{ \sum_{i,j} g^{ij} (D' - \nabla^{LC})_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right\} \qquad (4.11) \\
&= \iota_* \left\{ \sum_{i,j} g^{ij} (\nabla^{LC} - D)_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right\} \qquad (4.12) \\
&= \iota_* \left(\sum_i \alpha^{Ki} \frac{\partial}{\partial x^i} \right),
\end{aligned}$$

where ∇^{LC} is the Levi-Civita connection for g and Γ is the Christoffel symbol of ∇^{LC}, and where α^K is the Koszul form, i.e.,

$$\alpha^K = d \log |det[g_{ij}]|^{\frac{1}{2}}, \quad \alpha^K_i = \sum_r \Gamma^r_{ri}, \ \alpha^{Ki} = \sum_j g^{ij} \alpha^K_j$$

([13], p. 93 in [14]). We have the deformation from (4.11) to (4.12) by

$$\frac{D + D'}{2} = \nabla^{LC}.$$

Details of dual affine connections are described in later sections.

Let ∇^{LC*} be the Levi-Civita connection for the Hessian metric $g^* = D^* d\varphi^*$ on the dual domain $\Omega^* = \iota(\Omega)$, where $\varphi^* = \sum_i x^i (\partial\varphi/\partial x^i) - \varphi$ is the *Legendre transform* of φ. Then the term (4.12) is described as the follows:

$$\Delta_{(g,D^*)}\iota = \sum_{i,j} g^{ij}\{\nabla^{LC*}_{\frac{\partial}{\partial x^i}}(\iota_*\frac{\partial}{\partial x^j}) - \iota_*(D_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j})\}.$$

Moreover the γ-th component of $\Delta_{(g,D^*)}\iota$ is

$$(\Delta_{(g,D^*)}\iota)^\gamma = \sum_{i,j} g^{ij}\left(\frac{\partial^2\iota^\gamma}{\partial x^i\partial x^j} + \sum_{\delta,\beta}\Gamma^\gamma_{\delta\beta}\frac{\partial\iota^\delta}{\partial x^i}\frac{\partial\iota^\beta}{\partial x^j}\right), \quad \gamma = 1,\ldots,n+1$$

where $\iota^i(x) = x^*_i \circ \iota(x)$. Therefore, if the gradient mapping ι is a harmonic map with respect to (g, D^*), i.e., if $\Delta_{(g,D^*)}\iota \equiv 0$, we have

$$\sum_{i,j} g^{ij}\left(\frac{\partial^2\iota^\gamma}{\partial x^i\partial x^j} + \sum_{\delta,\beta}\Gamma^\gamma_{\delta\beta}\frac{\partial\iota^\delta}{\partial x^i}\frac{\partial\iota^\beta}{\partial x^j}\right) = 0, \quad \gamma = 1,\ldots,n+1. \tag{4.13}$$

Equation (4.13) is obtained by putting ι on ϕ of Eq. (4.3). Thus considering a Hessian domain Ω as a Kähler affine manifold, we have the next proposition.

Proposition 1 *The followings are equivalent:*

(i) *the gradient mapping ι is a harmonic map with respect to (g, D^*);*
(ii) *the gradient mapping $\iota : (\Omega, D) \to (\mathbf{A}^*_{n+1}, \nabla^{LC*})$ is an affine harmonic map.*

In [13, 14], Shima studied an affine harmonic map with the restriction of the gradient mapping ι to a level surface of a convex function φ.

The author does not clearly distinguish a phrase "relative to something" with a phrase "with respect to something".

4.5 α-Connections of Statistical Manifolds

We recall some definitions that are essential to the theory of statistical manifolds and relate α-connections to Hessian domains.

Given a torsion-free affine connection ∇ and a pseudo-Riemannian metric h on a manifold N, the triple (N, ∇, h) is said to be a *statistical manifold* if ∇h is symmetric. If the curvature tensor R of ∇ vanishes, (N, ∇, h) is said to be *flat*.

Let (N, ∇, h) be a statistical manifold and let ∇' be an affine connection on N such that

$$Xh(Y, Z) = h(\nabla_X Y, Z) + h(Y, \nabla'_X Z) \text{ for } X, Y \text{ and } Z \in \Gamma(TN),$$

where $\Gamma(TN)$ is the set of smooth tangent vector fields on N. The affine connection ∇' is torsion free and $\nabla' h$ is symmetric. Then ∇' is called the *dual connection* of

∇. The triple (N, ∇', h) is the *dual statistical manifold* of (N, ∇, h), and (∇, ∇', h) defines the *dualistic structure* on N. The curvature tensor of ∇' vanishes if and only if the curvature tensor of ∇ also vanishes. Under these conditions, (∇, ∇', h) becomes a *dually flat structure*.

Let N be a manifold with a dualistic structure (∇, ∇', h). For any $\alpha \in \mathbf{R}$, an affine connection defined by

$$\nabla^{(\alpha)} := \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla' \tag{4.14}$$

is called an α-*connection* of (N, ∇, h). The triple $(N, \nabla^{(\alpha)}, h)$ is also a statistical manifold, and $\nabla^{(-\alpha)}$ is the dual connection of $\nabla^{(\alpha)}$. The 1-connection $\nabla^{(1)}$, the (-1)-connection $\nabla^{(-1)}$, and the 0-connection $\nabla^{(0)}$ correspond to the ∇, ∇', and the Levi-Civita connection of (N, h), respectively. An α-connection does not need to be flat.

A Hessian domain is a flat statistical manifold. Conversely, a local region of a flat statistical manifold is a Hessian domain. For the dual connection D' defined by (4.10), (Ω, D', g) is the dual statistical manifold of (Ω, D, g) if a Hessian domain (Ω, D', g) is a statistical manifold [1, 13, 14].

4.6 α-Affine Harmonic Maps

In this section, we give a generalization of an affine harmonic map.

Considering a Hessian domain (Ω, D, g) as a statistical manifold, we have the α-connection of (Ω, D, g) by

$$D^{(\alpha)} = \frac{1+\alpha}{2}D + \frac{1-\alpha}{2}D'$$

for each $\alpha \in \mathbf{R}$. Let $D^{(\alpha)*}$ be an α-connection of (Ω^*, D^*, g^*) which is the dual statistical manifold of (Ω, D, g). Then the *Laplacian of the gradient mapping* ι with respect to $(g, D^{(\alpha)*})$ is given by

$$\begin{aligned}
\Delta_{(g,D^{(\alpha)*})}\iota &= \sum_{i,j} g^{ij}\left\{D^{(\alpha)*}_{\frac{\partial}{\partial x^i}}\left(\iota_*\frac{\partial}{\partial x^j}\right) - \iota_*\left(\nabla^{LC}_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}\right)\right\} \\
&= \iota_*\left\{\sum_{i,j} g^{ij}(D^{(-\alpha)} - \nabla^{LC})_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}\right\} \\
&= \iota_*\left\{\sum_{i,j} g^{ij}(\nabla^{LC} - D^{(\alpha)})_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}\right\}
\end{aligned}$$

$$= \sum_{i,j} g^{ij} \left\{ \nabla^{LC*}_{\frac{\partial}{\partial x^i}} (\iota_* \frac{\partial}{\partial x^j}) - \iota_* \left(D^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) \right\}.$$

If $\Delta_{(g, D^{(\alpha)*})} \iota \equiv 0$, we have

$$\sum_{i,j} g^{ij} \left\{ \frac{\partial^2 \iota^\gamma}{\partial x^i \partial x^j} - (1-\alpha) \sum_k \Gamma^k_{ij} \frac{\partial \iota^\gamma}{\partial x^k} + \sum_{\delta,\beta} \hat{\Gamma}^\gamma_{\delta\beta} \frac{\partial \iota^\delta}{\partial x^i} \frac{\partial \iota^\beta}{\partial x^j} \right\} = 0,$$

$$\gamma = 1, \ldots, n+1.$$

In general, we define the notion of α-affine harmonic maps as follows:

Definition 1 For a Kähler affine manifold (M, g) and a Riemannian manifold (N, h), a map $\phi : M \to N$ is said to be an α-affine harmonic map if

$$\sum_{i,j} g^{ij} \left(\frac{\partial^2 \phi^\gamma}{\partial x^i \partial x^j} - (1-\alpha) \sum_k \Gamma^k_{ij} \frac{\partial \phi^\gamma}{\partial x^k} + \sum_{\delta,\beta} \hat{\Gamma}^\gamma_{\delta\beta} \frac{\partial \phi^\delta}{\partial x^i} \frac{\partial \phi^\beta}{\partial x^j} \right) = 0, \tag{4.15}$$

$$\gamma = 1, \ldots, \dim\ N.$$

Then we obtain that the gradient mapping ι is a harmonic map with respect to $(g, D^{(\alpha)*})$ if and only if the map $\iota : (\Omega, D^{(\alpha)}) \to (\mathbf{A}^*_{n+1}, \nabla^*)$ is an α-affine harmonic map.

Remark 3 For $\alpha = 1$, a 1-affine harmonic map is an affine harmonic map.

Remark 4 For $\alpha = 0$, a 0-affine harmonic map is a harmonic map in the standard sense.

They are problems to find applications of α-affine harmonic maps and to investigate them.

4.7 Harmonic Maps for α-Conformal Equivalence

In this section, we describe harmonic maps with respect to α-conformal equivalence of statistical manifolds.

For a real number α, statistical manifolds (N, ∇, h) and $(N, \bar{\nabla}, \bar{h})$ are regarded as α-*conformally equivalent* if there exists a function ϕ on N such that

$$\bar{h}(X, Y) = e^{\phi} h(X, Y), \tag{4.16}$$

$$h(\bar{\nabla}_X Y, Z) = h(\nabla_X Y, Z) - \frac{1+\alpha}{2} d\phi(Z) h(X, Y) + \frac{1-\alpha}{2} \{d\phi(X) h(Y, Z) + d\phi(Y) h(X, Z)\} \tag{4.17}$$

for X, Y and $Z \in \Gamma(TN)$. Two statistical manifolds (N, ∇, h) and $(N, \bar{\nabla}, \bar{h})$ are α-conformally equivalent if and only if the dual statistical manifolds (N, ∇', h) and $(N, \bar{\nabla}', \bar{h})$ are $(-\alpha)$-conformally equivalent. A statistical manifold (N, ∇, h) is said to be *α-conformally flat* if (N, ∇, h) is locally α-conformally equivalent to a flat statistical manifold [19].

Let (N, ∇, h) and $(N, \bar{\nabla}, \bar{h})$ be α-conformally equivalent statistical manifolds of $\dim n \geq 2$, and $\{x^1, \ldots x^n\}$ a local coordinate system on N. Suppose that h and $\bar{h}$ are Riemannian metrices. We set $h_{ij} = h(\partial/\partial x^i, \partial/\partial x^j)$ and $[h^{ij}] = [h_{ij}]^{-1}$. Let $\pi_{id} : (N, \nabla, h) \to (N, \bar{\nabla}, \bar{h})$ be the identity map, i.e., $\pi_{id}(x) = x$ for $x \in N$, and π_{id*} the differential of π_{id}.

We define a harmonic map relative to $(h, \nabla, \bar{\nabla})$ as follows:

Definition 2 ([16, 18]) If a tension field $\tau_{(h,\nabla,\bar{\nabla})}(\pi_{id})$ vanishes on N, i.e.,

$$\tau_{(h,\nabla,\bar{\nabla})}(\pi_{id}) \equiv 0,$$

the map $\pi_{id} : (N, \nabla, h) \to (N, \bar{\nabla}, \bar{h})$ is said to be a *harmonic map relative to* $(h, \nabla, \bar{\nabla})$, where the tension field is defined by

$$\begin{aligned} \tau_{(h,\nabla,\bar{\nabla})}(\pi_{id}) &:= \sum_{i,j=1}^{n} h^{ij} \left\{ \bar{\nabla}_{\frac{\partial}{\partial x^i}} \left(\pi_{id*} \left(\frac{\partial}{\partial x^j} \right) \right) - \pi_{id*} \left(\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) \right\} \in \Gamma(\pi_{id}^{-1} TN) \\ &= \sum_{i,j=1}^{n} h^{ij} (\bar{\nabla}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} - \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}) \in \Gamma(TN). \end{aligned} \tag{4.18}$$

Then the next theorem holds.

Theorem 1 ([16, 18]) *For α-conformally equivalent statistical manifolds (N, ∇, h) and $(N, \bar{\nabla}, \bar{h})$ of $\dim N \geq 2$ satisfying Eqs. (4.16) and (4.17), if $\alpha = -(n-2)/(n+2)$ or ϕ is a constant function on N, the identity map $\pi_{id} : (N, \nabla, h) \to (N, \bar{\nabla}, \bar{h})$ is a harmonic map relative to $(h, \nabla, \bar{\nabla})$.*

Proof By Eqs. (4.17) and (4.18), for $k \in \{1, \ldots, n\}$ we have

$$\begin{aligned} & h \left(\tau_{(h,\nabla,\bar{\nabla})} (\pi_{id}), \frac{\partial}{\partial x^k} \right) \\ &= h \left(\sum_{i,j=1}^{n} h^{ij} \left(\bar{\nabla}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} - \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right), \frac{\partial}{\partial x^k} \right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i,j=1}^{n} h^{ij} \left\{ -\frac{1+\alpha}{2} d\phi\left(\frac{\partial}{\partial x^k}\right) h\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) \right. \\
&\quad + \frac{1-\alpha}{2} \left\{ d\phi\left(\frac{\partial}{\partial x^i}\right) h\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\right) \right. \\
&\quad \left. \left. + d\phi\left(\frac{\partial}{\partial x^j}\right) h\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^k}\right) \right\} \right\} \\
&= \sum_{i,j=1}^{n} h^{ij} \left\{ -\frac{1+\alpha}{2} \frac{\partial \phi}{\partial x^k} h_{ij} + \frac{1-\alpha}{2} \left(\frac{\partial \phi}{\partial x^i} h_{jk} + \frac{\partial \phi}{\partial x^j} h_{ik} \right) \right\} \\
&= \left\{ -\frac{1+\alpha}{2} \cdot n \cdot \frac{\partial \phi}{\partial x^k} + \frac{1-\alpha}{2} \left(\sum_{i=1}^{n} \frac{\partial \phi}{\partial x^i} \delta_{ik} + \sum_{j=1}^{n} \frac{\partial \phi}{\partial x^j} \delta_{jk} \right) \right\} \\
&= \left(-\frac{1+\alpha}{2} \cdot n + \frac{1-\alpha}{2} \cdot 2 \right) \frac{\partial \phi}{\partial x^k} \\
&= -\frac{1}{2} \left\{ (n+2)\,\alpha + (n-2) \right\} \frac{\partial \phi}{\partial x^k},
\end{aligned}
$$

where δ_{ij} is the Kronecker's delta. Therefore, if $\tau_{(h,\nabla,\bar{\nabla})}(\pi_{id}) \equiv 0$, it holds that $(n+2)\alpha + (n-2) = 0$ or $\partial\phi/\partial x^k = 0$ for all $k \in \{1, \ldots, n\}$ at each point in N. Thus we obtain Theorem 1. □

4.8 α-Conformal Equivalence of Level Surfaces

We show our previous results of α-conformal equivalence of level surfaces.

The next theorem holds for a 1-conformally flat statistical submanifold.

Theorem 2 ([19]) *Let M be a simply connected n-dimensional level surface of φ on an $(n+1)$-dimensional Hessian domain $(\Omega, D, g = Dd\varphi)$ with a Riemannian metric g, and suppose that $n \geq 2$. If (Ω, D, g) is a flat statistical manifold, then (M, D^M, g^M) is a 1-conformally flat statistical submanifold of (Ω, D, g), where D^M and g^M are the connection and the Riemannian metric on M induced by D and g, respectively.*

See in [15, 17–19] for realization problems related with α-conformal equivalence.

We now consider two simply connected level surfaces of dim $n \geq 2$ (M, D, g) and $(\hat{M}, \hat{D}, \hat{g})$, which are 1-conformally flat statistical submanifolds of (Ω, D, g). Let λ be a function on M such that $e^{\lambda(p)}\iota(p) \in \hat{\iota}(\hat{M})$ for $p \in M$, where $\hat{\iota}$ is the restriction of the gradient mapping ι to $\hat{M}$, and set $(e^{\lambda})(p) = e^{\lambda(p)}$. Note that the function e^{λ} projects M to $\hat{M}$ with respect to the dual affine coordinate system on Ω.

We define a mapping $\pi : M \to \hat{M}$ by

$$\hat{\iota} \circ \pi = e^{\lambda} \iota,$$

where ι (as denoted above) is the restriction of the gradient mapping ι to M. Let $\bar{D}'$ be an affine connection on M defined by

$$\pi_*(\bar{D}'_X Y) = \hat{D}'_{\pi_*(X)}\pi_*(Y) \quad \text{for } X, Y \in \Gamma(TM),$$

and $\bar{g}$ be a Riemannian metric on M such that

$$\bar{g}(X, Y) = e^{\lambda} g(X, Y) = \hat{g}(\pi_*(X), \pi_*(Y)).$$

The following theorem has been proposed elsewhere (cf. [9, 11]).

Theorem 3 ([20]) *For affine connections D' and $\bar{D}'$ on M, the following are true:*

(i) *D' and $\bar{D}'$ are projectively equivalent.*
(ii) *(M, D', g) and $(M, \bar{D}', \bar{g})$ are (-1)-conformally equivalent.*

Let $\bar{D}$ be an affine connection on M defined by

$$\pi_*(\bar{D}_X Y) = \hat{D}_{\pi_*(X)}\pi_*(Y) \quad \text{for } X, Y \in \Gamma(TM).$$

From the duality of $\hat{D}$ and $\hat{D}'$, $\bar{D}$ is the dual connection of $\bar{D}'$ on M. Then the next theorem holds (cf. [6, 9]).

Theorem 4 ([20]) *For affine connections D and $\bar{D}$ on M, we have that*

(i) *D and $\bar{D}$ are dual-projectively equivalent.*
(ii) *(M, D, g) and $(M, \bar{D}, \bar{g})$ are 1-conformally equivalent.*

For α-connections $D^{(\alpha)}$ and $\bar{D}^{(\alpha)} = D^{(-\alpha)}$ defined similarly to (4.14), we obtain the following corollary by Theorem 3, Theorem 4, and Eq. (4.17) with $\phi = \lambda$ [15].

Corollary 1 *For affine connections $D^{(\alpha)}$ and $\bar{D}^{(\alpha)}$ on M, $(M, D^{(\alpha)}, g)$ and $(M, \bar{D}^{(\alpha)}, \bar{g})$ are α-conformally equivalent.*

4.9 Harmonic Maps Relative to α-Connections on Level Surfaces

We denote $\hat{D}^{(\alpha)}_{\pi_*(X)}\pi_*(Y)$ by $\hat{D}^{(\alpha)}_X \pi_*(Y)$, considering it in the inverse-mapped section $\Gamma(\pi^{-1}T\hat{M})$. Let $\{x^1, \ldots, x^n\}$ be a local coordinate system on M. The notion of a harmonic map between two level surfaces $(M, D^{(\alpha)}, g)$ and $(\hat{M}, \hat{D}^{(\alpha)}, \hat{g})$ is defined as follows:

Definition 3 ([16, 18]) If a tension field $\tau_{(g, D^{(\alpha)}, \hat{D}^{(\alpha)})}(\pi)$ vanishes on M, i.e.,

$$\tau_{(g, D^{(\alpha)}, \hat{D}^{(\alpha)})}(\pi) \equiv 0,$$

the map $\pi : (M, D^{(\alpha)}, g) \to (\hat{M}, \hat{D}^{(\alpha)}, \hat{g})$ is said to be a harmonic map relative to $(g, D^{(\alpha)}, \hat{D}^{(\alpha)})$, where the tension field is defined by

$$\tau_{(g,D^{(\alpha)},\hat{D}^{(\alpha)})}(\pi) := \sum_{i,j=1}^{n} g^{ij} \left\{ \hat{D}^{(\alpha)}_{\frac{\partial}{\partial x^i}} \left(\pi_*(\frac{\partial}{\partial x^j}) \right) - \pi_* \left(D^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) \right\} \in \Gamma(\pi^{-1} T\hat{M}). \tag{4.19}$$

We now specify the conditions for harmonicity of a map $\pi : M \to \hat{M}$ relative to $(g, D^{(\alpha)}, \hat{D}^{(\alpha)})$.

Theorem 5 ([16, 18]) *Let $(M, D^{(\alpha)}, g)$ and $(\hat{M}, \hat{D}^{(\alpha)}, \hat{g})$ be simply connected n-dimensional level surfaces of an $(n+1)$-dimensional Hessian domain (Ω, D, g) with $n \geq 2$. If $\alpha = -(n-2)/(n+2)$ or λ is a constant function on M, a map $\pi : (M, D^{(\alpha)}, g) \to (\hat{M}, \hat{D}^{(\alpha)}, \hat{g})$ is a harmonic map relative to $(g, D^{(\alpha)}, \hat{D}^{(\alpha)})$, where*

$$\hat{\iota} \circ \pi = e^{\lambda} \iota, \ (e^{\lambda})(p) = e^{\lambda(p)}, \quad e^{\lambda(p)} \iota(p) \in \hat{\iota}(\hat{M}), \ p \in M,$$

and ι, $\hat{\iota}$ are the restrictions of the gradient mappings on Ω to M and $\hat{M}$, respectively.

Proof The tension field of the map π relative to $(g, D^{(\alpha)}, \hat{D}^{(\alpha)})$ is described by the pull-back of $(\hat{M}, \hat{D}^{(\alpha)}, \hat{g})$, namely $(M, \bar{D}^{(\alpha)}, \bar{g})$, as follows:

$$\begin{aligned}
\tau_{(g,D^{(\alpha)},\hat{D}^{(\alpha)})}(\pi) &= \sum_{i,j=1}^{n} g^{ij} \left\{ \hat{D}^{(\alpha)}_{\frac{\partial}{\partial x^i}} \left(\pi_* \left(\frac{\partial}{\partial x^j} \right) \right) - \pi_* \left(D^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) \right\} \\
&= \sum_{i,j=1}^{n} g^{ij} \left\{ \pi_* \left(\bar{D}^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) - \pi_* \left(D^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) \right\} \\
&= \pi_* \left(\sum_{i,j=1}^{n} g^{ij} \left(\bar{D}^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} - D^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right) \right)
\end{aligned}$$

Identifying $T_{\pi(x)}M$ with T_xM and considering the definition of π, we obtain

$$\tau_{(g,D^{(\alpha)},\hat{D}^{(\alpha)})}(\pi) = e^{\lambda} \sum_{i,j=1}^{n} g^{ij} \left(\bar{D}^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} - D^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right).$$

By Corollary 1, $(M, D^{(\alpha)}, g)$ and $(M, \bar{D}^{(\alpha)}, \bar{g})$ are α-conformally equivalent, so that Eq. (4.17) holds with $\phi = \lambda$, $h = g$, $\nabla = D^{(\alpha)}$, and $\bar{\nabla} = \bar{D}^{(\alpha)}$ for X, Y and $Z \in \Gamma(TM)$. Thus, for all $k \in \{1, \ldots, n\}$,

$$g \left(\tau_{(g,D^{(\alpha)},\hat{D}^{(\alpha)})}(\pi), \frac{\partial}{\partial x^k} \right)$$

$$
\begin{aligned}
&= g\left(e^{\lambda}\sum_{i,j=1}^{n} g^{ij}\left(\bar{D}^{(\alpha)}_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j} - D^{(\alpha)}_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}\right), \frac{\partial}{\partial x^k}\right)\\
&= e^{\lambda}\sum_{i,j=1}^{n} g^{ij}\left\{-\frac{1+\alpha}{2}d\lambda\left(\frac{\partial}{\partial x^k}\right)g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)\right.\\
&\qquad + \frac{1-\alpha}{2}\left\{d\lambda\left(\frac{\partial}{\partial x^i}\right)g\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\right)\right.\\
&\qquad \left.\left. + d\lambda\left(\frac{\partial}{\partial x^j}\right)g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^k}\right)\right\}\right\}\\
&= e^{\lambda}\sum_{i,j=1}^{n} g^{ij}\left\{-\frac{1+\alpha}{2}\frac{\partial\lambda}{\partial x^k}g_{ij} + \frac{1-\alpha}{2}\left(\frac{\partial\lambda}{\partial x^i}g_{jk} + \frac{\partial\lambda}{\partial x^j}g_{ik}\right)\right\}\\
&= e^{\lambda}\left\{-\frac{1+\alpha}{2}\cdot n\cdot\frac{\partial\lambda}{\partial x^k} + \frac{1-\alpha}{2}\left(\sum_{i=1}^{n}\frac{\partial\lambda}{\partial x^i}\delta_{ik} + \sum_{j=1}^{n}\frac{\partial\lambda}{\partial x^j}\delta_{jk}\right)\right\}\\
&= \left(-\frac{1+\alpha}{2}\cdot n + \frac{1-\alpha}{2}\cdot 2\right)e^{\lambda}\frac{\partial\lambda}{\partial x^k}\\
&= -\frac{1}{2}\left\{(n+2)\,\alpha + (n-2)\right\}\,e^{\lambda}\frac{\partial\lambda}{\partial x^k}.
\end{aligned}
$$

Therefore, if $\tau_{(g,D^{(\alpha)},\hat{D}^{(\alpha)})}(\pi) \equiv 0$, then $(n+2)\alpha + (n-2) = 0$ or $\partial\lambda/\partial x^k = 0$ for all $k \in \{1, \ldots, n\}$ at each point in N. Thus we obtain Theorem 5. □

Remark 5 If $n = 2$, harmonic maps π with non-constant functions λ exist if and only if $\alpha = 0$.

Remark 6 If $n \geq 3$, and a map π is a harmonic map with a non-constant function λ, then $-1 < \alpha < 0$.

Remark 7 For $\alpha \leq -1$ and $\alpha > 0$, harmonic maps π with non-constant functions λ do not exist.

Definition 3 and Theorem 5 are special cases of harmonic maps between α-conformally equivalent statistical manifolds discussed in our previous study [16].

We now provide specific examples of harmonic maps between level surfaces relative to α-connections.

Example 1 (Regular convex cone) Let Ω and ψ be a regular convex cone and its characteristic function, respectively. On the Hessian domain $(\Omega, D, g = Dd\log\psi)$, $d\log\psi$ is invariant under a 1-parameter group of dilations at the vertex p of Ω, i.e., $x \longrightarrow e^{t}(x-p) + p, t \in \mathbf{R}$ [5, 14]. Then, under these dilations, each map between level surfaces of $\log\psi$ is also a dilated map in the dual coordinate system. Hence, each dilated map between level surfaces of $\log\psi$ in the primal coordinate system is a harmonic map relative to an α-connection for any $\alpha \in \mathbf{R}$.

Example 2 (Symmetric cone) Let Ω and $\psi = Det$ be a symmetric cone and its characteristic function, respectively, where *Det* is the determinant of the Jordan algebra that generates the symmetric cone. Then, similar to Example 1, each dilated map at the origin between level surfaces of $\log \psi$ on the Hessian domain $(\Omega, D, g = Dd \log \psi)$ is a harmonic map relative to an α-connection for any $\alpha \in \mathbf{R}$

It is an important problem to find applications of non-trivial harmonic maps relative to α-connections.

Acknowledgments The author thanks the referees for their helpful comments.

References

1. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, Providence, Oxford University Press, Oxford (2000)
2. Cheng, S.Y., Yau, S.T.: The real Monge-Ampère equation and affine flat structures. In: Chern, S.S., Wu, W.T. (eds.) Differential Geometry and Differential Equations, Proceedings of the 1980 Beijing Symposium, Beijing, pp. 339–370 (1982)
3. Eelles, J., Lemaire, L.: Selected Topics in Harmonic Maps. American Mathematical Society, Providence (1983)
4. Grunau, H.C., Kühnel, M.: On the existence of Hermitian-harmonic maps from complete Hermitian to complete Riemannian manifolds. Math. Z. **249**, 297–327 (2005)
5. Hao, J.H., Shima, H.: Level surfaces of non-degenerate functions in r^{n+1}. Geom. Dedicata **50**, 193–204 (1994)
6. Ivanov, S.: On dual-projectively flat affine connections. J. Geom. **53**, 89–99 (1995)
7. Jost, J., Şimşir, F.M.: Affine harmonic maps. Analysis **29**, 185–197 (2009)
8. Jost, J., Yau, S.T.: A nonlinear elliptic system for maps from Hermitian to Riemannian manifolds and rigidity theorems in Hermitian geometry. Acta Math. **170**, 221–254 (1993)
9. Kurose, T.: On the divergence of 1-conformally flat statistical manifolds. Tôhoku Math. J. **46**, 427–433 (1994)
10. Ni, L.: Hermitian harmonic maps from complete Hermitian manifolds to complete Riemannian manifolds. Math. Z. **232**, 331–335 (1999)
11. Nomizu, K., Pinkal, U.: On the geometry and affine immersions. Math. Z. **195**, 165–178 (1987)
12. Nomizu, K., Sasaki, T.: Affine Differential Geometry: Geometry of Affine Immersions. Cambridge University Press, Cambridge (1994)
13. Shima, H.: Harmonicity of gradient mapping of level surfaces in a real affine space. Geom. Dedicata **56**, 177–184 (1995)
14. Shima, H.: The Geometry of Hessian Structures. World Scientific Publishing, Singapore (2007)
15. Uohashi, K.: On α-conformal equivalence of statistical submanifolds. J. Geom. **75**, 179–184 (2002)
16. Uohashi, K.: Harmonic maps relative to α-connections on statistical manifolds. Appl. Sci. **14**, 82–88 (2012)
17. Uohashi, K.: A Hessian domain constructed with a foliation by 1-conformally flat statistical manifolds. Int. Math. Forum **7**, 2363–2371 (2012)
18. Uohashi, K.: Harmonic maps relative to α-connections on Hessian domains. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information, First International Conference, GSI 2013, Paris, France, 28–30 August 2013. Proceedings, LNCS, vol. 8085, pp. 745–750. Springer, Heidelberg (2013)
19. Uohashi, K., Ohara, A., Fujii, T.: 1-Conformally flat statistical submanifolds. Osaka J. Math. **37**, 501–507 (2000)

20. Uohashi, K., Ohara, A., Fujii, T.: Foliations and divergences of flat statistical manifolds. Hiroshima Math. J. **30**, 403–414 (2000)
21. Urakawa, H.: Calculus of Variations and Harmonic Maps. Shokabo, Tokyo (1990) (in Japanese)

Chapter 5
A Riemannian Geometry in the q-Exponential Banach Manifold Induced by q-Divergences

Héctor R. Quiceno, Gabriel I. Loaiza and Juan C. Arango

Abstract In this chapter we consider a deformation of the nonparametric exponential statistical models, using the Tsalli's deformed exponentials, to construct a Banach manifold modelled on spaces of essentially bounded random variables. As a result of the construction, this manifold recovers the exponential manifold given by Pistone and Sempi up to continuous embeddings on the modeling space. The q-divergence functional plays two important roles on the manifold; on one hand, the coordinate mappings are in terms of the q-divergence functional; on the other hand, this functional induces a Riemannian geometry for which the Amari's α-connections and the Levi-Civita connections appears as special cases of the q-connections induced, $\nabla^{(q)}$. The main result is the flatness (zero curvature) of the manifold.

5.1 Introduction

The study of differential geometry structure of statistical models, specially its Riemannian structure, has been developed in two ways, the finite and infinite dimensional cases. The former due to Rao [20], and Jeffreys [11], where the Fisher information is given as a metric for a parametric statistical model $\{p(x,\theta), \theta = (\theta^1, \ldots, \theta^n)\}$ together with the non-flat Levi-Civita connection. Efron [7], introduced the concept of statistical curvature and implicitly used a new connection, known as the exponential connection which was deeply studied by Dawid [6]. The study on the finite dimensional case culminated with Amari [1] who defined a one-parameter family of α-connections which specializing to the exponential connection when $\alpha \to 1$, the essential concept of duality, and the notions of statistical divergence among others. This geometry is characterized by the facts that the exponential connection for the exponential family has vanishing components and zero curvature.

H. R. Quiceno (✉) · G. I. Loaiza · J. C. Arango
Universidad Eafit, Carrera 49 N° 7 Sur-50, Medellin, Colombia, Suramérica
e-mail: romanquiceno@gmail.com

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_5,

The infinite-dimensional case, was initially developed by Pistone and Sempi [19], constructing a manifold for the exponential family with the use of Orlicz spaces as coordinate space. Gibilisco and Pistone [10] defined the exponential connection as the natural connection induced by the use of Orlicz spaces and show that the exponential and mixture connections are in duality relationship with the α-connections just as in the parametric case. In these structures, every point of the manifold is a probability density on some sample measurable space, with particular focus on the exponential family since almost every model of the non-extensive physics belongs to this family. Nevertheless, some models (complex models, see [4]) do not fit in this representations so a new family of distributions must be defined to contain those models, one of these families is the q-exponential family based on a deformation of the exponential function which has been used in several applications using the Tsalli's index q, for references see [4, 21].

In order to give a geometric structure for this models, Amari and Ohara [3] studied the geometry of the q-exponential family in the finite dimensional setting and they found this family to have a dually flat geometrical structure derived from Legendre transformation and is understood by means of conformal geometry. In 2013, Loaiza and Quiceno [15] constructed for this family, a non-parametric statistical manifold modeled on essentially bounded function spaces, such that each q-exponential parametric model is identified with the tangent space and the coordinate maps are naturally defined in terms of relative entropies in the context of Tsallis; and in [14], the Riemannian structure is characterized.

The manifold constructed in [14, 15]; is characterized by the fact that when $q \to 1$ then the non-parametric exponential models are obtained and the manifold constructed by Pistone and Sempi, is recovered, up to continuous embeddings on the modeling space, which means that the manifolds are related by some map $L^{\Phi}(p \cdot \mu) \leftrightarrow L^{\infty}(p \cdot \mu)$; which should be investigated.

5.2 q-Exponential Functions and q-Exponential Families

As mentioned, some complex phenomena do not fit the Gibbs distribution but the power law, the Tsalli's q-entropy is an example capturing such systems. Based on the Tsalli's entropy index q, it has been constructed a generalized exponential family named q-exponential family, which is being presented, see [4, 16, 21].

Given a real number q, we consider the q-deformed exponential and logarithmic functions which are respectively defined by

$$e_q^x = [1 + (1-q)x]^{\frac{1}{1-q}} \quad \text{if} \quad \frac{-1}{1-q} \leq x \tag{5.1}$$

$$\ln_q(x) = \frac{x^{1-q} - 1}{1-q} \quad \text{if} \quad x > 0. \tag{5.2}$$

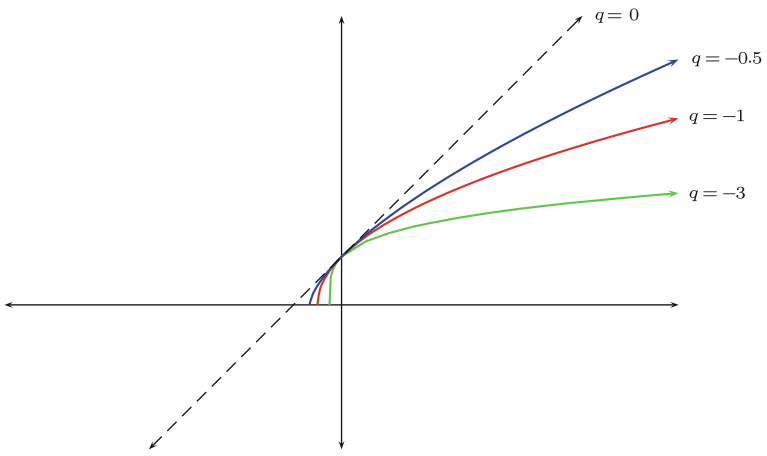


Fig. 5.1 $\exp_q(x)$ for $q < 0$

The above functions satisfy similar properties of the natural exponential and logarithmic functions (Fig. 5.1).

5.2.1 *q-Deformed Properties and Algebra*

It is necessary to show the basic properties of the q-deformed exponential and logarithm functions. Take the definitions given in (5.1) and (5.2) (Fig. 5.2).

Proposition 1 *1. For* $q < 0$, $x \in [0, \infty)$, $\exp_q(x)$ *is positive, continuous, increasing, concave and such that:*

$$\lim_{x \to \infty} \exp_q(x) = \infty.$$

2. For $0 < q < 1$, $x \in [0, \infty)$, $\exp_q(x)$ *is positive, continuous, increasing, convex and such that:*

$$\lim_{x \to \infty} \exp_q(x) = \infty.$$

3. For $1 < q$, $x \in \left[0, \frac{1}{q-1}\right)$, $\exp_q(x)$ *is positive, continuous, increasing, convex and such that:*

$$\lim_{x \to \left(\frac{1}{q-1}\right)^-} \exp_q(x) = \infty.$$

Some graphics are shown for different values of the index q, to illustrate the behavior of $\exp_q(x)$ (Fig. 5.3).

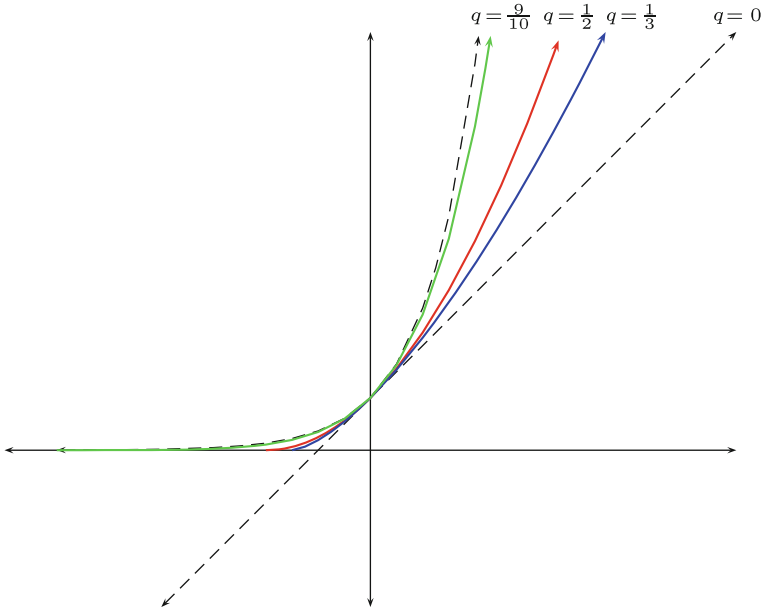


Fig. 5.2 $\exp_q(x)$ for $0 < q < 1$

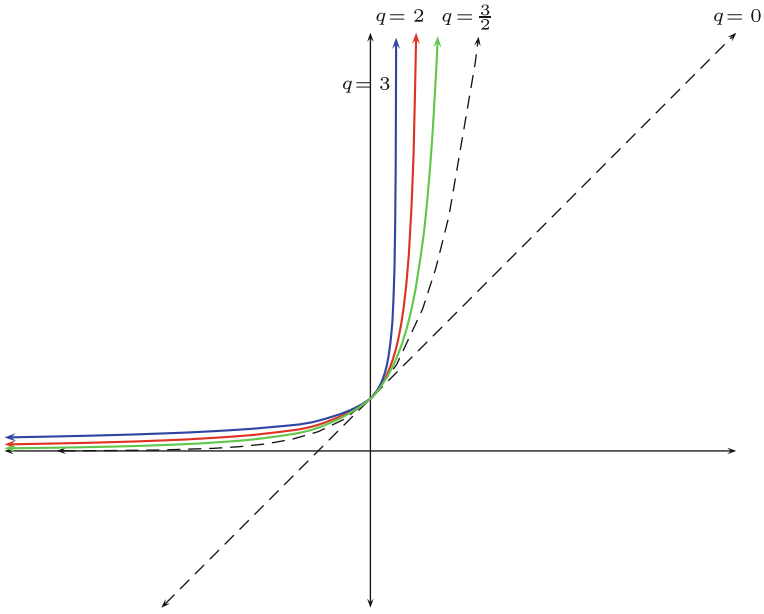


Fig. 5.3 $\exp_q(x)$ for $q > 1$

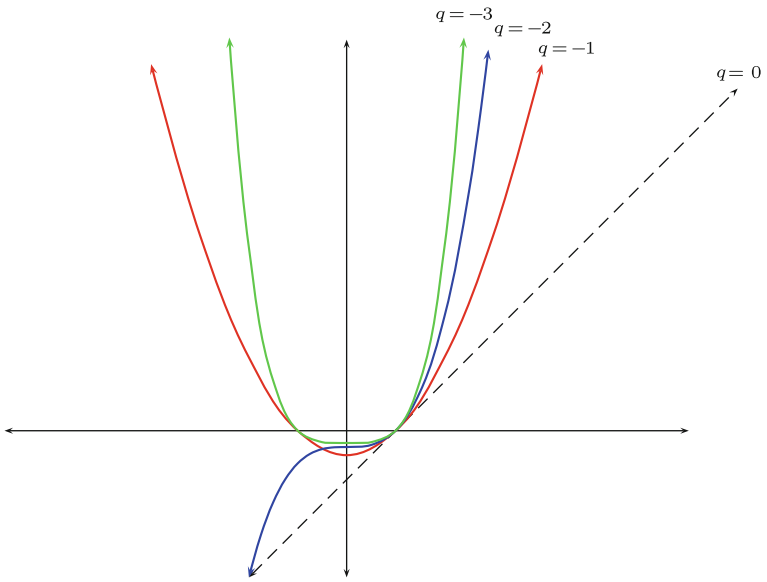


Fig. 5.4 $\ln_q(x)$ for $q < 0$

Proposition 2 *For the function* $\ln_q(x)$

1. *For* $q < 0$, $x \in [0, \infty)$, $\ln_q(x)$ *is continuous, increasing, convex and such that:*

$$\lim_{x \to \infty} \ln_q(x) = \infty.$$

2. *For* $0 < q < 1$, $x \in [0, \infty)$, $\ln_q(x)$ *is continuous, increasing, concave and such that:*

$$\lim_{x \to \infty} \ln_q(x) = \infty.$$

3. *For* $1 < q$, $x \in (0, \infty)$, $\ln_q(x)$ *is increasing, continuous, concave and such that:*

$$\lim_{x \to \infty} \ln_q(x) = \frac{1}{q-1}.$$

Some graphics are shown for different values of the index q, to illustrate the behavior of $\ln_q(x)$ (Fig. 5.4).

The following proposition shows that the deformed functions, share similar properties to the natural ones (Fig. 5.5).

Proposition 3 *1. Product*

$$\exp_q(x) \exp_q(y) = \exp_q(x + y + (1-q)xy). \tag{5.3}$$

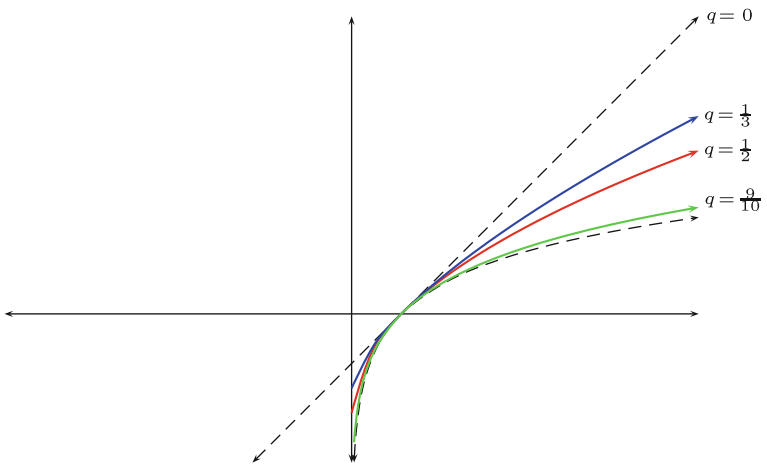


Fig. 5.5 $\ln_q(x)$ for $0 < q < 1$

2. *Quotient*

$$\frac{\exp_q(x)}{\exp_q(y)} = \exp_q\left(\frac{x-y}{1+(1-q)y}\right). \tag{5.4}$$

3. *Power law*

$$(\exp_q x)^n = \exp_{1-\frac{(1-q)}{n}}(nx). \tag{5.5}$$

4. *Inverse*

$$\left[\exp_q(x)\right]^{-1} = \exp_q\left(\frac{-x}{1+(1-q)x}\right) = \exp_{2-q}(-x). \tag{5.6}$$

5. *Derivative*

$$\frac{d}{dx}\left[\exp_q(x)\right] = (\exp_q(x))^q = \exp_{2-\frac{1}{q}}(qx). \tag{5.7}$$

6. *Integral*

$$\int \exp_q(nx)dx = \frac{1}{(2-q)n}(\exp_q(nx))^{2-q}. \tag{5.8}$$

7. *Product*

$$\ln_q(xy) = \ln_q(x) + \ln_q(y) - (1-q)\ln_q(x)\ln_q(y). \tag{5.9}$$

8. *Quotient*

$$\ln_q\left(\frac{x}{y}\right) = \frac{\ln_q(x) - \ln_q(y)}{1+(1-q)\ln_q(y)}. \tag{5.10}$$

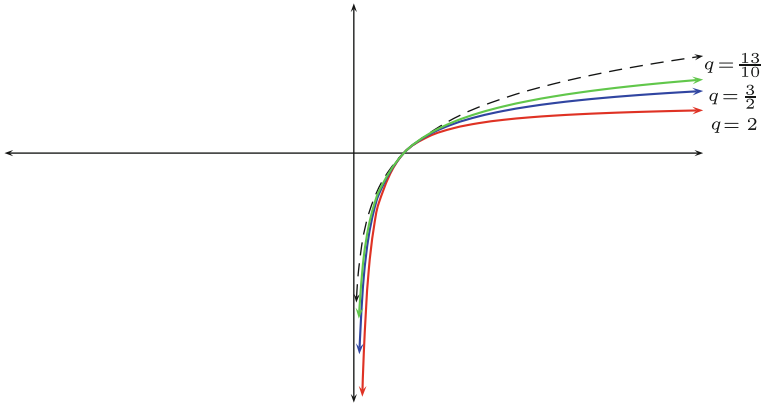


Fig. 5.6 $\ln_q(x)$ for $q > 1$

9. *Power law*

$$\ln_q(x^n) = \frac{n}{1-q} \ln_{1-n}(x^{1-q}). \tag{5.11}$$

10. *Inverse*

$$\ln_q\left(x^{-1}\right) = \frac{\ln_q(x)}{1 + (1-q)\ln_q(x)} = \frac{-1}{x^{1-q}} \ln_q(x). \tag{5.12}$$

11. *Derivative*

$$\frac{d}{dx}\left[\ln_q(x)\right] = \frac{1}{x^q}. \tag{5.13}$$

12. *Integral*

$$\int \ln_q(x)dx = \frac{x(\ln_q(x) - 1)}{2-q}. \tag{5.14}$$

5.2.2 q-Algebra

For real numbers, there are two binary operations given in terms of the index q, as follows (Fig. 5.6).

1. The q-sum $\oplus_q : \mathbb{R}^2 \to \mathbb{R}$, is given by:

$$x \oplus_q y = x + y + (1-q)xy. \tag{5.15}$$

2. The q-product $\otimes_q : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$, is given by:

$$x \otimes_q y = [x^{1-q} + y^{1-q} - 1]_+^{\frac{1}{1-q}} \qquad x > 0 \text{ and } y > 0. \tag{5.16}$$

Proposition 4 *The above operations are associative, modulative and commutative, being* 0 *and* 1 *the respective modules.*

Proposition 5 *Let* $x \neq \frac{1}{q-1}$*, then there exist an unique inverse b, for* $\oplus_q$*; denoted by* $b = \ominus_q x$ *and given by* $b = -\frac{x}{1+(1-q)x}$.

Then

$$x \oplus_q (\ominus_q x) = 0.$$

With this definition, the q-difference can be defined as:

$$\begin{aligned} x \ominus_q y = x \oplus_q (\ominus_q y) &= x - \frac{y}{1+(1-q)y} - \frac{(1-q)xy}{1+(1-q)y} \\ &= \frac{x-y}{1+(1-q)y}. \end{aligned} \tag{5.17}$$

And so, the following properties follow:

$$x \ominus_q y = \ominus_q y \oplus_q x. \tag{5.18}$$

$$x \ominus_q (y \ominus_q z) = (x \ominus_q y) \oplus_q z. \tag{5.19}$$

Proposition 6 *For* $x > 0$ *exist an unique inverse for* $\otimes_q$*, denoted* $\oslash_q x$ *and given by* $b = (2 - x^{1-q})^{\frac{1}{1-q}}$.

With the previous proposition, it is defined a q-quotient for reals x y, positive, such that $x^{1-q} \leq y^{1-q} + 1$, given by:

$$\begin{aligned} x \oslash_q y = x \otimes_q (\oslash_q y) &= (x^{1-q} + (2 - y^{1-q})_+ - 1)^{\frac{1}{1-q}} \\ &= (x^{1-q} - y^{1-q} + 1)_+^{\frac{1}{1-q}}. \end{aligned} \tag{5.20}$$

Note that for $0 < x \leq 2$ it follows $1 \oslash_q (1 \oslash_q x) = x$; and if $q < 1$, the expression $1 \oslash_q 0$ is convergent.

$$x \oslash_q y = 1 \oslash_q (y \oslash_q x) \quad \text{if} \quad x^{1-q} \leq y^{1-q} + 1. \tag{5.21}$$

$$\begin{aligned} x \oslash_q (y \oslash_q z) = (x \oslash_q y) \otimes_q z = (x \otimes_q z) \oslash_q y \\ \text{if} \quad z^{1-q} - 1 \leq y^{1-q} \leq x^{1-q} + 1. \end{aligned} \tag{5.22}$$

This definitions among with Proposition 5.3, allow to prove the next proposition.

Proposition 7 *1.*

$$\exp_q(x)\exp_q(y) = \exp_q(x \oplus_q y). \tag{5.23}$$

$$\frac{\exp_q(x)}{\exp_q(y)} = \exp_q(x \ominus_q y). \tag{5.24}$$

$$\exp_q(x) \otimes_q \exp_q(y) = \exp_q(x + y). \tag{5.25}$$

$$\exp_q(x) \oslash_q \exp_q(y) = \exp_q(x - y). \tag{5.26}$$

2.

$$\ln_q(xy) = \ln_q(x) \oplus_q \ln_q(y). \tag{5.27}$$

$$\ln_q\left(\frac{x}{y}\right) = \ln_q(x) \ominus_q \ln_q(y). \tag{5.28}$$

$$\ln_q(x \otimes_q y) = \ln_q(x) + \ln_q(y). \tag{5.29}$$

$$\ln_q(x \oslash_q y) = \ln_q(x) - \ln_q(y). \tag{5.30}$$

5.2.3 The q-Exponential Family

Some interesting models of statistical physics can be written into the following form,

$$f_\beta(x) = c(x)\exp_q(-\alpha(\beta) - \beta H(x)), \tag{5.31}$$

readers interested in this examples must see [16]. If the q-exponential in the r.h.s. diverges then $f_\beta(x) = 0$ is assumed. The function $H(x)$ is the Hamiltonian. The parameter β is usually the inverse temperature. The normalization $\alpha(\beta)$ is written inside the q-exponential. The function $c(x)$ is the prior distribution. It is a reference measure and must not depend on the parameter β.

If a model is the above form then it is said to belong to the q-exponential family. In the limit $q \to 1$ these are the models of the standard exponential family. In that case the expression (5.31) reduces to

$$f_\beta(x) = c(x)\exp(-\alpha(\beta) - \beta H(x))$$

which is known as the Boltzman-Gibbs distribution.

The convention that $f_\beta(x) = 0$ when the r.h.s. of (5.31) diverges may seem weird. However, one can argue that this is the right thing to do, so the integral of the model is always finite. Also, the example of the harmonic oscillator, given below, will clarify this point. A reformulation of (5.31) is therefore that either $f_\beta(x) = 0$ or

$$\ln_q\left(\frac{f_\beta(x)}{c(x)}\right) = -\alpha(\beta) - \beta H(x).$$

Models belonging to the such a family share a number of interesting properties. In particular, they all fit into the thermodynamic formalism. As a consequence, the probability density $f_\beta(x)$ may be considered to the be the equilibrium distribution of the model at the given value of the parameter β. Some basic examples of this family are developed in [16], as follows.

5.2.3.1 The q-Gaussian Distribution ($q < 3$)

Many of the distributions encountered in the literature on non-extensive statistical can be brought into the form (5.31). A prominent model encountered in this context is the q-Gaussian distribution

$$f(x) = \frac{1}{c_q \sigma} \exp_q\left(-\frac{x^2}{\sigma^2}\right) \tag{5.32}$$

with

$$\begin{aligned} c_q &= \int_{-\infty}^{\infty} \exp_q(-x^2)\, dx \\ &= \sqrt{\frac{\pi}{q-1}}\, \frac{\Gamma\left(-\frac{1}{2} + \frac{1}{q-1}\right)}{\Gamma\left(\frac{1}{q-1}\right)} \quad \text{if} \quad 1 < q < 3, \\ &= \sqrt{\frac{\pi}{q-1}}\, \frac{\Gamma\left(1 + \frac{1}{q-1}\right)}{\Gamma\left(\frac{3}{2} + \frac{1}{1-q}\right)} \quad \text{if} \quad q < 1. \end{aligned}$$

Take for instance $q = \frac{1}{2}$. Then (5.32) becomes

$$f(x) = \frac{15\sqrt{2}}{32\sigma}\left[1 - \frac{x^2}{\sigma^2}\right]_+^2.$$

Note that this distribution vanishes outside the interval $[-\sigma, \sigma]$. Then, the $q \to 1$ case, reproduces the conventional Gauss distribution. For $q = 2$ one obtains

$$f(x) = \frac{1}{\pi}\, \frac{\sigma}{x^2 + \sigma^2}. \tag{5.33}$$

This is known as the Cauchy distribution. The function (5.33) is also called a Lorentzian. In the range $1 \leq q < 3$ the q-Gaussian is strictly positive on the whole

line. For $q < 1$ it is zero outside an interval. For $q \geq 3$ the distribution cannot be normalized because

$$f(x) \sim \frac{1}{|x|^{\frac{2}{q-1}}} \quad \text{as} \quad |x| \to \infty \,.$$

5.2.3.2 Kappa-Distributions $1 < q < \frac{5}{3}$

The following distribution is known in plasma physics as the kappa-distribution

$$f(v) = \frac{1}{A(\kappa) v_0^3} \frac{v^2}{\left(1 + \frac{1}{\kappa - a} \frac{v^2}{v_0^2}\right)^{1+k}}. \tag{5.34}$$

This distribution is a modification of the Maxwell distribution, exhibiting a power law decay like $f(v) \simeq v^{-2\kappa}$ for large v.

Expression (5.34) can be written in the form of q-exponential with $q = 1 + \frac{1}{1+\kappa}$ and

$$f(v) = \frac{v^2}{A(\kappa) v_0^3} \exp_q \left(-\frac{1}{2 - q - (q-1)a} \frac{v^2}{v_0^2} \right). \tag{5.35}$$

However, in order to be of the form (5.31), the pre-factor of (5.35) should not depend on the parameter v_0. Introduce an arbitrary constant $c > 0$ with the dimensions of a velocity. The one can write

$$f(v) = \frac{4\pi v^2}{c^3} \exp_q \left(-\ln_{2-q} \left[4\pi A(\kappa) \left(\frac{v_0}{c} \right)^3 \right] - \left[4\pi A(\kappa) \left(\frac{v_0}{c} \right)^3 \right]^{q-1} h(q, v) \right).$$

where $h(q, v) = \frac{1}{2-q-(q-1)a} \frac{v^2}{v_0^2}$.

In the case $q \to 1$, one obtains the Maxwell-distribution.

5.2.3.3 Speed of Harmonic Oscillator ($q = 3$)

The distribution of velocities v of a classical harmonic oscillator is given by

$$f(v) = \frac{1}{\pi} \frac{1}{\sqrt{v_0^2 - v^2}}.$$

It diverges when $|v|$ approaches its maximal value v_0 and vanished for $|v| > v_0$. This distribution can e written into the form (5.31) of a q-exponential family with $q = 3$. To do so, let $x = v$ and

$$
\begin{aligned}
c(x) &= \frac{\sqrt{2}}{\pi|v|}, \\
\beta &= \frac{1}{2} m v_0^2, \\
H(x) &= \frac{1}{\frac{1}{2} m v^2} = \frac{1}{K}, \\
\alpha(\beta) &= -\frac{3}{2}.
\end{aligned}
$$

5.3 Amari and Ohara's q-Manifold

On 2011, Amari and Ohara [3], constructed a finite dimensional manifold for the family of q-exponential distributions with the properties of a dually flat geometry derived from Legendre transformations and such that the maximizer of the q-scort distribution is a Bayesian estimator.

Let $\mathcal{S}_n$ denote the family of discrete distributions over $(n+1)$ elements $X = \{x_0, x_1, \ldots, x_n\}$; put $p_i = \text{Prob}\{x = x_i\}$ and denote the probability distribution vector $\mathbf{p} = (p_0, p_1, \ldots, p_n)$.
The probability of x is written as

$$
p(x) = \sum_{i=0}^{n} p_i \delta_i(x).
$$

Proposition 8 *The family $\mathcal{S}_n$, has the structure of a q-exponential family for any q.*

The proof is based on the fact that the q-exponential family is written as

$$
p(x, \theta) = \exp_q\{\theta \cdot x - \psi_q(\theta)\}, \tag{5.36}
$$

where $\psi_q(\theta)$ is called the q-potential given by

$$
\psi_q(\theta) = -\ln_q(p_0), \text{ and } p_0 = 1 - \sum_{i=0}^{n} p_i \delta_i(x).
$$

It should be noted that the family in (5.36), is not in general of the form given in [16] since it does not have the pre factor (prior distribution) $c(x)$, as in (5.31).
The geometry is induced by a q-divergence defined by the q-potential. Since ψ is a convex function, it defines a divergence of the Bregman-type

$$
D_q[p(x, \theta_1) : p(x, \theta_2)] = \psi_q(\theta_2) - \psi_q(\theta_1) - \nabla\psi_q(\theta_1) \cdot (\theta_2 - \theta_1),
$$

which simplifies to

$$D_q[p:r] = \frac{1}{(1-q)h_q(p)}\left(1 - \sum_{i=0}^{n} p_i^q r_i^{1-q}\right),$$

where p, r are two discrete distributions and $h_q(p) = \sum_{i=0}^{n} p_i^q$.

For two points on the manifold, infinitesimally close, the divergence is

$$D_q[p(x,\theta) : p(x,\theta+d\theta)] = \sum g_{ij}^q(\theta) d\theta^i d\theta^j,$$

where $g_{ij}^q = \partial_i \partial_j \psi(\theta)$ is the q-metric tensor.

Note that, when $q = 1$, this metric reduces to the usual Fisher metric.

Proposition 9 *The q-metric is given by a conformal transformation of the Fisher information metric g_{ij}^F, as*

$$g_{ij}^q = \frac{q}{h_q(\theta)} g_{ij}^F.$$

With this Riemannian geometry, the geodesics curves of the manifold, given by

$$\ln_q[p(x,t)] = t \ln_q[p_1(x)] + (1-t)\ln_q[p_2(x)] - c(t),$$

allows to find the maximizer of the q-score function as a Bayesian estimator, see [3].

5.4 q-Exponential Statistical Banach Manifold

Let (Ω, Σ, μ) be a probability space and q a real number such that $0 < q < 1$. Denote by $\mathfrak{M}_\mu$ the set of strictly positive probability densities μ-a.e. For each $p \in \mathfrak{M}_\mu$ consider the probability space $(\Omega, \Sigma, p \cdot \mu)$, where $p \cdot \mu$ is the probability measure given by

$$(p \cdot \mu)(A) = \int_A p d\mu.$$

Thus, the space of essentially bounded functions $L^\infty(p \cdot \mu)$ is a Banach space respect to the essential supremum norm denoted by $\|\cdot\|_{p,\infty}$ and is included on the set of random variables $u \in L^1(p \cdot \mu)$ such that the maps $\hat{u}_p : \mathbb{R} \to [0, +\infty]$ given by

$$\hat{u}_p(t) := \int e^{(tu)} p d\mu = E_p[e^{(tu)}],$$

are finite in a neighborhood of the origin 0.

For each $p \in \mathfrak{M}_\mu$ consider

$$B_p := \{u \in L^\infty(p \cdot \mu) : \ E_p[u] = 0\},$$

which (with the essential supremum norm) is a closed normed subspace of the Banach space $L^\infty(p \cdot \mu)$, thus B_p is a Banach space.

Two probability densities $p, z \in \mathfrak{M}_\mu$ are connected by a one-dimensional q-exponential model if there exist $r \in \mathfrak{M}_\mu$, $u \in L^\infty(r \cdot \mu)$, a real function of real variable ψ and $\delta > 0$ such that for all $t \in (-\delta, \delta)$, the function f defined by

$$f(t) = e_q^{tu \ominus_q \psi(t)} r,$$

satisfies that there are $t_0, t_1 \in (-\delta, \delta)$ for which $p = f(t_0)$ and $z = f(t_1)$. The function f is called one-dimensional q-exponential model, since it is a deformation of the model $f(t) = e^{tu - \psi(t)} r$, see [18].
Define the mapping M_p by

$$M_p(u) = E_p[e_q^{(u)}],$$

denoting its domain by $D_{M_p} \subset L^\infty(p \cdot \mu)$. Also, define the mapping $K_p : B_{p,\infty}(0, 1) \to [0, \infty]$, for each $u \in B_{p,\infty}(0, 1)$, by

$$K_p(u) = \ln_q[M_p(u)].$$

Using results on convergence of series in Banach spaces, see [12], it can be proven that the domain D_{M_p} of M_p contains the open unit ball $B_{p,\infty}(0, 1) \subset L^\infty(p \cdot \mu)$. Also if restricting M_p to $B_{p,\infty}(0, 1)$, this function is analytic and infinitely Fréchet differentiable.

Let (Ω, Σ, μ) be a probability space and q a real number with $0 < q < 1$. Let be $\mathcal{V}_p := \{u \in B_p : \|u\|_{p,\infty} < 1\}$, for each $p \in \mathfrak{M}_\mu$. Define the maps

$$e_{q,p} : \mathcal{V}_p \to \mathfrak{M}_\mu$$

by

$$e_{q,p}(u) := e_q^{(u \ominus_q K_p(u))} p,$$

which are injective and their ranges are denoted by $\mathcal{U}_p$. For each $p \in \mathfrak{M}_\mu$ the map

$$s_{q,p} : \mathcal{U}_p \to \mathcal{V}_p$$

given by

$$s_{q,p}(z) := \ln_q\left(\frac{z}{p}\right) \ominus_q E_p\left[\ln_q\left(\frac{z}{p}\right)\right],$$

is precisely the inverse map of $e_{q,p}$. Maps $s_{q,p}$ are the coordinate maps for the manifold and the family of pairs $(\mathcal{U}_p, s_{q,p})_{p \in \mathfrak{M}_\mu}$ define an atlas on $\mathfrak{M}_\mu$; and the transition maps (change of coordinates), for each

$$u \in s_{q,p_1}(\mathcal{U}_{p_1} \bigcap \mathcal{U}_{p_2}),$$

are given by

$$s_{p_2}(e_{p_1}(u)) = \frac{u \oplus_q \ln_q(\frac{p_1}{p_2}) - E_{p_2}[u \oplus_q \ln_q(\frac{p_1}{p_2})]}{1 + (1-q)E_{p_2}[u \oplus_q \ln_q(\frac{p_1}{p_2})]}.$$

Given

$$u \in s_{q,p_1}(\mathcal{U}_{q,p_1} \bigcap \mathcal{U}_{q,p_2}),$$

the derivative of map $s_{q,p_2} \circ s_{q,p_1}^{-1}$ evaluated at u in the direction of $v \in L^\infty(p_1 \cdot \mu)$ is of form

$$D(s_{q,p_2} \circ s_{q,p_1}^{-1})(u) \cdot v = A(u) - B(u)E_{p_2}[A(u)],$$

where $A(u)$, $B(u)$ are functions depending on u. This, allows to establish the main result of [15] (Theorem 14), that is, the collection of pairs $\{(\mathcal{U}_p, s_{q,p})\}_{p \in \mathfrak{M}_\mu}$ is a C^∞-atlas modeled on B_p, and the corresponding manifold is called *$q-$exponential statistical Banach manifold.*

Finally, the tangent bundle of the manifold, is characterized, (Proposition 15) [15], by regular curves on the manifold, as follows.
Let $g(t)$ be a regular curve for $\mathfrak{M}_\mu$ where $g(t_0) = p$, and $u(t) \in \mathcal{V}_z$ be its coordinate representation over $s_{q,z}$.
Then

$$g(t) = e_q^{[u(t) \ominus_q K_z(u(t))]}{}_z$$

and:

1. $\frac{d}{dt} \ln_q \left(\frac{g(t)}{p}\right)_{t=t_0} = Tu'(t) - Q[M_p(u(t))]^{1-q} E_p[u'(t)]$ for some constants T and Q.
2. If $z = p$, i.e. the charts are centered in the same point; the tangent vectors are identified with the q-score function in t given by $\frac{d}{dt} \ln_q \left(\frac{g(t)}{p}\right)_{t=t_0} = Tu'(t_0)$.
3. Consider a two dimensional q-exponential model

$$f(t,q) = e_q^{(tu \ominus_q K_p(tu))} p \tag{5.37}$$

where if $q \to 1$ in (5.37), one obtains the one dimensional exponential models $e^{(tu - K_p(tu))} p$.
The q-score function

$$\frac{d}{dt} \ln_q \left(\frac{f(t)}{p}\right)_{t=t_0}$$

for a one dimensional q-exponential model (5.37) belongs to $span[u]$ at $t = t_0$, where $u \in \mathcal{V}_p$.

So the tangent space $T_p(\mathfrak{M}_\mu)$ is identified with the collection of q-score functions or the one dimensional $q-$exponential models (5.37), where the charts (trivializing

mappings) are given by

$$(g, u) \in \mathcal{T}(\mathcal{U}_p) \to (s_{q,p}(g),\, A(u) - B(u)E_p[A(u)]),$$

defined in the collection of open subsets $\mathcal{U}_p \times \mathcal{V}_p$ of $\mathfrak{M}_\mu \times L^\infty(p \cdot \mu)$.

5.5 Induced Geometry

In this section, we will find a metric and then the connections of the manifold, derived from the q-divergence functional and characterizing the geometry of the q-exponential manifold. For further details see [22].

The q-divergence functional is given as follows [15].
Let f be a function, defined for all $t \neq 0$ and $0 < q < 1$, by

$$f(t) = -t \ \ln_q\left(\frac{1}{t}\right)$$

and for $p, z \in \mathfrak{M}_\mu$, the q-divergence of z with respect to p is given by

$$I^{(q)}(z||p) := \int_\Omega p\ f\left(\frac{z}{p}\right) d\mu = \frac{1}{1-q}\left[1 - \int_\Omega z^q p^{1-q}\, d\mu\right], \tag{5.38}$$

which is the Tsallis' divergence functional [9]. Some properties of this functional are well known, for example that it is equal to the $\alpha-$divergence functional up to a constant factor where $\alpha = 1 - 2q$, satisfying the invariance criterion. Moreover, when $q \to 0$ then $I^{(q)}(z||p) = 0$ and if $q \to 1$ then $I^{(q)}(z||p) = K(z||p)$ which is the Kullback-Leibler divergence functional [13]. As a consequence of Proposition (17) in [15], the manifold is related with the q-divergence functional as

$$s_{q,p}(z) = \left(\frac{1}{1 + (q-1)I^{(q)}(p||z)}\right)\left(\ln_q\left(\frac{z}{p}\right) + I^{(q)}(p||z)\right).$$

The following result, is necessary to guarantee the existence of the Riemannian metric of the manifold.

Proposition 10 *Let $p, z \in \mathfrak{M}_\mu$ then*

$$(d_u)_z\ I^{(q)}(z||p)|_{z=p} = (d_v)_p\ I^{(q)}(z||p)|_{z=p} = 0,$$

where the subscript p, z means that the directional derivative is taken with respect to the first and the second arguments in $I^{(q)}(z||p)$, respectively, along the direction $u \in T_z(\mathfrak{M}_\mu)$ or $v \in T_p(\mathfrak{M}_\mu)$.

Proof Writing (5.38) as

$$I^{(q)}(z||p) = \int_{\Omega} \frac{\left[(1-q)p + (q)z - p^{(1-q)}z^{(q)}\right]}{1-q} d\mu,$$

it follows

$$(d_u)_z\, I^{(q)}(z||p) = \frac{1}{1-q}\int_{\Omega}\left[q - qp^{(1-q)}z^{(q-1)}\right] u d\mu$$

and

$$(d_v)_p\, I^{(q)}(z||p) = \frac{1}{1-q}\int_{\Omega}\left[(1-q) - (1-q)p^{(-q)}z^{(q)}\right] v d\mu;$$

when $z = p$ the desired result holds. □

According to Proposition (16) in [15], the functional $I^{(q)}(z||p)$ is bounded, since:

$$I^{(q)}(z||p) \geq 0 \text{ and equality holds iff } \; p = z$$

and

$$I^{(q)}(z||p) \leq \int_{\Omega} (z-p)\; f'\left(\frac{z}{p}\right) d\mu.$$

Then, together with previous proposition, the q-divergence functional induces a Riemannian metric g and a pair of connections, see Eguchi [8], given by:

$$g(u,v) = -(d_u)_z\,(d_v)_p\, I^{(q)}(z||p)|_{z=p} \tag{5.39}$$

$$\langle \nabla_w u, v\rangle = -(d_w)_z\,(d_u)_z\,(d_v)_p\, I^{(q)}(z||p)|_{z=p}, \tag{5.40}$$

where $v \in T_p(\mathfrak{M}_\mu)$, $u \in T_p(\mathfrak{M}_\mu)$ and w is a vector field.
Denote $\Sigma(\mathfrak{M}_\mu)$ the set of vector fields $u : \mathcal{U}_p \to T_p(\mathcal{U}_p)$, and $F(\mathfrak{M}_\mu)$ the set of $\mathcal{C}^\infty$ functions $f : \mathcal{U}_p \to \mathcal{R}$. The following result establish the metric.

Proposition 11 *Let $p, z \in \mathfrak{M}_\mu$ and v, u vector fields, the metric tensor (field) $g : \Sigma(\mathfrak{M}_\mu) \times \Sigma(\mathfrak{M}_\mu) \to F(\mathfrak{M}_\mu)$ is given by*

$$g(u,v) = q\int_{\Omega} \frac{uv}{p} d\mu.$$

Proof By direct calculation over $I^{(q)}(z||p)$, one gets

$$(d_v)_p I^{(q)}(z||p) = \frac{1}{1-q}\int_{\Omega}\left[(1-q) - (1-q)p^{(-q)}z^{(q)}\right] v d\mu$$

and

$$(d_u)_z\,(d_v)_p\, I^{(q)}(z||p) = -q\int_{\Omega}\left[p^{(-q)}z^{(q-1)}\right] uv d\mu,$$

so by (5.39), it follows

$$g(u,v) = q\int_{\Omega}\frac{uv}{p}\,d\mu.$$

□

Note that when $q \to 1$, this metric reduces to the one induced by the (α,β)-divergence functional which induces the Fisher metric on parametric models.

The connections are characterized as follows.

Proposition 12 *The family of covariant derivatives (connections)*

$$\nabla_w^{(q)}u : \Sigma(\mathfrak{M}_\mu)\times\Sigma(\mathfrak{M}_\mu)\to\Sigma(\mathfrak{M}_\mu),$$

are given as

$$\nabla_w^{(q)}u = d_w u - \left(\frac{1-q}{p}\right)uw.$$

Proof Considering $(d_u)_z\,(d_v)_p\,I^{(q)}(z||p)$ as in proof of Proposition 5.11, we get

$$-(d_w)_z(d_u)_z(d_v)_p I^{(q)}(z||p) = q\int_{\Omega}p^{(-q)}\left[(q-1)z^{(q-2)}uw + z^{(q-1)}d_w u\right]v d\mu.$$

By the previous proposition it follows that

$$g\left(\nabla_w^{(q)}u, v\right) = q\int_{\Omega}\frac{\nabla_w^{(q)}u}{p}\,v d\mu$$

and then

$$g\left(\nabla_w^{(q)}u, v\right) = \langle\nabla_w^{(q)}u, v\rangle.$$

Then,

$$q\;p^{-1}\left[(q-1)p^{-1}uw + d_w u\right] = q\;\frac{\nabla_w^{(q)}u}{p},$$

so

$$\nabla_w^{(q)}u = d_w u - \left(\frac{1-q}{p}\right)uw.$$

□

It is easy to prove that the associated conjugate connection is given by $\nabla_w^{*(q)}u = d_w u - \frac{q}{p}uw$. Notice that taking $q = \frac{1-\alpha}{2}$ yields to the Amaris's one-parameter family of $\alpha-$connections in the form

$$\nabla_w^{(\alpha)}u = d_w u - \left(\frac{1+\alpha}{2p}\right)uw;$$

and taking $q = \frac{1}{2}$ the Levi-Civita connection results.

Finally, the geometry of the manifold is characterized by calculating the curvature and torsion tensors, for which it will be proved that equals zero.

Proposition 13 *For the q-exponential manifold and the connection given in the previous proposition, the curvature tensor and the torsion tensor satisfy $R(u, v, w) = 0$ and $T(u, v) = 0$.*

Proof Remember that

$$R(u, v, w) = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w \tag{5.41}$$

$$T(u, v) = \nabla_u v - \nabla_v u - [u, v]. \tag{5.42}$$

Using the general form

$$\nabla_v w = d_v w + \Gamma(v, w),$$

where

$$\Gamma : \Sigma(\mathfrak{M}_\mu) \times \Sigma(\mathfrak{M}_\mu) \to \Sigma(\mathfrak{M}_\mu)$$

is a bilinear form (the counterpart of the Christoffel symbol); and since

$$d_u (\nabla_v w) = d_u (d_v w) + \Gamma (d_u v, w) + \Gamma (v, d_u w) + d_u \Gamma,$$

where $d_u \Gamma = \Gamma' u$ is the derivative of the bilinear form Γ, it follows that

$$\begin{aligned} \nabla_u \nabla_v w = {} & d_u (d_v w) + \Gamma (d_u v, w) + \Gamma (v, d_u w) + d_u \Gamma \\ & + \Gamma (u, d_v w) + \Gamma (u, \Gamma(v, w)) , \end{aligned}$$

and

$$\begin{aligned} \nabla_v \nabla_u w = {} & d_v (d_u w) + \Gamma (d_v u, w) + \Gamma (u, d_v w) + d_v \Gamma \\ & + \Gamma (v, d_u w) + \Gamma (v, \Gamma(u, w)) ; \end{aligned}$$

moreover,

$$\nabla_{[u,v]} w = d_{[u,v]} w + \Gamma (d_u v - d_v u, w) = d_{[u,v]} w + \Gamma (d_u v, w) - \Gamma (d_v u, w) ,$$

where $[u, v] = d_u v - d_v u$; and then

$$\nabla_{[u,v]} w = d_u (d_v w) - d_v (d_u w) + \Gamma (d_u v, w) - \Gamma (d_v u, w) .$$

Substituting in (5.41) and (5.42) it follows that

$$R(u, v, w) = \Gamma (u, \Gamma(v, w)) - \Gamma (v, \Gamma(u, w)) + d_u \Gamma(v, w) - d_v \Gamma(u, w),$$

and

$$T(u,v) = \Gamma(u,v) - \Gamma(v,u).$$

Since $\Gamma(u,v) = -\frac{1-q}{p}\, uv$, and $d_u\Gamma(v,w) = \frac{1-q}{p^2}\, uvw$, one gets:

$$\begin{aligned} R(u,v,w) &= \Gamma\left(u, -\frac{1-q}{p}vw\right) - \Gamma\left(v, -\frac{1-q}{p}\, uw\right) \\ &\quad + \frac{1-q}{p^2}\, uvw - \frac{1-q}{p^2}\, vuw \\ &= -\left[\frac{1-q}{p}\right] u\left[-\frac{1-q}{p}vw\right] + \left[\frac{1-q}{p}\right] v\left[-\frac{1-q}{p}uw\right] \\ &= 0 \end{aligned}$$

and

$$T(u,v) = -\frac{1-q}{p}\, uv + \frac{1-q}{p}\, vu = 0.$$

□

Since the mapping $\nabla^{(q)} \leftrightarrow \nabla^{(\alpha)}$ is smooth, it is expected that the geodesic curves and parallel transports obtained from the q-connections preserves a smooth isomorphism with the curves given by α-connections. Also, it must be investigated if the metric tensor field in Proposition 5.11 is given by a conformal transformation of the Fisher information metric.

References

1. Amari, S.: Differential-Geometrical Methods in Statistics. Springer, New York (1985)
2. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, RI: Providence (2000) (Translated from the 1993 Japanese original by Daishi Harada)
3. Amari, S., Ohara, A.: Geometry of q-exponential family of probability distributions. Entropy **13**, 1170–85 (2011)
4. Borges, E.P.: Manifestaões dinâmicas e termodinâmicas de sistemas não-extensivos. Tese de Dutorado, Centro Brasileiro de Pesquisas Fisicas, Rio de Janeiro (2004)
5. Cena, A., Pistone, G.: Exponential statistical manifold. Ann. Inst. Stat. Math. **59**, 27–56 (2006)
6. Dawid, A.P.: On the conceptsof sufficiency and ancillarity in the presence of nuisance parameters. J. Roy. Stat. Soc. B **37**, 248–258 (1975)
7. Efron, B.: Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Stat. **3**, 1189–242 (1975)
8. Eguchi, S.: Second order efficiency of minimum coontrast estimator in a curved exponential family. Ann. Stat. **11**, 793–03 (1983)
9. Furuichi, S.: Fundamental properties of Tsallis relative entropy. J. Math. Phys. **45**, 4868–77 (2004)
10. Gibilisco, P., Pistone, G.: Connections on non-parametric statistical manifolds by Orlicz space geometry. Inf. Dim. Anal. Quantum Probab. Relat. Top. **1**, 325–47 (1998)

11. Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. A **186**, 453–61 (1946)
12. Kadets, M.I., Kadets, V.M.: Series in Banach spaces. In: Conditional and Undconditional Convergence. Birkaaauser Verlang, Basel (1997) (Traslated for the Russian by Andrei Iacob)
13. Kulback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
14. Loaiza, G., Quiceno, H.: A Riemannian geometry in the q-exponential Banach manifold induced by q-divergences. Geometric science of information, In: Proceedings of First International Conference on GSI 2013, pp. 737–742. Springer, Paris (2013)
15. Loaiza, G., Quiceno, H.R.: A q-exponential statistical Banach manifold. J. Math. Anal. Appl. **398**, 446–6 (2013)
16. Naudts, J.: The q-exponential family in statistical physics. J. Phys. Conf. Ser. **201**, 012003 (2010)
17. Pistone, G.: k-exponential models from the geometrical viewpoint. Eur. Phys. J. B **70**, 29–37 (2009)
18. Pistone, G., Rogantin, M.-P.: The exponential statistical manifold: Parameters, orthogonality and space transformations. Bernoulli **4**, 721–760 (1999)
19. Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. **23**(5), 1543–1561 (1995)
20. Rao, C.R: Information and accuracy attainable in estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–91 (1945)
21. Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. **52**, 479–487 (1988)
22. Zhang, J.: Referential duality and representational duality on statistical manifolds. In: Proceedings of the 2nd International Symposium on Information Geometry and its Applications, pp. 58–67, Tokyo, (2005)

Chapter 6
Computational Algebraic Methods in Efficient Estimation

Kei Kobayashi and Henry P. Wynn

Abstract A strong link between information geometry and algebraic statistics is made by investigating statistical manifolds which are algebraic varieties. In particular it is shown how first and second order efficient estimators can be constructed, such as bias corrected Maximum Likelihood and more general estimators, and for which the estimating equations are purely algebraic. In addition it is shown how Gröbner basis technology, which is at the heart of algebraic statistics, can be used to reduce the degrees of the terms in the estimating equations. This points the way to the feasible use, to find the estimators, of special methods for solving polynomial equations, such as homotopy continuation methods. Simple examples are given showing both equations and computations.

6.1 Introduction

Information geometry gives geometric insights and methods for studying the statistical efficiency of estimators, testing, prediction and model selection. The field of algebraic statistics has proceeded somewhat separately but recently a positive effort is being made to bring the two subjects together, notably [15]. This paper should be seen as part of this effort.

A straightforward way of linking the two areas is to ask how far algebraic methods can be used when the statistical manifolds of information geometry are algebraic, that is algebraic varieties or derived forms, such as rational quotients. We call such models "algebraic statistical models" and will give formal definitions.

K. Kobayashi (✉)
The Institute of Statistical Mathematics, 10-3, Midori-cho,
Tachikawa, Tokyo, Japan
e-mail: kei@ism.ac.jp

H. P. Wynn
London School of Economics, London, UK
e-mail: H.Wynn@lse.ac.uk

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_6,

In the standard theory for non-singular statistical models, maximum likelihood estimators (MLEs) have first-order asymptotic efficiency and bias-corrected MLEs have second-order asymptotic efficiency. A short section covers briefly the basic theory of asymptotic efficiency using differential geometry, necessary for our development.

We shall show that for some important algebraic models, the estimating equations of MLE type become polynomial and the degrees usually become very high if the model has a high-dimensional parameter space. In this paper, asymptotically efficient algebraic estimators, a generalization of bias corrected MLE, are studied. By algebraic estimators we mean estimators which are the solution of algebraic equations. A main result is that for (algebraic) curved exponential family, there are second-order efficient estimators whose polynomial degree is at most two. These are computed by decreasing the degree of the estimating equations using Gröbner basis methods, the main tool of algebraic statistics. We supply some the basic Gröbner theory in Appendix A. See [22].

The reduction of the degree saves computational costs dramatically when we use computational methods for solving the algebraic estimating equations. Here we use homotopy continuation methods of [19, 24] to demonstrate this effect for a few simple examples, for which we are able to carry out the Gröbner basis reduction. Appendix B discusses homotopy continuation methods.

Although, as mentioned, the links between computational algebraic methods and the theory of efficient estimators based on differential geometry are recent, two other areas of statistics, not covered here, exploit differential geometry methods. The first is tube theory. The seminal paper by Weyl [26] has been used to give exact confidence level values (size of tests), and bounds, for certain Gaussian simultaneous inference problems: [17, 20]. This is very much related to the theory of up-crossings of Gaussian processes using expected Euler characteristic methods, see [1] and earlier papers. The second area is the use of the resolution of singularities (incidentally related to the tube theory) in which confidence levels are related to the dimension and the solid angle tangent of cones with apex at a singularity in parameters space [12, 25]. Moreover, the degree of estimating equations for MLE has been studied for some specific algebraic models, which are not necessarily singular [11]. In this paper we cover the non-singular case, for rather more general estimators than MLE, and show that algebraic methods have a part to play.

Most of the theories in the paper can be applied to a wider class of Multivariate Gaussian models with some restrictions on their covariance matrices, for example models studied in [6, 14]. Though the second-order efficient estimators proposed in the paper can be applied to them potentially, the cost for computing Gröbner basis prevents their direct application. Further innovation in the algebraic computation is required for real applications, which is a feature of several other areas of algebraic statistics.

Section 6.2 gives some basic background in estimation and differential geometry for it. Sections 6.3 and 6.4, which are the heart of the paper, give the algebraic developments and Sect. 6.5 gives some examples. Section 6.6 carries out some computation using homotopy continuation.

6.2 Statistical Manifolds and Efficiency of Estimators

In this section, we introduce the standard setting of statistical estimation theory, via information geometry. See [2, 4] for details. It is recognized that the ideas go back to at least the work of Rao [23], Efron [13] and Dawid [10]. The subject of information geometry was initiated by Amari and his collaborators [3, 5].

Central to this family of ideas is that the rates of convergence of statistical estimators and other test statistics depend on the metric and curvature of the parametric manifolds in a neighborhood of the MLE or the null hypothesis. In addition Amari realized the importance of two special models, the affine exponential model and the affine mixture model, *e* and *m* frame respectively. In this paper we concentrate on the exponential family model but also look at curved subfamilies. By extending the dimension of the parameter space of the exponential family, we are able to cover some classes of mixture models. The extension of the exponential model to infinite dimensions is covered by [21].

6.2.1 Exponential Family and Estimators

A full exponential family is a set of probability distributions $\{\mathrm{d}P(x|\theta) \mid \theta \in \Theta\}$ with a parameter space $\Theta \subset \mathbb{R}^d$ such that

$$\mathrm{d}P(x|\theta) = \exp(x_i\theta^i - \psi(\theta))\mathrm{d}\nu,$$

where $x \in \mathbb{R}^d$ is a variable representing a sufficient statistic and ν is a carrier measure on $\mathbb{R}^d$. Here $x_i\theta^i$ means $\sum_i x_i\theta^i$ (Einstein summation notation).

We call θ a natural parameter and $\eta = \eta(\theta) := E[x|\theta]$ an expectation parameter. Denote $E = E(\Theta) := \{\eta(\theta) \mid \theta \in \Theta\} \subset \mathbb{R}^d$ as the corresponding expectation parameter space. Note that the relation $\eta(\theta) = \nabla_\theta\psi(\theta)$ holds. If the parameter space is restricted to a subset $\mathcal{V}_\Theta \subset \Theta$, we obtain a *curved exponential family*

$$\{\mathrm{d}P(x|\theta) \mid \theta \in \mathcal{V}_\Theta\}.$$

The corresponding space of the expectation parameter is denoted by $\mathcal{V}_E := \{\eta(\theta) \mid \theta \in \mathcal{V}_\Theta\} \subset E$.

Figure 6.1 explains how to define an estimator by a local coordinate. Let $(u, v) \in \mathbb{R}^p \times \mathbb{R}^{d-p}$ with a dimension p of $\mathcal{V}_\Theta$ be a local coordinate system around the true parameter θ^* and define $\mathcal{U} \subset \mathbb{R}^p$ such that $\{\theta(u, 0)|u \in \mathcal{U}\} = \mathcal{V}_\Theta$. For a full exponential model with N samples obtained by composing a map $(X^{(1)}, \ldots, X^{(N)}) \mapsto \theta(\eta)|_{\eta=\bar{X}}$ and a coordinate projection map $\theta(u, v) \mapsto u$, we can define a (local) estimator $(X^{(1)}, \ldots, X^{(N)}) \mapsto u$. We define an estimator by $\eta(u, v)$ similarly. Since $\bar{X}$ is a sufficient statistic of θ (and η) in the full exponential family, every estimator can be computed by $\bar{X}$ rather than the original data $\{X_i\}$. Therefore in the rest of the paper, we write X as shorthand for $\bar{X}$.

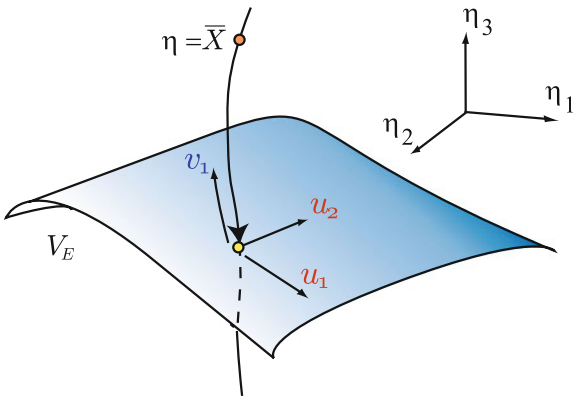


Fig. 6.1 A projection to the model manifold according to a local coordinate defines an estimator

6.2.2 Differential Geometrical Entities

Let $w := (u, v)$ and use indexes $\{i, j, \ldots\}$ for θ and η, $\{a, b, \ldots\}$ for u, $\{\kappa, \lambda, \ldots\}$ for v and $\{\alpha, \beta, \ldots\}$ for w. The following are used for expressing conditions for asymptotic efficiency of estimators, where Einstein notation is used.

Differential geometrical entities

- $\eta_i(\theta) = \dfrac{\partial}{\partial\theta^i}\psi(\theta)$,
- Fisher metric $\underline{G} = (g_{ij})$ w.r.t. θ: $g_{ij}(\theta) = \dfrac{\partial^2\psi(\theta)}{\partial\theta^i\partial\theta^j}$,
- Fisher metric $\bar{G} = (g^{ij})$ w.r.t. η: $\bar{G} = \underline{G}^{-1}$,
- Jacobian: $B_{i\alpha}(\theta) := \dfrac{\partial\eta_i(w)}{\partial w^\alpha}$,
- e-connection: $\Gamma^{(e)}_{\alpha\beta,\gamma} = \left(\dfrac{\partial^2}{\partial w^\alpha\partial w^\beta}\theta^i(w)\right)\left(\dfrac{\partial}{\partial w^\gamma}\eta_i(w)\right)$,
- m-connection: $\Gamma^{(m)}_{\alpha\beta,\gamma} = \left(\dfrac{\partial^2}{\partial w^\alpha\partial w^\beta}\eta_i(w)\right)\left(\dfrac{\partial}{\partial w^\gamma}\theta^i(w)\right)$,

6.2.3 Asymptotic Statistical Inference Theory

Under some regularity conditions on the carrier measure ν, potential function ψ and the manifolds $\mathcal{V}_\Theta$ or $\mathcal{V}_E$, the asymptotic theory below is available. These conditions are for guaranteeing the finiteness of the moments and the commuting of the

expectation and the partial derivative $\frac{\partial}{\partial\theta}E_\theta[f] = E_\theta[\frac{\partial f}{\partial\theta}]$. For more details of the required regularity conditions, see Sect. 2.1 of [4].

1. If $\hat{u}$ is a consistent estimator (i.e. $P(\|\hat{u}-u\| > \epsilon) \to 0$ as $N \to \infty$ for any $\epsilon > 0$), the squared error matrix of $\hat{u}$ is

$$\begin{aligned} E_u[(\hat{u}-u)(\hat{u}-u)^\top] &= E_u[(\hat{u}^a-u^a)(\hat{u}^b-u^b)] \\ &= N^{-1}[g_{ab}-g_{a\kappa}g^{\kappa\lambda}g_{b\lambda}]^{-1}+O(N^{-2}). \end{aligned}$$

   Here $[\cdot]^{-1}$ means the matrix inverse. Thus, if $g_{a\kappa}=0$ for all a and κ, the main term in the r.h.s. becomes minimum. We call such an estimator as a *first-order efficient* estimator.
2. The bias term becomes

$$E_u[\hat{u}^a-u^a] = (2N)^{-1}b^a(u)+O(N^{-2})$$

   for each a where $b^a(u) := \Gamma^{(m)a}_{\ cd}(u)g^{cd}(u)$. Then, the *bias corrected estimator* $\check{u}^a := \hat{u}^a - b^a(\hat{u})$ satisfies $E_u[\check{u}^a-u^a] = O(N^{-2})$.
3. Assume $g_{a\kappa}=0$ for all a and κ, then the square error matrix is represented by

$$E_u[(\check{u}^a-u^a)(\check{u}^b-u^b)] = \frac{1}{N}g^{ab}+\frac{1}{2N^2}\left(\overset{(m)}{\Gamma}{}^{2ab}_{M}+2\overset{(e)}{H}{}^{2ab}_{M}+\overset{(m)}{H}{}^{2ab}_{A}\right)+o(N^{-2}).$$

   See Theorem 5.3 of [4] and Theorem 4.4 of [2] for the definition of the terms in the r.h.s. Of the four dominating terms in the r.h.s., only

$$\overset{(m)}{H}{}^{2ab}_{A} := g^{\kappa\mu}g^{\lambda\nu}H^{(m)a}_{\ \ \kappa\lambda}H^{(m)b}_{\ \ \mu\nu}$$

   depends on the selection of the estimator.

   Here $H^{(m)a}_{\ \ \kappa\lambda}$ is an embedding curvature and equal to $\Gamma^{(m)a}_{\ \ \kappa\lambda}$ when $g_{a\kappa}=0$ for every a and κ. Since $H^{2ab(m)}_{A}$ is the square of $\Gamma^{(m)a}_{\ \ \kappa\lambda}$, the square error matrix attains the minimum in the sense of positive definiteness if and only if

$$\Gamma^{(m)}{}_{\kappa\lambda,a}(w)\Big|_{v=0} = \left(\frac{\partial^2}{\partial v^\kappa\partial v^\lambda}\eta_i(w)\right)\left(\frac{\partial}{\partial u^a}\theta^i(w)\right)\Big|_{v=0} = 0. \tag{6.1}$$

   Therefore the elimination of the m-connection (6.1) implies *second-order efficiency* of the estimator after a bias correction, i.e. it becomes optimal among the bias-corrected first-order efficient estimators up to $O(N^{-2})$.

6.3 Algebraic Models and Efficiency of Algebraic Estimators

This section studies asymptotic efficiency for statistical models and estimators which are defined algebraically. Many models in statistics are defined algebraically. Perhaps most well known are polynomial regression models and algebraic conditions on probability models such as independence and conditional independence. Recently there has been considerable interest in marginal models [7] which are typically linear restrictions on raw probabilities. In time series autoregressive models expressed by linear transfer functions induce algebraic restrictions on covariance matrices. Our desire is to have a definition of algebraic statistical model which can be expressed from within the curved exponential family framework but is sufficiently broad to cover cases such as those just mentioned. Our solution is to allow algebraic conditions in the natural parameter θ, mean parameter η or both. The second way in which algebra enters is in the form of the estimator.

6.3.1 Algebraic Curved Exponential Family

We say a curved exponential family is *algebraic* if the following two conditions are satisfied.

(C1) $\mathcal{V}_\Theta$ or $\mathcal{V}_E$ is represented by a real algebraic variety, i.e. $\mathcal{V}_\Theta := \mathcal{V}(f_1, \ldots, f_k) = \{\theta \in \mathbb{R}^d | f_1(\theta) = \cdots = f_k(\theta) = 0\}$ or similarly $\mathcal{V}_E := \mathcal{V}(g_1, \ldots, g_k)$ for $f_i \in \mathbb{R}[\theta^1, \ldots, \theta^d]$ and $g_i \in \mathbb{R}[\eta_1, \ldots, \eta_d]$.

(C2) $\theta \mapsto \eta(\theta)$ or $\eta \mapsto \theta(\eta)$ is represented by some algebraic equations, i.e. there are $h_1, \ldots, h_k \in \mathbb{R}[\theta, \eta]$ such that locally in $\mathcal{V}_\Theta \times \mathcal{V}_E$, $h_i(\theta, \eta) = 0$ iff $\eta(\theta) = \eta$ or $\theta(\eta) = \theta$.

Here $\mathbb{R}[\theta^1, \ldots, \theta^d]$ means a polynomial of $\theta^1, \ldots, \theta^d$ over the real number field $\mathbb{R}$ and $\mathbb{R}[\theta, \eta]$ means $\mathbb{R}[\theta^1, \ldots, \theta^d, \eta_1, \ldots, \eta_d]$. The integer k, the size of the generators, is not necessarily equal to $d - p$ but we assume $\mathcal{V}_\Theta$ (or $\mathcal{V}_E$) has dimension p around the true parameter. Note that if $\psi(\theta)$ is a rational form or the logarithm of a rational form, (C2) is satisfied.

6.3.2 Algebraic Estimators

The parameter set $\mathcal{V}_\Theta$ (or $\mathcal{V}_E$) is sometimes singular for algebraic models. But throughout the following analysis, we assume non-singularity around the true parameter $\theta^* \in \mathcal{V}_\Theta$ (or $\eta^* \in \mathcal{V}_E$ respectively).

Following the discussion at the end of Sect. 6.2.1. We call $\theta(u, v)$ or $\eta(u, v)$ an *algebraic estimator* if

(C3) $w \mapsto \eta(w)$ or $w \mapsto \theta(w)$ is represented algebraically.

We remark that the MLE for an algebraic curved exponential family is an algebraic estimator.

If conditions (C1), (C2) and (C3) hold, then all of the geometrical entities in Sect. 6.2.2 are characterized by special polynomial equations. Furthermore, if $\psi(\theta) \in \mathbb{R}(\theta) \cup \log \mathbb{R}(\theta)$ and $\theta(w) \in \mathbb{R}(w) \cup \log \mathbb{R}(w)$, then the geometrical objects have the additional property of being rational.

6.3.3 Second-Order Efficient Algebraic Estimators, Vector Version

Consider an algebraic estimator $\eta(u, v) \in \mathbb{R}[u, v]^d$ satisfying the following vector equation:

$$X = \eta(u, 0) + \sum_{i=p+1}^{d} v_{i-p} e_i(u) + c \cdot \sum_{j=1}^{p} f_j(u, v) e_j(u) \tag{6.2}$$

where, for each u, $\{e_j(u);\ j = 1, \ldots, p\} \cup \{e_i(u); i = p+1, \ldots, d\}$ is a complete basis of $\mathbb{R}^d$ such that $\langle e_j(u), (\nabla_u \eta) \rangle_g = 0$ and $f_j(u, v) \in \mathbb{R}[u][v]_{\geq 3}$, namely a polynomial whose degree in v is at least 3 with coefficients polynomial in u, for $j = 1, \ldots, p$. Remember we use a notation $X = \bar{X} = \frac{1}{N} \sum_i X_i$. The constant c is to control the perturbation (see below).

A straightforward computation of the m-connection in (6.1) at $v = 0$ for

$$\eta(w) = \eta(u, 0) + \sum_{i=p+1}^{d} v_{i-p} e_i(u) + c \cdot \sum_{j=1}^{p} f_j(u, v) e_j(u)$$

shows it to be zero. This gives

Theorem 1 *Vector equation (6.2) satisfies the second-order efficiency (6.1).*

We call (6.2) *a vector version of a second-order efficient estimator*. Note that if $c = 0$, (6.2) gives an estimating equation for the MLE. Thus the last term in (6.2) can be recognized as a perturbation from the MLE.

Figure 6.2 is a rough sketch of the second-order efficient estimators. Here the model is embedded in an m-affine space. Given a sample (red point), the MLE is an orthogonal projection (yellow point) to the model with respect to the Fisher metric. But a second-order efficient estimator maps the sample to the model along a "cubically" curved manifold (red curve).

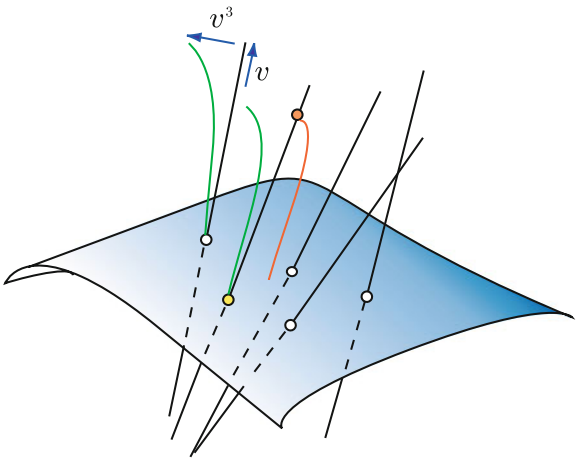


Fig. 6.2 Image of the vector version of the second-order efficient estimators

6.3.4 Second-Order Efficient Algebraic Estimators, Algebraic Version

Another class of second-order efficient algebraic estimators we call the *algebraic version*, which is defined by the following simultaneous polynomial equations with $\eta_u = \eta(u, 0)$.

$$(X - \eta_u)^\top \tilde{e}_j(u, \eta_u) + c \cdot h_j(X, u, \eta_u, X - \eta_u) = 0 \text{ for } j = 1, \ldots, p \qquad (6.3)$$

where $\{\tilde{e}_j(u, \eta_u) \in \mathbb{R}[u, \eta_u]^d; j = 1, \ldots, p\}$ span $((\nabla_u \eta(u, 0))^{\perp_{\bar{G}}})^{\perp_E}$ for every u and $h_j(X, u, \eta_u, t) \in \mathbb{R}[X, u, \eta_u][t]_3$ (degree $= 3$ in t) for $j = 1, \ldots, p$. The constant c is to control the perturbation. The notation $\bar{G}$ represents the Fisher metric on the full-exponential family with respect to η. The notation $(\nabla_u \eta(u, 0))^{\perp_{\bar{G}}}$ means the subspace orthogonal to $span(\partial_a \eta(u, 0))_{a=1}^p$ with respect to $\bar{G}$ and $(\cdot)^{\perp_E}$ means the orthogonal complement in the sense of Euclidean vector space. Here, the term "degree" of a polynomial means the maximum degree of its terms. Note that the case $(X - \eta_u)^\top \tilde{e}_j(u, \eta_u) = 0$ for $j = 1, \ldots, p$ gives a special set of the estimating equations of the MLE.

Theorem 2 *An estimator defined by a vector version (6.2) of the second-order efficient estimators is also represented by an algebraic version (6.3) where* $h_j(X, u, \eta_u, t) = \tilde{f}_j(u, (\tilde{e}_i^\top t)_{i=1}^p, (\tilde{e}_i^\top (X - \eta_u))_{i=1}^p)$ *with a function* $\tilde{f}_j(u, v, \tilde{v}) \in \mathbb{R}[u, \tilde{v}][v]_3$ *such that* $\tilde{f}(u, v, v) = f(u, v)$.

Proof Take the Euclidean inner product of both sides of (6.2) with each $\tilde{e}_j$ which is a vector Euclidean orthogonal to the subspace $span(\{e_i | i \neq j\})$ and obtain a system of polynomial equations. By eliminating variables v from the polynomial equations, an algebraic version is obtained. □

Theorem 3 *Every algebraic equation (6.3) gives a second-order efficient estimator (6.1).*

Proof Writing $X = \eta(u, v)$ in (6.3), we obtain

$$(\eta(u, v) - \eta(u, 0))^\top \tilde{e}_j(u) + c \cdot h_j(\eta(u, v), u, \eta(u, 0), \eta(u, v) - \eta(u, 0)) = 0.$$

Partially differentiate this by v twice, we obtain

$$\left(\frac{\partial^2 \eta(u, v)}{\partial v^\lambda \partial v^\kappa} \right)^\top \tilde{e}_j(u) \Big|_{v=0} = 0,$$

since each term of $h_j(\eta(u, v), u, \eta(u, 0), \eta(u, v) - \eta(u, 0))$ has degree more than 3 in its third component $(\eta_i(u, v) - \eta_i(u, 0))_{i=1}^d$ and $\eta(u, v) - \eta(u, 0)|_{v=0} = 0$. Since $\mathrm{span}\{\tilde{e}_j(u);\ j = 1, \ldots, p\} = ((\nabla_u \eta(u, 0))^{\perp_{\bar{G}}})^{\perp_E} = \mathrm{span}\{\bar{G}\partial_{u_a}\eta;\ a = 1, \ldots, p\}$, we obtain

$$\Gamma^{(m)}_{\kappa\lambda a}\Big|_{v=0} = \frac{\partial^2 \eta_i}{\partial v^\lambda \partial v^\kappa} g^{ij} \frac{\partial \eta_j}{\partial u^a}\Big|_{v=0} = 0.$$

This implies the estimator is second-order efficient. □

By Theorems 1, 2 and 3, the relationship between the three forms of the second-order efficient algebraic estimators is summarized as

$$(1) \Leftarrow (2) \Rightarrow (3) \Rightarrow (1).$$

Furthermore, if we assume the estimator has a form $\eta \in \mathbb{R}(u)[v]$, that is a polynomial in v with coefficients rational in u, every first-order efficient estimator satisfying (6.1) can be written in a form (6.2) after resetting coordinates v for the estimating manifold. In this sense, we can say (1) $\Rightarrow$ (2) and the following corollary holds.

Corollary 1 *If $\eta \in \mathbb{R}(u)[v]$, the forms (1), (2) and (3) are equivalent.*

6.3.5 Properties of the Estimators

The following theorem is a straightforward extension of the local existence of MLE. That is to say, the existence for sufficiently large sample size. The regularity conditions are essentially the same as for the MLE but with an additional condition referring to the control constant c.

Proposition 1 (Existence and uniqueness of the estimate) *Assume that the Fisher matrix is non-degenerate around $\eta(u^*) \in \mathcal{V}_E$. Then the estimate given by (6.2) or*

(6.3) locally uniquely exists for small c, i.e. there is a neighborhood $G(u^*) \subset \mathbb{R}^d$ of $\eta(u^*)$ and $\delta > 0$ such that for every fixed $X \in G(u^*)$ and $-\delta < c < \delta$, a unique estimate exists.

Proof Under the condition of the theorem, the MLE always exists locally. Furthermore, because of the nonsingular Fisher matrix, the MLE is locally bijective (by the implicit representation theorem). Thus $(u_1, \ldots, u_p) \mapsto (g_1(x-\eta_u), \ldots, g_p(x-\eta_u))$ for $g_j(x - \eta_u) := (X - \eta_u)^\top \tilde{e}_j(u, \eta_u)$ in (6.3) is locally bijective. Since $\{g_i\}$ and $\{h_i\}$ are continuous, we can select $\delta > 0$ for (6.3) to be locally bijective for every $-\delta < c < \delta$. □

6.3.6 Summary of Estimator Construction

We summarize how to define a second-order efficient algebraic estimator (vector version) and how to compute an algebraic version from it.

Input:
- a potential function ψ satisfying (C2),
- polynomial equations of η, u and v satisfying (C3),
- $m_1, \ldots, m_{d-p} \in \mathbb{R}[\eta]$ such that $\mathcal{V}_E = V(m_1, \ldots, m_{d-p})$ gives the model,
- $f_j \in \mathbb{R}[u][v]_{\geq 3}$ and $c \in \mathbb{R}$ for a vector version

Step 1 Compute ψ and $\theta(\eta)$, $G(\eta)$, $(\Gamma^{(m)}(\eta)$ for bias correction)

Step 2 Compute $f_{ai} \in \mathbb{R}[\eta][\xi_{11}, \ldots, \xi_{pd}]_1$ s.t. $f_{aj}(\xi_{11}, \ldots, \xi_{pd}) := \partial_{u^a} m_j$ for $\xi_{bi} := \partial_{u^b} \eta_i$.

Step 3 Find $e_{p+1}, \ldots, e_d \in (\nabla_u \eta)^{\perp_{\bar{G}}}$ by eliminating $\{\xi_{aj}\}$ from $\langle e_i, \partial_{u^a} \eta \rangle_{\bar{G}} = e_{ik}(\eta) g^{kj}(\eta) \xi_{aj} = 0$ and $f_{aj}(\xi_{11}, \ldots, \xi_{pd}) = 0$.

Step 4 Select $e_1, \ldots, e_p \in \mathbb{R}[\eta]$ s.t. $e_1(\eta), \ldots, e_d(\eta)$ are linearly independent.

Step 5 Eliminate v from

$$X = \eta(u, 0) + \sum_{i=p+1}^{d} v_{i-p} e_i + c \cdot \sum_{j=1}^{p} f_j(u, v) e_j$$

and compute $(X - \eta)^\top \tilde{e}_j$ and $h \in (\mathbb{R}[\eta][X - \eta]_3)^p$, given by Theorem 2.

Output(Vector version):

$$X = \eta(u, 0) + \sum_{i=p+1}^{d} v_{i-p} e_i(\eta) + c \cdot \sum_{j=1}^{p} f_j(u, v) e_j(\eta).$$

Output(Algebraic version):

$$(X-\eta)^\top \tilde{e} + c \cdot h(X-\eta) = 0.$$

6.3.7 Reduction of the Degree of the Estimating Equations

As we noted in Sect. 6.3.4, if we set $h_j = 0$ for all j, the estimator becomes the MLE. In this sense, ch_j can be recognized as a perturbation from the likelihood equations. If we select each $h_j(X, u, \eta_u, t) \in \mathbb{R}[X, u, \eta_u][t]_3$ tactically, we can reduce the degree of the polynomial estimating equation. For algebraic background, the reader refers to Appendix A.

Here, we assume $u \in \mathbb{R}[\eta_u]$. For example, we can set $u_i = \eta_i$. Then $\tilde{e}_j(u, \eta_u)$ is a function of η_u, so we write it as $\tilde{e}_j(\eta)$. Define an ideal $\mathcal{I}_3$ of $\mathbb{R}[X, \eta]$ as

$$\mathcal{I}_3 := \langle \{(X_i - \eta_i)(X_j - \eta_j)(X_k - \eta_k) \mid 1 \le i, j, k \le d\} \rangle.$$

Select a monomial order $\prec$ and set $\eta_1 \succ \cdots \succ \eta_d \succ X_1 \succ \cdots \succ X_d$. Let $G_\prec = \{g_1, \ldots, g_m\}$ be a Gröbner basis of $\mathcal{I}_3$ with respect to $\prec$. Then the remainder (normal form) r_j of $(X-\eta)^\top \tilde{e}_j(\eta)$, the first term of the l.h.s. of (6.3), with respect to $G_\prec$, is uniquely determined for each j.

Theorem 4 *If the monomial order $\prec$ is the pure lexicographic,*

1. *r_j for $j = 1, \ldots, p$ has degree at most 2 with respect to η, and*
2. *$r_j = 0$ for $j = 1, \ldots, p$ are the estimating equations for a second-order efficient estimator.*

Proof Assume r_j has a monomial term whose degree is more than 2 with respect to η and represent the term as $\eta_a\eta_b\eta_c q(\eta, X)$ with a polynomial $q \in \mathbb{R}(\eta, X)$ and a combination of indices a, b, c. Then $\{\eta_a\eta_b\eta_c + (X_a-\eta_a)(X_a-\eta_a)(X_a-\eta_a)\}q(\eta, X)$ has a smaller polynomial order than $\eta_a\eta_b\eta_c q(\eta, X)$ since $\prec$ is pure lexicographic satisfying $\eta_1 \succ \cdots \succ \eta_d \succ X_1 \succ \cdots \succ X_d$. Therefore by subtracting $(X_a-\eta_a)(X_a-\eta_a)(X_a-\eta_a)\}q(\eta, X) \in \mathcal{I}_3$ from r_j, the polynomial degree decreases. This contradicts the fact r_j is the normal form so each r_j has degree at most 2.

Furthermore each polynomial in $\mathcal{I}_3$ is in $\mathbb{R}[X, u, \eta_u][X - \eta]_3$ and therefore by taking the normal form, the condition for the algebraic version (6.3) of second-order efficiency still holds. □

The reduction of the degree is important when we use algebraic algorithms such as homotopy continuation methods [18] to solve simultaneous polynomial equations since computational cost depends highly on the degree of the polynomials.

6.4 First-Order Efficiency

It is not surprising that, for first-order efficiency, almost the same arguments hold as for second-order efficiency.

By Theorem 5.2 of [4], a consistent estimator is first-order efficient if and only if

$$g_{\kappa a} = 0. \tag{6.4}$$

Consider an algebraic estimator $\eta(u, v) \in \mathbb{R}[u, v]^d$ satisfying the following vector equation:

$$X = \eta(u, 0) + \sum_{i=p+1}^{d} v_{i-p} e_i(u) + c \cdot \sum_{j=1}^{p} f_j(u, v) e_j(u) \tag{6.5}$$

where, for each u, $\{e_j(u);\ j = 1, \ldots, p\} \cup \{e_i(u);\ i = p+1, \ldots, d\}$ is a complete basis of $\mathbb{R}^d$ s.t. $\langle e_j(u), (\nabla_u \eta)\rangle_g = 0$ and $f_j(u, v) \in \mathbb{R}[u][v]_{\geq 2}$, a polynomial whose degree of v is at least 2, for $j = 1, \ldots, p$. Similarly, $c \in \mathbb{R}$ is a constant for perturbation. Here, the only difference between (6.2) for the second-order efficiency and (6.5) for the first-order efficiency is the degree of the $f_j(u, v)$ with respect to v.

The algebraic version of the first-order efficient algebraic estimator is defined by the following simultaneous polynomial equalities with $\eta_u = \eta(u, 0)$.

$$(X - \eta_u)^\top \tilde{e}_j(u, \eta_u) + c \cdot h_j(X, u, \eta_u, X - \eta_u) = 0 \text{ for } j = 1, \ldots, p \tag{6.6}$$

where $\{\tilde{e}_j(u, \eta_u) \in \mathbb{R}[u, \eta_u]^d;\ j = 1, \ldots, p\}$ span $((\nabla_u \eta(u, 0))^{\perp_{\tilde{G}}})^{\perp_E}$ for every u and $h_j(X, u, \eta_u, t) \in \mathbb{R}[X, u, \eta_u][t]_2$ (degree $= 2$ w.r.t. t) for $j = 1, \ldots, p$. Here, the only difference between (6.3) for the second-order efficiency and (6.6) for the first-order efficiency is the degree of the $h_j(X, u, \eta_u, t)$ with respect to t.

Then the relation between the three different forms of first-order efficiency can be proved in the same way manner as for Theorem 1, 2 and 3.

Theorem 5 *(i) Vector version (6.5) satisfies the first-order efficiency.*
(ii) An estimator defined by a vector version (6.5) of the first-order efficient estimators is also represented by an algebraic version (6.6).
(iii) Every algebraic version (6.6) gives a first-order efficient estimator.

The relationship between the three forms of the first-order efficient algebraic estimators is summarized as (4) $\Leftarrow$ (5) $\Rightarrow$ (6) $\Rightarrow$ (4). Furthermore, if we assume the estimator has a form $\eta \in \mathbb{R}(u)[v]$, the forms (6.4), (6.5) and (6.6) are equivalent.

Let $\mathcal{R} := \mathbb{Z}[X, \eta]$ and define

$$\mathcal{I}_2 := \langle \{(X_i - \eta_i)(X_j - \eta_j) \mid 1 \leq i, j \leq d\} \rangle$$

as an ideal of $\mathcal{R}$. In a similar manner, let $\prec$ be a monomial order such that $\eta_1 \succ \cdots \succ \eta_d \succ X_1 \succ \cdots \succ X_d$. Let $G_\prec = \{g_1, \ldots, g_m\}$ be a Gröbner basis of $\mathcal{I}_2$ with respect to $\prec$. The properties of the normal form r_i of $(X - \eta(u, 0))^\top \tilde{e}_i(u)$ with respect to $G_\prec$ are then covered by the following:

Theorem 6 *If the monomial order $\prec$ is the pure lexicographic,*
(i) *r_i for $i = 1, \ldots, d$ has degree at most 1 with respect to η, and*
(ii) *$r_i = 0$ for $i = 1, \ldots, d$ are the estimating equations for a first-order efficient estimator.*

6.5 Examples

In this section, we show how to use the algebraic computation to design asymptotically efficient estimators for two simple examples. The examples satisfy the algebraic conditions (C1), (C2) and (C3) so it is verified that necessary geometric entities have an algebraic form as mentioned in Sect. 6.3.2.

6.5.1 Example: Periodic Gaussian Model

The following periodic Gaussian model shows how to compute second-order efficients estimators and their biases.

- Statistical Model:

$$X \sim N(\mu, \Sigma(a)) \text{ with } \mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \Sigma(a) = \begin{bmatrix} 1 & a & a^2 & a \\ a & 1 & a & a^2 \\ a^2 & a & 1 & a \\ a & a^2 & a & 1 \end{bmatrix} \text{ for } 0 \leq a < 1.$$

  Here, the dimension of the full exponential family and the curved exponential family are $d = 3$ and $p = 1$, respectively.
- Curved exponential family:

$$\log f(x|\theta) = 2\,(x_1x_2 + x_2x_3 + x_3x_4 + x_4x_1)\,\theta_2 + 2\,(x_3x_1 + x_4x_2)\,\theta_3 - \psi(\theta),$$

- Potential function:

$$\psi(\theta) = -1/2\,\log({\theta_1}^4 - 4\,{\theta_1}^2{\theta_2}^2 + 8\,\theta_1{\theta_2}^2\theta_3 - 2\,{\theta_1}^2{\theta_3}^2 - 4\,{\theta_2}^2{\theta_3}^2 + {\theta_3}^4) + 2\,\log(2\,\pi),$$

- Natural parameter:

$$\theta(a) = \left[\frac{1}{1 - 2a^2 + 4a^4}, -\frac{a}{1 - 2a^2 + 4a^4}, \frac{a^2}{1 - 2a^2 + 4a^4}\right]^\top,$$

- Expectation parameter: $\eta(a) = [-2, -4a, -2a^2]^\top$,
- Fisher metric with respect to η:

$$(g^{ij}) = \begin{bmatrix} 2a^4 + 4a^2 + 2 & 8a\left(1+a^2\right) & 8a^2 \\ 8a\left(1+a^2\right) & 4 + 24a^2 + 4a^4 & 8a\left(1+a^2\right) \\ 8a^2 & 8a\left(1+a^2\right) & 2a^4 + 4a^2 + 2 \end{bmatrix},$$

- A set of vectors $e_i \in \mathbb{R}^3$:

$$e_0(a) := [0, -1, a]^\top \in \partial_a \eta(a),$$

$$e_1(a) := [3a^2 + 1, 4a, 0]^\top,\ e_2(a) := [-a^2 - 1, 0, 2]^\top \in (\partial_a \eta(a))^{\perp_{\bar{G}}}.$$

- A vector version of the second-order efficient estimator is, for example,

$$x - \eta + v_1 \cdot e_1 + v_2 \cdot e_2 + c \cdot v_1^3 \cdot e_0 = 0.$$

- A corresponding algebraic version of the second-order efficient estimator: by eliminating v_1 and v_2, we get $g(a) + c \cdot h(a) = 0$ where

$$g(a) := 8(a-1)^2(a+1)^2(1+2a^2)^2(4a^5 - 8a^3 + 2a^3x_3 - 3x_2a^2 + 4a + 4ax_1 + 2ax_3 - x_2)$$

and

$$h(a) := (2a^4 + a^3x_2 - a^2x_3 + 2a^2 + ax_2 - 2x_1 - x_3 - 4)^3.$$

- An estimating equation for MLE:

$$4a^5 - 8a^3 + 2a^3x_3 - 3x_2a^2 + 4a + 4ax_1 + 2ax_3 - x_2 = 0.$$

- Bias correction term for an estimator $\hat{a}$: $\hat{a}(\hat{a}^8 - 4\hat{a}^6 + 6\hat{a}^4 - 4\hat{a}^2 + 1)/(1 + 2\hat{a}^2)^2$.

6.5.2 Example: Log Marginal Model

Here, we consider a log marginal model. See [7] for more on marginal models.

- Statistical model (Poisson regression):
  $X_{ij} \overset{i.i.d}{\sim} \mathrm{Po}(Np_{ij})$ s.t. $p_{ij} \in (0, 1)$ for $i = 1, 2$ and $j = 1, 2, 3$ with model constraints:

$$\begin{aligned} &p_{11} + p_{12} + p_{13} + p_{21} + p_{22} + p_{23} = 1, \\ &p_{11} + p_{12} + p_{13} = p_{21} + p_{22} + p_{23}, \\ &\frac{p_{11}/p_{21}}{p_{12}/p_{22}} = \frac{p_{12}/p_{22}}{p_{13}/p_{23}}. \end{aligned} \tag{6.7}$$

Condition (6.7) can appear in a statistical test of whether acceleration of the ratio p_{1j}/p_{2j} is constant.

In this case, $d = 6$ and $p = 3$.

- Log density w.r.t. the point mass measure on $\mathbb{Z}^6_{\geq 0}$:

$$\log f(x|p) = \log\left\{\prod_{ij} e^{-Np_{ij}}(Np_{ij})^{X_{ij}}\right\} = -N + \sum_{ij} X_{ij}\log(Np_{ij}).$$

- The full expectation family is given by

$$\begin{bmatrix} X_1 & X_2 & X_3 \\ X_4 & X_5 & X_6 \end{bmatrix} := \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix},$$

$$\begin{bmatrix} \eta_1 & \eta_2 & \eta_3 \\ \eta_4 & \eta_5 & \eta_6 \end{bmatrix} = N \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \end{bmatrix},$$

$\theta^i = \log(\eta_i)$ and $\psi(\theta) = N$.

- The Fisher metric w.r.t. θ: $g_{ij} = \frac{\partial^2\psi}{\partial\theta^i\partial\theta^j} = \delta_{ij}\eta_i$.
- Selection of the model parameters:

$$[u_1, u_2, u_3] := [\eta_1, \eta_3, \eta_5] \text{ and } [v_1, v_2, v_3] := [\eta_2, \eta_4, \eta_6].$$

- A set of vectors $e_i \in \mathbb{R}^6$:

$$e_0 := \begin{bmatrix} \eta_2^2(\eta_4 - \eta_6) \\ -\eta_2^2(\eta_4 - \eta_6) \\ 0 \\ -\eta_3\eta_5^2 - 2\eta_2\eta_4\eta_6 \\ 0 \\ \eta_3\eta_5^2 + 2\eta_2\eta_4\eta_6 \end{bmatrix} \in (\nabla_u\eta),$$

$$[e_1, e_2, e_3] := \left[\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \eta_1(-\eta_1\eta_5^2 + \eta_3\eta_5^2) \\ \eta_2(-\eta_1\eta_5^2 - 2\eta_2\eta_4\eta_6) \\ 0 \\ \eta_4(\eta_2^2\eta_4 - \eta_2^2\eta_6) \\ \eta_5(\eta_2^2\eta_4 + 2\eta_1\eta_3\eta_5) \\ 0 \end{bmatrix}, \begin{bmatrix} \eta_1(\eta_1\eta_5^2 - \eta_3\eta_5^2) \\ \eta_2(\eta_1\eta_5^2 + 2\eta_2\eta_4\eta_6) \\ 0 \\ \eta_4(2\eta_1\eta_3\eta_5 + \eta_2^2\eta_6) \\ 0 \\ \eta_6(\eta_2^2\eta_4 + 2\eta_1\eta_3\eta_5) \end{bmatrix}\right]$$
$$\in ((\nabla_u\eta)^{\perp_{\tilde{G}}})^3$$

- A vector version of the second-order efficient estimator is, for example,

$$X - \eta + v_1 \cdot e_1 + v_2 \cdot e_2 + v_3 \cdot e_3 + c \cdot v_1^3 \cdot e_0 = 0.$$

- The bias correction term of the estimator = 0.
- A set of estimating equations for MLE:

$\{x_1\eta_2{}^2\eta_4{}^2\eta_6 - x_1\eta_2{}^2\eta_4\eta_6{}^2 - x_2\eta_1\eta_2\eta_4{}^2\eta_6 + x_2\eta_1\eta_2\eta_4\eta_6{}^2 - 2\,x_4\eta_1\eta_2\eta_4\eta_6{}^2 - x_4\eta_1\eta_3\eta_5{}^2\eta_6 + 2\,x_6\eta_1\eta_2\eta_4{}^2\eta_6 + x_6\eta_1\eta_3\eta_4\eta_5{}^2,$
$-x_2\eta_2\eta_3\eta_4{}^2\eta_6 + x_2\eta_2\eta_3\eta_4\eta_6{}^2 + x_3\eta_2{}^2\eta_4{}^2\eta_6 - x_3\eta_2{}^2\eta_4\eta_6{}^2 - x_4\eta_1\eta_3\eta_5{}^2\eta_6 - 2\,x_4\eta_2\eta_3\eta_4\eta_6{}^2 + x_6\eta_1\eta_3\eta_4\eta_5{}^2 + 2\,x_6\eta_2\eta_3\eta_4{}^2\eta_6,$
$-2\,x_4\eta_1\eta_3\eta_5{}^2\eta_6 - x_4\eta_2{}^2\eta_4\eta_5\eta_6 + x_5\eta_2{}^2\eta_4{}^2\eta_6 - x_5\eta_2{}^2\eta_4\eta_6{}^2 + 2\,x_6\eta_1\eta_3\eta_4\eta_5{}^2 + x_6\eta_2{}^2\eta_4\eta_5\eta_6,$
$\eta_1\eta_3\eta_5{}^2 - \eta_2{}^2\eta_4\eta_6,\ \eta_1+\eta_2+\eta_3-\eta_4-\eta_5-\eta_6, -\eta_1-\eta_2-\eta_3-\eta_4-\eta_5-\eta_6+1\}$

The total degree of the equations is $5 \times 5 \times 5 \times 4 \times 1 \times 1 = 500$.

- A set of estimating equations for a 2nd-order efficient estimator with degree at most 2:

$\{-3\,x_1x_2x_4{}^2x_6\eta_2+6\,x_1x_2x_4{}^2x_6\eta_6+x_1x_2x_4{}^2\eta_2\eta_6-2\,x_1x_2x_4{}^2\eta_6{}^2+3\,x_1x_2x_4x_6{}^2\eta_2-6\,x_1x_2x_4x_6{}^2\eta_4+2\,x_1x_2x_4x_6\eta_2\eta_4-2\,x_1x_2x_4x_6\eta_2\eta_6-x_1x_2x_6{}^2\eta_2\eta_4+2\,x_1x_2x_6{}^2\eta_4{}^2+3\,x_1x_3x_4x_5{}^2\eta_6-2\,x_1x_3x_4x_5\eta_5\eta_6-3\,x_1x_3x_5{}^2x_6\eta_4+2\,x_1x_3x_5x_6\eta_4\eta_5+x_1x_4{}^2x_6\eta_2{}^2-2\,x_1x_4{}^2x_6\eta_2\eta_6 - x_1x_4x_5{}^2\eta_3\eta_6 - x_1x_4x_6{}^2\eta_2{}^2 + 2\,x_1x_4x_6{}^2\eta_2\eta_4 + x_1x_5{}^2x_6\eta_3\eta_4 + 3\,x_2{}^2x_4{}^2x_6\eta_1 - x_2{}^2x_4{}^2\eta_1\eta_6 - 3\,x_2{}^2x_4x_6{}^2\eta_1 - 2\,x_2{}^2x_4x_6\eta_1\eta_4 + 2\,x_2{}^2x_4x_6\eta_1\eta_6 + x_2{}^2x_6{}^2\eta_1\eta_4 - x_2x_4{}^2x_6\eta_1\eta_2 - 2\,x_2x_4{}^2x_6\eta_1\eta_6 + x_2x_4x_6{}^2\eta_1\eta_2 + 2\,x_2x_4x_6{}^2\eta_1\eta_4 - x_3x_4x_5{}^2\eta_1\eta_6 + x_3x_5{}^2x_6\eta_1\eta_4,$
$3\,x_1x_3x_4x_5{}^2\eta_6-2\,x_1x_3x_4x_5\eta_5\eta_6-3\,x_1x_3x_5{}^2x_6\eta_4+2\,x_1x_3x_5x_6\eta_4\eta_5-x_1x_4x_5{}^2\eta_3\eta_6+x_1x_5{}^2x_6\eta_3\eta_4 + 3\,x_2{}^2x_4{}^2x_6\eta_3 - x_2{}^2x_4{}^2\eta_3\eta_6 - 3\,x_2{}^2x_4x_6{}^2\eta_3 - 2\,x_2{}^2x_4x_6\eta_3\eta_4 + 2\,x_2{}^2x_4x_6\eta_3\eta_6 + x_2{}^2x_6{}^2\eta_3\eta_4 - 3\,x_2x_3x_4{}^2x_6\eta_2 + 6\,x_2x_3x_4{}^2x_6\eta_6 + x_2x_3x_4{}^2\eta_2\eta_6 - 2\,x_2x_3x_4{}^2\eta_6{}^2+3\,x_2x_3x_4x_6{}^2\eta_2-6\,x_2x_3x_4x_6{}^2\eta_4+2\,x_2x_3x_4x_6\eta_2\eta_4-2\,x_2x_3x_4x_6\eta_2\eta_6-x_2x_3x_6{}^2\eta_2\eta_4 + 2\,x_2x_3x_6{}^2\eta_4{}^2 - x_2x_4{}^2x_6\eta_2\eta_3 - 2\,x_2x_4{}^2x_6\eta_3\eta_6 + x_2x_4x_6{}^2\eta_2\eta_3 + 2\,x_2x_4x_6{}^2\eta_3\eta_4+x_3x_4{}^2x_6\eta_2{}^2-2\,x_3x_4{}^2x_6\eta_2\eta_6-x_3x_4x_5{}^2\eta_1\eta_6-x_3x_4x_6{}^2\eta_2{}^2+2\,x_3x_4x_6{}^2\eta_2\eta_4 + x_3x_5{}^2x_6\eta_1\eta_4,$
$6\,x_1x_3x_4x_5{}^2\eta_6-4\,x_1x_3x_4x_5\eta_5\eta_6-6\,x_1x_3x_5{}^2x_6\eta_4+4\,x_1x_3x_5x_6\eta_4\eta_5-2\,x_1x_4x_5{}^2\eta_3\eta_6+2\,x_1x_5{}^2x_6\eta_3\eta_4 + 3\,x_2{}^2x_4{}^2x_6\eta_5 - x_2{}^2x_4{}^2\eta_5\eta_6 - 3\,x_2{}^2x_4x_5x_6\eta_4 + 3\,x_2{}^2x_4x_5x_6\eta_6 + x_2{}^2x_4x_5\eta_4\eta_6 - x_2{}^2x_4x_5\eta_6{}^2 - 3\,x_2{}^2x_4x_6{}^2\eta_5 - x_2{}^2x_4x_6\eta_4\eta_5 + x_2{}^2x_4x_6\eta_5\eta_6 + x_2{}^2x_5x_6\eta_4{}^2 - x_2{}^2x_5x_6\eta_4\eta_6 + x_2{}^2x_6{}^2\eta_4\eta_5 - 2\,x_2x_4{}^2x_6\eta_2\eta_5 + 2\,x_2x_4x_5x_6\eta_2\eta_4 - 2\,x_2x_4x_5x_6\eta_2\eta_6 + 2\,x_2x_4x_6{}^2\eta_2\eta_5 - 2\,x_3x_4x_5{}^2\eta_1\eta_6 + 2\,x_3x_5{}^2x_6\eta_1\eta_4,$
$\eta_1\eta_3\eta_5{}^2 - \eta_2{}^2\eta_4\eta_6,\ \eta_1+\eta_2+\eta_3-\eta_4-\eta_5-\eta_6\ , -\eta_1-\eta_2-\eta_3-\eta_4-\eta_5-\eta_6+1\}.$

The total degree of the polynomial equations is 32.

- A set of estimating equations for a first-order-efficient estimator with degree at most 1:

Table 6.1 Computational time for each estimate by the homotopy continuation methods

Algorithm	Estimator	#Paths	Running time (s) (avg. ± std.)
Linear	MLE	500	1.137 ± 0.073
Homotopy	2nd eff.	32	$\mathbf{0.150 \pm 0.047}$
Polyhedral	MLE	64	0.267 ± 0.035
Homotopy	2nd eff.	24	$\mathbf{0.119 \pm 0.027}$

$$\{-x_5{}^2x_4\eta_6x_1x_3+x_5{}^2x_6\eta_4x_1x_3+2\,x_6{}^2\eta_4x_1x_2x_4-2\,x_4{}^2\eta_6x_1x_2x_6-x_6{}^2x_1x_2\eta_2x_4+ x_4{}^2x_1x_2\eta_2x_6+x_2{}^2x_6{}^2\eta_1x_4-x_4{}^2x_2{}^2\eta_1x_6,$$
$$-x_5{}^2x_4\eta_6x_1x_3+x_5{}^2x_6\eta_4x_1x_3+2\,x_6{}^2\eta_4x_2x_3x_4-2\,x_4{}^2\eta_6x_2x_3x_6-x_6{}^2x_2x_3\eta_2x_4+ x_4{}^2x_2x_3\eta_2x_6+x_2{}^2x_6{}^2\eta_3x_4-x_4{}^2x_2{}^2\eta_3x_6,$$
$$-2\,x_5{}^2x_4\eta_6x_1x_3+2\,x_5{}^2x_6\eta_4x_1x_3-x_4x_6x_5x_2{}^2\eta_6+x_4x_5x_2{}^2\eta_4x_6-x_4{}^2x_2{}^2\eta_5x_6+ x_4x_6{}^2x_2{}^2\eta_5,$$
$$\eta_1\eta_3\eta_5{}^2-\eta_2{}^2\eta_4\eta_6,\ \eta_1+\eta_2+\eta_3-\eta_4-\eta_5-\eta_6\,,-\eta_1-\eta_2-\eta_3-\eta_4-\eta_5-\eta_6+6\}.$$

The estimating equations for a second-order-efficient estimator above look much more complicated than the estimating equation for the MLE, but each term of the first three polynomials are at most degree 2. Thanks to this degree reduction, the computational costs for the estimates become much smaller as we will see in Sect. 6.6.

6.6 Computation

To obtain estimates based on the method of this paper, we need fast algorithms to find the solution of polynomial equations. The authors have carried out computations using homotopy continuation method (matlab program HOM4PS2 by Lee, Li and Tsuai [18]) for the log marginal model in Sect. 6.5.2 and a data $\bar{X} = (1, 1, 1, 1, 1, 1)$.

The run time to compute each estimate on a standard laptop (Intel(R) Core (TM) i7-2670QM CPU, 2.20 GHz, 4.00 GB memory) is given by Table 6.1. The computation is repeated 10 times and the averages and the standard deviations are displayed. Note the increasing of the speed for the second-order efficient estimators is due to the degree reduction technique. The term "path" in the table heading refers to a primitive iteration step within the homotopy method. In the faster polyhedron version, the solution region is subdivided into polyhedral domains.

Figure 6.3 shows the mean squared error and the computational time of the MLE, the first-order estimator and the second-order efficient estimator of Sect. 6.5.2. The true parameter is set $\eta^* = (1/6, 1/4, 1/12, 1/12, 1/4, 1/6)$, a point in the model manifold, and N random samples are generated i.i.d. from the distribution with the parameter. The computation is repeated for exponentially increasing sample sizes $N = 1, \ldots, 10^5$. In general, there are multiple roots for polynomial equations and here we selected the root closest to the sample mean by the Euclidean norm. Figure 6.3(1) also shows that the mean squared error is approximately the same for the three estimators, but (2) shows that the computational time is much more for the MLE.

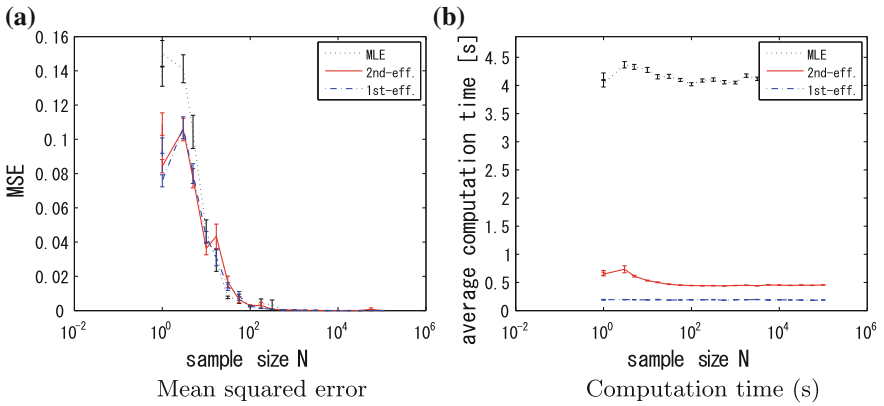


Fig. 6.3 The mean squared error and computation time for each estimate by the homotopy continuation method

6.7 Discussion

In this paper we have concentrated on reduction of the polynomial degree of the estimating equations and shown the benefits in computation of the solutions. We do not expect the estimators to be closed form, such as a rational polynomial form in the data. The most we can expect is they are algebraic, that is they are the solution of algebraic equations. They lie on a zero dimensional algebraic variety. It is clear that there is no escape from using mixed symbolic-numerical methods. In algebraic statistics the number of solution of the ML equation is called the ML degree. Given that we have more general estimating equations than pure ML equations this points to an extended theory or "quasi" ML degree of efficient estimator degree. The issue of exactly which solution to use as our estimators persists. In the paper we suggest taking the solution closest to the sufficient statistic in the Euclidian metric. We could use other metrics and more theory is needed.

Here we have put forward estimating equations with reduced degree and shown the benefits in terms of computation. But we could have used other criteria for choosing the equations, while remaining in the efficient class. We might prefer to choose an equation which reduces the bias further via decreasing the next order term. There may thus be some trade off between degree and bias.

Beyond the limited ambitions of this paper to look at second-order efficiency lie several other areas, notably hypothesis testing and model selection. But the question is the same: to what extent can we bring the algebraic methods to bear, for example by expressing additional differential forms and curvatures in algebraic terms. Although estimation typically requires a mixture of symbolic and numeric methods in some cases only the computation of the efficient estimate requires numeric procedures and the other computations can be carrying out symbolically.

Acknowledgments This paper has benefited from conversations with and advice from a number of colleagues. We should thank Satoshi Kuriki, Tomonari Sei, Wicher Bergsma and Wilfred Kendall. The first author acknowledges support by JSPS KAKENHI Grant 20700258, 24700288 and the second author acknowledges support from the Institute of Statistical Mathematics for two visits in 2012 and 2013 and from UK EPSRC Grant EP/H007377/1. A first version of this paper was delivered at the WOGAS3 meeting at the University of Warwick in 2011. We thank the sponsors. The authors also thank the referees of the short version in GSI2013 and the referees of the first long version of the paper for insightful suggestions.

A Normal Forms

A basic text for the materials in this section is [9]. The rapid growth of modern computational algebra can be credited to the celebrated Buchberger's algorithm [8].

A monomial ideal I in a polynomial ring $K[x_1, \ldots, x_n]$ over a field K is an ideal for which there is a collection of monomials $f_1, \ldots, f_m$ such that any $g \in I$ can be expressed as a sum

$$g = \sum_{i=1}^{m} g_i(x) f_i(x)$$

with some polynomials $g_i \in K[x_1, \ldots, x_n]$. We can appeal to the representation of a monomial $x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$ by its exponent $\alpha = (\alpha_1, \ldots, \alpha_n)$. If $\beta \geq 0$ is another exponent then

$$x^\alpha x^\beta = x^{\alpha+\beta},$$

and $\alpha + \beta$ is in the positive (shorthand for non-negative) "orthant" with corner at α. The set of all monomials in a monomial ideal is the union of all positive orthants whose "corners" are given by the exponent vectors of the generating monomial $f_1, \ldots, f_m$. A monomial ordering written $x^\alpha \prec x^\beta$ is a total (linear) ordering on monomials such that for $\gamma \geq 0$, $x^\alpha \prec x^\beta \Rightarrow x^{\alpha+\gamma} \prec x^{\beta+\gamma}$. Any polynomial $f(x)$ has a leading terms with respect to $\prec$, written $LT(f)$.

There are, in general, many ways to express a given ideal I as being generated from a basis $I = \langle f_1, \ldots, f_m \rangle$. That is to say, there are many choices of basis. Given an ideal I a set $\{g_1, \ldots g_m\}$ is called a Gröbner basis (G-basis) if:

$$\langle LT(g_1), \ldots, LT(g_m) \rangle \ = \ \langle LT(I) \rangle,$$

where $\langle LT(I) \rangle$ is the ideal generated by all the monomials in I. We sometimes refer to $\langle LT(I) \rangle$ as the *leading term ideal*. Any ideal I has a Gröbner basis and any Gröbner basis in the ideal is a basis of the ideal.

Given a monomial ordering and an ideal expressed in terms of the G-basis, $I = \langle g_1, \ldots, g_m \rangle$, any polynomial f has a unique remainder with respect the quotient operation $K[x_1, \ldots, x_k]/I$. That is

$$f = \sum_{i=1}^{m} s_i(x) g_i(x) + r(x).$$

We call the remainder $r(x)$ the *normal form* of f with respect to I and write $NF(f)$. Or, to stress the fact that it may depend on $\prec$, we write $NF(f, \prec)$. Given a monomial ordering $\prec$, a polynomial $f = \sum_{\alpha \in L} \theta_\alpha x^\alpha$ for some L is a normal form with respect to $\prec$ if $x^\alpha \notin \langle LT(f) \rangle$ for all $\alpha \in L$. An equivalent way of saying this is: given an ideal I and a monomial ordering $\prec$, for every $f \in K[x_1, \ldots, x_k]$ there is a unique normal form $NF(f)$ such that $f - NF(f) \in I$.

B Homotopy Continuation Method

Homotopy continuation method is an algorithm to find the solutions of simultaneous polynomial equations numerically. See, for example, [19, 24] for more details of the algorithm and theory.

We will explain the method briefly by a simple example of 2 equations with 2 unknowns

$$\text{Input: } f, g \in \mathbb{R}[x, y]$$

$$\text{Output: The solutions of } f(x, y) = g(x, y) = 0.$$

Step 1 Select arbitrary polynomials of the form:

$$\begin{aligned} f_0(x, y) &:= f_0(x) := a_1 x^{d_1} - b_1 = 0, \\ g_0(x, y) &:= g_0(y) := a_2 y^{d_2} - b_2 = 0 \end{aligned} \tag{6.8}$$

where $d_1 = \deg(f)$ and $d_2 = \deg(g)$. Polynomial equations in this form are easy to solve.

Step 2 Take the convex combinations:

$$\begin{aligned} f_t(x, y) &:= t f(x, y) + (1 - t) f_0(x, y), \\ g_t(x, y) &:= t g(x, y) + (1 - t) g_0(x, y) \end{aligned}$$

then our target becomes the solution for $t = 1$.

Step 3 Compute the solution for $t = \delta$ for small δ by the solution for $t = 0$ numerically.

Step 4 Repeat this until we obtain the solution for $t = 1$.

Figure 6.4 shows a sketch of the algorithm. This algorithm is called *the (linear) homotopy continuation method* and justified if the path connects $t = 0$ and $t = 1$

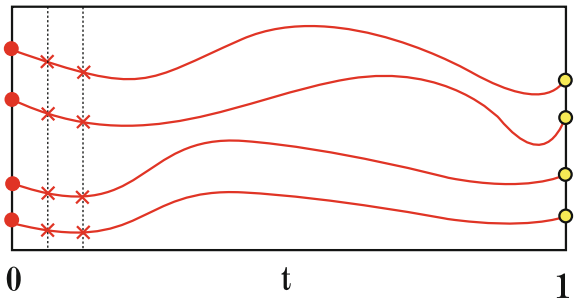


Fig. 6.4 Paths for the homotopy continuation method

continuously without an intersection. That can be proved for almost all a and b. See [19].

For each computation for the homotopy continuation method, the number of the paths is the number of the solutions of (6.8). In this case, the number of paths is $d_1 d_2$. In general case with m unknowns, it becomes $\prod_{i=1}^{m} d_i$ and this causes a serious problem for computational cost. Therefore decreasing the degree of second-order efficient estimators plays an important role for the homotopy continuation method.

Note that in order to solve this computational problem, the authors of [16] proposed the nonlinear homotopy continuation methods (or the polyhedral continuation methods). But as we can see in Sect. 6.5.2, the degree of the polynomials still affects the computational costs.

References

1. Adler, R.J., Taylor, J.E.: Random Fields and Geometry. Springer Monographs in Mathematics. Springer, New York (2007)
2. Amari, S., Nagaoka, H.: Methods of Information Geometry, vol. 191. American Mathematical Society, Providence (2007)
3. Amari, S.: Differential geometry of curved exponential families-curvatures and information loss. Ann. Stat. **10**, 357–385 (1982)
4. Amari, S.: Differential-Geometrical Methods in Statistics. Springer, New York (1985)
5. Amari, S., Kumon, M.: Differential geometry of Edgeworth expansions in curved exponential family. Ann. Inst. Stat. Math. **35**(1), 1–24 (1983)
6. Andersson, S., Madsen, J.: Symmetry and lattice conditional independence in a multivariate normal distribution. Ann. Stat. **26**(2), 525–572 (1998)
7. Bergsma, W.P., Croon, M., Hagenaars, J.A.: Marginal Models for Dependent, Clustered, and Longitudinal Categorical Data. Springer, New York (2009)
8. Buchberger, B.: Bruno Buchberger's PhD thesis 1965: an algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. J. Symbol. Comput. **41**(3), 475–511 (2006)
9. Cox, D.A., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra, 3/e (Undergraduate Texts in Mathematics). Springer, New York (2007)

10. Dawid, A.P.: Further comments on some comments on a paper by Bradley Efron. Ann. Stat. **5**(6), 1249 (1977)
11. Drton, M., Sturmfels B., Sullivant, S.: Lectures on Algebraic Statistics. Springer, New York (2009)
12. Drton, M.: Likelihood ratio tests and singularities. Ann. Stat. **37**, 979–1012 (2009)
13. Efron, B.: Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Stat. **3**, 1189–1242 (1975)
14. Gehrmann, H., Lauritzen, S.L.: Estimation of means in graphical Gaussian models with symmetries. Ann. Stat. **40**(2), 1061–1073 (2012)
15. Gibilisco, P., Riccomagno, E., Rogantin, M.P., Wynn, H.P.: Algebraic and Geometric Methods in Statistics. Cambridge University Press, Cambridge (2009)
16. Huber, B., Sturmfels, B.: A polyhedral method for solving sparse polynomial systems. Math. Comput. **64**, 1541–1555 (1995)
17. Kuriki, S., Takemura, A.: On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. Ann. Appl. Probab. **12**(2), 768–796 (2002)
18. Lee, T.L., Li, T.Y., Tsai, C.H.: HOM4PS2.0: a software package for solving polynomial systems by the polyhedral homotopy continuation method. Computing **83**(2–3), 109–133 (2008)
19. Li, T.Y.: Numerical solution of multivariate polynomial systems by homotopy continuation methods. Acta Numer **6**(1), 399–436 (1997)
20. Naiman, D.Q.: Conservative confidence bands in curvilinear regression. Ann. Stat. **14**(3), 896–906 (1986)
21. Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. **23**, 1543–1561 (1995)
22. Pistone, G., Wynn, H.P.: Generalised confounding with Gröbner bases. Biometrika **83**, 653–666 (1996)
23. Rao, R.C.: Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**(3), 81–91 (1945)
24. Verschelde, J.: Algorithm 795: PHCpack: a general-purpose solver for polynomial systems by homotopy continuation. ACM Trans. Math. Softw. (TOMS) **25**(2), 251–276 (1999)
25. Watanabe, S.: Algebraic analysis for singular statistical estimation. In: Watanabe, O., Yokomori, T. (eds.) Algorithmic Learning Theory. Springer, Berlin (1999)
26. Weyl, H.: On the volume of tubes. Am. J. Math. **61**, 461–472 (1939)

Chapter 7
Eidetic Reduction of Information Geometry Through Legendre Duality of Koszul Characteristic Function and Entropy: From Massieu–Duhem Potentials to Geometric Souriau Temperature and Balian Quantum Fisher Metric

Frédéric Barbaresco

Abstract Based on Koszul theory of sharp convex cone and its hessian geometry, Information Geometry metric is introduced by Koszul form as hessian of Koszul–Vinberg Characteristic function logarithm (KVCFL). The front of the Legendre mapping of this KVCFL is the graph of a convex function, the Legendre transform of this KVCFL. By analogy in thermodynamic with Dual Massieu–Duhem potentials (Free Energy and Entropy), the Legendre transform of KVCFL is interpreted as a "Koszul Entropy". This Legendre duality is considered in more general framework of Contact Geometry, the odd-dimensional twin of symplectic geometry, with Legendre fibration and mapping. Other analogies will be introduced with large deviation theory with Cumulant Generating and Rate functions (Legendre duality by Laplace Principle) and with Legendre duality in Mechanics between Hamiltonian and Lagrangian. In all these domains, we observe that the "Characteristic function" and its derivatives capture all information of random variable, system or physical model. We present two extensions of this theory with Souriau's Geometric Temperature deduced from covariant definition of thermodynamic equilibriums, and with Balian quantum Fisher metric defined and built as hessian of von Neumann entropy. Finally, we apply Koszul geometry for Symmetric/Hermitian Positive Definite Matrices cones, and more particularly for covariance matrices of stationary signal that are characterized by specific matrix structures: Toeplitz Hermitian Positive Definite Matrix structure (covariance matrix of a stationary time series) or Toeplitz-Block-Toeplitz Hermitian Positive Definite Matrix structure (covariance matrix of a stationary space–time series).

F. Barbaresco (✉)
Thales Air Systems, Voie Pierre-Gilles de Gennes, F91470 Limours, France
e-mail: frederic.barbaresco@thalesgroup.com

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_7,

By extension, we introduce a new geometry for non-stationary signal through Fréchet metric space of geodesic paths on structured matrix manifolds. We conclude with extensions towards two general concepts of "Generating Inner Product" and "Generating Function".

Keywords Koszul characteristic function · Koszul entropy · Koszul forms · Laplace principle · Massieu–Duhem potential · Projective legendre duality · Contact geometry · Information geometry · Souriau geometric temperature · Balian quantum information metric · Cartan–Siegel homogeneous bounded domains · Generating function

7.1 Preamble

The Koszul–Vinberg Characteristic Function (KVCF) is a dense knot in important mathematical fields such as Hessian Geometry, Kählerian Geometry, Affine Differential Geometry and Information Geometry. As essence of Information Geometry, this paper develops KVCF as transverse concept in Thermodynamics, in Probability, in Large Deviation Theory, in Statistical Physics and in Quantum Physics. From general KVCF definition, the paper defines Koszul Entropy, that coincides with the Legendre transform of minus the logarithm of KVCF, and interprets both by analogy in thermodynamic with Dual Massieu–Duhem potentials.

In a second step, the paper proposes to synthesize and unified the three following models: Koszul model (with Koszul entropy) in Hessian Geometry, Souriau model in statistical physics (with new notion of geometric temperature) and Balian model in quantum physics (with Fisher metric as hessian of von Neumann entropy). These three models are formally similar in the characteristic function, the entropy, Legendre transform, dual coordinate system, Hessian metric and Fisher metric. These facts suggest that there exist close inter-relation and beautiful inner principle through these fields of research.

Finally, the paper studies the space of symmetric/hermitian positive definite matrices from view point of the Koszul theory though KVCF tool and extend the theory for covariance matrices for stationary and non-stationary signals characterized by Toeplitz matrix structure.

The paper concludes with extensions towards two general concepts of "Generating Inner Product" and "Generating Function". This tutorial paper will help readers to have a synthetic view of different domains of science where Hessian metric manifold is the essence, emerging from Information Geometry Theory.

This paper is not only a survey of different fields of science unified by Koszul theory, but some new results are introduced:

- Computation of Koszul Entropy from general definition of Koszul Characteristic function $\psi_\Omega(x) = \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \; \forall x \in \Omega$ and Legendre transform:

$$\Phi^*(x^*) = \left\langle x, x^* \right\rangle - \Phi(x) \text{ with } x^* = D_x \Phi \text{ and } x = D_{x^*} \Phi^* \text{ where } \Phi(x) = -\log \psi_\Omega(x)$$

$$\Phi^*(x^*) = \Phi^* \left(\int_{\Omega^*} \xi \cdot p_x(\xi) d\xi \right) = -\int_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \text{ with } x^* = \int_{\Omega^*} \xi \cdot p_x(\xi) d\xi$$

- Definition of Koszul density by analogy of Koszul Entropy with Shannon Entropy

$$p_{\bar{\xi}}(\xi) = \frac{e^{-\left\langle \xi, \Theta^{-1}(\bar{\xi}) \right\rangle}}{\int\limits_{\Omega^*} e^{-\left\langle \xi, \Theta^{-1}(\bar{\xi}) \right\rangle} d\xi} \text{ with } x = \Theta^{-1}(\bar{\xi}) \text{ and } \bar{\xi} = \Theta(x) = \frac{d\Phi(x)}{dx}$$

  where $\bar{\xi} = \int_{\Omega^*} \xi \cdot p_{\bar{\xi}}(\xi) d\xi$ and $\Phi(x) = -\log \int_{\Omega^*} e^{-\langle x, \xi \rangle} d\xi$
- Property that Barycenter of Koszul Entropy is equal to Entropy of Barycenter $E\left[\Phi^*(\xi)\right] = \Phi^*\left(E\left[\xi\right]\right), \forall \xi \in \Omega^*$

$$\int\limits_{\Omega^*} \Phi^*(\xi) p_x(\xi) d\xi = \Phi^* \left(\int\limits_{\Omega^*} \xi . p_x(\xi) d\xi \right), \forall \xi \in \Omega^*$$

- Property that Koszul density is the classical Maximum Entropy solution

$$\underset{p_x(.)}{\text{Max}} \left[-\int\limits_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \right] \text{ such } \begin{cases} \int\limits_{\Omega^*} p_x(\xi) d\xi = 1 \\ \int\limits_{\Omega^*} \xi \cdot p_x(\xi) d\xi = x^* \end{cases}$$

$$\Leftrightarrow \begin{cases} p_x(\xi) = \frac{e^{-\langle x, \xi \rangle}}{\int\limits_{\Omega^*} e^{-\langle x, \xi \rangle} d\xi} \\ x = \Theta^{-1}(x^*), \Theta(x) = \frac{d\Phi(x)}{dx} \end{cases}$$

- Strict equivalence between Hessian Koszul metric and Fisher metric from Information Geometry

$$I(x) = -E_\xi \left[\frac{\partial^2 \log p_x(\xi)}{\partial x^2} \right] = \frac{\partial^2 \log \psi_\Omega(x)}{\partial x^2}$$

- Interpretation of Information Geometry Dual Potentials in the framework of Contact Geometry through the concept of Legendre fibration and Legendre mapping.
- Computation of Koszul Density for Symmetric/Hermitian Positive Definite Matrices

$$p_x(\xi) = c_{\bar{\xi}} \cdot \beta_K \det\left(\bar{\xi}^{-1}\right)^{\frac{n+1}{2}} e^{-\frac{n+1}{2} Tr\left(\bar{\xi}^{-1}\xi\right)} \text{ with } \beta_K = \left(\frac{n+1}{2}\right)^{\frac{n+1}{2}} \text{ and}$$

$$\bar{\xi} = \int_{\Omega^*} \xi \cdot p_x(\xi) \cdot d\xi$$

- Computation of Koszul Metric for (Toeplitz-Block) Toeplitz Hermitian Positive Definite covariante matrices of stationary signal by means of Trench/Verblunsky Theorem and Complex Autoregressive model; that proves that ***a stationary space–time state of an Electromagnetic Field is coded by one point in product space RxD^{m-1}xSD^{n-1} where D is the Poincaré unit disk, and SD is the Siegel unit disk.*** The metric is given by:

$$ds^2 = n \cdot m \cdot (d\log(P_0))^2 + n\sum_{i=1}^{m-1}(m-i)\frac{|d\mu_i|^2}{\left(1-|\mu_i|^2\right)^2} + \sum_{k=1}^{n-1}(n-k)Tr\left[\left(I - A_k^k A_k^{k+}\right)^{-1} dA_k^k \left(I - A_k^{k+} A_k^k\right)^{-1} dA_k^{k+}\right]$$

$$\text{with } \begin{cases} \left(R_0, A_1^1, \ldots, A_{n-1}^{n-1}\right) \in THPD_m \times SD^{n-1} \text{ with } SD = \left\{Z/ZZ^+ < I_m\right\} \\ R_0 \rightarrow (\log(P_0), \mu_1, \ldots, \mu_{m-1}) \in R \times D^{m-1} \text{ with } D = \{z/zz^* < 1\} \end{cases}$$

- New model of non-stationary signal where one time series can be split into several stationary signals on a shorter time scale, represented by time sequence of stationary covariance matrices or a geodesic polygon/path on covariance matrix manifold.
- Definition of distance between non-stationary signals as Fréchet distance between two geodesic paths in abstract metric spaces of covariance matrix manifold.

$$\begin{cases} d_{Fréchet}(R_1, R_2) = \underset{\alpha,\beta}{Inf} \underset{t\in[0,1]}{Max} \left\{d_{geo}(R_1(\alpha(t)), R_2(\beta(t)))\right\} \\ \text{with } d_{geo}^2(R_1(\alpha(t)), R_2(\beta(t)))) = \left\|\log\left(R_1^{-1/2}(\alpha(t))R_2(\beta(t))R_1^{-1/2}(\alpha(t))\right)\right\|^2 \end{cases}$$

- Introduction of a modified Fréchet distance to take into account close dependence of elements between points of time series paths

$$d_{geo\text{-}path}(R_1, R_2) = \underset{\alpha,\beta}{Inf}\left\{\int_0^1 d_{geo}(R_1(\alpha(t)), R_2(\beta(t)))\, dt\right\}$$

We will begin this exposition by a global view of all geometries that are interrelated through the cornerstone concept of Koszul–Vinberg characteristic function and metric.

7.2 Characteristic Function in Geometry/Statistic/ Thermodynamic

In the title of the paper "Eidetic Reduction" is a reference to Husserl phenomenology [149], but should be understood in the context of following Bergson's definition of "Eidos" (synonymous of "Idea" or "morphe" in greek philosophy), emerging from his philosophy during this time period called by Frédéric Worms [152–154] " Moment 1900" which is still decisive, a hundred years later. In 1907, Henri Bergson (Philosopher, the 1927 Nobel Prize in Literature, Professor at College de France, member of Académie Française, awarded in 1930 by highest French honour, the Grand-Croix de la Légion d'honneur) gave his definition of greek concept of "*Eidos*" in the book "*Creative Evolution*" [150, 151]:

> The word, eidos, which we translate here by "Idea", has, in fact, this threefold meaning. It denotes (1) the quality, (2) the form or essence, (3) the end or design (in the sense of intention) of the act being performed, that is to say, at bottom, the design (in the sense of drawing) of the act supposed accomplished. These three aspects are those of the adjective, substantive and verb, and correspond to the three essential categories of language. After the explanations we have given above, we might, and perhaps we ought to, translated eidos by "view" or rather by "moment." For eidos is the stable view taken of the instability of things: the quality, which is a moment of becoming; the form which is a moment of evolution; the essence, which is the mean form above and below which the other forms are arranged as alterations of the mean; finally, the intention or mental design which presides over the action being accomplished, and which is nothing else, we said, than the material design, traced out and contemplated beforehand, of the action accomplished. To reduce things to Ideas is therefore to resolve becoming into its principal moments, each of these being, moreover, by the hypothesis, screened from the laws of time and, as it were, plucked out of eternity. That is to say that we end in the philosophy of Ideas when we apply the cinematographical mechanism of the intellect to the analysis of the real.

The paper title should be then considered as echo of this Bergson's citation on "*Eidos*" with twofold meaning in our development: "*Eidos*" considered with meaning of "*equilibrium state*" as defined by Bergson "*Eidos is the stable view taken of the instability of things*", but also "*Eidos*" thought as "*Eidetic Reduction*" about "*Essence*" of Information Geometry from where we will try to favor the emergence of "*Koszul characteristic function*" concept. Eidetic reduction considered as the study of "*Information Geometry*" reduced into a necessary essence, where will be identified its basic and invariable components that will be declined in different domains of Science (mathematics, physics, information theory,…).

As essence of Information Geometry, we will develop the "*Koszul Characteristic Function*", transverse concept in Thermodynamics (linked with Massieu–Duhem Potentials), in Probability (linked with Poincaré generating moment function), in Large Deviations Theory (linked with Laplace Principle), in Mechanics (linked with Contact Geometry), in Geometric Physics (linked with Souriau Geometric temperature in Mechanical Symplectic geometry) and in Quantum Physics (linked with Balian Quantum hessian metric). We will try to explore close inter-relations between these domains through geometric tools developed by Jean-Louis Koszul (Koszul forms, …).

M. Massieu a montré que, si l'on fait choix pour variables indépendantes de v et de T ou de p et de T, il existe une fonction, d'ailleurs inconnue, de laquelle les trois fonctions des variables, p, U et S dans le premier cas, v, U et S dans le second, peuvent se déduire facilement. M. Massieu a donné à cette fonction, dont la forme dépend du choix des variables, le nom de *fonction caractéristique*.

Puisque des fonctions de M. Massieu on peut déduire les autres fonctions des variables, toutes les équations de la Thermodynamique pourront s'écrire de manière à ne plus renfermer que ces fonctions et leurs dérivées; il en résultera donc, dans certains cas, une grande simplification. Nous verrons bientôt une application importante de ces fonctions.

Fig. 7.1 Text of Poincaré lecture on thermodynamic with development of the concept of "Massieu characteristic function"

First, we will observe that derivatives of the Koszul–Vinberg Characteristic Function Logarithm (KVCFL) $\log \psi_\Omega(x) = \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi$ are invariant by the automorphisms of the convex cone Ω, and so KVCFL Hessian defines naturally a Riemannian metric, that is the basic root of development in Information Geometry.

In thermodynamic, François Massieu was the first to introduce the concept of characteristic function ϕ (or Massieu–Duhem potential). This characteristic function or thermodynamic potential is able to provide all body properties from their derivatives. Among all thermodynamic Massieu–Duhem potentials, Entropy S is derived from Legendre–Moreau transform of characteristic function logarithm ϕ: $S = \phi - \beta \cdot \frac{\partial \phi}{\partial \beta}$ with $\beta = \frac{1}{kT}$ the thermodynamic temperature.

The most popular notion of "*characteristic function*" was first introduced by Henri Poincaré in his lecture on probability [40], using the property that all moments of statistical laws could be deduced from its derivatives. Paul Levy has made a systematic use of this concept. We assume that Poincaré was influenced by his school fellow at Ecole des Mines de Paris, François Massieu, and his work on thermodynamic potentials (generalized by Pierre Duhem in an Energetic Theory). This assertion is corroborated by the observation that Poincaré had added in his lecture on thermodynamics in second edition [39], one chapter on "Massieu characteristic function" with many developments and applications (Fig. 7.1).

In Large Deviation Theory, Laplace Principle allows to introduce the (scaled cumulant) Generating Function $\phi(x) = \mathrm{Lim}_{n\to\infty} \frac{1}{n} \log \psi_n(x)$ with $\psi_n(x) = \int e^{nx\varsigma} p(\Sigma_n = \varsigma) d\varsigma$ as the more natural element to characterize behavior of statistical variable for large deviations. By use of Laplace principle and the fact that the asymptotic behavior of an integral is determined by the largest value of the integrand, it can be proved that the (scaled cumulant) Generating Function $\phi(x)$ is the Legendre transform of the rate Function $I(\varsigma) = \mathrm{Lim}_{n\to\infty} -\frac{1}{n} \log P(\Sigma_n \in d\varsigma)$.

For all these cases, we can observe that the "*Characteristic function*" and its derivatives capture all information of random variable, system or physical model. Furthermore, the general notion of Entropy could be naturally defined by the Legendre Transform of the Koszul characteristic function logarithm. In the general case, Legendre transform of KVCFL will be designated as "*Koszul Entropy*".

This general notion of "*characteristic function*" has been generalized in two different directions by physicists Jean-Marie Souriau and Roger Balian.

In 1970, Jean-Marie Souriau, that had been student of Elie Cartan, introduced the concept of coadjoint action of a group on its momentum space (or "*moment map*"),

based on the orbit method works, that allows to define physical observables like energy, momentum as pure geometrical objects. For Jean-Marie Souriau, equilibriums states are indexed by a geometric parameter β with values in the Lie algebra of the Lorentz–Poincaré group. Souriau approach generalizes the Gibbs equilibrium states, β playing the role of temperature. The invariance with respect to the group, and the fact that the entropy S is a convex function of β, imposes very strict conditions, that allow Souriau to interpreted β as a space–time vector (the temperature vector of Planck), giving to the metric tensor g a null Lie derivative. In our development, before exposing Souriau theory, we will introduce Legendre Duality between the variational Euler–Lagrange and the symplectic Hamilton–Jacobi formulations of the equations of motion and will introduce Cartan–Poincaré invariant.

In 1986, Roger Balian has introduced a natural metric structure for the space of states $\hat{D}$ in quantum mechanics, from which we can deduce the distance between a state and one of its approximations. Based on quantum information theory, Roger Balian has built on physical grounds this metric as hessian metric $ds^2 = d^2S(\hat{D})$ from $S(\hat{D})$, von Neumann's entropy $S(\hat{D}) = -Tr\left[\hat{D}\ln(\hat{D})\right]$. Balian has then recovered same relations than in classical statistical physics, with also a "*quantum characteristic function*" logarithm $F(\hat{X}) = \ln Tr \exp \hat{X}$, Legendre transform of von-Neuman Entropy $S(\hat{D}) = F\left(\hat{X}\right) - \left\langle \hat{D}, \hat{X} \right\rangle$. In this framework, Balian has introduced the notion of "*Relevant Entropy*".

We will synthetize all these analogies in a table for the three models of Koszul, Souriau and Balian (Information Geometry case being a particular case of Koszul geometry).

In last chapters, we will apply the Koszul theory for defining geometry of Symmetric/Hermitian Positive Definite Matrices, and more particularly for covariance matrix of stationary signal that are characterized by specific matrix structures: Toeplitz Hermitian Positive Definite Matrix structure (covariance matrix of a stationary time series) or Toeplitz-Block-Toeplitz Hermitian Positive Definite Matrix structure (covariance matrix of a stationary space–time series). We will see that "*Toeplitz*" matrix structure could be captured by complex autogressive model parameterization. This parameterization could be naturally introduced without arbitrary through Trench's theorem (or equivalently Verblunsky's theorem) or through Partial Iwasawa decomposition. By extension, we introduce a new geometry for non-stationary signal through Fréchet metric space of geodesic paths on structured matrix manifolds.

We conclude with two general concepts of "Generating Inner Product" and "Generating Function" that extend previous developments.

The Koszul–Vinberg characteristic function is a dense knot in mathematics and could be introduced in the framework of different geometries: Hessian Geometry (Jean-Louis Koszul work), Homogeneous convex cones geometry (Ernest Vinberg work), Homogeneous Symmetric Bounded Domains Geometry (Eli Cartan and Carl Ludwig Siegel works), Symplectic Geometry (Thomas von Friedrich and Jean-Marie Souriau work), Affine Geometry (Takeshi Sasaki and Eugenio Calabi works) and Information Geometry (Calyampudi Rao [193] and Nikolai Chentsov [194] works).

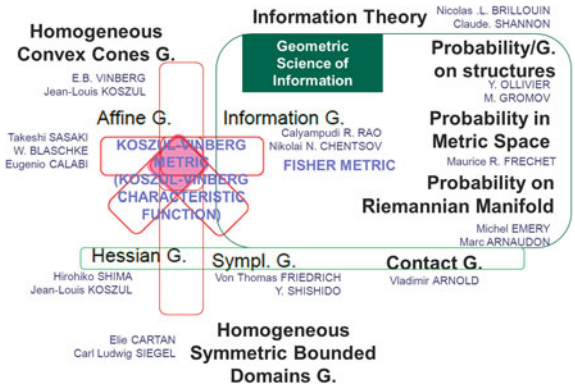


Fig. 7.2 Landscape of geometric science of information and key cornerstone position of "*Koszul–Vinberg characteristic function*"

Through Legendre duality, Contact Geometry (Vladimir Arnold work) is considered as the odd-dimensional twin of symplectic geometry and could be used to understand Legendre mapping in Information Geometry. Fisher metric of Information Geometry could be introduced as hessian metric from Koszul–Vinberg characteristic function logarithm or from Koszul Entropy (Legendre transform of Koszul–Vinberg characteristic function logarithm).

In a more general context, we can consider Information Geometry in the framework of "*Geometric Science of Information*" (see first SEE/SMF GSI'13 conference on this topic: http://www.gsi2013.org, organized at Ecole des Mines de Paris in August 2013). Geometric Science of Information also includes Probability in Metric Space (Maurice Fréchet work), Probability/Geometry on structures (Yann Ollivier and Misha Gromov works) and Probability on Riemannian Manifold (Michel Emery and Marc Arnaudon works).

In Fig. 7.2, we give the general landscape of "*Geometric Science of Information*" where "*Koszul–Vinberg characteristic function*" appears as a key cornerstone between different geometries.

7.3 Legendre Duality and Projective Duality

In following chapters, we will see that Logarithm of Characteristic function and Entropy will be related by Legendre transform, that we can consider in the context of projective duality. Duality is an old and very fruitful Idea ("*Eidos*") in Mathematics that has been constantly generalized [59–65]. A duality translates concepts, theorems or mathematical structures into other concepts, theorems or structures, in a one-to-one fashion, often by means of an involution operation and sometimes with fixed points.

Most simple duality is linear duality in the plane with points and lines (two different points can be joined by a unique line. Two different lines meet in one point unless they are parallel). By adding some points at infinity (to avoid particular case

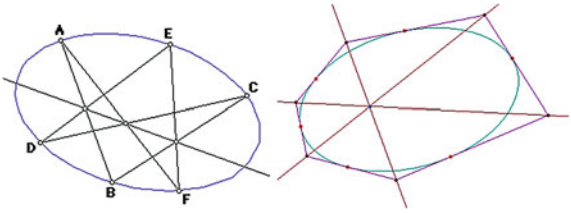


Fig. 7.3 (On the *left*) Pascal's theorem, (on the *right*) Brianchon's theorem

of parallel lines) then we obtain the projective plane in which the duality is given symmetrical relationship between points and lines, and led to the classical principle of projective duality, where the dual theorem is also a theorem.

Most Famous example is given by **Pascal's theorem** (the Hexagrammum Mysticum Theorem) stating that:

- If the vertices of a simple hexagon are points of a point conic, then its diagonal points are collinear: *If an arbitrary six points are chosen on a conic (i.e., ellipse, parabola or hyperbola) and joined by line segments in any order to form a hexagon, then the three pairs of opposite sides of the hexagon (extended if necessary) meet in three points which lie on a straight line, called the Pascal line of the hexagon*

The dual of Pascal's Theorem is known as ***Brianchon's Theorem***:

- If the sides of a simple hexagon are lines of a line conic, then the diagonal lines are concurrent (Fig. 7.3).

The Legendre(–Moreau) transform [158, 165] is an operation from convex functions on a vector space to functions on the dual space. The Legendre transform is related to projective duality and tangential coordinates in algebraic geometry, and to the construction of dual Banach spaces in analysis. Classical Legendre transform in Euclidean space is given by fixing a scalar product $\langle ., . \rangle$on R^n. For a function $F: R^n \to R \cup \{\pm\infty\}$ let:

$$G(y) = LF(y) = \operatorname*{Sup}_x \{\langle y, x\rangle - F(x)\} \tag{7.1}$$

This is an involution on the class of convex lower semi-continuous functions on R^n.

There are two dual possibilities to describe a function. We can either use a function, or we may regard the curve as the envelope of its tangent planes (Fig. 7.4).

In the framework of Information geometry, Bregman divergence appears as essentially the canonical divergence on a dually flat manifold equipped with a pair of biorthogonal coordinates induced from a pair of "potential functions" under the Legendre transform [159].

To illustrate the role of Legendre transform in Information Geometry, we provide a canonical example, with the relations for the Multivariate Normal Gaussian Law $N(m, R)$:

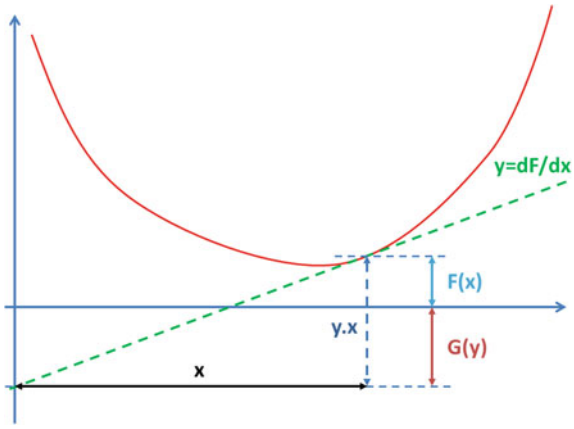


Fig. 7.4 Legendre transform $G(y)$ of $F(x)$

- Dual Coordinates systems:

$$\begin{cases} \tilde{\Theta} = (\theta, \Theta) = \left(R^{-1}m, (2R)^{-1}\right) \\ \tilde{H} = (\eta, H) = \left(m, -R + mm^T\right) \end{cases} \tag{7.2}$$

- Dual potential functions:

$$\begin{cases} \tilde{\Psi}\left(\tilde{\Theta}\right) = 2^{-}Tr\left(\Theta^{-1}\theta\theta^T\right) - 2^{-1}\log\left(\det\Theta\right) + 2^{-1}n\log(2\pi e) \\ \tilde{\Phi}\left(\tilde{H}\right) = -2^{-1}\log\left(1 + \eta^T H^{-1}\eta\right) - 2^{-1}\log\left(\det(-H)\right) - 2^{-1}n\log\left(2\pi e\right) \end{cases} \tag{7.3}$$

related by Legendre transform:

$$\tilde{\Phi}\left(\tilde{H}\right) = \left\langle\tilde{\Theta}, \tilde{H}\right\rangle - \tilde{\Psi}\left(\tilde{\Theta}\right) \text{ with } \left\langle\tilde{\Theta}, \tilde{H}\right\rangle = Tr\left(\theta\eta^T + \Theta H^T\right) \tag{7.4}$$

$$\begin{cases} \frac{\partial\tilde{\Psi}}{\partial\theta} = \eta \\ \frac{\partial\tilde{\Psi}}{\partial\Theta} = H \end{cases} \text{ and } \begin{cases} \frac{\partial\tilde{\Phi}}{\partial\eta} = \theta \\ \frac{\partial\tilde{\Phi}}{\partial H} = \Theta \end{cases} \tag{7.5}$$

with $\tilde{\Phi}\left(\tilde{H}\right) = E\left[\log p\right]$ the Entropy.

In the theory of Information Geometry introduced by Rao [193] and Chentsov [194], a Riemannian manifold is then defined by a metric tensor given by hessian of these dual potential functions:

$$g_{ij} = \frac{\partial^2\tilde{\Psi}}{\partial\tilde{\Theta}_i\partial\tilde{\Theta}_j} \text{ and } g^*_{ij} = \frac{\partial^2\tilde{\Phi}}{\partial\tilde{H}_i\partial\tilde{H}_j} \tag{7.6}$$

One of the important concepts in information geometry is mutually dual (conjugate) affine connections, and could be studied in the framework of Hessian or affine differential geometry. For more details and other developments on "dual affine connections" and "alpha-connections", we invite you to read the two books by Shun Ishi Amari [191, 192].

In this paper, we will not develop concepts of dual affine connections, but the "*hessian manifolds*" theory that was initially studied by Jean-Louis Koszul in a more general framework. In Sect. 7.4, we will expose theory of Koszul–Vinberg characteristic function on convex sharp cones that will be presented as a general framework of Information geometry.

7.4 Koszul Characteristic Function/Entropy by Legendre Duality

We define Koszul–Vinberg hessian metric on convex sharp cone, and observe that the Fisher information metric of Information Geometry coincides with the canonical Koszul Hessian metric (Koszul form) [1, 27–32]. We also observe, by Legendre duality (Legendre transform of Koszul characteristic function logarithm), that we are able to introduce a *Koszul Entropy*, that plays the role of general Entropy definition.

7.4.1 Koszul–Vinberg Characteristic Function and Metric of Convex Sharp Cone

Koszul [1, 27, 32] and Vinberg [33, 146] have introduced an affinely invariant Hessian metric on a sharp convex cone Ω through its characteristic function ψ. In the following, Ω is a sharp open convex cone in a vector space E of finite dimension on R (a convex cone is sharp if it does not contain any full straight line). In dual space E^* of E, Ω^* is the set of linear strictly positive forms on $\overline{\Omega} - \{0\}$ and Ω^* is the dual cone of Ω and is a sharp open convex cone. If $\xi \in \Omega^*$, then the intersection $\Omega \cap \{x \in E / \langle x, \xi \rangle = 1\}$ is bounded. $G = Aut(\Omega)$ is the group of linear transform of E that preserves Ω. $G = Aut(\Omega)$ operates on Ω^* by $\forall g \in G = Aut(\Omega)$, $\forall \xi \in E^*$ then $\tilde{g} \cdot \xi = \xi \circ g^{-1}$.

Koszul–Vinberg Characteristic function definition: Let $d\xi$ be the Lebesgue measure on E^*, the following integral:

$$\psi_\Omega(x) = \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi \ \ \forall x \in \Omega \tag{7.7}$$

with Ω^* the dual cone is an analytic function on Ω, with$\psi_\Omega(x) \in]0, +\infty[$, called the ***Koszul–Vinberg characteristic function*** of cone Ω, with the properties:

- The Bergman kernel of $\Omega + iR^{n+1}$ is written as $K_\Omega(\mathrm{Re}(z))$ up to a constant where K_Ω is defined by the integral:

$$K_\Omega(x) = \int_{\Omega^*} e^{-\langle \xi, x\rangle} \psi_{\Omega^*}(\xi)^{-1} d\xi \tag{7.8}$$

- ψ_Ω is analytic function defined on the interior of Ω and $\psi_\Omega(x) \to +\infty$ as $x \to \partial\Omega$ If $g \in Aut(\Omega)$ then $\psi_\Omega(gx) = |\det g|^{-1} \psi_\Omega(x)$ and since $tI \in G = Aut(\Omega)$ for any $t > 0$, we have

$$\psi_\Omega(tx) = \psi_\Omega(x)/t^n \tag{7.9}$$

- ψ_Ω is logarithmically strictly convex, and $\varphi_\Omega(x) = \log(\psi_\Omega(x))$ is strictly convex

Koszul 1-form α: The differential 1-form

$$\alpha = d\varphi_\Omega = d\log\psi_\Omega = d\psi_\Omega/\psi_\Omega \tag{7.10}$$

is invariant by all automorphisms $G = Aut(\Omega)$ of Ω. If $x \in \Omega$ and $u \in E$ then

$$\langle \alpha_x, u\rangle = -\int_{\Omega^*} \langle \xi, u\rangle \cdot e^{-\langle \xi, x\rangle} d\xi \text{ and } \alpha_x \in -\Omega^* \tag{7.11}$$

Koszul 2-form β: The symmetric differential 2-form

$$\beta = D\alpha = d^2 \log\psi_\Omega \tag{7.12}$$

is a positive definite symmetric bilinear form on E invariant under $G = Aut(\Omega)$.

$$D\alpha > 0 \text{ (Schwarz inequality and } d^2\log\psi_\Omega(u,v) = \int_{\Omega^*} \langle \xi, u\rangle \langle \xi, v\rangle e^{-\langle \xi, u\rangle} d\xi) \tag{7.13}$$

Koszul–Vinberg Metric: $D\alpha$ defines a Riemanian structure invariant by $Aut(\Omega)$, and then the Riemanian metric is given by

$$g = d^2 \log\psi_\Omega \tag{7.14}$$

$$d^2\log\psi(x) = d^2\left[\log\int \psi_u du\right] = \frac{\int \psi_u d^2\log\psi_u du}{\int \psi_u du} + \frac{1}{2}\frac{\iint \psi_u \psi_v (d\log\psi_u - d\log\psi_v)^2 du dv}{\iint \psi_u \psi_v du dv}$$

A diffeomorphism is used to define dual coordinate:

$$x^* = -\alpha_x = -d \log \psi_\Omega(x) \tag{7.15}$$

with $\langle df(x), u\rangle = D_u f(x) = \frac{d}{dt}\Big|_{t=0} f(x+tu)$. When the cone Ω is symmetric, the map $x^* = -\alpha_x$ is a bijection and an isometry with a unique fixed point (the manifold is a Riemannian Symmetric Space given by this isometry): $(x^*)^* = x$, $\langle x, x^*\rangle = n$ and $\psi_\Omega(x)\psi_{\Omega^*}(x^*) = cste \cdot x^*$ is characterized by $x^* = \arg\min\{\psi(y)/y \in \Omega^*, \langle x, y\rangle = n\}$ and x^* is the center of gravity of the cross section $\{y \in \Omega^*, \langle x, y\rangle = n\}\, of\, \Omega^*$:

$$\begin{aligned} x^* &= \int_{\Omega^*} \xi \cdot e^{-\langle \xi, x\rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi, \\ &\langle -x^*, h\rangle = d_h \log \psi_\Omega(x) \\ &= -\int_{\Omega^*} \langle \xi, h\rangle\, e^{-\langle \xi, x\rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \end{aligned} \tag{7.16}$$

7.4.2 Koszul Entropy and Its Barycenter

From this last equation, we can deduce "***Koszul Entropy***" ***defined as Legendre Transform of minus logarithm of Koszul–Vinberg characteristic function*** $\Phi(x)$:

$$\Phi^*(x^*) = \langle x, x^*\rangle - \Phi(x) \text{ with } x^* = D_x\Phi \text{ and } x = D_{x^*}\Phi^* \text{ where } \Phi(x) = -\log \psi_\Omega(x) \tag{7.17}$$

$$\Phi^*(x^*) = \left\langle (D_x\Phi)^{-1}(x^*), x^*\right\rangle - \Phi\left[(D_x\Phi)^{-1}(x^*)\right] \ \forall x^* \in \{D_x\Phi(x)/x \in \Omega\} \tag{7.18}$$

By (7.11), and using that $-\langle \xi, x\rangle = \log e^{-\langle \xi, x\rangle}$ we can write:

$$-\langle x^*, x\rangle = \int_{\Omega^*} \log e^{-\langle \xi, x\rangle} \cdot e^{-\langle \xi, x\rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \tag{7.19}$$

and

$$\begin{aligned} \Phi^*(x^*) &= \langle x, x^*\rangle - \Phi(x) = -\int_{\Omega^*} \log e^{-\langle \xi, x\rangle} \cdot e^{-\langle \xi, x\rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi + \log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \\ \Phi^*(x^*) &= \left[\left(\int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi\right) \cdot \log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi - \int_{\Omega^*} \log e^{-\langle \xi, x\rangle} \cdot e^{-\langle \xi, x\rangle} d\xi\right] \Big/ \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \end{aligned} \tag{7.20}$$

We can then consider this Legendre transform as an Entropy, named ***Koszul Entropy*** if we rewrite equation (7.20) in the new form

$$\Phi^* = -\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \tag{7.21}$$

with

$$p_x(\xi) = e^{-\langle \xi, x\rangle}\int_{\Omega^*} e^{-\langle \xi, x\rangle}d\xi = e^{-\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi} = e^{-\langle x,\xi\rangle+\Phi(x)} \text{ and}$$

$$x^* = \int_{\Omega^*} \xi \cdot p_x(\xi)d\xi \tag{7.22}$$

We will call $p_x(\xi) = \frac{e^{-\langle \xi,x\rangle}}{\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi}$ the Koszul Density, with the property that:

$$\log p_x(\xi) = -\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi = -\langle x,\xi\rangle + \Phi(x) \tag{7.23}$$

and

$$E_\xi\left[-\log p_x(\xi)\right] = \left\langle x, x^*\right\rangle - \Phi(x) \tag{7.24}$$

We can observe that:

$$\begin{aligned}\Phi(x) = -\log\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi = -\log\int_{\Omega^*} e^{-[\Phi^*(\xi)+\Phi(x)]}d\xi = \Phi(x) - \log\int_{\Omega^*} e^{-\Phi^*(\xi)}d\xi \\ \Rightarrow \int_{\Omega^*} e^{-\Phi^*(\xi)}d\xi = 1\end{aligned} \tag{7.25}$$

To make appear x^* in $\Phi^*(x^*)$, we have to write:

$$\begin{aligned}\log p_x(\xi) = \log e^{-\langle x,\xi\rangle+\Phi(x)} = \log e^{-\Phi^*(\xi)} = -\Phi^*(\xi) \\ \Rightarrow \Phi^* = -\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi = \int_{\Omega^*} \Phi^*(\xi)p_x(\xi)d\xi = \Phi^*(x^*)\end{aligned} \tag{7.26}$$

The last equality is true if and only if:

$$\int_{\Omega^*} \Phi^*(\xi)p_x(\xi)d\xi = \Phi^*\left(\int_{\Omega^*} \xi \cdot p_x(\xi)d\xi\right) \text{ as } x^* = \int_{\Omega^*} \xi \cdot p_x(\xi)d\xi \tag{7.27}$$

This condition could be written:

$$E\left[\Phi^*(\xi)\right] = \Phi^*\left(E\left[\xi\right]\right), \xi \in \Omega^* \tag{7.28}$$

The meaning of this relation is that "***Barycenter of Koszul Entropy is Koszul Entropy of Barycenter***".
This condition is achieved for $x^* = D_x\Phi$ taking into account Legendre Transform property:

$$\begin{aligned}\text{Legendre Transform: } \Phi^*(x^*) &= \sup_x\left[\left\langle x, x^*\right\rangle - \Phi(x)\right] \\ &\Rightarrow \begin{cases}\Phi^*(x^*) \geq \langle x, x^*\rangle - \Phi(x) \\ \Phi^*(x^*) \geq \int\limits_{\Omega^*} \Phi^*(\xi) p_x(\xi) d\xi\end{cases} \\ &\Rightarrow \begin{cases}\Phi^*(x^*) \geq E\left[\Phi^*(\xi)\right] \\ \text{equality for } x^* = \frac{d\Phi}{dx}\end{cases}\end{aligned} \tag{7.29}$$

7.4.3 Relation with Maximum Entropy Principle

Classically, the density given by Maximum Entropy Principle [188–190] is given by:

$$\max_{p_x(.)}\left[-\int\limits_{\Omega^*} p_x(\xi)\log p_x(\xi) d\xi\right] \text{ such } \begin{cases}\int\limits_{\Omega^*} p_x(\xi) d\xi = 1 \\ \int\limits_{\Omega^*} \xi \cdot p_x(\xi) d\xi = x^*\end{cases} \tag{7.30}$$

If we take $q_x(\xi) = e^{-\langle \xi, x\rangle} / \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi = e^{-\langle x, \xi\rangle - \log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi}$ such that:

$$\begin{cases}\int\limits_{\Omega^*} q_x(\xi) \cdot d\xi = \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \Big/ \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi = 1 \\ \log q_x(\xi) = \log e^{\overbrace{-\langle x, \xi\rangle - \log \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi}} = -\langle x, \xi\rangle - \log \int\limits_{\Omega^*} e^{-\langle x, \xi\rangle} d\xi\end{cases} \tag{7.31}$$

Then by using the fact that $\log x \geq \left(1 - x^{-1}\right)$ with equality if and only if $x = 1$, we find the following:

$$-\int\limits_{\Omega^*} p_x(\xi) \log \frac{p_x(\xi)}{q_x(\xi)} d\xi \leq -\int\limits_{\Omega^*} p_x(\xi)\left(1 - \frac{q_x(\xi)}{p_x(\xi)}\right) d\xi \tag{7.32}$$

We can then observe that:

$$\int\limits_{\Omega^*} p_x(\xi)\left(1 - \frac{q_x(\xi)}{p_x(\xi)}\right) d\xi = \int\limits_{\Omega^*} p_x(\xi) d\xi - \int\limits_{\Omega^*} q_x(\xi) d\xi = 0 \tag{7.33}$$

because $\int_{\Omega^*} p_x(\xi)d\xi = \int\limits_{\Omega^*} q_x(\xi)d\xi = 1$.
We can then deduce that:

$$-\int\limits_{\Omega^*} p_x(\xi)\log\frac{p_x(\xi)}{q_x(\xi)}d\xi \leq 0 \Rightarrow -\int\limits_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \leq -\int\limits_{\Omega^*} p_x(\xi)\log q_x(\xi)d\xi \tag{7.34}$$

If we develop the last inequality, using expression of $q_x(\xi)$:

$$-\int\limits_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \leq -\int\limits_{\Omega^*} p_x(\xi)\left[-\langle x,\xi\rangle - \log\int\limits_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi\right]d\xi \tag{7.35}$$

$$-\int\limits_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \leq \left\langle x, \int\limits_{\Omega^*} \xi\cdot p_x(\xi)d\xi\right\rangle + \log\int\limits_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi \tag{7.36}$$

If we take $x^* = \int_{\Omega^*} \xi\cdot p_x(\xi)d\xi$ and $\Phi(x) = -\log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi$, then we deduce that the Koszul density $q_x(\xi) = e^{-\langle \xi,x\rangle}/\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi = e^{-\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi}$ is the Maximum Entropy solution constrained by $\int_{\Omega^*} p_x(\xi)d\xi = 1$ and $\int_{\Omega^*} \xi\cdot p_x(\xi)d\xi = x^*$:

$$-\int\limits_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \leq \langle x, x^*\rangle - \Phi(x) \tag{7.37}$$

$$-\int\limits_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \leq \Phi^*(x^*) \tag{7.38}$$

We have then proved that Koszul Entropy provides density of Maximum Entropy:

$$p_{\bar{\xi}}(\xi) = \frac{e^{-\langle \xi,\Theta^{-1}(\bar{\xi})\rangle}}{\int\limits_{\Omega^*} e^{-\langle \xi,\Theta^{-1}(\bar{\xi})\rangle}d\xi} \text{ with } x = \Theta^{-1}(\bar{\xi}) \text{ and } \bar{\xi} = \Theta(x) = \frac{d\Phi(x)}{dx} \tag{7.39}$$

where

$$\bar{\xi} = \int\limits_{\Omega^*} \xi\cdot p_{\bar{\xi}}(\xi)d\xi \text{ and } \Phi(x) = -\log\int\limits_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi \tag{7.40}$$

We can then deduce Maximum Entropy solution without solving Classical variational problem with Lagrangian hyperparameters, but only by inversing function $\Theta(x)$ .This remark was made by Jean-Marie Souriau in the paper [97], if we take vector with tensor components $\xi = \begin{pmatrix} z \\ z\otimes z \end{pmatrix}$, components of $\bar{\xi}$ will provide moments of first and second order of the density of probability $p_{\bar{\xi}}(\xi)$, that is defined by Gaussian

law. In this particular case, we can write:

$$\langle \xi, x \rangle = a^T z + \frac{1}{2} z^T H z \tag{7.41}$$

with $a \in R^n$ and $H \in Sym(n)$. By change of variable given by $z' = H^{1/2} z + H^{-1/2} a$, we can then compute the logarithm of the Koszul characteristic function:

$$\Phi(x) = -\frac{1}{2}\left[a^T H^{-1} a + \log\det\left[H^{-1}\right] + n\log(2\pi)\right] \tag{7.42}$$

We can prove that first moment is equal to $-H^{-1}a$ and that components of variance tensor are equal to elements of matrix H^{-1} , that induces the second moment. The Koszul Entropy, its Legendre transform, is then given by:

$$\Phi^*(\bar{\xi}) = \frac{1}{2}\left[\log\det\left[H^{-1}\right] + n\log(2\pi \cdot e)\right] \tag{7.43}$$

7.4.4 Crouzeix Relation and Its Consequences

The duality between dual potential functions is recovered by this relation:

$$\Phi^*(x^*) + \Phi(x) = \left\langle x, x^* \right\rangle \text{with} \, x^* = \frac{d\Phi}{dx} \text{and} \, x = \frac{d\Phi^*}{dx^*} \text{where} \, \Phi(x) = -\log \psi_\Omega(x)$$

$$\left\{ \begin{array}{l} \frac{d\Phi}{dx} = x^* \\ \frac{d\Phi^*}{dx^*} = x \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \frac{d^2\Phi}{dx^2} = \frac{dx^*}{dx} \\ \frac{d^2\Phi^*}{dx^{*2}} = \frac{dx}{dx^*} \end{array} \right. \Rightarrow \frac{d^2\Phi}{dx^2} \cdot \frac{d^2\Phi^*}{dx^{*2}} = 1 \Rightarrow \frac{d^2\Phi}{dx^2} = \left[\frac{d^2\Phi^*}{dx^{*2}}\right]^{-1}$$
$$\Rightarrow ds^2 = -\frac{d^2\Phi}{dx^2} dx^2 = -\left[\frac{d^2\Phi^*}{dx^{*2}}\right]^{-1} \cdot \left[\frac{d^2\Phi^*}{dx^{*2}} \cdot dx^*\right]^2 = -\frac{d^2\Phi^*}{dx^{*2}} \cdot dx^{*2} \tag{7.44}$$

The relation $\frac{d^2\Phi}{dx^2} = \left[\frac{d^2\Phi^*}{dx^{*2}}\right]^{-1}$ has been established by J.P. Crouzeix in 1977 in a short communication [176] for convex smooth functions and their Legendre transforms. This result has been extended for non-smooth function by Seeger [177] and Hiriart-Urruty [178], using polarity relationship between the second-order subdifferentials. This relation was mentioned in texts of variational calculus and theory of elastic materials (with work potentials) [178].

This last relation has also be used in the framework of the Monge-Ampere measure associated to a convex function, to prove equality with Lebesgue measure λ:

$$m_\Phi(\Lambda) = \int_\Lambda \varphi(x)dx = \lambda\left(\{\nabla\phi(\mathrm{x})/\mathrm{x} \in \Lambda\}\right) \tag{7.45}$$

$$\forall \Lambda \in B_\Omega \text{ (Borel set in } \Omega) \text{ and } \phi(x) = \det\left[\nabla^2 \Phi(x)\right]$$

That is proved using Crouezix relation: $\nabla^2\Phi(x) = \nabla^2\Phi\left(\nabla\Phi^*(y)\right) = \left[\nabla^2\Phi^*(y)\right]^{-1}$

$$m_\Phi(\Lambda) = \int_\Lambda \varphi(x)dx = \int_\Lambda \det\left[\nabla^2\Phi(x)\right]\cdot dx$$

$$m_\Phi(\Lambda) = \int_{(\nabla\Phi^*)^{-1}(A)} \det\left[\nabla^2\Phi\left(\nabla\Phi^*(y)\right)\right]\cdot\det\left[\nabla^2\Phi^*(y)\right]dy$$

$$= \int_{\nabla\Phi(\Lambda)} 1\cdot dy = \lambda\left(\{\nabla\phi(\mathrm{x})/\mathrm{x}\in\Lambda\}\right) \tag{7.46}$$

7.4.5 Koszul Metric and Fisher Metric

To make the link with Fisher metric given by Fisher Information matrix $I(x)$, we can observe that the second derivative of $\log p_x(\xi)$ is given by:

$$p_x(\xi) = e^{-\langle\xi,x\rangle}/\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi = e^{-\langle x,\xi\rangle-\log\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi}$$

$$\Rightarrow \log p_x(\xi) = -\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi \tag{7.47}$$

with $\Phi(x) = -\log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi = -\log\Psi_\Omega(x)$

$$\frac{\partial^2\log p_x(\xi)}{\partial x^2} = \frac{\partial^2\Phi(x)}{\partial x^2} \Rightarrow I(x) = -E_\xi\left[\frac{\partial^2\log p_x(\xi)}{\partial x^2}\right] = -\frac{\partial^2\Phi(x)}{\partial x^2} = \frac{\partial^2\log\psi_\Omega(x)}{\partial x^2} \tag{7.48}$$

$$I(x) = -E_\xi\left[\frac{\partial^2\log p_x(\xi)}{\partial x^2}\right] = \frac{\partial^2\log\psi_\Omega(x)}{\partial x^2} \tag{7.49}$$

We could then deduce the close interrelation between Fisher metric and hessian of Koszul–Vinberg characteristic logarithm, that are totally equivalent.

7.4.6 Extended Results by Koszul, Vey and Sasaki

Koszul [1] and Vey [2, 160] have developed these results with the following theorem for connected Hessian manifolds:

Koszul–Vey Theorem: Let M be a connected Hessian manifold with Hessian metric g. Suppose that admits a closed 1-form α such that $D\alpha = g$ and there exists a group G of affine automorphisms of M preserving α:

- If M/G is quasi-compact, then the universal covering manifold of M is affinely isomorphic to a convex domain Ω real affine space not containing any full straight line.
- If M/G is compact, then Ω is a sharp convex cone.

On this basis, Jean-Louis Koszul has given a Lie Group construction of homogeneous cone that has been developed and applied in Information Geometry by Hirohiko Shima [163, 164] in the framework of hessian geometry.

Sasaki has developed the study of Hessian manifolds in Affine Geometry [49, 50]. He has denoted by S_c the level surface of ψ_Ω: $S_c = \{\psi_\Omega(x) = c\}$ which is a non-compact submanifold in Ω, and by ω_c the induced metric of $d^2 \log \psi_\Omega$ on S_c, then assuming that the cone Ω is homogeneous under $G(\Omega)$, he proved that S_c is a homogeneous hyperbolic affine hypersphere and every such hyperspheres can be obtained in this way .Sasaki also remarks that ω_c is identified with the affine metric and S_c is a global Riemannian symmetric space when Ω is a self-dual cone. He conclude that, let Ω be a regular convex cone and let $g = d^2 \log \psi_\Omega$ be the canonical Hessian metric, then each level surface of the characteristic function ψ_Ω is a minimal surface of the Riemannian manifold (Ω, g) .

7.4.7 Geodesics for Koszul Hessian Metric

Last contributor is Rothaus [147] that has studied the construction of geodesics for this hessian metric geometry, using the following property:

$$\Gamma^i_{jk} = \frac{1}{2} g^{il} \left(\frac{\partial g_{kl}}{\partial x_j} + \frac{\partial g_{jl}}{\partial x_k} - \frac{\partial g_{jk}}{\partial x_l} \right) = \frac{1}{2} g^{il} \frac{\partial^3 \log \psi_\Omega(x)}{\partial x_j \partial x_k \partial x_l} \text{with } g_{ij} = \frac{\partial^2 \log \psi_\Omega(x)}{\partial x_i \partial x_j} \tag{7.50}$$

or expressed also according the Christoffel symbol of the first kind:

$$[i, jk] = \frac{1}{2} \left(\frac{\partial g_{kl}}{\partial x_j} + \frac{\partial g_{jl}}{\partial x_k} - \frac{\partial g_{jk}}{\partial x_l} \right) = \frac{1}{2} \frac{\partial^3 \log \psi_\Omega(x)}{\partial x_j \partial x_k \partial x_l} \tag{7.51}$$

Then geodesic is given by:

$$\frac{d^2 x_k}{ds^2} + \Gamma^k_{ij} \frac{dx_i}{ds} \frac{dx_j}{ds} = g_{lk} \frac{d^2 x_k}{ds^2} + [k, ij] \frac{dx_i}{ds} \frac{dx_j}{ds} = 0 \tag{7.52}$$

that could be developed with previous relation:

$$\frac{d^2x_k}{ds^2}\frac{\partial^2 \log \psi_\Omega}{\partial x_k \partial x_l} + \frac{1}{2}\frac{dx_i}{ds}\frac{dx_j}{ds}\frac{\partial^3 \log \psi_\Omega}{\partial x_l \partial x_i \partial x_j} = 0 \tag{7.53}$$

We can then observe that:

$$\frac{d^2}{ds^2}\left[\frac{\partial \log \psi_\Omega}{\partial x_l}\right] = \frac{dx_i}{ds}\frac{dx_j}{ds}\frac{\partial^3 \log \psi_\Omega}{\partial x_i \partial x_j \partial x_l} + \frac{d^2x_k}{ds^2}\frac{\partial^2 \log \psi_\Omega}{\partial x_k \partial x_l} \tag{7.54}$$

The geodesic equation can then be rewritten:

$$\frac{d^2x_k}{ds^2}\frac{\partial^2 \log \psi_\Omega}{\partial x_k \partial x_l} + \frac{d^2}{ds^2}\left[\frac{\partial \log \psi_\Omega}{\partial x_l}\right] = 0 \tag{7.55}$$

That we can put in vector form using notations $x^* = -d \log \psi_\Omega$ and Fisher matrix $I(x) = d^2 \log \psi_\Omega$:

$$I(x)\frac{d^2x}{ds^2} - \frac{d^2x^*}{ds^2} = 0 \tag{7.56}$$

7.4.8 Koszul Forms and Metric for Symmetric Positive Definite Matrices

Using Koszul results of Sect. 7.4.2, let v be the volume element of g. We define a closed 1-form α and β a symmetric bilinear form by: $D_X v = \alpha(X)v$ and $\beta = D\alpha$. The forms α and β, called the first Koszul form and the second Koszul form for a Hessian structure *(D; g)* respectively are given by:

$$v = \left(\det\left[g_{ij}\right]\right)^{1/2} dx^1 \wedge \cdots dx^n \Rightarrow \alpha_i = \frac{\partial}{\partial x^i}\log\left(\det\left[g_{ij}\right]\right)^{\frac{1}{2}} v \text{ and}$$

$$\beta_{ij} = \frac{\partial \alpha_i}{\partial x^j} = \frac{1}{2}\frac{\partial^2 \log \det\left[g_{kl}\right]}{\partial x^i \partial x^j}$$

The pair *(D; g)* of flat connection D and Hessian metric g define the Hessian structure. As seen previously, Koszul studied flat manifolds endowed with a closed 1-form α such that $D\alpha$ is positive definite, whereupon $D\alpha$ is a Hessian metric. The Hessian structure *(D; g)* is said to be of Koszul type, if there exists a closed 1-form α such that $g = D\alpha$. The second Koszul form β plays a role similar to the Ricci tensor for Kählerian metric.

We can then apply this Koszul geometry framework for cones of Symmetric Positive Definite Matrices.

Let the inner product $\langle x, y\rangle = Tr\,(xy)\,, \forall x, y \in Sym_n(R)$, Ω be the set of symmetric positive definite matrices is an open convex cone and is self-dual $\Omega^* = \Omega$.

$$\psi_\Omega(x) = \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \quad \begin{matrix} = \\ \langle x, y\rangle = Tr(xy) \\ \Omega^* = \Omega \text{ self-dual} \end{matrix} \quad \det x^{-\frac{n+1}{2}} \psi(I_n) \tag{7.57}$$

$$g = d^2 \log \psi_\Omega = -\frac{n+1}{2} d^2 \log \det x \text{ and } x^* = -d \log \psi_\Omega = \frac{n+1}{2} d \log \det x$$
$$= \frac{n+1}{2} x^{-1} \tag{7.58}$$

Sasaki approach could also be used for the regular convex cone consisting of all positive definite symmetric matrices of degree n. $(D, Dd \log \det x)$ is a Hessian structure on Ω, and each level surface of $\det x$ is a minimal surface of the Riemannian manifold $(\Omega, g = -Dd \log \det x)$.

Koszul [27] has introduced another 1-form definition for homogeneous bounded domain given by:

$$\alpha = -\frac{1}{4} d\Psi(X) \text{ with } \Psi(X) = Tr_{g/b}\left[ad(JX) - Jad(X)\right] \ \forall X \in g \tag{7.59}$$

We can illustrate this new Koszul expression for Poincaré's Upper Half Plane $V = \{z = x + iy/y > 0\}$ (most simple symmetric homogeneous bounded domain). Let vector fields $X = y\frac{d}{dx}$ and $Y = y\frac{d}{dy}$, and J tensor of complex structure V defined by

$$JX = Y. \text{ As } [X, Y] = -Y \text{and } ad(Y) \,.\, Z = [Y, Z] \text{ then } \begin{cases} Tr\left[ad(JX) - Jad(X)\right] = 2 \\ Tr\left[ad(JY) - Jad(Y)\right] = 0 \end{cases} \tag{7.60}$$

The Koszul 1-form and then the Koszul/Poincaré metric is given by:

$$\Psi(X) = 2\frac{dx}{y} \Rightarrow \alpha = -\frac{1}{4} d\Psi = -\frac{1}{2}\frac{dx \wedge dy}{y^2} \Rightarrow ds^2 = \frac{dx^2 + dy^2}{2y^2} \tag{7.61}$$

This could be also applied for Siegel's Upper Half Plane $V = \{Z = X + iY \,/\, Y > 0\}$ (more natural extension of Poincaré Upper-half plane, and general notion of symmetric bounded homogeneous domains studied by Elie Cartan and Carl-Luwid siegel):

$$\begin{cases} SZ = (AZ + B)D^{-1} \\ A^T D = I, B^T D = D^T B \end{cases} \text{ with } S = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} \text{ and } J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{7.62}$$

$$\Psi(dX + idY) = \frac{3p+1}{2} Tr\left(Y^{-1} dX\right) \Rightarrow \begin{cases} \alpha = -\frac{1}{4} d\Psi = \frac{3p+1}{8} Tr\left(Y^{-1} dZ \wedge Y^{-1} d\bar{Z}\right) \\ ds^2 = \frac{(3p+1)}{8} Tr\left(Y^{-1} dZ Y^{-1} d\bar{Z}\right) \end{cases} \tag{7.63}$$

To recover the metric of Symmetric Positive Definite (HPD) matrices, we take $Z = iR$ (with $X = 0$), the metric $ds^2 = Tr\left[\left(R^{-1}dR\right)^2\right]$. In the context of Information Geometry, this metric is metric for multivariate Gaussian law of covariance matrix R and zero mean. This metric will be studied in Sect. 7.7

7.4.9 *Koszul–Vinberg Characteristic Function as Universal Barrier in Convex Programming*

Convex Optimization theory has developed the notion of Universal barrier that could be also interpreted as Koszul–Vinberg Characteristic function. Homogeneous cones that have been studied by Elie Cartan, Carl Ludwig Siegel, Ernest Vinberg and Jean-Louis Koszul are very specific cones because of their invariance properties under linear transformations. They were classified and algebraic constructed by Vinberg using Siegel domains. Homogeneous cones [41, 42, 53–55] have also been studied for developing interior-point algorithms in optimization, through the notion of self-concordant barriers. Nesterov and Nemirovskii [145] have shown that any cone in R^n admits a logarithmically homogeneous universal barrier function, defined as a volume integral. Güler and Tunçel [37, 38, 143, 148] have used a recursive scheme to construct a barrier through Siegel domains referring to a more general construction of Nesterov and Nemirovskii. More recently, Olena Shevchenko [142] has defined a recursive formula for optimal dual barrier functions on homogeneous cones, based on the primal construction of Güler and Tunçel by means of the dual Siegel cone construction of Rothaus [144, 147].

Vinberg and Gindikin have proved that every homogeneous cones can be obtained by a Siegel construction. To illustrate Siegel construction, we can consider K an homogeneous cone in R^k and $B(u, u) \in K$ a K-bilinear symmetric form, Siegel Cone is then defined by:

$$C_{Siegel}(K, B) = \left\{(x, u, t) \in R^k \mathrm{x} R^p \mathrm{x} R / t > 0, tx - B(u, u) \in K\right\} \tag{7.64}$$

and is homogeneous.

Rothaus has considered dual Siegel domain construction by means of a symmetric linear mapping $U(y)$ (positive definite for $y \in \operatorname{int} K^*$) defined by $\langle U(y)u, v\rangle = \langle B(u, v), y\rangle$, from which we can define dual Siegel cone:

$$C^*_{Siegel}(K, B) = \left\{(y, v, s) \in K^* \mathrm{x} R^p \mathrm{x} R / t > 0, s > \left\langle U(y)^{-1}v, v\right\rangle\right\} \tag{7.65}$$

With respect to the inner product $\langle (x, u, t), (y, v, s)\rangle = \langle x, y\rangle + 2\langle u, v\rangle + st$.
With the condition $s > \left\langle U(y)^{-1}v, v\right\rangle$ that is equivalent by means of Schur complement technique to:

$$\begin{bmatrix} s & v^T \\ v & U(y) \end{bmatrix} > 0 \tag{7.66}$$

7.4.10 Characteristic Function and Laplace Principle of Large Deviations

In probability, the "characteristic function" has been introduced by Poincaré by analogy with Massieu work in thermodynamic. For the Koszul–Vinberg characteristic function, if we replace the Lebesgue measure by Borel measure, then we recover the classical definition of characteristic function in Probability, and the previous KVCF could be compared by analogy with:

$$\psi_X(z) = \int_{-\infty}^{+\infty} e^{izx} dF(x) = \int_{-\infty}^{+\infty} e^{izx} p(x) \cdot dx = E\left[e^{izx}\right] \tag{7.67}$$

Let μ be a positive Borel Measure on euclidean space V. Assume that the following integral is finite for all x in an open set

$$\Omega \subset V \colon \psi_x(y) = \int e^{-\langle y,x\rangle} d\mu(x) \tag{7.68}$$

For $x \in \Omega$, consider the probability measure:

$$p\,(y, dx) = \frac{1}{\psi_x(y)} e^{-\langle y,x\rangle} d\mu(x) \tag{7.69}$$

then mean is given by:

$$m(y) = \int xp(y, dx) = -\nabla \log \psi_x(y) \tag{7.70}$$

and covariance

$$\langle V(y)u, v\rangle = \int \langle x - m(y), u\rangle \langle x - m(y), v\rangle \, p(y, dx) = D_u D_v \log \psi_x(y) \tag{7.71}$$

In large deviation theory, the objective is to study the asymptotic behavior of remote tails of sequences of probability distributions and was initialized by Laplace [10] and unified formally by Sanov [44] and Varadhan [43]. Large deviations theory concerns itself with the exponential decline of the probability measures of extreme, tail or rare events, as the number of observations grows arbitrarily large [45, 46].

Large deviation principle: Let $\{\Sigma_n\}$ be a sequence of random variables indexed by the positive integer n, and let $P(\Sigma_n \in d\varsigma) = P(\Sigma_n \in [\varsigma, \varsigma + d\varsigma])$ denote the probability measure associated with these random variables. We say that $\Sigma_n$or $P(\Sigma_n \in d\varsigma)$ satisfies a large deviation principle if the limit $I(\varsigma) = \text{Lim}_{n\to\infty} -\frac{1}{n}\log P(\Sigma_n \in d\varsigma)$ exists The function $I(\varsigma)$ defined by this limit is called the rate function; the parameter n of decay is called in large deviation theory the speed. The existence of a large deviation principle for Σ_n means concretely that the dominant behavior of $P(\Sigma_n \in d\varsigma)$ is a decaying exponential with n, with rate $I(\varsigma) : P(\Sigma_n \in d\varsigma) \approx e^{-n\cdot I(\varsigma)} d\varsigma$ or $p(\Sigma_n = \varsigma) \approx e^{-n\cdot I(\varsigma)}$.

If we write the Legendre-Fenchel transform of a function h(x) defined by $g(k) = \underset{x}{\text{Sup}}\{k \cdot x - h(x)\}$, we can express the following principle and Laplace's method:

Laplace Principle [10]: Let $\{(\Omega_n, F_n, P_n), n \in N\}$ be a sequence of probability spaces, Θ a complete separable metric space, $\{Y_n, n \in N\}$ a sequence of random variables such that Y_n maps Ω_n into Θ, and I a rate function on Θ. Then, Y_n satisfies the Laplace principle on Θ with rate function I if for all bounded, continuous functions f mapping Θ into R:

$$\lim_{n\to\infty} \frac{1}{n} \log E_{P_n}\left[e^{n\cdot f(Y_n)}\right] = \sup_{x\in\Theta}\{f(x) - I(x)\} \tag{7.72}$$

If Y_n satisfies the large deviation principle on Θ with rate function I, then $P_n(Y_n \in dx) \approx e^{-n\cdot I(x)} dx$:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log E_{P_n}\left[e^{n\cdot f(Y_n)}\right] &= \lim_{n\to\infty} \frac{1}{n} \log \int_{\Omega_n} e^{nf(Y_n)} \cdot dP_n \\
&= \lim_{n\to\infty} \frac{1}{n} \log \int_{\Theta} e^{n\cdot f(x)} P_n(Y_n \in dx) \\
&\approx \lim_{n\to\infty} \frac{1}{n} \log \int_{\Theta} e^{n\cdot f(x)} e^{-n\cdot I(x)} \cdot dx \\
&= \lim_{n\to\infty} \frac{1}{n} \log \int_{\Theta} e^{n\cdot[f(x)-I(x)]} \cdot dx
\end{aligned} \tag{7.73}$$

The asymptotic behavior of the last integral is determined by the largest value of the integrand [10]:

$$\lim_{n\to\infty} \frac{1}{n} \log E_{P_n}\left[e^{n\cdot f(Y_n)}\right] = \frac{1}{n} \log\left[\sup_{x\in\Theta}\left\{e^{n\cdot[f(x)-I(x)]}\right\}\right] = \sup_{x\in\Theta}\{f(x) - I(x)\} \tag{7.74}$$

The generating function of Σ_n is defined as:

$$\psi_n(x) = E\left[e^{nx\Sigma_n}\right] = \int e^{nx\varsigma} P(\Sigma_n \in d\varsigma) \tag{7.75}$$

In terms of the density $p(\Sigma_n)$, we have instead

$$\psi_n(x) = \int e^{nx\varsigma} p(\Sigma_n = \varsigma)\, d\varsigma \tag{7.76}$$

In both expressions, the integral is over the domain of Σ_n. The function $\phi(x)$ defined by the limit: $\phi(x) = \underset{n\to\infty}{\text{Lim}} \frac{1}{n} \log \psi_n(x)$ is called the scaled cumulant generating function of Σ_n . It is also called the log-generating function or free energy function of Σ_n. In the following in thermodynamic, it will be called Massieu Potential. The existence of this limit is equivalent to writing $\psi_n(x) \approx e^{n\cdot\phi(x)}$.

Gärtner-Ellis Theorem: If $\phi(x)$ is differentiable, then Σ_n satisfies a large deviation principle with rate function $I(\varsigma)$ by the Legendre–Moreau [158] transform of $\phi(x)$:

$$I(\varsigma) = \underset{x}{\text{Sup}}\{x \cdot \varsigma - \varphi(x)\} \tag{7.77}$$

Varadhan's Theorem: If Σ_n satisfies a large deviation principle with rate function $I(\varsigma)$, then its scaled cumulant generating function $\phi(x)$ is the Legendre–Fenchel transform of $I(\varsigma)$:

$$\varphi(x) = \underset{\varsigma}{\text{Sup}}\{x \cdot \varsigma - I(\varsigma)\} \tag{7.78}$$

Properties of $\varphi(x)$: statistical moments of Σ_n are given by derivatives of $\phi(x)$:

$$\left.\frac{d\varphi(x)}{dx}\right|_{x=0} = \underset{n\to\infty}{\text{Lim}} E[\Sigma_n] \text{ and } \left.\frac{d^2\varphi(x)}{dx^2}\right|_{x=0} = \underset{n\to\infty}{\text{Lim}} \text{Var}[\Sigma_n] \tag{7.79}$$

7.4.11 Legendre Mapping by Contact Geometry in Mechanics

Legendre Transform is also classically used to derive the Hamiltonian formalism of mechanics from the Lagrangian formulation. Lagrangian is a convex function of the tangent space. The Legendre transform gives the Hamiltonian *H(p,q)* as a function of the coordinates *(p,q)* of the cotangent bundle, where the inner product used to define the Legendre transform is inherited from the pertinent canonical symplectic structure. More recently, Vladimir Arnold has demonstrated that this Legendre duality could be interpreted in the framework of Contact Geometry with notion of Legendre mapping [96] (Fig. 7.5).

As described by Vladimir Arnold, in the general case, we can define the Hamiltonian *H* as the fiberwise Legendre transformation of the Lagrangian L:

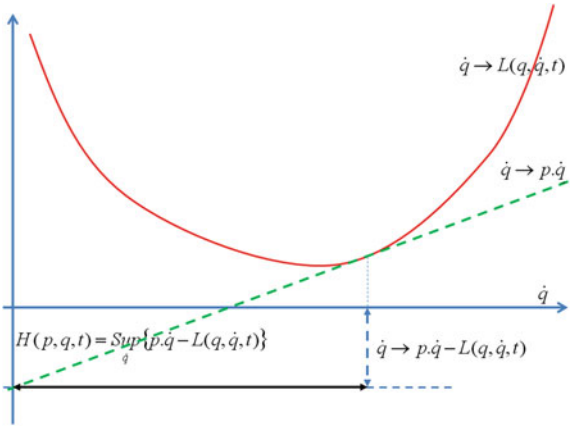


Fig. 7.5 Legendre transform between Hamiltonian and Lagrangian

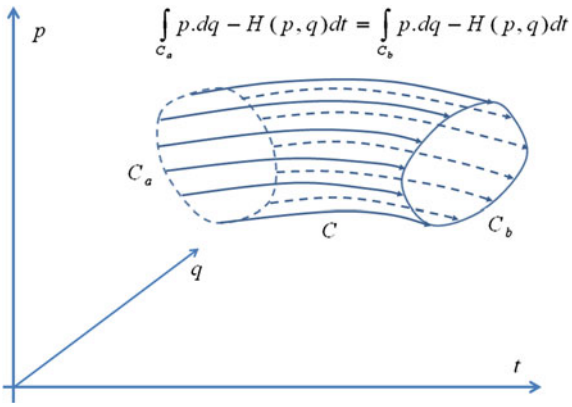


Fig. 7.6 Poincaré–Cartan invariant

$$H(p,q,t) = \underset{\dot{q}}{\text{Sup}}\,(p \cdot \dot{q} - L(q,\dot{q},t)) \tag{7.80}$$

Due to strict convexity, $H(p,q,t) = p \cdot \dot{q} - L(q,\dot{q},t)$ supremum is reached in a unique point $\dot{q}$ such that $p = \partial_{\dot{q}} L(q,\dot{q},t)$, and $\dot{q} = \partial_p H(p,q,t)$.

If we consider total differential of Hamiltonian:

$$\left.\begin{aligned} dH &= \dot{q}dp + pd\dot{q} - \partial_q L dq - \partial_{\dot{q}} L d\dot{q} - \partial_t L dt = \dot{q}dp - \partial_q L dq - \partial_t L dt \\ &= \partial_p H dp + \partial_q H dq + \partial_t H dt \end{aligned}\right\} \tag{7.81}$$
$$\Rightarrow \begin{cases} \dot{q} = \partial_p H \\ -\partial_q L = \partial_q H \end{cases}$$

Euler–Lagrange equation $\partial_t \partial_{\dot{q}} L - \partial_q L = 0$ with $p = \partial_{\dot{q}} L$ and $-\partial_q L = \partial_q H$ provides the second Hamilton equation $\dot{p} = -\partial_q H$ with $\dot{q} = \partial_p H$ in Darboux coordinates.

Considering Pfaffian form $\omega = p \cdot dq - H \cdot dt$ related to Poincaré–Cartan integral invariant [26], based on $\omega = \partial_{\dot{q}} L \cdot dq - \left(\partial_{\dot{q}} L \cdot \dot{q} - L\right) \cdot dt = L \cdot dt + \partial_{\dot{q}} L \varpi$ with

$\varpi = dq - \dot{q} \cdot dt$, Dedecker [15] has observed, that the property that among all forms $\theta \equiv L \cdot dt \bmod \varpi$ the form $\omega = p \cdot dq - H \cdot dt$ is the only one satisfying $d\theta \equiv 0 \bmod \varpi$, is a particular case of more general Lepage congruence [16] related to transversally condition (Fig. 7.6).

Contact geometry was used in Mechanic [11, 12] and in Thermodynamic [13, 14, 17, 18, 24, 25], where integral submanifolds of dimension n in $2n + 1$ dimensional contact manifold are called Legendre submanifolds. A smooth fibration of a contact manifold, all of whose are Legendre, is called a Legendre Fibration. In the neighbourhood of each point of the total space of a Legendre Fibration there exist contact Darboux coordinates *(z, q, p)* in which the fibration is given by the projection $(z, q, p) => (z, q)$. Indeed, the fibres *(z, q)* = *cst* are Legendre subspaces of the standard contact space. A Legendre mapping is a diagram consisting of an embedding of a smooth manifold as a Legendre submanifold in the total space of a Legendre fibration, and the projection of the total space of the Legendre fibration onto the base. Let us consider the two Legendre fibrations of the standard contact space R^{2n+1} of 1—jets of functions on R^n: $(u, p, q) \mapsto (u, q)$ and $(u, p, q) \mapsto (p \cdot q - u, p)$, the projection of the l-graph of a function $u = S(q)$ onto the base of the second fibration gives a Legendre mapping:

$$q \mapsto \left(q \frac{\partial S}{\partial q} - S(q), \frac{\partial S}{\partial q} \right) \tag{7.82}$$

If S is convex, the front of this mapping is the graph of a convex function, the Legendre transform of the function S:

$$\left(S^*(p), p \right) \tag{7.83}$$

with $S^*(p) = q \cdot p - S(q)$ where $p = \frac{\partial S}{\partial q}$.

More generally as considered by Vladimir Arnold, all Symplectic geometry of even-dimensional phase spaces has an odd-dimensional twin contact geometry. The relation between contact geometry and symplectic geometry is similar to the relation between linear algebra and projective geometry. Any fact in symplectic geometry can be formulated as a contact geometry fact and vice versa. The calculations are simpler in the symplectic setting, but their geometric content is better seen in the contact version. The functions and vector fields of symplectic geometry are replaced by hypersurfaces and line fields in contact geometry.

Each contact manifold has a symplectization, which is a symplectic manifold whose dimension exceeds that of the contact manifold by one. Inversely, Symplectic manifolds have contactizations whose dimensions exceed their own dimensions by one. If a manifold has a serious reason to be odd dimensional it usually carries a natural contact structure. V. Arnold said that "symplectic geometry is all geometry," but that he prefer to formulate it in a more geometrical form: "contact geometry is all geometry".

Relation between contact structures and Legendre submanifolds are defined by:

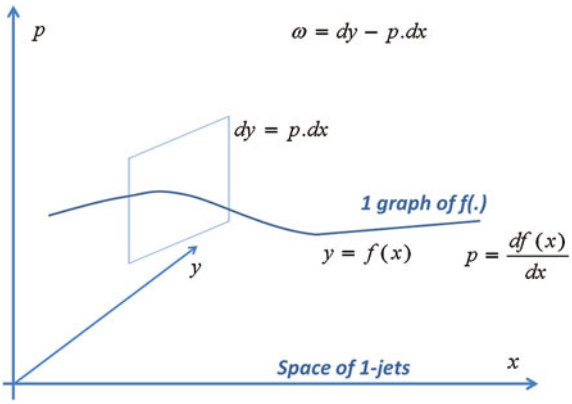


Fig. 7.7 Contact geometry by V. Arnold

- A contact structure on an odd-dimensional manifold M^{2n+1} is a field of hyperplanes (of linear subspaces of codimension 1) in the tangent spaces to M at all its points.
- All the generic fields of hyperplanes of a manifold of a fixed dimension are locally equivalent. They define the (local) contact structures.

As example, a 1-jet of a function:

$$y = f(x_1, \ldots, x_n) \tag{7.84}$$

at point x of manifold V^n is defined by the point

$$(x, y, p) \in R^{2n+1} \text{ where } p_i = \partial f / \partial x_i \tag{7.85}$$

The natural contact structure of this space is defined by the following condition: the 1-graphs $\{x, y = f(x), p = \partial f / \partial x\} \subset J^1(V^n, R)$ of all the functions on V should be the tangent structure hyper-plane at every point. In coordinates, this conditions means that the 1-form $dy - p.dx$ should vanish on the hyper-planes of the contact field (Fig. 7.7).

In thermodynamics, the Gibbs 1-form $dy - p \cdot dx$ will be given by:

$$dE = T \cdot dS - p \cdot dV \tag{7.86}$$

A contact structure on a manifold is a non-degenerate field of tangent hyper-planes:

- The manifold of contact elements in projective space coincides with the manifold of contact elements of the dual projective space
- A contact element in projective space is a pair, consisting of a point of the space and of a hyper-plane containing this point. The hyper-plane is a point of the dual projective space and the point of the original space defines a contact element of the dual space.

The manifold of contact elements of the projective space has two natural contact structures:

- The first is the natural contact structure of the manifold of contact elements of the original projective space.
- The second is the natural contact structure of the manifold of contact elements of the dual projective space.

The dual of the dual hyper-surface is the initial hyper-surface (at least if both are smooth for instance for the boundaries of convex bodies). The affine or coordinate version of the projective duality is called the Legendre transformation. Thus contact geometry is the geometrical base of the theory of Legendre transformation.

7.4.12 *Massieu Characteristic Function and Duhem Potentials in Thermodynamic*

In 1869, François Massieu, French Engineer from Corps des Mines, has presented two papers to French Science Academy on "*characteristic function*" in Thermodynamic [3–5]. Massieu demonstrated that some mechanical and thermal properties of physical and chemical systems could be derived from two potentials called "*characteristic functions*". The infinitesimal amount of heat dQ received by a body produces external work of dilatation, internal work, and an increase of body sensible heat. The last two effects could not be identified separately and are noted dE (function E accounted for the sum of mechanical and thermal effects by equivalence between heat and work). The external work $P \cdot dV$ is thermally equivalent to $A \cdot P \cdot dV$ (with A the conversion factor between mechanical and thermal measures). The first principle provides $dQ = dE + A \cdot P \cdot dV$. For a closed reversible cycle (Joule/Carnot principles) $\int \frac{dQ}{T} = 0$ that is the complete differential dS of a function S of $dS = \frac{dQ}{T}$.
If we select volume V and temperature T as independent variables:

$$T \cdot dS = dQ \Rightarrow T \cdot dS - dE = A \cdot P \cdot dV \Rightarrow d\,(TS) - dE = S \cdot dT + A \cdot P \cdot dV \quad (7.87)$$

If we set $H = TS - E$, then we have $dH = S \cdot dT + A \cdot P \cdot dV = \dfrac{\partial H}{\partial T} \cdot dT + \dfrac{\partial H}{\partial V} \cdot dV$ (7.88)

Massieu has called H the "characteristic function" because all body characteristics could be deduced of this function: $S = \frac{\partial H}{\partial T}$, $P = \frac{1}{A}\frac{\partial H}{\partial V}$ and $E = TS - H = T\frac{\partial H}{\partial T} - H$

If we select pressure P and temperature T as independent variables: Massieu characteristic function is then given by

$$H' = H - AP \cdot V \quad (7.89)$$

We have

$$dH' = dH - AP \cdot dV - AV \cdot dP = S \cdot dT - AV \cdot dP = \frac{\partial H'}{\partial T} \cdot dT + \frac{\partial H'}{\partial P} \cdot dP \quad (7.90)$$

And we can deduce:

$$S = \frac{\partial H'}{\partial T} \text{ and } V = -\frac{1}{A}\frac{\partial H'}{\partial P} \quad (7.91)$$

And inner energy:

$$E = TS - H = TS - H' - AP \cdot V \Rightarrow E = T\frac{\partial H'}{\partial T} - H' + P \cdot \frac{\partial H'}{\partial P} \quad (7.92)$$

The most important result of Massieu consists in deriving all body properties dealing with thermodynamics from characteristic function and its derivatives: "***je montre, dans ce mémoire, que toutes les propriétés d'un corps peuvent se déduire d'une fonction unique, que j'appelle la fonction caractéristique de ce corps***" [5].

Massieu results were extended by Gibbs that shown that Massieu functions play the role of potentials in the determination of the states of equilibrium in a given system, and Duhem [6–9] in a more general presentation had introduced Thermodynamic potentials in putting forward the analytic development of the mechanical Theory of heat.

In thermodynamics, the Massieu potential is the Legendre transform of the Entropy, and depends on the inverse temperature $\beta = 1/kT$:

$$\phi(\beta) = -k\beta \cdot F = S - E/T \text{ where } F \text{ is the Free Energy } F = E - TS \quad (7.93)$$

$$-\frac{\partial \phi(\beta)}{\partial(k\beta)} = \frac{\partial(k\beta F)}{\partial(k\beta)} = F + k\beta\frac{\partial F}{\partial(k\beta)} = F + k\beta\left(\frac{\partial F}{\partial T}\right)_V\left(\frac{\partial T}{\partial(k\beta)}\right) \quad (7.94)$$

$$-\frac{\partial \phi(\beta)}{\partial(k\beta)} = F + k\beta(-S)\left(-\frac{1}{(k\beta)^2}\right) = F + \frac{S}{k\beta} = (E - TS) + TS = E \quad (7.95)$$

with E the inner energy. The Legendre transform of the Massieu potential provides the Entropy S:

$$L(\phi) = k\beta \cdot \frac{\partial \phi(\beta)}{\partial(k\beta)} - \phi(\beta) = k\beta \cdot (-E) - k\beta \cdot F = k\beta(F - E) = -S \quad (7.96)$$

On these bases, Duhem has founded Mechanics on the principles of Thermodynamics. Duhem has defined a more general potential function $\Omega = G(E - TS) + W$ with G mechanic equivalent of heat. In case of constant volume, $W = 0$, the potential Ω becomes Helmoltz Free Energy and in the case of constant pressure, $W = P \cdot V$, the potential Ω becomes Gibbs–Duhem Free Enthalpy. Duhem has written: "***Nous avons***

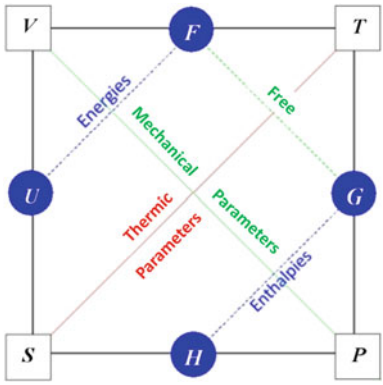


Fig. 7.8 Thermodynamic square of Max Born (*V* volume, *P* pressure, *T* temperature, *S* entropy, *U* energy, *F* free "Helmoltz" energy, *H* enthalpy, *G* free "Gibbs–Duhem" enthalpy)

fait de la Dynamique un cas particulier de la Thermodynamique, une Science qui embrasse dans des principes communs tous les changements d'état des corps, aussi bien les changements de lieu que les changements de qualités physiques" [6].

Classically, thermodynamic potentials and relations are synthetized in the thermodynamic square introduced by Max Born The corners represent common conjugate variables while the sides represent thermodynamic potentials. The placement and relation among the variables serves as a key to recall the relations they constitute (Fig. 7.8).

In his "*Thermodynamique générale*", four scientists were credited by Duhem with having carried out "the most important researches on that subject":

- F. Massieu that had derived Thermodynamics from a "characteristic function and its partial derivatives"
- J. W. Gibbs that shown that Massieu's functions "could play the role of potentials in the determination of the states of equilibrium" in a given system.
- H. von Helmholtz that had put forward "similar ideas"
- A. von Oettingen that had given "an exposition of Thermodynamics of remarkable generality" based on general duality concept in "Die thermodynamischen Beziehungen antithetisch entwickelt", St. Petersburg 1885.

Arthur Joachim Von Oettingen has developed duality in parallel in Thermodynamic and in Music [66–76]. In Music, he has introduced Harmonic Dualism: Categories of music-theoretical work that accept the absolute structural equality of major and minor triads as objects derived from a single, unitary process that structurally contains the potential for twofold, or binary, articulation. The tonic fundamental of the major triad is the structural parallel of the phonic overtone of the minor triad: in each case these tones are consonant with their respective chord. The Phonic overtone of the major triad and the tonic fundamental of the minor triad are dissonant with their respective chord. Oettingen introduced Topographic version of major–minor opposition: "All pure consonant triads stand in the form of right triangles, whose hypotenuses all form a diagonal minor third. In the major klang, the right angle is

oriented to the top (of the diagram); in the minor klang, the right angle is oriented to the bottom".

To conclude this chapter on "characteristic function" in thermodynamics, we introduce recent works of Pavlov and Sergeev [111] from Steklov Mathematical Institute, that have used ideas of Koszol [112], Berezin [113], Poincaré [114] and Caratheodory [115] to analyze the geometry of the space of thermodynamical states and demonstrate that the differential-geometric structure of this space selects entropy as the Lagrange function in mechanics in the sense that determining the entropy completely characterizes a thermodynamic system (this development will make one link with our Sect. 7.4.11 related to Legendre transform in Mechanics). They have demonstrated that thermodynamics is not separated from other branches of theoretical physics on the level of the mathematical apparatus.

Technically, both analytic mechanics and thermodynamics are based on the same dynamical principle, which we can use to obtain the thermodynamic equations as equations of motion of a system with non-holonomic constraints. The special feature of thermodynamics is that the number of these constraints coincides with the number of degrees of freedom.

Using notation, $\beta = 1/k \cdot T$, the laws of equilibrium thermodynamics is based on the Gibbs distribution:

$$p_{Gibbs}(H) = \frac{1}{Z} e^{-\beta H} \tag{7.97}$$

where $H(p, q, a)$ the Hamilton function depending on the external parameters $a = (a_1, \ldots, a_k)$ and the normalizing factor:

$$Z(\beta, a) = \sum_N \int e^{-\beta H} d\omega_N \tag{7.98}$$

with the internal Energy:

$$E = \hat{H} = -\frac{\partial \log Z}{\partial \beta} = \sum_N \int H \cdot \frac{e^{-\beta H}}{Z} d\omega_N = \sum_N \int H \cdot p_{Gibbs}(H) \cdot d\omega_N \tag{7.99}$$

Gibbs distribution to derive thermodynamic states is based on the theorem formulated by Berezin [113] (entropy functional subject to the internal energy and normalization has a unique maximum at the "point" for the Gibbs distribution).

P.V. Pavlov and V.M. Sergeev define the generalized forces by the formula:

$$A = -\sum_N \int \frac{\partial H}{\partial a} \cdot p_{Gibbs}(H) \cdot d\omega_N = \frac{1}{\beta} \frac{\partial \log Z}{\partial a} \tag{7.100}$$

The 1-form of the heat source is given by:

$$\omega_Q = dE + A \cdot da = -d\left(\frac{\partial \log Z}{\partial \beta}\right) + \frac{1}{\beta} \frac{\partial \log Z}{\partial a} \cdot da \tag{7.101}$$

Using the relation:

$$\beta^{-1} d\left(\beta \frac{\partial \log Z}{\partial \beta}\right) = \beta^{-1} \frac{\partial \log Z}{\partial \beta} d\beta + d\left(\frac{\partial \log Z}{\partial \beta}\right) \tag{7.102}$$

They obtain correspondence with the Shannon definition of the entropy

$$\begin{aligned} \beta \omega_Q &= d\left(\log Z - \beta \frac{\partial \log Z}{\partial \beta}\right) = d\left(-\sum_N \int p_{Gibbs}(H) \cdot \log p_{Gibbs}(H) \cdot d\mu_N\right) \\ &= dS \end{aligned} \tag{7.103}$$

providing:

- the second law of thermodynamics:

$$\omega_Q = T \cdot dS \tag{7.104}$$

- the Gibbs equation:

$$dE = T \cdot dS - A \cdot da \text{ or } dF = -S \cdot dT - A \cdot da \text{ with } F = E - TS \tag{7.105}$$

From the last Gibbs equation, the coefficients of the 1-forms are gradients

$$S = -\frac{\partial F}{\partial T} \text{ and } A = -\frac{\partial F}{\partial a} \tag{7.106}$$

and the free energy is the first integral of system of differential equation.

Pavlov and Seeger note that the matrix of 2-form $\Omega_Q = d\omega_Q$ is degenerate and has the rank 2 (nonzero elements of the 2-form matrix Ω_Q are concentrated in the first row and first column):

$$\Omega_Q = \frac{\partial A}{\partial T} dT \wedge da = -\frac{\partial^2 F}{\partial a \partial T} dT \wedge da \tag{7.107}$$

They observe that from the standpoint of the analogy with mechanical dynamical systems, the state space of a thermodynamic system (with local coordinates) is a configuration space. By physical considerations (temperature, volume, mean particle numbers, and other characteristics are bounded below by zero), it is a manifold with a boundary. They obtain the requirement natural for foliation theory that the fibers be transverse to the two-dimensional underlying manifold.

In Pavlov and Seeger formulation, the thermodynamic analysis of a dynamical system begins with a heuristic definition of the system entropy as a function of dynamical variables constructed based on general symmetry properties and the geometric structure of the space of thermodynamic states is fixed to be a foliation of codimension two.

7.4.13 Souriau's Geometric Temperature and Covariant Definition of Thermodynamic Equilibriums

Jean-Marie Souriau in [97–106], student of Elie Cartan at ENS Ulm, has given a covariant definition of thermodynamic equilibriums and has formulated statistical mechanics [161, 162] and thermodynamics in the framework of Symplectic Geometry by use of symplectic moments and distribution-tensor concepts, giving a geometric status for temperature and entropy. This work has been extended by Vallée and Saxcé [107, 110, 156], Iglésias [108, 109] and Dubois [157].

The first general definition of the "*moment map*" (constant of the motion for dynamical systems) was introduced by J. M. Souriau during 1970s, with geometric generalizion such earlier notions as the Hamiltonian and the invariant theorem of Emmy Noether describing the connection between symmetries and invariants (it is the moment map for a one-dimensional Lie group of symmetries). In symplectic geometry the analog of Noether's theorem is the statement that the moment map of a Hamiltonian action which preserves a given time evolution is itself conserved by this time evolution. The conservation of the moment of a Hamilotnian action was called by Souriau the "*Symplectic or Geometric Noether theorem*" (considering phases space as symplectic manifold, cotangent fiber of configuration space with canonical symplectic form, if Hamiltonian has Lie algebra, moment map is constant along system integral curves. Noether theorem is obtained by considering independently each component of momemt map).

In previous approach based on Koszul work, we have defined two convex functions $\Phi(x)$ and $\Phi^*(x^*)$ with dual system of coordinates x and x^* on dual cones Ω and Ω^*:

$$\begin{aligned}\Phi(x) &= -\log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \ \forall x \in \Omega \text{ and } \Phi^*(x^*) = \langle x, x^*\rangle - \Phi(x) \\ &= -\int_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \end{aligned} \tag{7.108}$$

where

$$\begin{aligned} x^* &= \int_{\Omega^*} \xi \cdot p_x(\xi) d\xi \text{ and } p_x(\xi) = e^{-\langle \xi, x\rangle} / \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \\ &= e^{-\langle x, \xi\rangle - \log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi} = e^{-\langle x, \xi\rangle + \Phi(x)} \end{aligned} \tag{7.109}$$

with

$$x^* = \frac{\partial \Phi(x)}{\partial x} \text{ and } x = \frac{\partial \Phi^*(x^*)}{\partial x^*} \tag{7.110}$$

Souriau introduced these relations in the framework of variational problems to extend them with a covariant definition. Let M a differentiable manifold with

a continuous positive density $d\omega$ and let E a finite vector space and $U(\xi)$ a continuous function defined on M with values in E, continuous positive function $p(\xi)$ solution of this variational problem:

$$\underset{p(\xi)}{\text{ArgMin}} \left[s = -\int_M p(\xi) \log p(\xi) d\omega \right] \text{ such that } \begin{cases} \int_M p(\xi) d\omega = 1 \\ \int_M U(\xi) p(\xi) d\omega = Q \end{cases} \tag{7.111}$$

is given by:

$$\begin{aligned} p(\xi) &= e^{\Phi(\beta) - \beta \cdot U(\xi)} \text{ with } \Phi(\beta) = -\log \int_M e^{-\beta \cdot U(\xi)} d\omega \text{ and } Q \\ &= \frac{\int_M U(\xi) e^{-\beta \cdot U(\xi)} d\omega}{\int_M e^{-\beta \cdot U(\xi)} d\omega} \end{aligned} \tag{7.112}$$

Entropy $s = -\int_M p(\xi) \log p(\xi) d\omega$ can be stationary only if there exist a scalar Φ and an element β belonging to the dual of E, where Φ and β are Lagrange parameters associated to the previous constraints. Entropy appears naturally as Legendre transform of Φ:

$$s(Q) = \beta \cdot Q - \Phi(\beta) \tag{7.113}$$

This value is a strict minimum of s, and the equation $Q = \frac{\int_M U(\xi) e^{-\beta \cdot U(\xi)} d\omega}{\int_M e^{-\beta \cdot U(\xi)} d\omega}$ has a maximum of one solution for each value of Q. The function $\Phi(\beta)$ is differentiable and we can write $d\Phi = d\beta \cdot Q$ and identifying E with its bidual:

$$Q = \frac{\partial \Phi}{\partial \beta} \tag{7.114}$$

Uniform convergence of $\int_M U(\xi) \otimes U(\xi) e^{-\beta \cdot U(\xi)} d\omega$ proves that $-\frac{\partial^2 \Phi}{\partial \beta^2} > 0$ and that $-\Phi(\beta)$ is convex. Then, $Q(\beta)$ and $\beta(Q)$ are mutually inverse and differentiable, and $ds = \beta \cdot dQ$. Identifying E with its bidual:

$$\beta = \frac{\partial s}{\partial Q} \tag{7.115}$$

Classically, if we take $U(\xi) = \begin{pmatrix} \xi \\ \xi \otimes \xi \end{pmatrix}$, components of Q will provide moments of first and second order of the density of probability $p(\xi)$, that is defined by Gaussian law.

Souriau has applied this approach for classical statistical mechanic system. Considering a mechanical system with n parameters $q_1, \ldots, q_n$, its movement

could be defined by its phase at arbitrary time t on a manifold of dimension $2n$: $q_1, \ldots, q_n, p_1, \ldots, p_n$.

Liouville theorem shows that coordinate changes have a Jacobian equal to unity, and a Liouville density could be defined on manifold M: $d\omega = dq_1 \ldots dq_n \; dp_1 \ldots dp_n$ that will not depend of choice for t.

A system state is one point on $2n$-Manifold M and a statistical state is a law of probability defined on M such that $\int_M p(\xi) d\omega = 1$, and its time evolution is driven by: $\frac{\partial p}{\partial t} = \sum \frac{\partial p}{\partial p_j} \frac{\partial H}{\partial q_j} - \frac{\partial p}{\partial q_j} \frac{\partial H}{\partial p_j}$ where H is the Hamiltonian.

A thermodynamic equilibrium is a statistical state that maximizes the entropy:

$$s = -\int_M p(\xi) \log p(\xi) d\omega \tag{7.116}$$

among all states giving the mean value of energy Q:

$$\int_M H(\xi) \cdot p(\xi) d\omega = Q \tag{7.117}$$

Apply for free particles, for ideal gas, equilibrium is given for $\beta = \frac{1}{kT}$ (with k Boltzmann constant) and if we set $S = k \cdot s$, previous relation $ds = \beta \cdot dQ$ provides: $dS = \frac{dQ}{T}$ and $S = \int \frac{dQ}{T}$ and $\Phi(\beta)$ is identified with Massieu–Duhem Potential.

We recover also the Maxwell Speed law:

$$p(\xi) = cste \cdot e^{-\frac{H}{kT}} \tag{7.118}$$

Main discovery of Jean-Marie Souriau is that previous thermodynamic equilibrium is not covariant on a relativity point of view. Then, he has proposed a covariant definition of thermodynamic equilibrium where the previous definition is a particular case. In previous formalization, Manifold M was solution of the variationnal problem:

$$d \int_{t_0}^{t_1} l\left(t, q_j, \frac{dq_j}{dt}\right) dt = 0 \text{ with } p_j = \frac{\partial l}{\partial q_j} \tag{7.119}$$

We can then consider the time variable t as other variables q_j through an arbitrary parameter τ, and defined the new variationnal problem by:

$$d \int_{t_0}^{t_1} L(q_J, \dot{q}_J) d\tau = 0 \text{ with } t = q_{n+1}, \dot{q}_J = \frac{dq_J}{d\tau} \text{ and } J = 1, 2, \ldots, n+1 \tag{7.120}$$

where

$$L\left(q_J, \dot{q}_J\right) = l\left(t, q_j, \frac{\dot{q}_j}{\dot{t}}\right)\dot{t} \tag{7.121}$$

Variables p_j are not changed and we have the relation:

$$p_{n+1} = l - \sum_j p_j \cdot \frac{dq_j}{dt} \tag{7.122}$$

If we compare with Sect. 7.4.11 on classical mechanic, we have:

$$p_{n+1} = -H \text{ with } H = \sum_j p_j \cdot \frac{dq_j}{dt} - l \,(H \text{ is Legendre transform of } l) \tag{7.123}$$

H is energy of system that is conservative if the Lagragian doesn't depend explicitly of time t. It is a particular case of Noether Theorem: If Lagrangian L is invariant by an infinitesimal transform $dQ_J = F_J(Q_K)$, $u = \sum_J p_J dQ_J$ is first integral of variations equations. As Energy is not the conjugate variable of time t, or the value provided by Noether theorem by system invariance to time translation, the thermodynamical equilibrium is not covariant. Then, J. M. Souriau proposes a new covariant definition of thermodynamic equilibrium:

Let a mechanical system with a Lagragian invariant by a Lie Group G. Equilibrium states by Group G are statistical states that maximizes the Entropy, while providing given mean values to all variables associated by Noether theorem to infinitesimal transforms of group G.

Noether theorem allows associating to all system movement ξ, a value $U(\xi)$ belonging to the vector dual A of Lie Algebra of group G. $U(\xi)$ is called *the moment* of the group.
For each derivation δ of this Lie algebra, we take:

$$U(\xi)(\delta) = \sum_J p_J \cdot \delta Q_J \tag{7.124}$$

With previous development, as E is dual of A, value β belongs to this Lie algebra A, geometric generalization of thermodynamic temperature. Value Q is a generalization of heat and belongs to the dual of A.

An Equilibrium state exists having the largest entropy, with a distribution function $p(\xi)$ that is the exponential of an affine function of U:

$$p(\xi) = e^{\Phi(\beta) - \beta \cdot U(\xi)} \text{ with } \Phi(\beta) = -\log \int_M e^{-\beta \cdot U(\xi)} d\omega \text{ and } Q$$

$$= \frac{\int\limits_M U(\xi) e^{-\beta \cdot U(\xi)} d\omega}{\int\limits_M e^{-\beta \cdot U(\xi)} d\omega} \tag{7.125}$$

with

$$s(Q) = \beta \cdot Q - \Phi(\beta), d\Phi = d\beta \cdot Q \text{ and } ds = \beta \cdot dQ \tag{7.126}$$

A statistical state $p(\xi)$ is invariant by δ if $\delta\left[p(\xi)\right] = 0$ for all ξ (*then* $p(\xi)$ is invariant by finite transform of G generated by δ).
J. M. Souriau gave the following theorem:

An equilibrium state allowed by a group G is invariant by an element δ of Lie Algebra A, if and only if $[\delta, \beta] = 0$ *(with [.], the Lie Bracket), with* β *the generalized equilibrium temperature.*
with $\beta = \frac{\partial s}{\partial Q}$ and $\int_M H(\xi) \cdot p(\xi) d\omega = Q$ where $s = -\int_M p(\xi) \log p(\xi) d\omega$ and $\int_M p(\xi) d\omega = 1$.

For classical thermodynamic, where G is abelian group of translation with respect to time t, all equilibrium states are invariant under G.
For Group of transformation of Space–Time, elements of Lie Algebra of G could be defined as vector fields in Space–Time. The generalized temperature β previously defined, would be also defined as a vector field. For each point of Manifold M , we could then define:

- Vector temperature:

$$\beta_M = \frac{V}{kT} \tag{7.127}$$

with

- Unitary Mean Speed:

$$V = \frac{\beta_M}{|\beta_M|} \text{ with } |V| = 1 \tag{7.128}$$

- Eigen Absolute Temperature:

$$T = \frac{1}{k \cdot |\beta_M|}. \tag{7.129}$$

Classical formula of thermodynamics are thus generalized, but variables are defined with a geometrical status, like the geometrical temperature β_M an element of the Lie algebra of the Galileo or Poincaré groups, interpreted as the field of space–time vectors. Souriau proved that in relativistic version β_M is a *time like vector* with an orientation that characterizes *the arrow of time*.

7.4.14 Quantum Characteristic Function: Von Neumann Entropy and Balian Quantum Fisher Metric

Information Geometry metric has been extended in Quantum Physics. In 1986, Roger Balian has introduced a natural metric structure for the space of states $\hat{D}$ in quantum mechanics, from which we can deduce the distance between a state and one of its approximations [88–90, 92–95]. Based on quantum information theory, Roger Balian has built on physical grounds this metric as hessian metric:

$$ds^2 = d^2 S(\hat{D}) \tag{7.130}$$

from $S(\hat{D})$, von Neumann's entropy:

$$S(\hat{D}) = -Tr\left[\hat{D}\log(\hat{D})\right] \tag{7.131}$$

Quantum mechanics uses concepts of

- "**observable**" $\hat{O}$: Random variables, elements of a non-commutative C*-algebra of operators describing a physical quantity of the system like angular momentum. For finite systems, "observables" $\hat{O}$ are represented by Hermitian matrices acting in a Hilbert space H.
- "**state**" $E(\hat{O})$: Quantum Expected Value considered as Information on the system allowing to make predictions about it. A quantum state is on the same footing as a probability distribution in classical physics and a state in classical statistical mechanics, except for the non-commutation of the observables. (this "expectation" cannot be given simultaneously for two non-commuting observables which cannot be measured with the same apparatus).

When observables are represented by matrices in a Hilbert space H, the state can be implemented in terms of a matrix $\hat{D}$ in the space H:

$$E(\hat{O}) = Tr\left[\hat{D}\hat{O}\right] \tag{7.132}$$

where $\hat{D}$ is the density matrix, which is a representation of the state, is

- ***Hermitian***: as $E(\hat{O})$ is real for $\hat{O}^+ = \hat{O}$
- ***Normalized***: as $E(\hat{I}) = 1$ for the unit observable $\hat{I}$
- ***Non-negative***: as $E(\hat{O}^2) \geq 0$ for any $\hat{O}$.

Through $E(\hat{O}) = Tr\left[\hat{D}\hat{O}\right]$, each state could be considered as a linear mapping $E(\hat{O}) = \left\langle \hat{D}, \hat{O} \right\rangle$ of the vector space of observables $\hat{O}$ onto real numbers, characterized by $\hat{D}$ that appears as elements of the dual vector space of observables $\hat{O}$. Liouville representations of quantum mechanics uses that the mathematical representation of quantum mechanics can be changed, without any incidence on physics,

by performing simultaneously in both vector spaces of observables and states a linear change of coordinates that preserves the scalar product $\left\langle \hat{D},\hat{O} \right\rangle$, that leaves invariant the expectation values

$$E(\hat{O}) = \left\langle \hat{D},\hat{O} \right\rangle. \tag{7.133}$$

Roger Balian has looked for a physically meaningful metric in the space of states, that depend solely on this space and not on the two dual spaces. Then, he considers the Entropy S associated with a state $\hat{D}$ that gathers all the information available on an ensemble E of systems. The Entropy S quantity measures the amount of missing information because of probabilistic character of predictions.

Von Neumann has introduced, in 1932, an extension of Boltzmann–Gibbs Entropy of classical statistical mechanics:

$$S(\hat{D}) = -Tr\left[\hat{D}\log(\hat{D})\right] \tag{7.134}$$

used by Jaynes to assign a state to a quantum system when only partial information about it is available based on:

- ***Maximum Entropy Principle***: among all possible states that are compatible with the known expectation values, the one which yields the maximum entropy under constraints on these data is selected.

that is derived from Laplace's principle of insufficient reason:

- Identification of expectation values in the Bayesian sense with averages over a large number of similar systems
- The expectation value of each observable for a given system is equal to its average over E showing that equi-probability for E implies maximum entropy for each system.

Then von Neumann's entropy has been interpreted by Jaynes and Brillouin as:

- ***Entropy = measure of missing information***: the state found by maximizing this entropy can be regarded as the least biased one, as it contains no more information than needed to account for the given data.

Different authors have shown that von Neumann Entropy is the only one that is physically satisfactory, since it is characterized by the properties required for a measure of disorder:

- ***Additivity*** for uncorrelated systems
- ***Subadditivity*** for correlated systems (suppressing correlations raises the uncertainty)
- ***Extensivity*** for a macroscopic system
- ***Concavity*** (putting together two different statistical ensembles for the same system produces a mixed ensemble which has a larger uncertainty than the average of the uncertainties of the original ensembles).

These properties are also requested to ensure that the maximum entropy criterion follows from Laplace's principle, since the latter principle provides an unambiguous construction of von Neumann's entropy.

Roger Balian has then developed the Geometry derived from von Neumann's entropy. using two physically meaningful scalar quantities:

- $Tr\left[\hat{D}\hat{O}\right]$ expectation values across the two dual spaces of observables & states
- $S = -Tr\left[\hat{D}\log(\hat{D})\right]$ entropy within the space of states.

The entropy S could be written as a scalar product $S = -\left\langle \hat{D}, \log(\hat{D}) \right\rangle$ where $\log(\hat{D})$ is an element of the space of observables, allowing physical geometric structure in these spaces. The second differential d^2S is a negative quadratic form of the coordinates of $\hat{D}$ that is induced by concavity of von Neumann entropy S. Then, Roger Balian has introduced the distance ds between the state $\hat{D}$ and a neighboring state $\hat{D} + d\hat{D}$ as the square root of

$$ds^2 = -d^2S = Tr\left[d\hat{D} \cdot d\log\hat{D}\right] \tag{7.135}$$

where the Riemannian metric tensor is the Hessian of $-S(\hat{D})$ as function of a set of independent coordinates of $\hat{D}$.

We recover the algebraic/geometric duality between $\hat{D}$ and $\ln \hat{D}$ through a Legendre transform:

$$S(\hat{D}) = F(\hat{X}) - \left\langle \hat{D}, \hat{X} \right\rangle \text{ with } S(\hat{D}) = -Tr\hat{D}\log\hat{D} = -\left\langle \hat{D}, \log\hat{D} \right\rangle \tag{7.136}$$

with $\hat{X}$ and $\hat{D}$ conjugate variables in the Legendre transform and where:

$$F(\hat{X}) = \log Tr \exp\hat{X} \tag{7.137}$$

an operators generalization of the Massieu potential. $F(\hat{X})$ Characterizes the canonical thermodynamic equilibrium states to which it reduces for $\hat{X} = \beta \cdot \hat{H}$ where $\hat{H}$ is the Hamiltonian.

$$dF = Tr\hat{D}d\hat{X} \text{ with } \hat{D} = \frac{\exp\hat{X}}{Tr\exp\hat{X}} \tag{7.138}$$

dF the partial derivatives of $F(\hat{X})$ with respect to the coordinates of $\hat{X}$. $\hat{D}$ is Hermitian, normalized and positive and can then be interpreted as a density matrix.
Legendre transformation appears naturally with:

$$S(\hat{D}) = -Tr\hat{D}\log\hat{D} = -Tr\left(\hat{D}\left(\hat{X} - \log Tr\exp\hat{X}\right)\right)$$
$$= -Tr\hat{D}\hat{X} + Tr\left(\hat{D}\right)\cdot\log Tr\exp\hat{X}$$
$$Tr\left(\hat{D}\right) = 1 \Rightarrow S(\hat{D}) = F\left(\hat{X}\right) - \left\langle\hat{D},\hat{X}\right\rangle \tag{7.139}$$

Roger Balian can then define Hessian metric from F, $ds^2 = d^2$ F in the conjugate space $\hat{X}$:

$$ds^2 = -dS^2 = Trd\hat{D}d\hat{X} = d^2F \tag{7.140}$$

Note that the normalization of $\hat{D}$ implies $Trd\hat{D} = 0$ and $Trd^2\hat{D} = 0$.

For time-dependent problems, predictions of $E(\hat{O}) = O = Tr\left[\hat{D}\hat{O}\right]$, given at some time, is transformed at later times

- ***Schrodinger Picture***: the observables $\hat{O}$ remain fixed, while the evolution of $\hat{D}(t)$ is governed by the ***Liouville-von Neumann equation of motion***:

$$i\hbar\frac{d\hat{D}}{dt} = \left[\hat{H},\hat{D}\right] \tag{7.141}$$

- ***Heisenberg picture***: the density operator $\hat{D}$ remains fixed while the observables $\hat{O}$ change in time according to ***Heisenberg's equation of motion***:

$$i\hbar\frac{d\hat{O}}{dt} = \left[\hat{O},\hat{H}\right] \tag{7.142}$$

The Liouville-von-Neumann equation describes the transfer of information from some observables to other ones generated by a completely known dynamics and can be modified to account for losses of information during a dynamical process. The von-Neumann entropy measures in dimensionless units our uncertainty when summarizes our statistical knowledge on the system, the quantum analogue of the Shannon entropy.

In the classical limit:

- observables $\hat{O}$ are replaced by commuting random variables, which are functions of the positions and momenta of the N particles.
- density operators $\hat{D}$ are replaced by probability densities D in the 6N-dimensional phase space.
- the trace by an integration over this space.
- the evolution of D is governed by the ***Liouville equation***.

At this step, we make a digression observing that the Liouville equation has been introduced first by Henri Poincaré in 1906, in a paper entitled "*Réflexions sur la théorie cinétique des gaz*" with this following first sentence:

> La théorie cinétique des gaz laisse encore subsister bien des points embarassants pour ceux qui sont accoutumés à la rigueur mathématique… L'un des points qui m'embarassaient le plus

était le suivant: il s'agit de démontrer que l'entropie va en diminuant, mais le raisonnement de Gibbs semble supposer qu'après avoir fait varier les conditions extérieures on attend que le régime soit établi avant de les faire varier à nouveau. Cette supposition est-elle essentielle, ou en d'autres termes, pourrait-on arriver à des résultats contraire au principe de Carnot en faisant varier les conditions extérieures trop vite pour que le régime permanent ait le temps de s'établir?

In this paper, Poincaré introduced Liouville equation for the time evolution of the phase-space probability density $p\,(x_i, t)$: $\frac{\partial p}{\partial t} + \sum_i \frac{\partial (pX_i)}{\partial x_i} = 0$ corresponding to a dynamical system defined by the ordinary differential equations $\frac{dx_i}{dt} = X_i$. For a system obeying Liouville theorem $\sum_i \frac{\partial X_i}{\partial x_i} = 0$, Poincaré gave the form $\frac{\partial p}{\partial t} + \sum_i X_i \frac{\partial p}{\partial x_i} = 0$ that can also be written as: $\frac{\partial p}{\partial t} = \{H, p\} = Lp$ in terms of the Poisson bracket $\{.,.\}$ of the Hamiltonian H with the probability density p, which defines the Liouvillian operator L.

Back to Balian theory, we can notice that usually, only a small set of data are controlled.

- $\hat{O}_i$ will be the observables that are controlled
- O_i their expectation values for the considered set of repeated experiments,
- $\hat{D}$ a density operator of the system, relying on the sole knowledge of the set

$$O_i = Tr\hat{O}_i\hat{D} \tag{7.143}$$

The ***maximum entropy criterion*** consists in:

- Selecting, among all the density operators subject to the constraints $O_i = Tr\hat{O}_i\hat{D}$, the one $\hat{D}_R$, which renders the von Neumann entropy $S(\hat{D}) = -Tr\hat{D}\log\hat{D}$ maximum.

Roger Balian said that:

An intuitive justication is the following: for any other $\hat{D}$ compatible with $O_i = Tr\hat{O}_i\hat{D}$, we have by construction $S(\hat{D}) < S(\hat{D}_R)$. The choice of $\hat{D}_R$ thus ensures that our description involves no more information than the minimum needed to account for the only available information $O_i = Tr\hat{O}_i\hat{D}$. The difference $S(\hat{D}_R) - S(\hat{D})$ measures some extra information included in $\hat{D}$, but not in $\hat{D}_R$, and hence, not in the only known expectation values O_i. Selecting $\hat{D}_R$ rather than any other $\hat{D}$ which satisfies $O_i = Tr\hat{O}_i\hat{D}$ is therefore the *least biased choice*, the one that allows the most reasonable predictions drawn from the known set O_i about other arbitrary quantities.

At this stage, Roger Balian has introduced the notion of ***Relevant Entropy*** as generalization of Gibbs entropy. For any given set of relevant observables $\hat{O}_i$, the expectation values O_i of which are known, the maximum of the von Neumann entropy $S(\hat{D}) = -Tr\hat{D}\log\hat{D}$ under the constraints $O_i = Tr\hat{O}_i\hat{D}$ is found by introducing Lagrangian multipliers, γ_i associated with each equation $O_i = Tr\hat{O}_i\hat{D}$ and Φ associated with the normalization of $\hat{D}$:

$$\hat{D}_R = \underset{\hat{D}}{\text{ArgMin}} \left[S(\hat{D}) + \gamma_i \left(O_i - Tr\hat{O}_i\hat{D} \right) \right] \text{ such that } Tr\hat{D} = 1 \tag{7.144}$$

Its value, $S(\hat{D}_R)$ is reached for a density operator of the form:

$$\hat{D}_R = \exp\left(\Phi - \sum_i \gamma_i \hat{O}_i \right) \tag{7.145}$$

where the multipliers are determined by $Tr\hat{D}_R\hat{O}_i = O_i$ and $Tr\hat{D}_R = 1$.

This least biased density operator has an exponential form, which generalizes the usual Gibbs distributions and the concavity of the von Neumann entropy ensures the unicity of $\hat{D}_R$.

The equations $Tr\hat{D}_R\hat{O}_i = O_i$ and $Tr\hat{D}_R = 1$ can be written by introducing a generalized thermodynamic potential $\Phi(\gamma_i)$, defined as function of the other multipliers γ_i by:

$$\Phi(\gamma_i) = -\log Tr \exp\left[-\sum_i \gamma_i \hat{O}_i \right] \tag{7.146}$$

where $\Phi(\gamma_i)$ is identified with a Massieu thermodynamic potential.

The relations between the data O_i and the multipliers γ_i are then given by:

$$\frac{\partial \Phi(\gamma_i)}{\partial \gamma_i} = O_i \tag{7.147}$$

The corresponding entropy $S(\hat{D}_R) = S_R(O_i)$ is a function of the variables O_i (or equivalently γ_i), as $\Phi(\gamma_i)$:

$$S_R(O_i) = \sum_i \gamma_i O_i - \Phi(\gamma_i) \tag{7.148}$$

Roger Balian gave the name of ***Relevant Entropy*** $S(\hat{D}_R) = S_R(O_i)$ associated with the set $\hat{O}_i$ of relevant observables selected in the considered situation, measuring the amount of information which is missing when only the data O_i are available.

The relations $\frac{\partial \Phi(\gamma_i)}{\partial \gamma_i} = O_i$ between the variables γ_i and the variables O_i can therefore be inverted as:

$$\frac{\partial S_R(O_i)}{\partial O_i} = \gamma_i \tag{7.149}$$

Roger Balian has given an orthogonal projection interpretation of maximum entropy solution, where in non-equilibrium statistical mechanics; a central problem consists in predicting the values at the time t of some set of variables O_i from the knowledge of their values at the initial time t_0. As remarked by Roger Balian, a general solution

of this inference problem is provided by the projection method of Nakajima [116] and Zwanzig [117].

The equations of motion of O_i (t) should be generated from:

$$i\hbar\frac{d\hat{D}}{dt} = \left[\hat{H}, \hat{D}\right] \text{ (the motion equation)} \tag{7.150}$$

$$O_i = Tr\hat{O}_i\hat{D} \tag{7.151}$$

At initiation, we have:

- $O_i(t_0)$: initial conditions are transformed into initial conditions on $\hat{D}(t_0)$
- $\hat{D}(t_0)$: the least biased choice given by the maximum entropy criterion:

$$\hat{D}_R = \exp\left(\Phi - \sum_i \gamma_i\hat{O}_i\right) \text{ with } Tr\hat{D}_R\hat{O}_i = O_i \text{ and } Tr\hat{D}_R = 1 \tag{7.152}$$

The von Neumann entropy $S(\hat{D}) = -Tr\hat{D}\log\hat{D}$ associated with $\hat{D}(t)$ remains constant in time, from the unity of the evolution $i\hbar\frac{d\hat{D}}{dt} = \left[\hat{H}, \hat{D}\right]$ (no information is lost during the evolution), but in general, $\hat{D}(t)$ does not keep the exponential form, which involves only the relevant observables O_i .

The lack of information associated with the knowledge of the variables $O_i(t)$ only is measured by the ***Relevant Entropy***:

$$S_R(O_i) = \sum_i \gamma_i O_i - \Phi(\gamma_i) \tag{7.153}$$

where the multipliers $\gamma_i(t)$ are time-dependent:

$$\frac{\partial\Phi(\gamma_i)}{\partial\gamma_i} = O_i \text{ or } \frac{\partial S_R(O_i)}{\partial O_i} = \gamma_i. \tag{7.154}$$

Generally, $S_R(O_i(t)) > S_R(O_i(t_0))$ because a part of the initial information on the set $\hat{O}_i$ has leaked at the time t towards irrelevant variables., due to dissipation in the evolution of the variables $O_i(t)$.

A reduced density operator $\hat{D}_R(t)$ can be built at each time with the set of relevant variables $O_i(t)$. As regards these variables, $\hat{D}_R(t)$ is equivalent to $\hat{D}(t)$:

$$Tr\hat{D}_R(t)\hat{O}_i = Tr\hat{D}(t)\hat{O}_i = O_i(t) \tag{7.155}$$

$\hat{D}_R(t)$ has the maximum entropy form retaining any information about the irrelevant variables, and is parameterized by the set of multipliers $\gamma_i(t)$ (in one-to-one correspondence with the set of relevant variables $O_i(t)$).

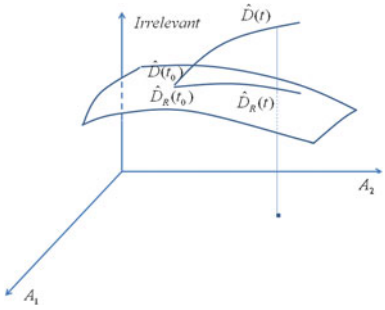


Fig. 7.9 Projection of $\hat{D}(t), \hat{D}_R(t)$ that has the maximum entropy form retaining any information about the irrelevant variables

Regarding density operators as points in a vector space, the correspondence from $\hat{D}(t)$ to $\hat{D}_R(t)$ is interpreted by Roger Balian as a projection onto the manifold R of reduced states (the space of states $\hat{D}$ endowed with the previous natural metric: $ds^2 = -dS^2 = Trd\hat{D} \cdot d\log\hat{D}$).

The correspondence $\hat{D} \rightarrow \hat{D}_R$ appears as an orthogonal projection [116, 117] where the projection from $\hat{D}$ to $\hat{D}_R = P\hat{D}$ is implemented by means of the projection superoperator:

$$P = \hat{D}_R \otimes I + \frac{\partial \hat{D}_R}{\partial O_i} \otimes \left(\hat{O}_i - O_i I\right) \tag{7.156}$$

where I is the unit superoperator in the Liouville space. P is defined as a sum of elementary projections (the tensor product of a state-like vector and an observable-like vector), lying in the conjugate Liouville spaces. The complementary projection superoperator Q onto the irrelevant space is defined as:

$$Q = I - P \tag{7.157}$$

with the following properties:

$$P^2 = P, Q^2 = Q, PQ = 0 \text{ and } IP = I \tag{7.158}$$

The density operator $\hat{D}$ can thereby be split at each time into its reduced relevant part $\hat{D}_R = P\hat{D}$ and its irrelevant part $\hat{D}_{IR} = Q\hat{D}$ (Fig. 7.9):

$$\hat{D} = P\hat{D} + Q\hat{D} = P\hat{D} + (I - Q)\hat{D} = \hat{D}_R + \hat{D}_{IR} \tag{7.159}$$

In Liouville representation, the Liouville-von Neumann equation could be written with Liouvillian L superoperator (tensor with 2×2 indices in Hilbert Space):

$$\frac{d\hat{D}}{dt} = L\hat{D} \Rightarrow \frac{d\hat{D}_{\alpha\beta}}{dt} = \sum_{\gamma\beta} L_{\alpha\beta,\gamma\delta}\hat{D}_{\delta\gamma} = \sum_{\gamma\beta} \frac{1}{i\hbar}\left(H_{\alpha\delta}\delta_{\beta\gamma} - H_{\gamma\beta}\delta_{\alpha\delta}\right)\hat{D}_{\delta\gamma} \tag{7.160}$$

The time-dependence of $\hat{D}_R$ is given by time-derivative:

$$\frac{d\hat{D}_R}{dt} = \frac{dP}{dt}\hat{D}_R + PL\hat{D}_R + PL\hat{D}_{IR} \tag{7.161}$$
$$\frac{d\hat{D}_{IR}}{dt} = \frac{dP}{dt}\hat{D}_{IR} - QL\hat{D}_R - QL\hat{D}_{IR}$$

Roger Balian solves these equations by introduction a Superoperator Green function.

7.5 Synthesis of Analogy Between Koszul Information Geometry Model, Souriau Statistical Physics Model and Balian Quantum Physics Model

We will synthetize in the following Table 7.1 results of Sect. 7.4 with Koszul Hessian structure of Information Geometry, Souriau model of Statistical Physics with general concept of geometric temperature, and Balian model of Quantum Physics with notion of Quantum Fisher Metric deduced from hessian of Von Neumann Entropy. Analogies between models will deal with Characteristic function, Entropy, Legendre Transform, Density of probability, Dual Coordinate System, Hessian Metric and Fisher metric.

7.6 Applications of Koszul Theory to Cones for Covariance Matrices of Stationary/Non-stationary Time Series and Stationary Space–Time Series

In this chapter, we apply theory developed in first chapters to define geometry of cones associated to Symmetric/Hermitian Positive Definite Matrices, that correspond to the case of covariance matrices for stationary signal [35, 36] with very specific matrix structures [77–79]:

- Toeplitz Hermitian Positive Definite Matrix structure (covariance matrix of a stationary time series)
- Toeplitz-Block-Toeplitz Hermitian Positive Definite Matrix structure (covariance matrix of a stationary space–time series).

We will prove that "Toeplitz" structure could be captured by complex autogressive model parameterization. This parameterization could be naturally introduced through Trench's theorem, Verblunsky's theorem or Partial Iwasawa decomposition theorem.

Table 7.1 Synthesis of Koszul, Souriau and Balian models respectively in hessian geometry, physical statistics and quantum physics

	Koszul model	Souriau model	Balian model
Characteristic function	$\Phi(x) = -\log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \forall x \in \Omega$	$\Phi(\beta) = -\log \int_M e^{-\beta \cdot U(\xi)} d\omega$	$\Phi(\gamma_i) = -\log Tr \exp\left[-\sum_i \gamma_i \hat{O}_i\right]$
Entropy	$\Phi^*(x^*) = -\int_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi$	$s = -\int_M p(\xi) \log p(\xi) d\omega$	$S(\hat{D}_R) = S_R(O_i)$ $S(\hat{D}_R) = -Tr\left(\hat{D}_R \log \hat{D}_R\right)$
Legendre transform	$\Phi^*(x^*) = \langle x, x^*\rangle - \Phi(x)$	$s(Q) = \beta \cdot Q - \Phi(\beta)$	$S_R(O_i) = \sum_i \gamma_i O_i - \Phi(\gamma_i)$
Density of probability	$p_x(\xi) = e^{-\langle x, \xi\rangle + \Phi(x)}$ $p_x(\xi) = e^{-\langle \xi, x\rangle} / \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi$	$p(\xi) = e^{\Phi(\beta) - \beta \cdot U(\xi)}$	$\hat{D}_R = \exp\left(\Phi - \sum_i \gamma_i \hat{O}_i\right)$
Dual coordinate	$x^* = \int_{\Omega^*} \xi \cdot p_x(\xi) d\xi$	$Q = \frac{\int_M U(\xi) e^{-\beta \cdot U(\xi)} d\omega}{\int_M e^{-\beta \cdot U(\xi)} d\omega}$	$Tr\hat{D}_R(t)\hat{O}_i = Tr\hat{D}(t)\hat{O}_i$ $Tr\hat{D}_R(t)\hat{O}_i = O_i(t)$
Dual coordinate system	$x^* = \frac{\partial \Phi(x)}{\partial x}$ and $x = \frac{\partial \Phi^*(x^*)}{\partial x^*}$	$Q = \frac{\partial \Phi}{\partial \beta}$ and $\beta = \frac{\partial s}{\partial Q}$	$\frac{\partial \Phi(\gamma_i)}{\partial \gamma_i} = O_i$ and $\frac{\partial S_R(O_i)}{\partial O_i} = \gamma_i$
Hessian metric	$ds^2 = -d^2\Phi^*(x^*)$	$ds^2 = -d^2 s(Q)$	$ds^2 = -dS^2 = Trd\hat{D}d\hat{X} = d^2F$
Fisher metric	$I(x) = -E_\xi\left[\frac{\partial^2 \log p_x(\xi)}{\partial x^2}\right]$ $I(x) = -\frac{\partial^2 \Phi(x)}{\partial x^2}$	$I(\beta) = -\frac{\partial^2 \Phi(\beta)}{\partial \beta^2}$	$I(\gamma) = \left[-\frac{\partial^2 \Phi}{\partial \gamma_i \partial \gamma_j}\right]_{i,j}$

7.6.1 Koszul Density for Symmetric Positive Definite (SPD) Matrix

(a) Classical approach based on Maximum Entropy
We can construct a probability model for symmetric positive-definite real random matrices using the entropy optimization principle [58]. Then, we will compare by the result obtained by Koszul theory. Let x be a random matrix with values in $Sym^{+}(n)$ whose probability distribution $p(x)\,\tilde{d}x$ is defined by a probability density function $p(x)$.This probability density function is such that:

$$\int_{x>0} p(x)\,\tilde{d}x = 1 \tag{7.162}$$

Available information for construction of the probability model

$$\int_{x>0} p(x)\,\tilde{d}x = 1 \text{ with } E(x) = \int_{x>0} x \cdot p(x)\,\tilde{d}x = \bar{x} \tag{7.163}$$

$$\text{and } E(\log\det x) = \int_{x>0} \log(\det x)\,.p(x)\,\tilde{d}x = v \text{ with } |v| < +\infty \tag{7.164}$$

$$\text{where } E\left[\log\det x\right] = v \text{ with } |v| < +\infty \Rightarrow E\left[\left|x^{-1}\right|_F^{\gamma}\right] < +\infty \tag{7.165}$$

Probability model using the maximum entropy principle: construct probability density function and characteristic function of random matrix x using the maximum entropy principle and the available information.

$$\begin{aligned} L(p) = &-\int_{x>0} p(x)\log p(x)\cdot\tilde{d}x - (\xi_0 - 1)\cdot\left(\int_{x>0} p(x)\tilde{d}x - 1\right) \\ &- \xi\cdot\left(\int_{x>0} \log(\det x)\cdot p(x)\tilde{d}x - v\right) \end{aligned} \tag{7.166}$$

$$p = \operatorname{Arg}\operatorname*{Max}_{p} L(p) \Rightarrow p(x) = c_0\cdot(\det x)^{\lambda-1}\,e^{-\langle\xi,x\rangle} \tag{7.167}$$

In order to perform this calculation, we need results concerning the Siegel integral for a positive-definite symmetric real matrix, given by Siegel in hi paper "Über der analytische Theorie der quadratischen Formen":

$$J(\lambda,\xi) = \int_{x>0} e^{-\langle\xi,x\rangle}\,[\det x]^{\lambda-1}\tilde{d}x \tag{7.168}$$

That could be computed:

$$\begin{aligned} J(\lambda,\xi) &= (2\pi)^{\frac{M(M-1)}{2}} \left[\prod_{l=1}^{M} \Gamma\left(\frac{M-l+2\lambda}{2}\right)\right] (\det x)^{-\frac{(M-1+2\lambda)}{2}} \\ \Gamma(\lambda) &= \int_0^{+\infty} t^{\lambda-1} e^{-t} dt \end{aligned} \tag{7.169}$$

If we compute the Characteristic function of x is given by:

$$\Phi_x(\theta) = E\left[e^{i\langle\theta,x\rangle}\right] = \int_{x>0} e^{i\langle\theta,x\rangle} p(x) \cdot \tilde{d}x \tag{7.170}$$

$$\Phi_x(\theta) = \left[\det\left(I_M - i\xi^{-1}\theta\right)\right]^{-\frac{(M-1+2\lambda)}{2}} \tag{7.171}$$

Considering the analytic extension of the mapping $w \mapsto J(\lambda, \xi - w)$ by writing $w = \mathrm{Re}(w) + i\theta$. Taking $\mathrm{Re}(w) = 0$, we deduce that:

$$\Phi_x(\theta) = c_0 \cdot J(\lambda, \xi - i\theta) \text{ and } \Phi_x(0) = 1 \Rightarrow c_0 = \frac{1}{J(\lambda,\xi)} \tag{7.172}$$

Finally, characteristic function of x is given by:

$$\Phi_x(\theta) = J(\lambda, \xi - i\theta) \cdot J(\lambda,\xi)^{-1} \tag{7.173}$$

First and second moments derived from characteristic function

$$\begin{aligned} E\left[x_{jk}\right] &= \frac{-i}{\left(2-\delta_{jk}\right)} \left.\frac{\partial \Phi_x(\theta)}{\partial \theta_{jk}}\right|_{\theta=0} \text{ and } E\left[x_{jk} x_{j'k'}\right] \\ &= \frac{-i}{\left(2-\delta_{jk}\right)\left(2-\delta_{j'k'}\right)} \left.\frac{\partial^2 \Phi_x(\theta)}{\partial \theta_{jk} \partial \theta_{j'k'}}\right|_{\theta=0} \end{aligned} \tag{7.174}$$

Using:

$$\det(I+h) = 1 + tr(h) + 0\left(|h|_F^2\right) \text{ and } \frac{\partial \det b}{\partial b_{jk}} = (\det b) \cdot b^{-1} \tag{7.175}$$

$$E\left[x_{jk}\right] = \frac{(M-1+2\lambda)}{2} \cdot \xi_{jk}^{-1} \tag{7.176}$$

$$E\left[x_{jk}x_{j'k'}\right] = E\left[x_{jk}\right]E\left[x_{j'k'}\right] + \frac{1}{(M-1+2\lambda)}\left(E\left[x_{j'k}\right]E\left[x_{jk'}\right] + E\left[x_{jj'}\right]E\left[x_{kk'}\right]\right) \tag{7.177}$$

First moment derived from characteristic function are given by: $\xi = \frac{(M-1+2\lambda)}{2}\bar{x}^{-1}$.

Characteristic function and probability density function of positive-definite random matrix:

$$\Phi_x(\theta) = \left[\det\left(I_M - \frac{2i}{(M-1+2\lambda)}\bar{x}\theta\right)\right]^{-\frac{(M-1+2\lambda)}{2}} \tag{7.178}$$

We deduce first two moments:

$$p(x) = c_x \cdot (\det x)^{\lambda-1} \cdot e^{-\frac{(M-1+2\lambda)}{2}\langle\bar{x}^{-1},x\rangle} = c_x \cdot (\det x)^{\lambda-1} \cdot e^{-\frac{(M-1+2\lambda)}{2}Tr(\bar{x}^{-1}x)}$$
$$\text{with } c_x = \frac{(2\pi)^{-\frac{M(M-1)}{4}}\left(\frac{M-1+2\lambda}{2}\right)^{\frac{M(M-1+2\lambda)}{2}}}{\left[\prod_{l=1}^{M}\Gamma\left(\frac{M-l+2\lambda}{2}\right)\right]\cdot(\det\bar{x})^{\frac{(M-1+2\lambda)}{2}}} \tag{7.179}$$

If we Maximum Entropy Density with $\lambda = 1$

$$p_x(\xi) = \beta_{ME}\cdot\det\left(\bar{\xi}^{-1}\right)^{\frac{n+1}{2}} e^{-\frac{(n+1)}{2}Tr(\bar{\xi}^{-1}\xi)} \text{ with}$$
$$\beta_{ME} = \left[\frac{(2\pi)^{-\frac{n(n-1)}{4}}\left(\frac{n+1}{2}\right)^{\frac{n(n+1)}{2}}}{\left[\prod_{l=1}^{n}\Gamma\left(\frac{n-l+2}{2}\right)\right]}\right] \tag{7.180}$$

(b) Method based on Koszul Density

Another approach is to consider directly Koszul theory with following definition of the density:

$$p_{\bar{\xi}}(\xi) = \frac{e^{-\langle\xi,\Theta^{-1}(\bar{\xi})\rangle}}{\int_{\Omega^*} e^{-\langle\xi,\Theta^{-1}(\bar{\xi})\rangle}d\xi} \text{with } x = \Theta^{-1}(\bar{\xi}) \text{ and } \bar{\xi} = \Theta(x) = \frac{d\Phi(x)}{dx} \tag{7.181}$$

$$\text{where } \bar{\xi} = \int_{\Omega^*}\xi\cdot p_{\bar{\xi}}(\xi)d\xi \text{ and } \Phi(x) = -\log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi \tag{7.182}$$

To compute this density, we have to define first the inner product $\langle .,. \rangle$ for our application considering Symmetric or Hermitian Positive Definite Matrices, that is a reductive homogeneous space. A homogeneous space *G/H* is said to be reductive if there exists a decomposition $g = m + h$, such that $Ad_H(m) =$

$\left\{hAh^{-1}: h \in H, A \in m\right\} \subset m$, where g and h are the Lie algebras of G and H. Given a bilinear form on m, there are associated G-invariant metric and G-invariant affine connection on G/H. The natural metric on G/H corresponds to a restriction of the Cartan–Kiling form on g (Lie Algebra of G). For covariance matrices, $g = gl(n)$ is the Lie algebra of nxn matrices, that could be written as the direct sum $gl(n) = m+h$, with m Symmetric or Hermitian matrices and $h = so(n)$ sub-Lie algebra of skew-symmetric or $h = u(n)$ skew-hermitian matrices: $A = 1/2 \cdot \left(A + A^T\right) + 1/2 \cdot \left(A - A^T\right)$. The symmetric matrices m are $Ad_{O(n)}$ -invariant because for any symmetric matrix S and orthogonal matrix Q, $Ad_{O(n)}(S) = QSQ^{-1} = QSQ^T$, that is symmetric.

As the set of covariance symmetric matrices is related to the quotient space $GL(n)/O(n)$ with $O(n)$ orthogonal Lie Group (because for any covariance matrix R, there is an equivalence class $R^{1/2}O(n)$), and the set of covariance hermitian matrices is related to the quotient space $GL(n)/U(n)$ with $U(n)$ Lie Group of Unitary matrices, covariance matrices admit a $GL(n)$ -invariant metric and connection corresponding to the following Cartan–Killing bilinear form (Elie Cartan has introduced this form in his Ph.D.):

$$\langle X, Y\rangle_I = Tr\,(XY) \text{ at } R = I \tag{7.183}$$

$$g_R(X, Y) = \langle X, Y\rangle_R = Tr\left(XR^{-1}YR^{-1}\right) \text{at arbitrary } R \tag{7.184}$$

With this inner product given by Cartan–Killing bilinear form, we can then developed the expression of the Koszul density:

$$\begin{cases} \langle x, y\rangle = Tr\,(xy)\,, \forall x, y \in Sym_n(R) \\ \psi_\Omega(x) = \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \quad = \quad \det x^{-\frac{n+1}{2}}\,\psi(I_n) \\ \qquad\qquad\qquad \langle x, y\rangle = Tr(xy) \\ \qquad\qquad\qquad \Omega^* = \Omega \text{ self-dual} \\ x^* = \bar{\xi} = -d\log\psi_\Omega = \frac{n+1}{2} d\log\det x = \frac{n+1}{2} x^{-1} \end{cases} \tag{7.185}$$

By Koszul formula, we have then directly

$$p_x(\xi) = e^{-Tr(x\xi) + \frac{n+1}{2}\log\det x} = \left[\det\left(\alpha\bar{\xi}^{-1}\right)\right]^\alpha e^{-Tr\left(\alpha\bar{\xi}^{-1}\xi\right)} \text{ with}$$
$$\bar{\xi} = \int\limits_{\Omega^*} \xi \cdot p_x(\xi) \cdot d\xi \tag{7.186}$$

But if we use normalization constraint:

$$\int\limits_{\xi} p_x(\xi)\tilde{d}\xi = 1 \Rightarrow \det(\nu)^{\frac{n+1}{2}} \int\limits_{\xi>0} e^{-Tr(\nu\xi)} = c_{\bar{\xi}}^{-1} \text{ with } \nu = \frac{n+1}{2}\bar{\xi}^{-1}$$
$$c_{\bar{\xi}}^{-1} = (2\pi)^{\frac{n(n-1)}{2}} \left[\prod_{l=1}^{n} \Gamma\left(\frac{n-l+2}{2}\right)\right] \tag{7.187}$$

That we could write:

$$p_x(\xi) = e^{-\langle x,\xi\rangle + \frac{n+1}{2}\log\det x} = c_{\bar{\xi}} \cdot \beta_K \det\left(\bar{\xi}^{-1}\right)^{\frac{n+1}{2}} e^{-\frac{n+1}{2}Tr\left(\bar{\xi}^{-1}\xi\right)} \text{ with}$$

$$\beta_{\mathrm{K}} = \left(\frac{n+1}{2}\right)^{\frac{n+1}{2}} \tag{7.188}$$

7.7 Geometry of Toeplitz and Toeplitz-Block-Toeplitz Hermitian Positive Definite Matrices and Representation in Product Space $RxD^{m-1}xSD^{n-1}$

The geometry of Toeplitz and Toeplitz-Block-Toeplitz Hermitian Positive Definite Matrices (THPD and TBTHPD matrices) is addressed in this chapter. In order to introduce with no arbitrary, the relevant metric, we will jointly introduce it through:

- Information Geometry (Fisher metric that is invariant by all changes of parameterization)
- Cartan–Siegel Homogenous Bounded Domains geometry [47, 48] (Siegel metric, deduced from Symplectic geometry , that is invariant by all automorphisms of these bounded domains).

Both geometries will provide the same metric, with representation of Toeplitz-Block-Toeplitz Hermitian Positive Definite Matrices in Product Space $R\mathrm{x}D^{m-1}\mathrm{x}SD^{n-1}$ (with D Poincaré unit disk and SD Siegel unit disk).

To deduce metric of Hermitian Postive Definite Matrices with an additional Toeplitz matrix structure, we use the Trench [51] theorem proving that there is a diffeomorphism between THPD matrices and a specific parameterization deduced from the Toeplitz matrix structure. By analogy, if we consider this THPD, as a covariance matrix of a stationary signal, this parameterization could be exactly identified with Complex Auto-Regressive (CAR) model of this signal. All THPD matrices are then diffeomorphic to $(r_0, \mu_1, \ldots, \mu_n) \in R^+\mathrm{x}D^n$ (r_0 is a real "scale" parameter, μ_k are called reflection/Verblunsky coefficients of CAR model in D the complex unit Poincare disk, and are "shape" parameters). This result has been found previously by Samuel Verblunsky in 1936 [52]. We have observed that this CAR parameterization of the THPD matrix could be also interpreted as Partial Iwasawa decomposition of the matrix in Lie Group Theory. At this step, to introduce the "natural" metric of this model, we use jointly Burbea/Rao results in Information Geometry and Koszul results in Hessian geometry, where the conformal metric is given by the Hessian of the Entropy of the CAR model. This metric has all good properties of invariances and is conformal. To regularize the CAR inverse problem, we have used a "regularized" Burg reflection coefficient avoiding prior selection of AR model order.

For TBTHPD matrices, we used matrix extension of Verblunsky theorem (given the diffeomorphism of TBTHPD matrix with $(R_0, M_1, \ldots, M_n) \in Herm(n)^+ \text{x} SD^n$) and Matrix-Burg like algorithm to compute a Matrix CAR model, where Verblunsky coefficients M_k are no longer in unit Poincare disk but in unit Siegel disk SD.

7.7.1 Metric of Toeplitz Hermitian Positive Definite Covariance Matrices

Classicaly, when we consider covariance matrix of stationary signal, correlation coefficient are given by $r_{n,n-k} = r_k = E\left[z_n z_{n-k}^*\right]$ and the covariance matrix R_n has a Toeplitz structure. To take into account this matrix structure constraint, Partial Iwasawa decomposition should be considered. This is equivalent for time or space signal to Complex AutoRegressive (CAR) Model decomposition (see Trench Theorem):

$$\Omega_n = (\alpha_n \cdot R_n)^{-1} = W_n \cdot W_n^+ = \left(1 - |\mu_n|^2\right) \cdot \begin{bmatrix} 1 & A_{n-1}^+ \\ A_{n-1} & \Omega_{n-1} + A_{n-1} \cdot A_{n-1}^+ \end{bmatrix} \tag{7.189}$$

$$W_{\text{n}} = \sqrt{1 - |\mu_n|^2} \begin{bmatrix} 1 & 0 \\ A_{n-1} & \Omega_{n-1}^{1/2} \end{bmatrix} \text{with } \Omega_{n-1} = \Omega_{n-1}^{1/2} \cdot \Omega_{n-1}^{1/2+} \tag{7.190}$$

where

$$\alpha_n^{-1} = \left[1 - |\mu_n|^2\right] \cdot \alpha_{n-1}^{-1}, A_n = \begin{bmatrix} A_{n-1} \\ 0 \end{bmatrix} + \mu_n \begin{bmatrix} A_{n-1}^{(-)} \\ 1 \end{bmatrix} \text{ and } V^{(-)} = J \cdot V^* \tag{7.191}$$

with J antidiagonal matrix

In the framework of Information Geometry, Information metric could be introduced as Kählerian metric where Kähler potential is given by the Entropy $\tilde{\Phi}(R_n, P_0)$:

$$\tilde{\Phi}(R_n, P_0) = \log\left(\det R_n^{-1}\right) - \log(\pi \cdot e) = -\sum_{k=1}^{n-1} (n-k) \cdot \log\left[1 - |\mu_k|^2\right] - n \cdot \log[\pi \cdot e \cdot P_0] \tag{7.192}$$

Information metric is then given by hessian of Entropy in Iwasawa parameterization:

$$g_{ij} \equiv \frac{\partial^2 \tilde{\Phi}}{\partial \theta_i^{(n)} \partial \theta_j^{(n)*}} \text{ where } \theta^{(n)} = \left[P_0 \ \mu_1 \cdots \mu_{n-1}\right]^T \tag{7.193}$$

with $\{\mu_k\}_{k=1}^{n-1}$ Regularized Burg reflection coefficient [132] and $P_0 = \alpha_0^{-1}$ mean signal Power. Kählerian metric (from Information Geometry) is finally:

$$ds_n^2 = d\theta^{(n)+}\left[g_{ij}\right]d\theta^{(n)} = n \cdot \left(\frac{dP_0}{P_0}\right)^2 + \sum_{i=1}^{n-1}(n-i)\frac{|d\mu_i|^2}{\left(1-|\mu_i|^2\right)^2} \tag{7.194}$$

7.7.2 Metric of Toeplitz-Block-Toeplitz Hermitian Positive Definite Covariance Matrices

Previous approach can be extended to Toeplitz-Block-Toeplitz Hermitian Positive Definite Covariance Matrices (TBTHPD). Based on "matrix" generalization of Trench Algorithm, if we consider TBTHPD matrix [82, 83, 86, 87]:

$$R_{p,n+1} = \begin{bmatrix} R_0 & R_1 & \cdots & R_n \\ R_1^+ & R_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & R_1 \\ R_n^+ & \cdots & R_1^+ & R_0 \end{bmatrix} = \begin{bmatrix} R_{p,n} & \tilde{R}_n \\ \tilde{R}_n^+ & R_0 \end{bmatrix} \tag{7.195}$$

$$\text{with } \tilde{R}_n = V\begin{bmatrix} R_1 \\ \vdots \\ R_n \end{bmatrix}^* \text{ and } V = \begin{bmatrix} 0 & \cdots & 0 & J_p \\ \vdots & \iddots & \iddots & 0 \\ 0 & J_p & \iddots & \vdots \\ J_p & 0 & \cdots & 0 \end{bmatrix} \tag{7.196}$$

From Burg-like parameterization, we can deduced this inversion of Toeplitz-Block-Toeplitz matrix:

$$R_{p,n+1}^{-1} = \begin{bmatrix} \alpha_n & \alpha_n \cdot \widehat{A}_n^+ \\ \alpha_n \cdot \widehat{A}_n & R_{p,n}^{-1} + \alpha_n \cdot \widehat{A}_n \cdot \widehat{A}_n^+ \end{bmatrix} \tag{7.197}$$

and

$$R_{p,n+1} = \begin{bmatrix} \alpha_n^{-1} + \widehat{A}_n^+ \cdot R_{p,n} \cdot \widehat{A}_n & -\widehat{A}_n^+ \cdot R_{p,n} \\ -R_{p,n} \cdot \widehat{A}_n & R_{p,n} \end{bmatrix} \tag{7.198}$$

with

$$\alpha_n^{-1} = \left[1 - A_n^n A_n^{n+}\right] \cdot \alpha_{n-1}^{-1}, \alpha_0^{-1} = R_0$$

and

$$\widehat{A}_n = \begin{bmatrix} A_1^1 \\ \vdots \\ A_n^n \end{bmatrix} = \begin{bmatrix} \widehat{A}_{n-1} \\ 0_p \end{bmatrix} + A_n^n \cdot \begin{bmatrix} J_p A_{n-1}^{n-1*} J_p \\ \vdots \\ J_p A_1^{n-1*} J_p \\ I_p \end{bmatrix} \tag{7.199}$$

Where we have the following Burg-like generalized forward and backward linear prediction:

$$\begin{cases} \varepsilon_{n+1}^f(k) = \sum_{l=0}^{n+1} A_l^{n+1}(k) Z(k-l) = \varepsilon_n^f(k) + A_{n+1}^{n+1} \varepsilon_n^b(k-1) \\ \varepsilon_{n+1}^b(k) = \sum_{l=0}^{n} JA_l^{n+1}(k)^* JZ(k-n+l) = \varepsilon_n^b(k-1) + JA_{n+1}^{n+1*} J \varepsilon_n^f(k) \end{cases}$$

with

$$\begin{cases} \varepsilon_0^f(k) = \varepsilon_0^b(k) = Z(k) \\ A_0^{n+1} = I_p \end{cases}$$

$$A_{n+1}^{n+1} = -2 \left[\sum_{k=1}^{N+n} \varepsilon_n^f(k) \varepsilon_n^b(k-1)^+ \right] \left[\sum_{k=1}^{N+n} \varepsilon_n^f(k) \varepsilon_n^f(k)^+ + \sum_{k=1}^{N+n} \varepsilon_n^b(k) \varepsilon_n^b(k)^+ \right]^{-1} \tag{7.200}$$

Using Schwarz's inequality, it is easily to prove that A_{n+1}^{n+1} Burg-Like reflection coefficient matrix lies in Siegel Disk $A_{n+1}^{n+1} \in SD_p$.

Then as we have observed, metric and geometry of Toeplitz-Block-Toeplitz matrices are related to geometry of Siegel Disk considered as an Hermitian Homogeneous bounded domain [34].

Siegel Disk has been introduced by Carl Ludwig Siegel [56, 84] through Symplectic Group Sp_{2n} R that is one possible generalization of the group $SL_2R = Sp_2R$ (group of invertible matrices with determinant 1) to higher dimensions. This generalization goes further; since they act on a symmetric homogeneous space, the Siegel upper half plane, and this action has quite a few similarities with the action of SL_2 R on the Poincaré's hyperbolic plane. Let F be either the real or the complex field, the Symplectic Group is the group of all matrices $M \in GL_{2n}$ F satisfying:

$$Sp(n,F) \equiv \left\{ M \in GL(2n,F) / M^T J M = J \right\} , J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} \in SL(2n,R) \tag{7.201}$$

or

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in Sp(n,F) \Leftrightarrow A^T C \text{ and } B^T D \text{ symmetric} \tag{7.202}$$

and

$$A^T D - C^T B = I_n$$

The Siegel upper half plane is the set of all complex symmetric $n \times n$ matrices with positive definite imaginary part:

$$SH_n = \{Z = X + iY \in Sym(n, C)/\mathrm{Im}(Z) = Y > 0\} \tag{7.203}$$

The action of the Symplectic Group on the Siegel upper half plane is transitive. The group $PSp(n, R) \equiv Sp(n, R)/\{\pm I_{2n}\}$ is group of SH_n biholomorphisms via generalized Möbius transformations:

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \Rightarrow M(Z) = (AZ + B)\,(CZ + D)^{-1} \tag{7.204}$$

$PSp(n, R)$ acts as a sub-group of isometries. Siegel has proved that Symplectic transformations are isometries for the Siegel metric in SH_n. It can be defined on SH_n using the distance element at the point $Z = X + iY$, as defined by:

$$ds^2_{Siegel} = Tr\left(Y^{-1}\,(dZ)\,Y^{-1}\left(dZ^+\right)\right) \text{ with } Z = X + iY \tag{7.205}$$

with associated volume form:

$$\Omega = Tr\left(Y^{-1}dZ \wedge Y^{-1}dZ^+\right) \tag{7.206}$$

C. L. Siegel has proved that distance in Siegel Upper-Half Plane is given by:

$$d^2_{Siegel}\,(Z_1, Z_2) = \left(\sum_{k=1}^{n} \log^2\left(\frac{1 + \sqrt{r_k}}{1 - \sqrt{r_k}}\right)\right) \text{ with } Z_1, Z_2 \in SH_n \tag{7.207}$$

and r_k eigenvalues of the cross-ratio:

$$R\,(Z_1, Z_2) = (Z_1 - Z_2)\left(Z_1 - Z_2^+\right)^{-1}\left(Z_1^+ - Z_2^+\right)\left(Z_1^+ - Z_2\right)^{-1} \tag{7.208}$$

This is deduced from the second derivative of $Z \to R\,(Z_1, Z)$ in $Z_1 = Z$ given by:

$$D^2R = 2dZ\left(Z - Z^+\right)^{-1} dZ^+(Z^+ - Z)^{-1} = (1/2) \cdot dZY^{-1}dZ^+Y^{-1} \tag{7.209}$$

and

$$ds^2 = Tr\left(Y^{-1}dZY^{-1}dZ^+\right) = 2 \cdot Tr\left(D^2R\right) \tag{7.210}$$

In parallel, in China in 1945, Hua Lookeng has given the equations of geodesic in Siegel upper-half plane:

$$\frac{d^2Z}{ds^2} + i\frac{dZ}{ds}Y^{-1}\frac{dZ}{ds} = 0 \tag{7.211}$$

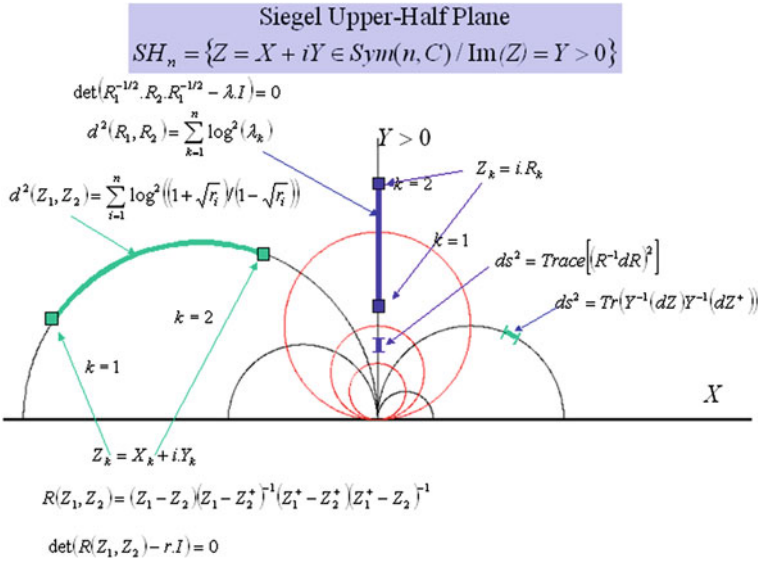


Fig. 7.10 Geometry of Siegel upper half-plane

Using generalized Cayley transform $W = (Z - iI_n)(Z + iI_n)^{-1}$, Siegel Upper-half Plane SH_n is transformed in unit Siegel disk $SD_n = \{W/WW^+ < I_n\}$ where the metric in Siegel Disk is given by:

$$ds^2 = Tr\left[\left(I_n - WW^+\right)^{-1} dW \left(I_n - W^+W\right)^{-1} dW^+\right] \tag{7.212}$$

The differential equation of the geodesics in Siegel Unit disk is then given by:

$$\frac{d^2W}{ds^2} + 2WW^+\left(I - WW^+\right)^{-1} W = 0 \tag{7.213}$$

Contour of Siegel Disk is called its Shilov boundary $\partial SD_n = \{W/WW^+ - I_n = 0_n\}$. We can also defined horosphere. Let $U \in \partial SD_n$ and $k \in R_*^+$, the following set is called horosphere in siegel disk (Fig. 7.10):

$$\begin{aligned} H(k,U) &= \left\{Z/0 < k\left(I - Z^+Z\right) - \left(I - Z^+U\right)\left(I - U^+Z\right)\right\} \\ &= \left\{Z/\left\|Z - \frac{1}{k+1}U\right\|\right\} < \frac{k}{k+1} \end{aligned} \tag{7.214}$$

Hua Lookeng and Siegel have proved that the previous positive definite quadratic differential is invariant under the group of automorphisms of the Siegel Disk. Considering $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ such that $M^* \begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix} M = \begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix}$:

$$\begin{aligned} V = M(W) &= (AZ+B)(CZ+D)^{-1} \\ &\Rightarrow \left(I_n - VV^+\right)^{-1} dV \left(I_n - V^+V\right)^{-1} dV^+ \\ &= \left(BW^+ + A\right)\left(I_n - WW^+\right)^{-1} dW \left(I_n - W^+W\right)^{-1} dW^+ \left(BW^+ + A\right)^{-1} \\ &\Rightarrow ds_V^2 = ds_W^2 \end{aligned} \tag{7.215}$$

To go further to study Siegel Disk geometry, we need now to define what are the automorphisms of Siegel Disk SD_n. They are all defined by:

$$\forall \Psi \in Aut(SD_n), \exists U \in U(n,C) \,/\, \Psi(Z) = U\Phi_{Z_0}(Z)U^t \tag{7.216}$$

$$\text{with } \Sigma = \Phi_{Z_0}(Z) = \left(I - Z_0 Z_0^+\right)^{-1/2} (Z - Z_0)\left(I - Z_0^+ Z\right)^{-1}\left(I - Z_0^+ Z_0\right)^{1/2} \tag{7.217}$$

and its inverse:

$$\begin{aligned} G &= \left(I - Z_0 Z_0^+\right)^{1/2} \Sigma \left(I - Z_0^+ Z_0\right)^{-1/2} = (Z - Z_0)\left(I - Z_0^+ Z\right)^{-1} \\ &\Rightarrow \begin{cases} Z = \Phi_{Z_0}^{-1}(\Sigma) = (GZ_0^+ + I)^{-1}(G + Z_0) \\ \text{with } G = \left(I - Z_0 Z_0^+\right)^{1/2} \Sigma \left(I - Z_0^+ Z_0\right)^{-1/2} \end{cases} \end{aligned} \tag{7.218}$$

By analogy with Poincaré's unit Disk, C. L. Siegel has deduced geodesic distance in SD_n:

$$\forall Z, W \in SD_n, d(Z,W) = \frac{1}{2}\log\left(\frac{1 + |\Phi_Z(W)|}{1 - |\Phi_Z(W)|}\right) \tag{7.219}$$

Information metric will be then introduced as a Kähler potential defined by Hessian of multi-channel entropy $\tilde{\Phi}\left(R_{p,n+1}\right)$:

$$\tilde{\Phi}\left(R_{p,n}\right) = \log\left(\det R_{p,n}^{-1}\right) + cte = -Tr\left(\log R_{p,n}\right) + cste \Rightarrow g_{i\bar{j}} = Hess\left[\phi\left(R_{p,n}\right)\right] \tag{7.220}$$

Using partitioned matrix structure of Toeplitz-Block-Toeplitz matrix $R_{p,n+1}$, recursively parameterized by Burg-Like reflection coefficients matrix $\left\{A_k^k\right\}_{k=1}^{n-1}$ with $A_k^k \in SD_n$, we can give Multi-variate entropy, matrix extension of previous Entropy:

$$\tilde{\Phi}(R_{p,n}) = -\sum_{k=1}^{n-1}(n-k)\cdot \log\det\left[I - A_k^k A_k^{k+}\right] - n\cdot\log\left[\pi\cdot e\cdot \det R_0\right] \tag{7.221}$$

Paul Malliavin has proved that this form is a Kähler Potential of an invariant Kähler metric (Information Geometry metric in our case). The metric is given by matrix extension as Hessian of this entropic potential $ds^2 = d^2\tilde{\Phi}(R_{p,n})$:

$$ds^2 = nTr\left[\left(R_0^{-1}dR_0\right)^2\right] + \sum_{k=1}^{n-1}(n-k)Tr\left[\left(I - A_k^k A_k^{k+}\right)^{-1} dA_k^k \left(I - A_k^{k+}A_k^k\right)^{-1} dA_k^{k+}\right] \quad (7.222)$$

Then finally, we have proved that the geometry of Toeplit-Block-Toeplitz covariance matrix of a space–time time serie is given by the following conformal metric:

$$\begin{aligned} ds^2 &= n\cdot m\cdot (d\log(P_0))^2 + n\cdot \sum_{i=1}^{m-1}(m-i)\frac{|d\mu_i|^2}{\left(1-|\mu_i|^2\right)^2} \\ &+ \sum_{k=1}^{n-1}(n-k)Tr\left[\left(I - A_k^k A_k^{k+}\right)^{-1} dA_k^k \left(I - A_k^{k+}A_k^k\right)^{-1} dA_k^{k+}\right] \end{aligned} \quad (7.223)$$

Where parameters are in Poincaré disk and Siegel disk:

$$\begin{aligned} &\left(R_0, A_1^1, \ldots, A_{n-1}^{n-1}\right) \in THPD_m \times SD^{n-1} \text{ with } SD = \left\{Z/ZZ^+ < I_m\right\} \\ &R_0 \rightarrow (\log(P_0), \mu_1, \ldots, \mu_{m-1}) \in R \times D^{m-1} \text{ with } D = \{z/zz^* < 1\} \end{aligned} \quad (7.224)$$

We have finally proved that for a space—time state of a stationary time series the TBTHPD covariance matrix is coded in $R \times D^{m-1} \times SD^{n-1}$

This result could be intuitively recovered and illustrated in the case for **m = 2 and n = 1** (Fig. 7.11).

$$R = \begin{bmatrix} h & a-ib \\ a+ib & h \end{bmatrix} > 0, \det R > 0 \Leftrightarrow h^2 > a^2 + b^2 \quad (7.225)$$

where we can rewrite the matrix:

$$R = h\cdot \begin{bmatrix} 1 & \mu^* \\ \mu & 1 \end{bmatrix} > 0, h \in R_+^*, \mu = \frac{a+ib}{h} \in D = \{z/|z| < 1\} \quad (7.226)$$

R is then parameterized by $h \in R_+^*$ and by $\mu \in D$.

We can observe that this parameterization $\{\log(h), \mu\} \in R \times D$ is linked with Hadamard compactification of $R \times D$ described in [131] (Fig. 7.12).

7.7.3 *Extension to Non-stationary Signal: Geometry of THPD Matrices Time Series Considered as Paths on Structured Matrix Manifolds*

In his book on *Creative Evolution*, Bergson has written about "*Form and Becoming*": "*life is an evolution. We concentrate a period of this evolution in a stable view which*

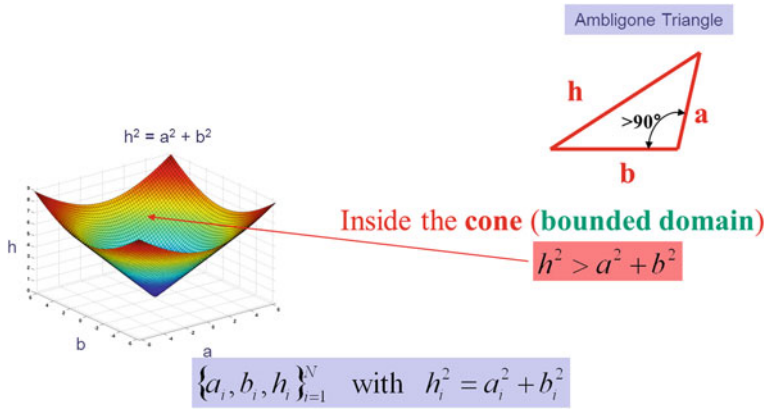


Fig. 7.11 Bounded homogeneous cone associated to a 2×2 symmetric positive definite matrix

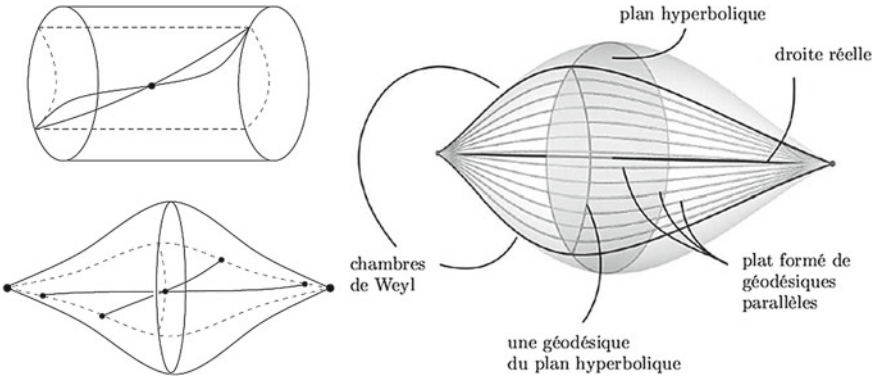


Fig. 7.12 Hadamard compactification (illustration from "Géométrie des bords: compactifications différentiables et remplissages holomorphes", Benoît Kloeckner) [131]

we call a form, and, when the change has become considerable enough to overcome the fortunate inertia of our perception, we say that the body has changed its form. But in reality the body is changing form at every moment; or rather, there is no form, since form is immobile and the reality is movement. What is real is the continual change of form: form is only a snapshot view of a transition".

In Sect. 7.7.1, we have considered stationary signal, but we could envisage to extend the application for non-stationary signal, defined as a continual change of covariance matrices. We will then extend approach of Sect. 7.7.1 to non-stationary time series. Many methods have been explored to address model of non-stationary time series [134–141]. We propose to extend the previous geometric approach for non-stationary signal corresponding to fast time variation of a time series. We will assume that each non-stationary signal in one time series can be split into several stationary signals on a shorter time scale, represented by time sequence of stationary

covariance matrices or a geodesic polygon on covariance matrix manifold. We adapt the Fréchet distance [125, 126, 166, 167] between two curves in the plane with a natural extension to more general geodesic curves in abstract metric spaces used for covariance matrix manifold.

7.7.3.1 General Definition of Fréchet Surface Metric

A Fréchet surface is an equivalence class of parameterized surfaces in a metric space, where surfaces are considered independently of how they are parameterized [125].

Maurice Fréchet has introduced a distance for Surfaces.

Let *(X, d)* be a metric space, M^2 a compact two-dimensional manifold, f a continuous mapping $f: M^2 \rightarrow X$, called a parameterized surface, and $\sigma: M^2 \rightarrow M^2$ a homeomorphism of M^2 onto itself. Two such surfaces f_1 and f_2 are equivalent if:

$$\underset{\sigma}{Inf} \underset{p \in M^2}{\text{Max}}\, d\left[f_1(p), f_2(\sigma(p))\right] = 0 \tag{7.227}$$

where the infimum is taken over all possible σ. A class f^* of parameterized surfaces, equivalent to f, is called a Fréchet surface. It is a generalization of the notion of a surface in Euclidean space to the case of an arbitrary metric space (X, d).

The Fréchet surface metric between Fréchet surfaces and f_2^* is:

$$\underset{\sigma}{Inf} \underset{p \in M^2}{\text{Max}}\, d\left[f_1(p), f_2(\sigma(p))\right] \tag{7.228}$$

where the infimum is taken over all possible homeomorphisms σ.

Let (X, d) be a metric space. Consider a set F of all continuous mappings $f: A \rightarrow X, g: B \rightarrow X, \ldots$, where A,B, ... are in R^n ,are homeomorphic to $[0, 1]^n$ for a fixed dimension n.

The Fréchet semimetric on F is:

$$\underset{\sigma}{Inf} \underset{x \in A}{\text{Sup}}\, d\left[f(x), g(\sigma(x))\right] \tag{7.229}$$

where the infimum is taken over all orientation preserving homeomorphisms σ: $A \rightarrow B$. It becomes the Fréchet metric on the set of equivalence classes:

$$f^* = \{g / d_F(g, f) = 0\} \tag{7.230}$$

For $n = 1$ and (X, d) is Euclidean space R^n, this metric is the original Fréchet distance introduced in 1906 between parametric curves $f, g: [0, 1] \rightarrow R^n$. For more details on Fréchet Surface Metric, read the "Dictionary of Distances" of Elena and Michel Deza [187].

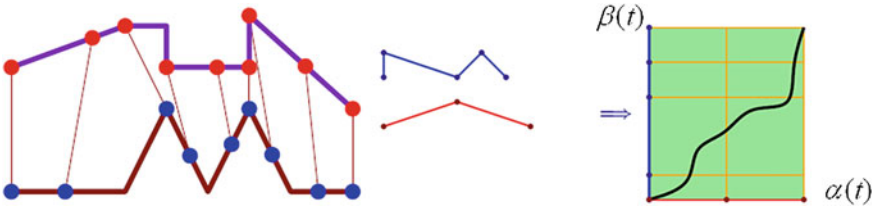


Fig. 7.13 Fréchet distance between two polygonal *curves* (and indexing all matching of *points*)

More recently, other distance have been introduced between opened curves [195–199] based on diffeomorphism theory, but for our application, we will only used the Fréchet distance and its extension on manifold.

7.7.3.2 Fréchet Distance Between Geodesic Paths on Manifold

In this section, we recall classical definition of distance between paths in Euclidean space. If we consider two curves, we can define similarity between each other. Classically, Hausdorff distance, that is the maximum distance between a point on one curve and its nearest neighbor on the other curve, is classically used but it does not take into account the flow and orientation of the curves. The Fréchet distance [125, 126] between two curves is defined as the minimum length of a leash required to connect a dog and its owner as they walk without backtracking along their respective curves from one endpoint to the other. The Fréchet metric takes the flow of the two curves into account; the pairs of points whose distance contributes to the Fréchet distance sweep continuously along their respective curves (Fig. 7.13).

Let P and Q be two given curves, the Fréchet distance between P and Q is defined as the infimum over all reparameterizations α and β of $[0, 1]$ of the maximum over all $t \in [0, 1]$ of the distance in between $P(\alpha(t))$ and $Q(\beta(t))$. In mathematical notation, the Fréchet distance $d_{Fréchet}(P, Q)$ is:

$$\begin{cases} d_{Fréchet}(P, Q) = \underset{\alpha,\beta}{Inf} \underset{i\in[0,1]}{\text{Max}} \{d(P(\alpha(t)), Q(\beta(t)))\} \\ \alpha \text{ and } \beta : [0,1] \rightarrow [0, 1] \text{ Non decreasing and surjective} \end{cases} \tag{7.231}$$

Alt and Godau [127] have introduced a polynomial-time algorithm to compute the Fréchet distance between two polygonal curves in Euclidean space. For two polygonal curves with m and n segments, the computation time is $O(mn \log(mn))$. Alt and Godau have defined the free-space diagram between two curves for a given distance threshold ε is a two-dimensional region in the parameter space that consist of all point pairs on the two curves at distance at most ε as given in the following equation: $D_\varepsilon(P, Q) = \left\{(\alpha, \beta) \in [0, 1]^2 / d_{Fréchet}(P(\alpha(t)), Q(\beta(t))) \leq \varepsilon\right\}$.

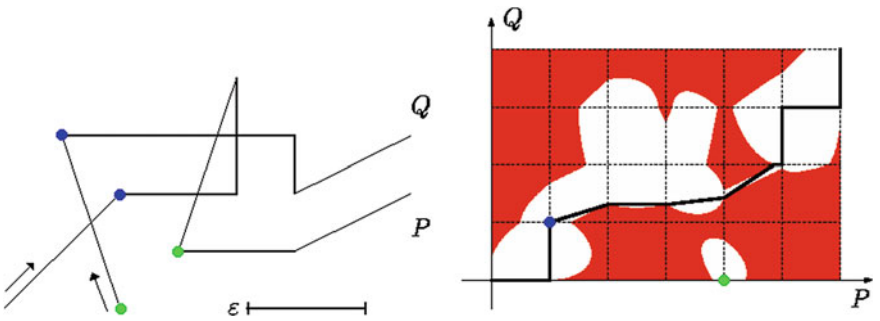


Fig. 7.14 Fréchet free-space diagram for two polygonal *curves P* and *Q* with monotonicity in both directions

The Fréchet distance $d_{Fréchet}(P, Q)$ is at most ε if and only if the free-space diagram $D_\varepsilon(P, Q)$ contains a path which from the lower left corner to the upper right corner which is monotone both in the horizontal and in the vertical direction.

In an $n \times m$ free-space diagram, shown in Fig. 7.14, the horizontal and vertical directions of the diagram correspond to the natural parameterizations of P and Q respectively. Therefore, if there is a monotone increasing curve from the lower left to the upper right corner of the diagram (corresponding to a monotone mapping), it generates a monotonic path that defines a matching between point-sets P and Q.

7.7.3.3 Fréchet Distance of Geodesic Paths for THPD Time Series on Structured Matrix Manifolds

We extend previous distance between paths to Information Geometry Manifold of Toeplitz Hermitian Positive Definite matrices. When the two curves are embedded in a more complex metric space [128], the distance between two points on the curves is most naturally defined as the geodesic length of the shortest path between them. If we consider N subsets of M pulses in the burst, the burst can then be described by a poly-geodesic lines on Information Geometry Manifold. The set of N covariances matrices $\{R(t_1), R(t_2), \ldots, R(t_N)\}$ describe a discrete "polygonal" geodesic path on Information Geometry Manifold, and we can extend previous Fréchet Distance but with Geodesic distance previously introduced in Sect. 7.7.2

$$\begin{cases} d_{Fréchet}(R_1, R_2) = \underset{\alpha,\beta}{Inf}\ \underset{t\in[0,1]}{Max}\left\{d_{geo}(R_1(\alpha(t)), R_2(\beta(t)))\right\} \\ \text{with } d^2_{geo}(R_1(\alpha(t)), R_2(\beta(t)))) = \left|\log\left(R_1^{-1/2}(\alpha(t))R_2(\beta(t))R_1^{-1/2}(\alpha(t))\right)\right|^2 \end{cases} \tag{7.232}$$

As classical Fréchet distance doesn't take into account with Inf[Max] close dependence of elements between points of time series paths, we propose to define a new distance given by:

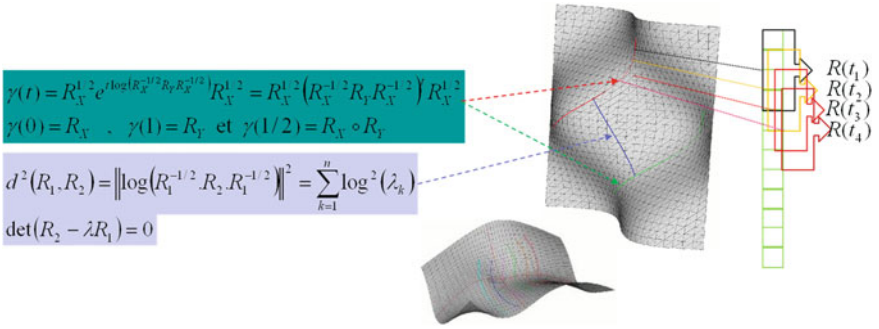


Fig. 7.15 Geodesic path on information geometry manifold where non stationary burst is decomposed on a sequence of stationary covariance matrices on THPD matrix manifold

$$d_{geo\text{-}path}(R_1, R_2) = \underset{\alpha,\beta}{Inf} \left\{ \int_0^1 d_{geo}(R_1(\alpha(t)), R_2(\beta(t)))\, dt \right\} \tag{7.233}$$

We have then to find the solution for computing the geodesic minimal path [133] on the Fréchet free-space diagram. The length of the path is not given by euclidean metric $ds^2 = dt^2$ (where $L = \int_L ds$) but geodesic metric weighted by d(.,.) of the free-space diagram:

$$L_g = \int_L g \cdot ds = \int_L ds_g \text{with } ds_g = d(R_1(\alpha(t)), R_2(\beta(t))) \cdot dt \tag{7.234}$$

This optimal shortest path could be computed by classical "Fast Marching method" (Fig. 7.15).

7.8 New Prospects: From Characteristic Functions to Generating Function and Generating Inner Product

We have shown that "Information geometry" should be explored as a particular application domain of Hessian geometry through Jean-Louis Koszul work (Koszul–Vinberg metric deduced from the associated Characteristic function having the main property to be invariant to all automorphisms of the convex cone).

Should we deduce that the "Eidos" or the "essence" of Information Geometry is limited to "*Koszul Characteristic Function*"? This notion seems not to be the more general one, and we would like to explore two new directions of research:

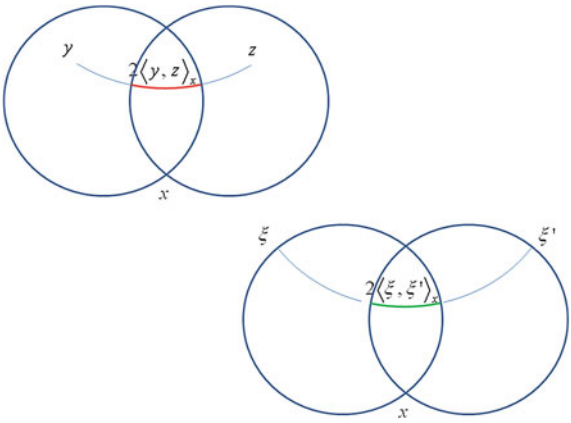


Fig. 7.16 Gromov inner product in homogeneous bounded domains and its Shilov boundary

- *Generating Inner Product*: In Koszul Geometry, we have two convex dual functions $\Phi(x)$ and $\Phi^*(x^*)$ with dual system of coordinates x and x^* defined on dual cones Ω and Ω^*: $\Phi(x) = -\log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi \forall x \in \Omega$ and $\Phi^*(x^*) = \langle x, x^*\rangle - \Phi(x)$. We can then remark that if we can define an Inner Product $\langle ., . \rangle$, we will be able to build convex characteristic function and its dual by Legendre transform because both are only dependent of the Inner product , and dual coordinate is also defined by $x^* = \arg\min\{\psi_\Omega(y)/y \in \Omega^*, \langle x, y\rangle = n\} = \int_{\Omega^*} \xi . e^{-\langle \xi, x\rangle} d\xi / \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi$ where x^* is also the center of gravity of the cross section $\{y \in \Omega^*, \langle x, y\rangle = n\}$ *of* Ω^* (with notation: $\Phi(x) = -\log \psi_\Omega(x)$).

  It is not possible to define an inner product for any two elements of a Lie Algebra, but a symmetric bilinear form, called "Cartan–Killing form", could be introduced. This form has been introduced first by Elie Cartan in 1894 in his PhD report. This form is defined according to the adjoint endomorphism Ad_x of g that is defined for every element x of g with the help of the Lie bracket:

$$Ad_x(y) = [x, y] \tag{7.235}$$

  The trace of the composition of two such endomorphisms defines a bilinear form, the Cartan–Killing form:

$$B(x, y) = Tr\left(Ad_x Ad_y\right) \tag{7.236}$$

  The Cartan–Killing form is symmetric:

$$B(x, y) = B(y, x) \tag{7.237}$$

  and has the associativity property:

$$B\left(\left[x,y\right],z\right)=B\left(x,\left[y,z\right]\right) \tag{7.238}$$

given by:

$$\begin{aligned} B\left(\left[x,y\right],z\right) &= Tr\left(Ad_{[x,y]}Ad_z\right)=Tr\left(\left[Ad_x,Ad_y\right]Ad_z\right)\\ &= Tr\left(Ad_x\left[Ad_y,Ad_z\right]\right)=B\left(x,\left[y,z\right]\right)\end{aligned}$$

Elie Cartan has proved that if g is a simple Lie algebra (the Killing form is non-degenerate) then any invariant symmetric bilinear form on g is a scalar multiple of the Cartan–Killing form. ***The Cartan–Killing form is invariant under automorphisms*** $\sigma \in Aut(g)$ of the algebra g:

$$B\left(\sigma(x),\sigma(y)\right)=B\left(x,y\right) \tag{7.239}$$

To prove this invariance, we have to consider:

$$\begin{cases}\sigma\left[x,y\right]=\left[\sigma(x),\sigma(y)\right]\\ z=\sigma(y)\end{cases}\Rightarrow\sigma\left[x,\sigma^{-1}(z)\right]=\left[\sigma(x),z\right]$$

rewritten

$$Ad_{\sigma(x)}=\sigma\circ Ad_x\circ\sigma^{-1}$$

Then

$$\begin{aligned} B\left(\sigma(x),\sigma(y)\right) &= Tr\left(Ad_{\sigma(x)}Ad_{\sigma(y)}\right)=Tr\left(\sigma\circ Ad_xAd_y\circ\sigma^{-1}\right)\\ &= Tr\left(Ad_xAd_y\right)=B(x,y)\end{aligned}$$

A natural G-invariant inner product could be then introduced by Cartan–Killing form:

$$\langle x,y\rangle=-B\left(x,\theta(y)\right) \tag{7.240}$$

where $\theta \in g$ is a Cartan involution (An involution on g is a Lie algebra automorphism θ of g whose square is equal to the identity).

As other generalization of inner product, we can also consider the case of *CAT(-1)-space* (generalization of simply connected Riemannian Manifold of negative curvature lower than unity) or of an Homogeneous Symmetric Bounded domains, and then define a "generating" Gromov Inner Product between three points x, y and z (relatively to x) that is defined by the distance [118]:

$$\langle y,z\rangle_x=\frac{1}{2}\left(d(x,y)+d(x,z)-d(y,z)\right) \tag{7.241}$$

with *d(.,.)* the distance in *CAT(-1)*. Intuitively, this inner product measures the distance of x to the geodesics between y to z.
This Inner product could be also defined for points on the Shilov Boundary of the domain through Busemann distance:

$$\left\langle \xi, \xi' \right\rangle_x = \frac{1}{2}\left(B_\xi(x,p) + B_{\xi'}(x,p)\right) \tag{7.242}$$

Independent of p, Where $B_\xi(x,y) = \underset{t\to+\infty}{\text{Lim}}\left[|x-r(t)| - |y-r(t)|\right]$ is horospheric distance, from x to y relatively to ξ , with $r(t)$ geodesic ray. We have the property that:

$$\left\langle \xi, \xi' \right\rangle_x = \underset{\substack{y\to\xi \\ y'\to\xi'}}{\text{Lim}} \left\langle y, y' \right\rangle_x \tag{7.243}$$

We can then define a visual metric on the Shilov boundary by (Fig. 7.16);

$$\begin{array}{l} d_x\left(\xi,\xi'\right) = e^{-\left\langle \xi,\xi' \right\rangle_x} \quad \text{if } \xi \neq \xi' \\ d_x\left(\xi,\xi'\right) = 0 \quad \text{otherwise} \end{array} \tag{7.244}$$

We can then define the characteristic function according to the origin 0:

$$\Phi(x) = -\log \int_{\Omega^*} e^{-\langle x,\gamma\rangle_0} d\gamma \text{ or } \Phi_\Omega(x) = -\log \int_{\Omega^*} e^{-\frac{1}{2}(d(0,x)+d(0,\gamma)-d(x,\gamma))} d\gamma \tag{7.245}$$

and

$$\Phi^*(x^*) = \left\langle x, x^* \right\rangle_0 - \Phi(x) = \frac{1}{2}\left(d(0,x) + d(0,x^*) - d(x,x^*)\right) - \Phi(x) \tag{7.246}$$

$$d(x,x^*) = \left(d(0,x^*) - 2\Phi^*(x^*)\right) + \left(d(0,x) - 2\Phi(x)\right) \tag{7.247}$$

with the center of gravity

$$x^* = \int_{\Omega^*} \gamma \cdot e^{-\langle x,\gamma\rangle_0} d\gamma \Big/ \int_{\Omega^*} e^{-\langle x,\gamma\rangle_0} d\gamma \tag{7.248}$$

All this relations are also true on the Shilov Boundary:

$$\Phi(\xi) = -\log \int_{\partial\Omega^*} e^{-\left\langle \xi,\xi' \right\rangle_0} d\xi' = -\log \int_{\partial\Omega^*} d_0\left(\xi,\xi'\right) \cdot d\xi' \tag{7.249}$$

where

$$\int_{\partial\Omega^*} d_0\left(\xi, \xi'\right) \cdot d\xi' \tag{7.250}$$

is the functional of Busemann barycenter on the Shilov Boundary $\partial\Omega^*$ (existence and unicity of this barycenter have been proved by Elie Cartan for Cartan–Hadamard Spaces) [123, 124].

- ***Generating Functions***: Characteristic function is a particular case of "generating function". In Mathematics, first idea to represent Lagrange manifolds through their generating functions was given by Hörmander [121] and then was generalized by Claude Viterbo [119, 120, 122]. A generating function for the Lagrange submanifold L is a function G from E to R defined on a vector bundle E, and such that:

$$L = \left\{ \left(x, \frac{\partial G}{\partial x} \right) / \frac{\partial G}{\partial \eta} = 0 \right\} \tag{7.251}$$

η is in the fibre. We have then to ,study in the framework of Information Geometry, families of generating functions more general than potential functions deduced from characteristic function.

In future work, it could be interesting to explore new direction related to polarized surface in the framework developed by Donaldson, Guillemin and Abreu, in which invariant Kähler metrics correspond to convex functions on the moment polytope of a toric variety [170, 171, 173–175] based on precursor work of Atiyah and Bott [172] on moment map and its convexity by Bruguières [181], Condevaux [182], Delzant [183], Guillemin and Sternberg [184, 185] and Kirwan [186]. Recently, Mikhail Kapranov has given a thermodynamical interpretation of the moment map for toric varieties [200].

We hope that this tutorial paper and new results will help readers to have s synthetic view of different domains of Science where Koszul hessian metric manifold is the essence emerging from Information Geometry Theory.

Das ist fast so, als ob ich Bergson wäre
Edmund Husserl (1917)
(after reading PhD first chapter of his student Roman Ingarden dedicated to Bergson Philosophy)

Création signifie, avant tout émotion. C'est alors seulement que l'esprit se sent ou se croit créateur. Il ne part plus d'une multiplicité d'éléments tout faits pour aboutir à une unité composite où il y aura un nouvel arrangement de l'ancien. Il s'est transporté tout d'un coup à quelque chose qui paraît à la fois un et unique, qui cherchera ensuite à s'étaler tant bien que mal en concepts multiples et communs, donnés d'avance dans des mots.
Henri Bergson
In "Les deux sources de la morale et de la religion"

References

1. Koszul, J.L.: Variétés localement plates et convexité. Osaka J. Math. **2**, 285–290 (1965)
2. Vey, J.: Sur les automorphismes affines des ouverts convexes saillants. Annali della Scuola Normale Superiore di Pisa, Classe di Science, 3e série, Tome **24**(4), 641–665 (1970)
3. Massieu, F.: Sur les fonctions caractéristiques des divers fluides. C. R. Acad. Sci. **69**, 858–862 (1869)
4. Massieu, F.: Addition au précédent Mémoire sur les fonctions caractéristiques. C. R. Acad. Sci. **69**, 1057–1061 (1869)
5. Massieu, F.: Thermodynamique: mémoire sur les fonctions caractéristiques des divers fluides et sur la théorie des vapeurs, 92 p. Académie des Sciences (1876)
6. Duhem, P.: Sur les équations générales de la thermodynamique. Annales Scientifiques de l'Ecole Normale Supérieure, 3e série, Tome **8**, 231 (1891)
7. Duhem, P.: Commentaire aux principes de la thermodynamique. Première partie, Journal de Mathématiques pures et appliquées, 4e série, Tome **8**, 269 (1892)
8. Duhem, P.: Commentaire aux principes de la thermodynamique—troisième partie. Journal de Mathématiques pures et appliquées, 4e série, Tome **10**, 203 (1894)
9. Duhem, P.: Les théories de la chaleur. Duhem **1992**, 351–1 (1895)
10. Laplace, P.S.: Mémoire sur la probabilité des causes sur les évènements. Mémoires de Mathématique et de Physique, Tome Sixième (1774)
11. Arnold, V.I., Givental, A.G.: Symplectic geometry. In: Encyclopedia of Mathematical Science, vol. 4. Springer, New York (translated from Russian) (2001)
12. Fitzpatrick, S.: On the geometric quantization of contact manifolds. http://arxiv.org/abs/0909.2023v3 (2013). Accessed Feb 2013
13. Rajeev, S.G.: Quantization of contact manifolds and thermodynamics. Ann. Phys. **323**(3), 768–82 (2008)
14. Gibbs, J.W.: Graphical methods in the thermodynamics of fluids. In: Bumstead, H.A., Van Name, R.G. (eds.) Scientific Papers of J Willard Gibbs, 2 vols. Dover, New York (1961)
15. Dedecker, P.: A property of differential forms in the calculus of variations. Pac. J. Math. **7**(4), 1545–9 (1957)
16. Lepage, T.: Sur les champs géodésiques du calcul des variations. Bull. Acad. Roy. Belg. Cl. Sci. **27**, 716–729, 1036–1046 (1936)
17. Mrugala, R.: On contact and metric structures on thermodynamic spaces. RIMS Kokyuroku **1142**, 167–81 (2000)
18. Ingarden R.S., Kossakowski A.: The poisson probability distribution and information thermodynamics. Bull. Acad. Pol. Sci. Sér. Sci. Math. Astron. Phys. **19**, 83–85 (1971)
19. Ingarden, R.S.: Information geometry in functional spaces of classical and quantum finite statistical systems. Int. J. Eng. Sci. **19**(12), 1609–33 (1981)
20. Ingarden, R.S., Janyszek, H.: On the local Riemannian structure of the state space of classical information thermodynamics. Tensor, New Ser. **39**, 279–85 (1982)
21. Ingarden, R.S., Kawaguchi, M., Sato, Y.: Information geometry of classical thermodynamical systems. Tensor, New Ser. **39**, 267–78 (1982)
22. Ingarden R.S.: Information geometry of thermodynamics. In: Transactions of the Tenth Prague Conference Czechoslovak Academy of Sciences, vol. 10A–B, pp. 421–428 (1987)
23. Ingarden, R.S.: Information geometry of thermodynamics, information theory, statistical decision functions, random processes. In: Transactions of the 10th Prague Conference, Prague/Czechoslovakia 1986, vol. A, pp. 421–428 (1988)
24. Ingarden, R.S., Nakagomi, T.: The second order extension of the Gibbs state. Open Syst. Inf. Dyn. **1**(2), 243–58 (1992)
25. Arnold V.I.: Contact geometry: the geometrical method of Gibbs's thermodynamics. In: Proceedings of the Gibbs Symposium, pp. 163–179. American Mathematical Society, Providence, RI (1990)
26. Cartan, E.: Leçons sur les Invariants Intégraux. Hermann, Paris (1922)

27. Koszul J.L.: Exposés sur les espaces homogènes symétriques. Publicação da Sociedade de Matematica de São Paulo (1959)
28. Koszul J.L.: Sur la forme hermitienne canonique des espaces homogènes complexes. Can. J. Math. **7**(4), 562–576 (1955)
29. Koszul, J.L.: Lectures on Groups of Transformations. Tata Institute of Fundamental Research, Bombay (1965)
30. Koszul, J.L.: Domaines bornées homogènes et orbites de groupes de transformations affines. Bull. Soc. Math. Fr. **89**, 515–33 (1961)
31. Koszul, J.L.: Ouverts convexes homogènes des espaces affines. Math. Z. **79**, 254–9 (1962)
32. Koszul, J.L.: Déformations des variétés localement plates. Ann. Inst. Fourier **18**, 103–14 (1968)
33. Vinberg, E.: Homogeneous convex cones. Trans. Moscow Math. Soc. **12**, 340–363 (1963)
34. Vesentini E.: Geometry of Homogeneous Bounded Domains. Springer, Berlin (2011). Reprint of the 1st Edn. C.I.M.E., Ed. Cremonese, Roma (1968)
35. Barbaresco F.: Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Fréchet median. In: Bhatia, R., Nielsen, F. (eds.) Matrix Information Geometry, pp. 199–256. Springer, New York (2012)
36. Arnaudon M., Barbaresco F., Le, Y.: Riemannian medians and means with applications to radar signal processing. IEEE J. Sel. Top. Sig. Process. **7**(4), 595–604 (2013)
37. Dorfmeister, J.: Inductive construction of homogeneous cones. Trans. Am. Math. Soc. **252**, 321–49 (1979)
38. Dorfmeister, J.: Homogeneous siegel domains. Nagoya Math. J. **86**, 39–83 (1982)
39. Poincaré, H.: Thermodynamique, Cours de Physique Mathématique. G. Carré, Paris (1892)
40. Poincaré, H.: Calcul des Probabilités. Gauthier-Villars, Paris (1896)
41. Faraut, J., Koranyi, A.: Analysis on Symmetric Cones. The Clarendon Press, New York (1994)
42. Faraut, J., Koranyi, A.: Oxford Mathematical Monographs. Oxford University Press, New York (1994)
43. Varadhan, S.R.S.: Asymptotic probability and differential equations. Commun. Pure Appl. Math. **19**, 261–86 (1966)
44. Sanov, I.N.: On the probability of large deviations of random magnitudes. Mat. Sb. **42**(84), 11–44 (1957)
45. Ellis, R.S.: The Theory of Large Deviations and Applications to Statistical Mechanics. Lecture Notes for Ecole de Physique Les Houches, France (2009)
46. Touchette, H.: The large deviation approach to statistical mechanics. Phys. Rep. **478**(1—-3), 1–69 (2009)
47. Cartan, E.: Sur les domaines bornés de l'espace de n variables complexes. Abh. Math. Semin. Hamburg **1**, 116–62 (1935)
48. Lichnerowicz, A.: Espaces homogènes Kähleriens. In: Collection Géométrie Différentielle, pp. 171–84, Strasbourg (1953)
49. Sasaki T.: A note on characteristic functions and projectively invariant metrics on a bounded convex domain. Tokyo J. Math. **8**(1), 49–79 (1985)
50. Sasaki, T.: Hyperbolic affine hyperspheres. Nagoya Math. J. **77**, 107–23 (1980)
51. Trench W.F.: An algorithm for the inversion of finite Toeplitz matrices. J. Soc. Ind. Appl. Math. **12**, 515–522 (1964)
52. Verblunsky, S.: On positive harmonic functions. Proc. London Math. Soc. **38**, 125–57 (1935)
53. Verblunsky, S.: On positive harmonic functions. Proc. London Math. Soc. **40**, 290–20 (1936)
54. Hauser, R.A., Güler, O.: Self-scaled barrier functions on symmetric cones and their classification. Found. Comput. Math. **2**(2), 121–43 (2002)
55. Vinberg, E.B.: Structure of the group of automorphisms of a homogeneous convex cone. Tr. Mosk. Mat. O-va **13**, 56–83 (1965)
56. Siegel, C.L.: Über der analytische theorie der quadratischen Formen. Ann. Math. **36**, 527–606 (1935)
57. Duan, X., Sun, H., Peng, L.: Riemannian means on special euclidean group and unipotent matrices group. Sci. World J. **2013**, ID 292787 (2013)

58. Soize C.: A nonparametric model of random uncertainties for reduced matrix models in structural dynamics. Probab. Eng. Mech. **15**(3), 277–294 (2000)
59. Bennequin, D.: Dualités de champs et de cordes. Séminaire N. Bourbaki, exp. no. 899, pp. 117–148 (2001–2002)
60. Bennequin D.: Dualité Physique-Géométrie et Arithmétique, Brasilia (2012)
61. Chasles M.: Aperçu historique sur l'origine et le développement des méthodes en géométrie (1837)
62. Gergonne, J.D.: Polémique mathématique. Réclamation de M. le capitaine Poncelet (extraite du bulletin universel des annonces et nouvelles scientifiques); avec des notes. Annales de Gergonne, vol. 18, pp. 125–125. http://www.numdam.org (1827–1828)
63. Poncelet, J.V.: Traité des propriétés projectives des figures (1822)
64. André, Y.: Dualités. Sixième séance, ENS, Mai (2008)
65. Atiyah, M.F.: Duality in mathematics and physics, lecture Riemann's influence in geometry. Analysis and Number Theory at the Facultat de Matematiques i Estadıstica of the Universitat Politecnica de Catalunya (2007)
66. Von Oettingen, A.J.: Harmoniesystem in dualer Entwicklung. Studien zur Theorie der Musik, Dorpat und Leipzig (1866)
67. Von Oettingen, A.J.: Das duale system der harmonie. In: Annalen der Naturphilosophie, vol. 1 (1902)
68. Von Oettingen, A.J.: Das duale system der harmonie. In: Annalen der Naturphilosophie, vol. 2, pp. 62–75 (1903/1904)
69. Von Oettingen, A.J.: Das duale system der harmonie. In: Annalen der Naturphilosophie, vol. 3, pp. 375–403 (1904)
70. Von Oettingen, A.J.: Das duale system der harmonie. In: Annalen der Naturphilosophie, vol. 4, pp. 241–269 (1905)
71. Von Oettingen, A.J.: Das duale system der harmonie. In: Annalen der Naturphilosophie, vol. 5, pp. 116–152, 301–338, 449–503 (1906)
72. Von Oettingen, A.J.: Das duale Harmoniesystem, Leipzig (1913)
73. Von Oettingen, A.J.: Die Grundlagen der musikwissenschaft und das duale reinistrument. In: Abhand-lungen der mathematisch-physikalischen Klasse der Königlich Sächsischen Gesellschaft der Wissenschaften, vol. 34, pp. S.I–XVI, 155–361 (1917)
74. D'Alembert, J.R.: Éléments de musique, théorique et pratique, suivant les principes de M. Rameau, Paris (1752)
75. Rameau, J.P.: Traité de l'harmonie Réduite à ses Principes Naturels. Ballard, Paris (1722)
76. Rameau, J.P.: Nouveau système de Musique Théorique. Ballard, Paris (1726)
77. Yang, L.: Médianes de mesures de probabilité dans les variétés riemanniennes et applications à la détection de cibles radar. Thèse de l'Université de Poitiers, tel-00664188, 2011, Thales PhD Award (2012)
78. Barbaresco, F.: Algorithme de Burg Régularisé FSDS. Comparaison avec l'algorithme de Burg MFE, pp. 29–32 GRETSI conference (1995)
79. Barbaresco, F.: Information geometry of covariance matrix. In: Nielsen, F., Bhatia, R. (eds.) Matrix Information Geometry Book. Springer, Berlin (2012)
80. Émery, M., Mokobodzki, G.: Sur le barycentre d'une probabilité dans une variété. Séminaire de probabilité Strasbourg **25**, 220–233 (1991)
81. Friedrich, T.: Die Fisher-information und symplektische strukturen. Math. Nachr. **153**, 273–96 (1991)
82. Bingham N.H.: Szegö's Theorem and Its Probabilistic Descendants. http://arxiv.org/abs/1108.0368v2 (2012)
83. Landau, H.J.: Maximum entropy and the moment problem. Bull. Am. Math. Soc. **16**(1), 47–77 (1987)
84. Siegel, C.L.: Symplectic geometry. Am. J. Math. **65**, 1–86 (1943)
85. Libermann, P., Marle, C.M.: Symplectic Geometry and Analytical Mechanics. Reidel, Dordrecht (1987)

86. Delsarte, P., Genin, Y.V.: Orthogonal polynomial matrices on the unit circle. IEEE Trans. Comput. Soc. **25**(3), 149–160 (1978)
87. Kanhouche, R.: A modified burg algorithm equivalent. In: Results to Levinson algorithm. http://hal.archives-ouvertes.fr/ccsd-00000624
88. Douady, C.J., Earle, C.J.: Conformally natural extension of homeomorphisms of circle. Acta Math. **157**, 23–48 (1986)
89. Balian, R.: A metric for quantum states issued from von Neumann's entropy. In: Nielsen, F., Barbaresco, F. (Eds.) Geometric Science of Information. Lecture Notes in Computer Science, vol. 8085, pp. 513–518
90. Balian, R.: Incomplete descriptions and relevant entropies. Am. J. Phys. **67**, 1078–90 (1989)
91. Balian, R.: Information in statistical physics. Stud. Hist. Philos. Mod. Phys. **36**, 323–353 (2005)
92. Allahverdyan, A., Balian, R., Nieuwenhuizen, T.: Understanding quantum measurement from the solution of dynamical models. Phys. Rep. **525**, 1–166 (2013) (ArXiv: 1107, 2138)
93. Balian, R.: From Microphysics to Macrophysics: Methods and Applications of Statistical Physics, vol. 1–2. Springer (2007)
94. Balian, R., Balazs, N.: Equiprobability, information and entropy in quantum theory. Ann. Phys. (NY) **179**, 97–144 (1987)
95. Balian, R., Alhassid, Y., Reinhardt, H.: Dissipation in many-body systems: a geometric approach based on information theory. Phys. Rep. **131**, 1–146 (1986)
96. Barbaresco F.: Information/contact geometries and Koszul entropy. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information. Lecture Notes in Computer Science, vol. 8085, pp. 604–611. Springer, Berlin (2013)
97. Souriau J.M.: Définition covariante des équilibres thermodynamiques. Suppl. Nuovo Cimento **1**, 203–216 (1966)
98. Souriau, J.M.: Thermodynamique et Géométrie **676**, 369–397 (1978)
99. Souriau, J.M.: Géométrie de l'espace de phases. Commun. Math. Phys. **1**, 374 (1966)
100. Souriau, J.M.: On geometric mechanics. Discrete Continuous Dyn. Syst. **19**(3), 595–607 (2007)
101. Souriau, J.M.: Structure des Systèmes Dynamiques. Dunod, Paris (1970)
102. Souriau, J.M.: Structure of dynamical systems. Progress in Mathematics, vol. 149. Birkhäuser Boston Inc., Boston. A symplectic view of physics (translated from the French by Cushman-de Vries, C.H.) (1997)
103. Souriau, J.M.: Thermodynamique relativiste des fluides. Rend. Sem. Mat. Univ. e Politec. Torino, 35:21–34 (1978), 1976/77
104. Souriau, J.M., Iglesias, P.: Heat cold and geometry. In: Cahen, M., et al. (eds.) Differential Geometry and Mathematical Physics, pp. 37–68 (1983)
105. Souriau, J.M.: Thermodynamique et géométrie. In: Differential Geometrical Methods in Mathematical Physics, vol. 2 (Proceedings of the International Conference, University of Bonn, Bonn, 1977). Lecture Notes in Mathematics, vol. 676, pp. 369–397. Springer, Berlin (1978)
106. Souriau, J.M.,: Dynamic systems structure (Chap. 16 Convexité, Chap. 17 Mesures, Chap. 18 Etats Statistiques, Chap. 19 Thermodynamique), unpublished technical notes, available in Souriau archive (document sent by Vallée, C.)
107. Vallée, C.: Lois de comportement des milieux continus dissipatifs compatibles avec la physique relativiste, thèse, Poitiers University (1978)
108. Iglésias P., Equilibre statistiques et géométrie symplectique en relativité générale. Ann. l'Inst. Henri Poincaré, Sect. A, Tome **36**(3), 257–270 (1982)
109. Iglésias, P.: Essai de thermodynamique rationnelle des milieux continus. Ann. l'Inst. Henri Poincaré, **34**, 1–24 (1981)
110. Vallée, C.: Relativistic thermodynamics of continua. Int. J. Eng. Sci. **19**(5), 589–601 (1981)
111. Pavlov, V.P., Sergeev, V.M.: Thermodynamics from the differential geometry standpoint. Theor. Math. Phys. **157**(1), 1484–1490 (2008)
112. Kozlov, V.V.: Heat Equilibrium by Gibbs and Poincaré. RKhD, Moscow (2002)

113. Berezin, F.A.: Lectures on Statistical Physics. Nauka, Moscow (2007) (English trans., World Scientific, Singapore, 2007)
114. Poincaré, H.: Réflexions sur la théorie cinétique des gaz. J. Phys. Theor. Appl. **5**, 369–403 (1906)
115. Carathéodory, C.: Math. Ann. **67**, 355–386 (1909)
116. Nakajima, S.: On quantum theory of transport phenomena. Prog. Theor. Phys. **20**(6), 948–959 (1958)
117. Zwanzig, R.: Ensemble method in the theory of irreversibility. J. Chem. Phys. **33**(5), 1338–1341 (1960)
118. Bourdon, M.: Structure conforme au bord et flot géodésique d'un CAT(-1)-espace. L'Enseignement Math. **41**, 63–102 (1995)
119. Viterbo, C.: Generating functions, symplectic geometry and applications. In: Proceedings of the International Congress Mathematics, Zurich (1994)
120. Viterbo, C.: Symplectic topology as the geometry of generating functions. Math. Ann. **292**, 685–710 (1992)
121. Hörmander, L.: Fourier integral operators I. Acta Math. **127**, 79–183 (1971)
122. Théret, D.: A complete proof of Viterbo's uniqueness theorem on generating functions. Topology Appl. **96**, 249–266 (1999)
123. Pansu, P.: Volume, courbure et entropie. Séminaire Bourbaki **823**, 83–103 (1996)
124. Besson, G., Courtois, G., Gallot, S.: Entropies et rigidités des espaces localement symétriques de courbure strictement négative. Geom. Funct. Anal. **5**, 731–799 (1995)
125. Fréchet, M.: Sur quelques points du calcul fonctionnel. Rend. Circolo Math. Palermo **22**, 1–74 (1906)
126. Fréchet, M.: L'espace des courbes n'est qu'un semi-espace de Banach. General Topology and Its Relation to Modern Analysis and Algebra, pp. 155–156, Prague (1962)
127. Alt, H., Godau, M.: Computing the Fréchet distance between two polygonal curves. Int. J. Comput. Geom. Appl. **5**, 75–91 (1995)
128. Fréchet, M.R.: Les éléments aléatoires de nature quelconque dans un espace distancié. Ann. l'Inst. Henri Poincaré **10**(4), 215–310 (1948)
129. Marle, C.M.: On mechanical systems with a Lie group as configuration space. In: M. de Gosson (ed.) Jean Leray '99 Conference Proceedings: the Karlskrona Conference in the Honor of Jean Leray, Kluwer, Dordrecht, pp. 183–203 (2003)
130. Marle, C.M.: On Henri Poincaré's note "Sur une forme nouvelle des équations de la mécanique". JGSP **29**, 1–38 (2013)
131. Kloeckner, B.: Géométrie des bords: compactifications différentiables et remplissages holomorphes. Thèse Ecole Normale Supérieure de Lyon. http://tel.archives-ouvertes.fr/tel-00120345 (2006). Accessed Dec 2006
132. Barbaresco, F.: Super Resolution Spectrum Analysis Regularization: Burg, Capon and Ago-antagonistic Algorithms, EUSIPCO-96, pp. 2005–2008, Trieste (1996)
133. Barbaresco, F.: Computation of most threatening radar trajectories areas and corridors based on fast-marching and level sets. In: IEEE CISDA Symposium, Paris (2011)
134. Michor, P.W., Mumford, D.: An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach. Appl. Comput. Harm. Anal. **23**(1), 74–113 (2007)
135. Chouakria-Douzal, A., Nagabhusha, P.N.: Improved Fréchet distance for time series. In: Data Sciences and Classification, pp. 13–20. Springer, Berlin (2006)
136. Bauer, M., et al.: Constructing reparametrization invariant metrics on spaces of plane curves. Preprint http://arxiv.org/abs/1207.5965
137. Fréchet, M.: L'espace dont chaque élément est une courbe n'est qu'un semi-espace de Banach. Ann. Sci. l'ENS, 3ème série. Tome **78**(3), 241–272 (1961)
138. Fréchet, M.: L'espace dont chaque élément est une courbe n'est qu'un semi-espace de Banach II. Ann.Sci. l'ENS, 3ème série. Tome **80**(2), pp. 135–137 (1963)
139. Chazal, F., et al.: Gromov-Hausdorff stable signatures for shapes using persistence. In: Eurographics Symposium on Geometry Processing 2009, Marc Alexa and Michael Kazhdan (Guest editors), vol. 28, no. 5 (2009)

140. Cagliari, F., Di Fabio B., Landi, C.: The natural pseudo-distance as a quotient pseudo metric, and applications. Preprint http://amsacta.unibo.it/3499/1/Forum_Submission.pdf
141. Frosini, P., Landi, C.: No embedding of the automorphisms of a topological space into a compact metric space endows them with a composition that passes to the limit. Appl. Math. Lett. **24**(10), 1654–1657 (2011)
142. Shevchenko, O.: Recursive construction of optimal self-concordant barriers for homogeneous cones. J. Optim. Theor. Appl. **140**(2), 339–354 (2009)
143. Güler, O., Tunçel, L.: Characterization of the barrier parameter of homogeneous convex cones. Math. Program. **81**(1), Ser. A, 55–76 (1998)
144. Rothaus, O.S.: Domains of positivity. Bull. Am. Math. Soc. **64**, 85–86 (1958)
145. Nesterov, Y., Nemirovskii, A.: Interior-point polynomial algorithms. In: Convex Programming, SIAM Studies in Applied Mathematics, vol. 13 (1994)
146. Vinberg, E.B.: The theory of homogeneous convex cones. Tr. Mosk. Mat. O-va. **12**, 303–358 (1963)
147. Rothaus, O.S.: The construction of homogeneous convex cones. Ann. Math. Ser. 2, **83**, 358–376 (1966)
148. Güler, O.: Barrier functions in interior point methods. Math. Oper. Res. **21**(4), 860–885 (1996)
149. Uehlein, F.A.: Eidos and Eidetic Variation in Husserl's Phenomenology. In: Language and Schizophrenia, Phenomenology, pp. 88–102. Springer, New York (1992)
150. Bergson, H.: L'évolution créatrice. Les Presses universitaires de France, Paris (1907). http://classiques.uqac.ca/classiques/bergson_henri/evolution_creatrice/evolution_creatrice.pdf
151. Riquier, C.: Bergson lecteur de Platon: le temps et l'eidos, dans interprétations des idées platoniciennes dans la philosophie contemporaine (1850–1950), coll. Tradition de la pensée classique, Paris, Vrin (2011)
152. Worms, F.: Bergson entre Russel et Husserl: un troisième terme? In: Rue Descartes, no. 29, Sens et phénomène, philosophie analytique et phénoménologie, pp. 79–96, Presses Universitaires de France, Sept. 2000
153. Worms, F.: Le moment 1900 en philosophie. Presses Universitaires du Septentrion, premier trimestre, Etudes réunies sous la direction de Frédéric Worms (2004)
154. Worms, F.: Bergson ou Les deux sens de la vie: étude inédite, Paris, Presses universitaires de France, Quadrige. Essais, débats (2004)
155. Bergson, H., Poincaré, H.: Le matérialisme actuel. Bibliothèque de Philosophie Scientifique, Paris, Flammarion (1920)
156. de Saxcé G., Vallée C.: Bargmann group, momentum tensor and Galilean invariance of Clausius-Duhem inequality. Int. J. Eng. Sci. **50**, 216–232 (2012)
157. Dubois, F.: Conservation laws invariants for Galileo group. CEMRACS preliminary results. ESAIM Proc. **10**, 233–266 (2001)
158. Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien. C. R. l'Acad. des Sci. Série A, Tome **255**, 2897–2899 (1962)
159. Nielsen, F.: Hypothesis testing, information divergence and computational geometry. In: GSI'13 Conference, Paris, pp. 241–248 (2013)
160. Vey, J.: Sur une notion d'hyperbolicité des variables localement plates. Faculté des sciences de l'université de Grenoble, Thèse de troisième cycle de mathématiques pures (1969)
161. Ruelle, D.: Statistical mechanics. In: Rigorous Results (Reprint of the 1989 edition). World Scientific Publishing Co., Inc, River Edge. Imperial College Press, London (1999)
162. Ruelle, D.: Hasard et Chaos. Editions Odile Jacob, Aout (1991)
163. Shima, H.: Geometry of Hessian Structures. In: Nielsen, F., Barbaresco, F. (eds.) Lecture Notes in Computer Science, vol. 8085, pp. 37–55. Springer, Berlin (2013)
164. Shima, H.: The Geometry of Hessian Structures. World Scientific, London (2007)
165. Zia, R.K.P., Redish Edward F., McKay Susan, R.: Making Sense of the Legendre Transform (2009), arXiv:0806.1147, June 2008
166. Fréchet, M.: Sur l'écart de deux courbes et sur les courbes limites. Trans. Am. Math. Soc. **6**(4), 435–449 (1905)

167. Taylor, A.E., Dugac, P.: Quatre lettres de Lebesgue à Fréchet. Rev. d'Hist. Sci. Tome **34**(2), 149–169 (1981)
168. Jensen, J.L.W.: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math. **30**(1), 175–193 (1906)
169. Needham, T.: A visual explanation of Jensen's inequality. Am. Math. Mon. **8**, 768–77 (1993)
170. Donaldson, S.K.: Scalar curvature and stability of toric variety. J. Differ. Geom. **62**, 289–349 (2002)
171. Abreu, M.: Kähler geometry of toric varieties and extremal metrics. Int. J. Math. **9**, 641–651 (1998)
172. Atiyah, M., Bott, R.: The moment map and equivariant cohomology. Topology **23**, 1–28 (1984)
173. Guan, D.: On modified Mabuchi functional and Mabuchi moduli space of kahler metrics on toric bundles. Math. Res. Lett. **6**, 547–555 (1999)
174. Guillemin, V.: Kaehler structures on toric varieties. J. Differ. Geom. **40**, 285–309 (1994)
175. Guillemin, V.: Moment maps and combinatorial invariants of Hamiltonian T^n-spaces, Birkhauser (1994)
176. Crouzeix, J.P.: A relationship between the second derivatives of a convex function and of its conjugate. Math. Program. **3**, 364–365 (1977) (North-Holland)
177. Seeger, A.: Second derivative of a convex function and of its Legendre-Fenchel transformate. SIAM J. Optim. **2**(3), 405–424 (1992)
178. Hiriart-Urruty, J.B.: A new set-valued second-order derivative for convex functions. Mathematics for Optimization, Mathematical Studies, vol. 129. North Holland, Amsterdam (1986)
179. Berezin, F.: Lectures on Statistical Physics (Preprint 157). Max-Plank-Institut für Mathematik, Bonn (2006)
180. Hill, R., Rice, J.R.: Elastic potentials and the structure of inelastic constitutive laws. SIAM J. Appl. Math. **25**(3), 448–461 (1973)
181. Bruguières, A.: Propriétés de convexité de l'application moment, séminaire N. Bourbaki, exp. no. 654, pp. 63–87 (1985–1986)
182. Condevaux, M., Dazord, P., Molino, P.: Géométrie du moment. Trav. Sémin. Sud-Rhodanien Géom. Univ. Lyon **1**, 131–160 (1988)
183. Delzant, T.: Hamiltoniens périodiques et images convexes de l'application moment. Bull. Soc. Math. Fr. **116**, 315–339 (1988)
184. Guillemin, V., Sternberg, S.: Convexity properties of the moment mapping. Inv. Math. **67**, 491–513 (1982)
185. Guillemin, V., Sternberg, S.: Convexity properties of the moment mapping. Inv. Math. **77**, 533–546 (1984)
186. Kirwan, F.: Convexity properties of the moment mapping. Inv. Math. **77**, 547–552 (1984)
187. Deza, E., Deza, M.M.: Dictionary of Distances. Elsevier, Amsterdam (2006)
188. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. II **106**(4), 620–630 (1957)
189. Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. II **108**(2), 171–190 (1957)
190. Jaynes, E.T.: Prior probabilities. IEEE Trans. Syst. Sci. Cybern. **4**(3), 227–241 (1968)
191. Amari, S.I., Nagaoka, H.: Methods of Information Geometry (Translation of Mathematical Monographs), vol. 191. AMS, Oxford University Press, Oxford (2000)
192. Amari, S.I.: Differential Geometrical Methods in Statistics. Lecture Notes in Statistics, vol. 28. Springer, Berlin (1985)
193. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–89 (1945)
194. Chentsov, N.N.: Statistical decision rules and optimal inferences. In: Transactions of Mathematics Monograph, vol. 53. American Mathematical Society, Providence (1982) (Published in Russian in 1972)
195. Trouvé, A., Younes, L.: Diffeomorphic matching in 1d: designing and minimizing matching functionals. In: Vernon, D. (ed.) Proceedings of ECCV (2000)

196. Trouvé, A., Younes, L.: On a class of optimal matching problems in 1 dimension. SIAM J. Control Opt. **39**(4), 1112–1135 (2001)
197. Younes, L.: Computable elastic distances between shapes. SIAM J. Appl. Math **58**, 565–586 (1998)
198. Younes, L.: Optimal matching between shapes via elastic deformations. Image Vis. Comput. **17**, 381–389 (1999)
199. Younes, L., Michor, P.W., Shah, J., Mumford, D.: A metric on shape space with explicit geodesics. Rend. Lincei Mat. Appl. **9**, 25–57 (2008)
200. Kapranov, M.: Thermodynamics and the Moment Map (preprint), arXiv:1108.3472, Aug 2011

Chapter 8
Distances on Spaces of High-Dimensional Linear Stochastic Processes: A Survey

Bijan Afsari and René Vidal

Abstract In this paper we study the geometrization of certain spaces of stochastic processes. Our main motivation comes from the problem of pattern recognition in high-dimensional time-series data (e.g., video sequence classification and clustering). In the first part of the paper, we provide a rather extensive review of some existing approaches to defining distances on spaces of stochastic processes. The majority of these distances are, in one way or another, based on comparing power spectral densities of the processes. In the second part, we focus on the space of processes generated by (stochastic) linear dynamical systems (LDSs) of fixed size and order, for which we recently introduced a class of group action induced distances called the alignment distances. This space is a natural choice in some pattern recognition applications and is also of great interest in control theory, where it is often convenient to represent LDSs in state-space form. In this case the space (more precisely manifold) of LDSs can be considered as the base space of a principal fiber bundle comprised of state-space realizations. This is due to a Lie group action symmetry present in the state-space representation of LDSs. The basic idea behind the alignment distance is to compare two LDSs by first aligning a pair of their realizations along the respective fibers. Upon a standardization (or bundle reduction) step this alignment process can be expressed as a minimization problem over orthogonal matrices, which can be solved efficiently. The alignment distance differs from most existing distances in that it is a structural or generative distance, since in some sense it compares how two processes are generated. We also briefly discuss averaging LDSs using the alignment distance via minimizing a sum of the squares of distances (namely, the so-called Fréchet mean).

B. Afsari (✉) · R. Vidal
Center for Imaging Science, Johns Hopkins University, Baltimore MD 21218, USA
e-mail: bijan@cis.jhu.edu

R. Vidal
e-mail: rvidal@cis.jhu.edu

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_8,

Keywords Stochastic processes · Pattern recognition · Linear dynamical systems · Extrinsic and intrinsic geometries · Principal fiber bundle · Generalized dynamic factor model · Minimum phase · Spectral factorization · All-pass filter · Hellinger distance · Itakura-Saito divergence · Fréchet mean

8.1 Introduction and Motivation

Pattern recognition (e.g., classification and clustering) of time-series data is important in many real world data analysis problems. Early applications include the analysis of one-dimensional data such as speech and seismic signals (see, e.g., [48] for a review). More recently, applications in the analysis of video data (e.g., activity recognition [1]), robotic surgery data (e.g., surgical skill assessment [12]), or biomedical data (e.g., analysis of multichannel EEG signals) have motivated the development of statistical techniques for the analysis of high-dimensional (or vectorial) time-series data.

The problem of pattern recognition for time-series data, in its full generality, needs tools from the theory of statistics on stochastic processes or function spaces. Thus it bears relations with the general problem of inference on (infinite dimensional) spaces of stochastic processes, which requires a quite sophisticated mathematical theory [30, 59]. However, at the same time, the pattern recognition problem is more complicated since, in general, it involves not only inference but also learning. Learning and inference on infinite dimensional spaces obviously can be daunting tasks. In practice, there have been different grand strategies proposed to deal with this problem (e.g., see [48] for a review). In certain cases it is reasonable and advantageous from both theoretical and computational points of view to simplify the problem by assuming that the observed processes are generated by models from a specific finite-dimensional class of models. In other words, one could follow a parametric approach based on *modeling* the observed time series and then performing statistical analysis and inference on a finite dimensional *space of models* (instead of the space of the observed *raw* data). In fact, in many real-world instances (e.g., video sequences [1, 12, 22, 60] or econometrics [7, 20, 24]), one could model the observed high-dimensional time series with low-order Linear Dynamical Systems (LDSs). In such instances the mentioned strategy could prove beneficial, e.g., in terms of implementation (due to significant compression achieved in high dimensions), statistical inference, and synthesis of time series. For 1-dimensional time-series data the success of Linear Predictive Coding (i.e., auto-regressive (AR) modeling) modeling and its derivatives in modeling speech signals is a paramount example [26, 49, 58]. These motivations lead us to state the following prototype problem:

Problem 1 (*Statistical analysis on spaces of LDSs*) Let $\{\boldsymbol{y}^i\}_{i=1}^N$ be a collection of p-dimensional time series indexed by time t. Assume that each time series $\boldsymbol{y}^i = \{\boldsymbol{y}_t^i\}_{t=1}^\infty$ can be approximately modeled by a (stochastic) LDS M_i of output-input size (p, m) and order n[1] realized as

[1] Typically in video analysis: $p \approx 1000$–10000, $m, n \approx 10$ (see e.g., [1, 12, 60]).

$$\begin{aligned} \boldsymbol{x}_t^i &= A_i \boldsymbol{x}_{t-1}^i + B_i \boldsymbol{v}_t, \\ \boldsymbol{y}_t^i &= C_i \boldsymbol{x}_t^i + D_i \boldsymbol{v}_t, \quad (A_i, B_i, C_i, D_i) \in \widetilde{\mathcal{SL}}_{m,n,p} = \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m} \end{aligned} \tag{8.1}$$

where $\boldsymbol{v}_t$ is a common stimulus process (e.g., white Gaussian noise with identity covariance)[2] and where the realization $R_i = (A_i, B_i, C_i, D_i)$ is learnt and assumed to be known. The problem is to: (1) Choose an appropriate space $\mathcal{S}$ of LDSs containing the learnt models $\{M_i\}_{i=1}^N$, (2) geometrize $\mathcal{S}$, i.e., equip it with an appropriate geometry (e.g., define a distance on $\mathcal{S}$), (3) develop tools (e.g., probability distributions, averages or means, variance, PCA) to perform statistical analysis (e.g., classification and clustering) in a computationally efficient manner.

The first question to ask is: why model the processes using the state-space model (representation) (8.1)? Recall that processes have equivalent ARMA and state-space representations. Moreover, model (8.1) is quite general and with n large enough it can approximate a large class of processes. More importantly, state-space representations (especially in high dimensions) are often more suitable for parameter learning or system identification. In important practical cases of interest such models conveniently yield more parsimonious parametrization than vectorial ARMA models which suffer from the *curse of dimensionality* [24]. The curse of dimensionality in ARMA models stems from the fact that for p-dimensional time series if p is very large the number of parameters of an ARMA model is roughly proportional to p^2, which could be much larger than the number of data samples available pT, where T is the observation time period (note that the autoregressive coefficient matrices are very large $p \times p$ matrices). However, in many situations encountered in real world examples, state-space models are more effective in overcoming the curse of dimensionality [20, 24]. The intuitive reason, as already alluded to, is that often (very) high-dimensional time series can be well approximated as being generated by a *low order* but high-dimensional dynamical system (which implies *small* n despite *large* p in the model (8.1)). This can be attributed to the fact that the components of the observed time series exhibit correlations (cross sectional correlation). Moreover, the contaminating noises also show correlation across different components (see [20, 24] for examples of exact and detailed assumptions and conditions to formalize these intuitive facts). Therefore, overall the number of parameters in the state-space model is small compared with p^2 and this is readily reflected in (or encoded by) the small size of the dynamics matrix A_i and the thinness of the observation matrix C_i in (8.1).[3]

[2] Note that in a different or more general setting the noise at the output could be a process $\boldsymbol{w}_t$ different (independent) from the input noise $\boldsymbol{v}_t$. This does not cause major changes in our developments. Since the output noise usually represents a perturbation which *cannot* be modeled, as far as Problem 1 is concerned, one could usually assume that $D_i = 0$.

[3] Note that we are not implying that ARMA models are incapable of modeling such time series. Rather the issue is that general or unrestricted ARMA models suffer from the curse of dimensionality in the identification problem, and the parametrization of a restricted class of ARMA models with a small number of parameters is complicated [20]. However, at the same time, by using state-space models it is easier to overcome the curse of dimensionality and this approach naturally leads to simple and effective identification algorithms [20, 22].

Also, in general, state-space models are more convenient for computational purposes than vectorial ARMA models. For example, in the case of high-dimensional time series most effective estimation methods are based on state-space domain system identification rooted in control theory [7, 41, 51]. Nevertheless, it should be noted that, in general, the identification of multi-input multi-output (MIMO) systems is a subtle problem (see Sect. 8.4 and e.g., [11, 31, 32]). However, for the case where $p > n$, there are efficient system identification algorithms available for finding the state-space parameters [20, 22].

Notice that in Problem 1 we are assuming that all the LDSs have the same order n (more precisely the minimal order, see Sect. 8.3.3.1). Such an assumption might seem rather restrictive and a more realistic assumption might be that all systems be of order not larger than n (see Sect. 8.5.1). Note that since in practice real data can be only *approximately* modeled by an LDS of fixed order, if n is not chosen too large, then gross over-fitting of n is less likely to happen. From a practical point of view (e.g., implementation) fixing the order for all systems results in great simplification in implementation. Moreover, in classification or clustering problems one might need to combine (e.g., average) such LDSs for the goal of replacing a class of LDSs with a representative LDS. Ideally one would like to define an average in a such a way that LDSs of the same order have an average of the same order and not higher, otherwise the problem can become intractable. In fact, most existing approaches tend to dramatically increase the order of the average LDS, which is certainly undesirable. Therefore, intuitively, we would like to consider a space $\mathcal{S}$ in which the order of the LDSs is fixed or limited. From a theoretical point of view also this assumption allows us to work with nicer mathematical spaces namely smooth manifolds (see Sect. 8.4).

Amongst the most widely used classification and clustering algorithms for static data are the k-nearest neighborhood and k-means algorithms, both of which rely on a notion of distance (in a feature space) [21]. These algorithms enjoy certain universality properties with respect to the probability distributions of the data; and hence in many practical situations where one has little prior knowledge about the nature of the data, they prove to be very effective [21, 35]. In view of this fact, in this paper we focus on the notion of distance between LDSs and the stochastic processes they generate. Hence, a natural question is what space we should use and what type of distance we should define on it. In Problem 1, obviously, the first two steps (which are the focus of this paper) have significant impacts on the third one. One has different choices for the space $\mathcal{S}$, as well as, for geometries on that space. The gamut ranges from an *infinite dimensional linear* space to a *finite dimensional (non-Euclidean) manifold*, and the geometry can be either *intrinsic* or *extrinsic*. By an intrinsic geometry we mean one in which a shortest path between two points in a space stays in the space, and by an extrinsic geometry we mean one where the distance between the two points is measured in an *ambient* space. In the second part of this paper, we study our recently developed approach, which is somewhere in between: to design an *easy-to-compute* extrinsic distance, while keeping the ambient space *not* too large.

This paper is organized as follows: In Sect. 8.2, we review some existing approaches in geometrization of spaces of stochastic processes. In Sect. 8.3, we focus

on processes generated by LDSs of fixed order, and in Sect. 8.4, we study smooth fiber bundle structures over spaces of LDSs generating such processes. Finally, in Sect. 8.5, we introduce our class of group action induced distances namely the *alignment* distances. The paper is concluded in Sect. 8.6. To avoid certain technicalities and just to convey the main ideas the proofs are omitted and will appear elsewhere. We should stress that the theory of alignment distances on spaces of LDSs is still under development; however, its basics have appeared in earlier papers [1–3]. This paper for most parts is an extended version of [3].

8.2 A Review of Existing Approaches to Geometrization of Spaces of Stochastic Processes

This review, in particular, since the subject appears in a range of disciplines is non-exhaustive. Our emphasis is on the core ideas in defining distances on spaces of stochastic processes rather than enumerating all such distances. Other sources to consult may include [9, 10, 25]. In view of Problem 1, our main interest is in the finite dimensional spaces of LDSs of fixed order and the processes they generate. However, since such a space can be embedded in the larger infinite dimensional space of "virtually all processes," first we consider the latter.

Remark 1 We shall discuss several "distance-like" measures some of which are known as "distance" in the literature. We will try to use the term *distance* exclusively for a true distance namely one which is symmetric, positive definite and obeys the triangle inequality. Due to convention or convenience, we still may use the term distance for something which is not a true distance, but the context will be clear. A distance-like measures is called a divergence it is only positive definite and it is called pseudo-distance, if it is symmetric and obeys the triangle inequality but it is only positive semi-definitive (i.e., a zero distance between two processes does not imply that they are the same). As mentioned above, our review is mainly to show different schools of thought and theoretical approaches in defining distances. Obviously, when it comes to comparing these distances and their effectiveness (e.g., in terms of recognition rate in a pattern recognition problem) ultimately things very much depend on the specific application at hand. Although we should mention that for certain 1D spectral distances there has been some research about their relative discriminative properties, especially for applications in speech processing, the relation between such distances and the human auditory perception system has been studied (see e.g., [9, 25, 26, 29, 49, 54]). Perhaps one aspect that one can judge rather comfortably and independently of the specific problem is the associated computational costs of calculating the distance and other related calculations (e.g., calculating a notion of average). In that regard, for Problem 1, when the time-series dimension p is very large (e.g., in video classification problems) our introduced alignment distance (see Sect. 8.5) is cheaper to calculate relative to most other distances and also renders itself quite effective in defining a notion of average [1].

Remark 2 Throughout the paper, unless otherwise stated, by a *process* we mean a (real-valued) discrete-time wide-sense (or second order) stationary zero mean Gaussian regular stochastic process (i.e., one with no deterministic component). Some of the language used in this paper is borrowed from the statistical signal processing and control literature for which standard references include [40, 56]. Since we use the Fourier and z-transforms often and there are some disparities between the definitions (or notations) in the literature we review some terminologies and establish some notations. The z-transform of a matrix sequence $\{\mathbf{h}_t\}_{-\infty}^{+\infty}(\mathbf{h}_t \in \mathbb{R}^{p\times m})$ is defined as $H(z) = \sum_{-\infty}^{+\infty} \mathbf{h}_t z^{-t}$ for z in the complex plane $\mathbb{C}$. By evaluating $H(z)$ on the unit circle in the complex plane $\mathbb{C}$ (i.e., by setting $z = e^{i\omega}, \omega \in [0, 2\pi]$) we get $H(e^{i\omega})$, the Fourier transform of $\{\mathbf{h}_t\}_{-\infty}^{+\infty}$, which sometimes we denote by $H(\omega)$. Note that the z-transform of $\{\mathbf{h}_{-t}\}_{-\infty}^{+\infty}$ is $H(z^{-1})$ and its Fourier transform is $H(e^{-i\omega})$, and since we deal with real sequences it is the same as $\overline{H(e^{i\omega})}$, the complex conjugate of $H(e^{i\omega})$. Also any matrix sequence $\{\mathbf{h}_t\}_0^{+\infty}$ defines (causal) a linear filter via the convolution operation $\boldsymbol{y}_t = h_t * e_t = \sum_{\tau=0}^{\infty} h_\tau \boldsymbol{\epsilon}_{t-\tau}$ on the m-dimensional sequence $\boldsymbol{\epsilon}_t$. In this case, we call $H(\omega)$ or $H(z)$ the *transfer function* of the *filter* and $\{\mathbf{h}_t\}_0^{+\infty}$ the impulse response of the filter. We also say that $\boldsymbol{\epsilon}_t$ is filtered by H to generate $\boldsymbol{y}_t$. If $H(z)$ is an analytic function of z outside the unit disk in the complex plane, then the filter is called asymptotically stable. If the transfer function $H(z)$ is a *rational* matrix function of z (meaning that each entry of $H(z)$ is a rational function of z), then the filter has a *finite* order state-space (LDS) realization in the form (8.1). The smallest (*minimal*) order of such an LDS can be determined as the sum of the orders of the denominator polynomials (in z) in the entries appearing in a specific representation (factorization) of $H(z)$, known as the *Smith-McMillan* form [40]. For a square transfer function this number (known as the *McMillan degree*) is, generically, equal to the order of the denominator polynomial in the determinant of $H(z)$. The roots of these denominators are the eigenvalues of the A matrix in the minimal state-space realization of $H(z)$ and the system is asymptotically stable if all these eigenvalues are inside the unit disk in $\mathbb{C}$.

8.2.1 Geometrizing the Space of Power Spectral Densities

A p-dimensional process $\{\boldsymbol{y}_t\}$ can be *identified* with its $p \times p$ *covariance sequence* sequences $C_{\boldsymbol{y}}(\tau) = \mathbb{E}\{\boldsymbol{y}_t \boldsymbol{y}_{t-\tau}^\top\}$ $(\tau \in \mathbb{Z})$, where $^\top$ denotes matrix transpose and $\mathbb{E}\{\cdot\}$ denotes the expectation operation under the associated probability measure. Equivalently, the process can be identified by the Fourier (or z) transform of its covariance sequence, namely the *power spectral density* (PSD) $P_{\boldsymbol{y}}(\omega)$, which is a $p \times p$ Hermitian positive semi-definite matrix for every $\omega \in [0, 2\pi]$.[4] We denote the space of all $p \times p$ PSD matrices by $\mathcal{P}_p$ and its subspace consisting of elements

[4] Strictly speaking, in order to be the PSD matrix of a regular stationary process, a matrix function on $[0, 2\pi]$ must satisfy other mild technical conditions (see [62] for details).

that are full-rank for almost every $\omega \in [0, 2\pi]$ by $\mathcal{P}_p^+$. Most of the literature prior to 2000 is devoted to geometrization of $\mathcal{P}_1^+$.

Remark 3 It is worth mentioning that the distances we discuss below here are blind to correlations, meaning that two processes might be correlated but their distance can be large or they can be uncorrelated but their distance can be zero. For us the starting point is the identification of a zero-mean (Gaussian) process with its probability distribution and hence its PSD. Consider the 1D case for convenience. Then in the Hilbert space geometry a distance between processes $\boldsymbol{y}_t^1$ and $\boldsymbol{y}_t^2$ can be defined as $\mathbb{E}\{(\boldsymbol{y}_t^1 - \boldsymbol{y}_t^2)^2\}$ in which case the correlation appears in the distance and a zero distance means almost surely equal sample paths, whereas in PSD-induced distances $\boldsymbol{y}_t$ and $-\boldsymbol{y}_t$ which have completely different sample paths have zero distance. In a more technical language, the topology induced by the PSD-induced distances on stochastic processes is coarser than the Hilbert space topology. Hence, perhaps to be more accurate we should further qualify the distances in this paper by the qualifier "PSD-induced". Obviously, the Hilbert space topology may be too restrictive in some practical applications. Interestingly, in the derivation of the Hellinger distance (see below) based on the optimal transport principle the issue of correlation shows up and there optimality is achieved when the two processes are uncorrelated (hence the distance is computed as if the processes were uncorrelated, see [27, p. 292] for details). In fact, this idea is also present in our approach (and most of the other approaches), where in order to compare two LDSs we assume that they are stimulated with *the same* input process, meaning uncorrelated input processes with identical probability distributions (see Sect. 8.3).

The space $\mathcal{P}_p$ is an *infinite dimensional* cone which also has a convex *linear* structure coming from matrix addition and multiplication by nonnegative reals. The most immediate distance on this space is the standard Euclidean distance:

$$d_{\mathrm{E}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \| P_{\boldsymbol{y}^1}(\omega) - P_{\boldsymbol{y}^2}(\omega) \|^2 \mathrm{d}\omega, \tag{8.2}$$

where $\| \cdot \|$ is a matrix norm (e.g., the Frobenius norm $\| \cdot \|_F$). In the 1-dimensional case (i.e., $\mathcal{P}_1$) one could also define a distance based on the principle of *optimal decoupling* or *optimal (mass) transport* between the probability distributions of the two processes [27, p. 292]. This approach results in the formula:

$$d_{\mathrm{H}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \left| \sqrt{P_{\boldsymbol{y}^1}(\omega)} - \sqrt{P_{\boldsymbol{y}^2}(\omega)} \right|^2 \mathrm{d}\omega, \tag{8.3}$$

This distance is derived in [28] and is also called the $\bar{d}_2$-distance (see also [27, p. 292]). In view of the Hellinger distance between probability measures [9], the above distance, in the literature, is also called the Hellinger distance [23]. Interestingly, d_{H} remains valid as the optimal transport-based distance for certain non-Gaussian processes, as well [27, p. 292]. The extension of the optimal transport-based definition to higher dimensions is not straightforward. However, note that in $\mathcal{P}_1$, d_{H} can be

thought of as a square root version of d_{E}. In fact, the square root based definition can be easily extended to higher dimensions, e.g., in (8.3) one could simply replace the scalar square roots with the (matrix) Hermitian square roots of $P_{\boldsymbol{y}^i}(\omega)$, $i = 1, 2$ (at each frequency ω) and use a matrix norm. Recall that the Hermitian square root of the Hermitian matrix Y is the unique Hermitian solution of the equation $Y = XX^H$, where H denotes conjugate transpose. We denote the Hermitian square root of Y as $Y^{1/2}$. Therefore, we could define the Hellinger distance in higher dimensions as

$$d^2_{\mathrm{H}}(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \| P^{1/2}_{\boldsymbol{y}^1}(\omega) - P^{1/2}_{\boldsymbol{y}^2}(\omega)\|^2_F \mathrm{d}\omega, \tag{8.4}$$

However note that, for any unitary matrix U, $X = Y^{1/2}U$ is also a solution to $Y = XX^H$ (but not Hermitian if U differs from the intensity). This suggests that, one may be able to do better by finding the best unitary matrix $U(\omega)$ to minimize $\| P^{1/2}_{\boldsymbol{y}^1}(\omega) - P^{1/2}_{\boldsymbol{y}^2}(\omega)U(\omega)\|_F$ (at each frequency ω). In [23] this idea has been used to define the (improved) Hellinger distance on $\mathcal{P}_p$, which can be written in closed-form as

$$d^2_{\mathrm{H}'}(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \| P^{1/2}_{\boldsymbol{y}^1} - P^{1/2}_{\boldsymbol{y}^2}\big(P^{1/2}_{\boldsymbol{y}^2} P_{\boldsymbol{y}^1} P^{1/2}_{\boldsymbol{y}^2}\big)^{-1/2} P^{1/2}_{\boldsymbol{y}^2} P^{1/2}_{\boldsymbol{y}^1}\|^2_F \mathrm{d}\omega, \tag{8.5}$$

where dependence of the terms on ω has been dropped. Notice that the matrix $U(\omega) = \big(P^{1/2}_{\boldsymbol{y}^2} P_{\boldsymbol{y}^1} P^{1/2}_{\boldsymbol{y}^2}\big)^{-1/2} P^{1/2}_{\boldsymbol{y}^2} P^{1/2}_{\boldsymbol{y}^1}$ is unitary for every ω and in fact it is a transfer function of an all-pass possibly infinite dimensional linear filter [23]. Here, by an *all-pass* transfer function or filter $U(\omega)$ we mean one for which $U(\omega)U(\omega)^H = I_p$. Also note that (8.5) seemingly breaks down if either of the PSDs is not full-rank. However, solving the related optimization shows that by continuity the expression remains valid. We should point out that recently a class of distances on $\mathcal{P}_1$ has been introduced by Georgiou et al. based on the notion of optimal mass transport or morphism between PSDs (rather than probability distributions, as above) [25]. Such distances enjoy some nice properties, e.g., in terms of robustness with respect to multiplicative and additive noise [25]. An extension to $\mathcal{P}_p$ also has been proposed [53]; however, the extension is no longer a distance and it is not clear if it inherits the robustness property.

Another (possibly deeper) aspect of working with the square root of the PSD is related to the ideas of *spectral factorization* and the *innovations process*. We review some basics, which can be found, e.g., in [6, 31, 32, 38, 62, 65]. The important fact is that the PSD $P_{\boldsymbol{y}}(\omega)$ of a regular process $\boldsymbol{y}_t$ in $\mathcal{P}_p$ is of constant rank $m \leq p$ almost everywhere in $[0, 2\pi]$. Moreover, it admits a factorization of the form $P_{\boldsymbol{y}}(\omega) = P_{l\boldsymbol{y}}(\omega)P_{l\boldsymbol{y}}(\omega)^H$, where $P_{l\boldsymbol{y}}(\omega)$ is $p \times m$-dimensional and uniquely determines its *analytic* extension $P_{l\boldsymbol{y}}(z)$ outside the unit disk in $\mathbb{C}$. In this factorization, $P_{l\boldsymbol{y}}(\omega)$, itself, is not determined uniquely and any two such factors are related by an $m \times m$-dimensional all-pass filter. However, if we require the extension $P_{l\boldsymbol{y}}(z)$ to be in the class of *minimum phase* filters, then the choice of the factor $P_{l\boldsymbol{y}}(\omega)$ becomes unique up to a constant unitary matrix. A $p \times m$

($m \leq p$) transfer function matrix $H(z)$ is called minimum phase if it is analytic outside the unit disk and of constant rank m there (including at $z = \infty$). Such a filter has an inverse filter, which is asymptotically stable. We denote this particular factor of $P_{\boldsymbol{y}}$ by $P_{+\boldsymbol{y}}$ and call it the *canonical spectral* factor. The canonical factor is still not unique, but the ambiguity is only in a constant $m \times m$ unitary matrix. The consequence is that $\boldsymbol{y}_t$ can be written as $\boldsymbol{y}_t = \sum_{\tau=0}^{\infty} \mathbf{p}_{+\tau} \boldsymbol{\epsilon}_{t-\tau}$, where the $p \times m$ matrix sequence $\{\mathbf{p}_{+t}\}_{t=0}^{\infty}$ is the inverse Fourier transform of $P_{+\boldsymbol{y}}(\omega)$ and $\boldsymbol{\epsilon}_t$ is an m-dimensional *white* noise process with covariance equal to the identity matrix I_m. This means that $\boldsymbol{y}_t$ is the output of a linear filter (i.e., an LDS of possibly *infinite* order) excited by a white noise process with standard covariance. The process $\boldsymbol{\epsilon}_t$ is called the *innovations process* or *fundamental process* of $\boldsymbol{y}_t$. Under the Gaussian assumption the innovation process is determined uniquely, otherwise it is determined up to an $m \times m$ unitary factor. The important case is when $P_{\boldsymbol{y}}(z)$ is full-rank outside the unit disk, in which case the inverse filter $P_{+\boldsymbol{y}}^{-1}$ is well-defined and asymptotically stable, and one could recover the innovations process by filtering $\boldsymbol{y}_t$ by its *whitening* filter $P_{+\boldsymbol{y}}^{-1}$.

Now, to compare two processes, one could somehow compare their canonical spectral factors[5] or if they are in $\mathcal{P}_p^+$ their whitening filters. In [38] a large class of divergences based on the idea of comparing associated whitening filters (in the frequency domain) have been proposed. For example, let $P_{+\boldsymbol{y}_i}$ be the canonical factor of $P_{\boldsymbol{y}_i}, i = 1, 2$. If one filters $\boldsymbol{y}_t^i, i = 1, 2$, with $P_{+\boldsymbol{y}^j}^{-1}, j = 1, 2$, then the output PSD is $P_{+\boldsymbol{y}^j}^{-1} P_{\boldsymbol{y}^i} P_{+\boldsymbol{y}^j}^{-H}$. Note that when $i = j$ then the output PSD is I_p across every frequency. It can be shown that $d_I(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \mathrm{tr}(P_{+\boldsymbol{y}^1}^{-1} P_{\boldsymbol{y}^2} P_{+\boldsymbol{y}^1}^{-H} - I_p) + \mathrm{tr}(P_{+\boldsymbol{y}^2}^{-1} P_{\boldsymbol{y}^1} P_{+\boldsymbol{y}^2}^{-H} - I_p) \mathrm{d}\omega$ is a symmetric divergence [38]. Note that $d_I(\boldsymbol{y}^1, \boldsymbol{y}^2)$ is independent of the unitary ambiguity in the canonical factor and in fact

$$d_I(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \mathrm{tr}(P_{\boldsymbol{y}^1}^{-1} P_{\boldsymbol{y}^2} + P_{\boldsymbol{y}^2}^{-1} P_{\boldsymbol{y}^1} - 2I_p) \mathrm{d}\omega. \tag{8.6}$$

Such divergences enjoy certain invariance properties, e.g., if we filter both processes with a common minimum phase filter, then the divergence remains unchanged. In particular, it is scale-invariant. Such properties are shared by the distances or divergences that are based on the ratios of PSDs (see below for more examples). Scale invariance in the case of 1D PSDs has been advocated as a desirable property, since in many cases the shape of the PSDs rather than their relative scale is the discriminative feature (see e.g., [9, 26]).

One can arrive at similar distances from other geometric or probabilistic paths. One example is the famous Itakura-Saito divergence (sometimes called distance)

[5] In fact, our approach (in Sects. 8.3–8.5) is also based on the idea of comparing the minimum phase (i.e., canonical) filters or factors in the case of processes with rational spectra. However, instead of comparing the associated transfer functions or impulse responses, we try to compare the associated state-space realizations (in a specific sense). This approach, therefore, is in some sense *structural* or *generative*, since it tries to compare how the processes are generated (according to the state-space representation) and the model order plays an explicit role in it.

between PSDs in $\mathcal{P}_1^+$ which is defined as

$$d_{\text{IS}}(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \left(\frac{P_{\boldsymbol{y}^1}}{P_{\boldsymbol{y}^2}} - \log \frac{P_{\boldsymbol{y}^1}}{P_{\boldsymbol{y}^2}} - 1 \right) \mathrm{d}\omega. \tag{8.7}$$

This divergence has been used in practice, at least, since the 1970s (see [48] for references). The Itakura-Saito divergence can be derived from the Kullback-Leibler divergence between (infinite dimensional) probability densities of the two processes (The definition is a time-domain based definition, however, the final result is readily expressible in the frequency domain).[6] On the other hand, Amari's information geometry-based approach [5, Chap. 5] allows to geometrize $\mathcal{P}_1^+$ in various ways and yields different distances including the Itakura-Saito distance (8.7) or a Riemannian distance such as

$$d_{\text{R}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \left(\log \left(\frac{P_{\boldsymbol{y}^1}}{P_{\boldsymbol{y}^2}} \right) \right)^2 \mathrm{d}\omega. \tag{8.8}$$

Furthermore, in this framework one can define geodesics between two processes under various Riemannian or non-Riemannian *connections*. The high-dimensional version of the Itakura-Saito distance has also been known since the 1980s [42] but is less used in practice:

$$d_{\text{IS}}(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \left(\text{trace}(P_{\boldsymbol{y}^2}^{-1} P_{\boldsymbol{y}^1}) - \log(\det(P_{\boldsymbol{y}^2}^{-1} P_{\boldsymbol{y}^1})) - p \right) \mathrm{d}\omega. \tag{8.9}$$

Recently, in [38] a Riemannian framework for geometrization of $\mathcal{P}_p^+$ for $p \geq 1$ has been proposed, which yields Riemannian distances such as:

$$d_{\text{R}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \| \log \left(P_{\boldsymbol{y}^1}^{-1/2} P_{\boldsymbol{y}^2} P_{\boldsymbol{y}^1}^{-1/2} \right) \|_F^2 \mathrm{d}\omega, \tag{8.10}$$

where log is the standard matrix logarithm. In general, such approaches are not suited for large p due to computational costs and the full-rankness requirement. We should stress that in (very) high dimensions the assumption of full-rankness of PSDs is not a viable one, in particular because usually not only the actual time series are highly correlated but also the contaminating noises are correlated, as well. In fact, this has lead to the search for models capturing this quality. One example is the class of *generalized linear dynamic factor models*, which are closely related to the tall, full rank LDS models (see Sect. 8.3.3 and [20, 24]).

Letting the above mentioned issues aside, for the purposes of Problem 1, the space $\mathcal{P}_p$ (or even $\mathcal{P}_p^+$) is *too large*. The reason is that it includes, e.g., ARMA processes of arbitrary large orders, and it is not clear, e.g., how an *average* of some ARMA models

[6] Notice that defining distances between probability densities in the time domain is a more general approach than the PSD-based approaches, and it can be employed in the case of nonstationary as well as non-Gaussian processes. However, such an approach, in general, is computationally difficult.

or processes of equal order might turn out. As mentioned before, it is convenient or reasonable to require the average to be of the same order.[7]

8.2.2 Geometrizing the Spaces of Models

Any distance on $\mathcal{P}_p$ (or $\mathcal{P}_p^+$) induces a distance, e.g., on a subspace corresponding to AR or ARMA models of a fixed order. This is an example of an extrinsic distance induced from an *infinite dimensional ambient* space to a *finite dimensional subspace*. In general, this framework is not ideal and we might try to, e.g., define an *intrinsic* distance on the finite dimensional subspace. In fact, Amari's original paper [4] lays down a framework for this approach, but lacks actual computations. For the one-dimensional case in [61], based on Amari's approach, distances between models in the space of ARMA models of fixed order are derived. For high order models or in high dimensions, such calculations are, in general, computationally difficult [61]. The main reason is that the dependence of PSD-based distances on state-space or ARMA parameters is, in general, highly nonlinear (the important exception is for parameters of AR models, especially in 1D).

Alternative approaches have also been pursued. For example, in [57] the main idea is to compare (based on the ℓ^2 norm) the coefficients of the infinite order AR models of two processes. This is essentially the same as comparing (in the time domain) the whitening filters of the two processes. This approach is limited to $\mathcal{P}_p^+$ and computationally demanding for large p. See [19] for examples of classification and clustering of 1D time-series using this approach. In [8], the space of 1D AR processes of a fixed order is geometrized using the geometry of positive-definite Toeplitz matrices (via the reflection coefficients parameterization), and, moreover, L^p averaging on that space is studied. In [50] a (pseudo)-distance between two processes is defined through a weighted ℓ^2 distance between the (infinite) sequences of the *cepstrum* coefficients of the two processes. Recall that the cepstrum of a 1D signal is the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of the signal. In the frequency domain this distance (known as the Martin distance) can be written as (up to a multiplicative constant)

$$d_{\mathrm{M}}^2(\boldsymbol{y}_1, \boldsymbol{y}_2) = \int \left(\mathfrak{D}^{\frac{1}{2}} \log \left(\frac{P_{\boldsymbol{y}_1}}{P_{\boldsymbol{y}_2}} \right) \right)^2 \mathrm{d}\omega, \tag{8.11}$$

where $\mathfrak{D}^\lambda$ is the fractional derivative operator in the frequency domain interpreted as multiplication of the corresponding Fourier coefficients in the time domain by $e^{\pi i \lambda/2} n^\lambda$ for $n \geq 0$ and by $e^{-\pi i \lambda/2}(-n)^\lambda$ for $n < 0$. Notice that d_M is scale-invariant in the sense described earlier and also it is a pseudo-distance since it is zero if the PSDs are multiple of each other (this is a true scale-invariance property, which in

[7] Interestingly, for an average defined based on the Itakura-Saito divergence in the space of 1D AR models this property holds [26], see also [5, Sect. 5.3].

certain applications is highly desirable).[8] Interestingly, in the case of 1D ARMA models, d_M can be expressed conveniently in closed form in terms of the poles and zeros of the models [50]. Moreover, in [18] it is shown that d_M can be calculated quite efficiently in terms of the parameters of the state-space representation of the ARMA processes. In fact, the Martin distance has a simple interpretation in terms of the subspace angles between the extended observability matrices (cf. Sect. 8.4.3) of the state-space representations [18]. This brings about important computational advantages and has allowed to extend a form of Martin distance to higher dimensions (see e.g., [16]). However, it should be noted that the extension of the Martin distance to higher dimensions in such a way that all its desirable properties carry over has proven to be difficult [13].[9] Nevertheless, some extensions have been quite effective in certain high-dimensional applications, e.g., video classification [16]. In [16], the approach of [18] is shown to be a special case of the family of Binet-Cauchy kernels introduced in [64], and this might explain the effectiveness of the extensions of the Martin distance to higher dimensions.

In summary, we should say that the extensions of the geometrical methods discussed in this section to $\mathcal{P}_p$ for $p > 1$ do not seem obvious or otherwise they are computationally very expensive. Moreover, these approaches often yield extrinsic distances induced from infinite dimensional ambient spaces, which, e.g., in the case of averaging LDSs of *fixed* order can be problematic.

8.2.3 Control-Theoretic Approaches

More relevant to us are [33, 46], where (*intrinsic*) state-space based Riemannian distances between LDSs of fixed size and fixed order have been studied. Such approaches ideally suit Problem 1, but they are computationally demanding. More recently, in [1] and subsequently in [2, 3], we introduced group action induced distances on certain spaces of LDSs of *fixed size* and *order*. As it will become clear in the next section, an important feature of this approach is that the LDS order is *explicit* in the construction of the distance, and the state-space parameters appear in the distance in a simple form. These features make certain related calculations (e.g., optimization) much more convenient (compared with other methods). Another aspect of our approach is that, contrary to most of the distances discussed so far, which compare the PSDs or the canonical factors directly, our approach amounts to comparing the

[8] It is interesting to note that by a simple modification some of the spectral-ratio based distances can attain this property, e.g., by modifying d_{R} in (8.8) as $d^2_{\mathrm{RI}}(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \left(\log\left(\frac{P_{y^1}}{P_{y^2}}\right)\right)^2 \mathrm{d}\omega - \left(\int \log\left(\frac{P_{y^1}}{P_{y^2}}\right)\mathrm{d}\omega\right)^2$ (see also [9, 25, 49]).

[9] This and the results in [53] underline the fact that defining distances on $\mathcal{P}_p$ for $p > 1$ may be challenging, not only from a computational point of view but also from a theoretical one. In particular, certain nice properties in 1D do not automatically carry over to higher dimensions by a simple extension of the definitions in 1D.

generative or the structural models of the processes or how they are generated. This feature also could be useful in designing more application-specific or structure-aware distances.

8.3 Processes Generated by LDSs of Fixed Order

Consider an LDS, M, of the form (8.1) with *a* realization $R = (A, B, C, D) \in \widetilde{\mathcal{SL}}_{m,n,p}$.[10] In the sequel, for various reasons, we will restrict ourselves to increasingly smaller submanifolds of $\widetilde{\mathcal{SL}}_{m,n,p}$, which will be denoted by additional superscripts. Recall that the $p \times m$ matrix transfer function is $T(z) = D + C(I_n - z^{-1}A)^{-1}B$, where $z \in \mathbb{C}$ and I_n is the n-dimensional identity matrix. *We assume that all LDSs are excited by the standard white Gaussian process.* Hence, the output PSD matrix (in the z-domain) is the $p \times p$ matrix function $P(z) = T(z)T^\top(z^{-1})$. The PSD is a rational matrix function of z whose rank (a.k.a. *normal rank*) is constant almost everywhere in $\mathbb{C}$. Stationarity of the output process is guaranteed if M is asymptotically stable. We denote the submanifold of such realizations by $\widetilde{\mathcal{SL}}^{\mathrm{a}}_{m,n,p} \subset \widetilde{\mathcal{SL}}_{m,n,p}$.

8.3.1 Embedding Stochastic Processes in LDS Spaces

Two (stochastic) LDSs are indistinguishable if their output PSDs are equal. Using this equivalence on the entire set of LDSs is not useful, because, as mentioned earlier two transfer functions which differ by an all-pass filter result in the same PSD. Therefore, the equivalence relation could induce a complicated many-to-one correspondence between the LDSs and the subspace of stochastic processes they generate. However, if we restrict ourselves to the subspace of minimum phase LDSs the situation improves. Let us denote the subspace of minimum-phase realizations by $\widetilde{\mathcal{SL}}^{\mathrm{a,mp}}_{m,n,p} \subset \widetilde{\mathcal{SL}}^{\mathrm{a}}_{m,n,p}$. This is clearly an open submanifold of $\widetilde{\mathcal{SL}}^{\mathrm{a}}_{m,n,p}$. In $\widetilde{\mathcal{SL}}^{\mathrm{a,mp}}_{m,n,p}$, the canonical spectral factorization of the output PSD is unique up to an orthogonal matrix [6, 62, 65]: let $T_1(z)$ and $T_2(z)$ have realizations in $\widetilde{\mathcal{SL}}^{\mathrm{a,mp}}_{m,n,p}$ and let $T_1(z)T_1^\top(z^{-1}) = T_2(z)T_2^\top(z^{-1})$, then $T_1(z) = T_2(z)\Theta$ for a unique $\Theta \in O(m)$, where $O(m)$ is the Lie group of $m \times m$ orthogonal matrices. Therefore, any p-dimensional processes with PSD of normal rank m can be identified with a simple equivalent class of stable and minimum-phase transfer functions and the corresponding LDSs.[11]

[10] It is crucial to have in mind that we explicitly distinguish between the LDS, M, and its realization R, which is not unique. As it will become clear soon, an LDS has an equivalent class of realizations.

[11] These rank conditions, interestingly, have differential geometric significance in yielding nice quotient spaces, see Sect. 8.4.

8.3.2 Equivalent Realizations Under Internal and External Symmetries

A fundamental fact is that there are *symmetries* or *invariances* due to certain Lie *group actions* in the model (8.1). Let $GL(n)$ denote the Lie group of $n \times n$ non-singular (real) matrices. We say that the Lie group $GL(n) \times O(m)$ *acts* on the realization space $\widetilde{\mathcal{SL}}_{m,n,p}$ (or its subspaces) via the action $\bullet$ defined as[12]

$$(P, \Theta) \bullet (A, B, C, D) = (P^{-1}AP, P^{-1}B\Theta, CP, D\Theta). \tag{8.12}$$

One can easily verify that under this action the output covariance sequence (or PSD) remains invariant. In general, the *converse* is not true. That is, two output covariance sequences might be equal while their corresponding realizations are not related via $\bullet$ (due to non-minimum phase and the action not being *free* [47], also see below). Recall that the action of a group on a set is called free if every element of the set is fixed only by the identity element of the group. For the converse to hold we need to impose further *rank* conditions, as we will see next.

8.3.3 From Processes to Realizations (The Rank Conditions)

Now, we study some rank conditions (i.e., submanifolds of $\widetilde{\mathcal{SL}}_{m,n,p}$ on) under which $\bullet$ is a free action.

8.3.3.1 Observable, Controllable, and Minimal Realizations

Recall that the *controllability* and *observability* matrices of order k associated with a realization $R = (A, B, C, D)$ are defined as $\mathcal{C}_k = [B, AB, \ldots, A^{k-1}B]$ and $\mathcal{O}_k = [C^\top, (CA)^\top, \ldots, (CA^{k-1})^\top]^\top$, respectively. A realization is called *controllable* (resp. *observable*) if $\mathcal{C}_k$ (resp. $\mathcal{O}_k$) is of rank n for $k = n$. We denote the subspace of controllable (resp. observable) realizations by $\widetilde{\mathcal{SL}}^{\text{co}}_{m,n,p}$ (resp. $\widetilde{\mathcal{SL}}^{\text{ob}}_{m,n,p}$). The space $\widetilde{\mathcal{SL}}^{\text{min}}_{m,n,p} = \widetilde{\mathcal{SL}}^{\text{co}}_{m,n,p} \cap \widetilde{\mathcal{SL}}^{\text{ob}}_{m,n,p}$ is called the space of *minimal* realizations. An important fact is that we cannot reduce the order (i.e., the size of A) of a minimal realization without changing its input-output behavior.

8.3.3.2 Tall, Full Rank LDSs

Another (less studied) rank condition is when C is of rank n (here $p \geq n$ is required). Denote by $\widetilde{\mathcal{SL}}^{\text{tC}}_{m,n,p} \subset \widetilde{\mathcal{SL}}^{\text{ob}}_{m,n,p}$ the subspace of such realizations and call a corresponding LDS *tall and full-rank*. Such LDSs are closely related to generalized

[12] Strictly speaking $\bullet$ is a right action; however, it is notationally convenient to write it as a left action in (8.12).

linear dynamic factor models for (very) high-dimensional time series [20] and also appear in video sequence modeling [1, 12, 60]. It is easy to verify that all the above realization spaces are smooth open submanifolds of $\widetilde{\mathcal{SL}}_{m,n,p}$. Their corresponding submanifolds of stable or minimum-phase LDSs (e.g., $\widetilde{\mathcal{SL}}^{\mathrm{a,mp,co}}_{m,n,p}$) are defined in an obvious way.

The following proposition forms the basis of our approach to defining distances between processes: any distance on the space of LDSs with realizations in the above submanifolds (with rank conditions) can be used to define a distance on the space of processes generated by those LDSs.

Proposition 1 *Let* $\tilde{\Sigma}_{m,n,p}$ *be* $\widetilde{\mathcal{SL}}^{a,mp,co}_{m,n,p}$, $\widetilde{\mathcal{SL}}^{a,mp,ob}_{m,n,p}$, $\widetilde{\mathcal{SL}}^{a,mp,min}_{m,n,p}$, *or* $\widetilde{\mathcal{SL}}^{a,mp,tC}_{m,n,p}$. *Consider two realizations* $R_1, R_2 \in \tilde{\Sigma}_{m,n,p}$ *excited by the standard white Gaussian process. Then we have:*

1. *If* $(P, \Theta) \bullet R_1 = R_2$ *for some* $(P, \Theta) \in GL(n) \times O(m)$, *then the two realizations generate the same (stationary) output process (i.e., outputs have the same PSD matrices).*
2. *Conversely, if the outputs of the two realizations are equal (i.e., they have the same PSD matrices), then there exists a unique* $(P, \Theta) \in GL(n) \times O(m)$ *such that* $(P, \Theta) \bullet R_1 = R_2$.

8.4 Principal Fiber Bundle Structures over Spaces of LDSs

As explained above, an LDS, M, has an equivalent class of realizations related by the action $\bullet$. Hence, M sits naturally in a *quotient* space, namely $\widetilde{\mathcal{SL}}_{m,n,p}/(GL(n) \times O(m))$. However, this quotient space is not smooth or even Hausdorff. Recall that if a Lie group G acts on a manifold *smoothly*, *properly*, and *freely*, then the quotient space has the structure of a *smooth manifold* [47]. Smoothness of $\bullet$ is obvious. In general, the action of a *non-compact* group such as $GL(n) \times O(m)$ is *not* proper. However, one can verify that the rank conditions we imposed in Proposition 1 are enough to make $\bullet$ both a proper and free action on the realization submanifolds (see [2] for a proof). The resulting quotient manifolds are denoted by dropping the superscript $\tilde{}$, e.g., $\mathcal{SL}^{\mathrm{a,mp,min}}_{m,n,p}$. The next theorem, which is an extension of existing results, e.g., in [33] shows that, in fact, we have a principal fiber bundle structure.

Theorem 1 *Let* $\tilde{\Sigma}_{m,n,p}$ *be as in Proposition 1 and* $\Sigma_{m,n,p} = \tilde{\Sigma}_{m,n,p}/(GL(n) \times O(m))$ *be the corresponding quotient LDS space. The realization-system pair* $(\tilde{\Sigma}_{m,n,p}, \Sigma_{m,n,p})$ *has the structure of a smooth principal fiber bundle with structure group* $GL(n) \times O(m)$. *In the case of* $\mathcal{SL}^{a,mp,tC}_{m,n,p}$ *the bundle is trivial (i.e., diffeomorphic to a product), otherwise it is trivial only when* $m = 1$ *or* $n = 1$.

The last part of the theorem has an important consequence. Recall that a principal bundle is trivial if it diffeomorphic to global product of its base space and its structure

group. Equivalently, this means that a trivial bundle admits a global smooth cross section or what is known as a smooth canonical form in the case of LDSs, i.e., a globally smooth mapping $s\colon \Sigma_{m,n,p} \to \widetilde{\Sigma}_{m,n,p}$ which assigns to every system a unique realization. This theorem implies that the minimality condition is a complicated nonlinear constraint, in the sense that it makes the bundle twisted and nontrivial for which no continuous canonical form exists. Establishing this obstruction put an end to control theorists' search for canonical forms for MIMO LDSs in the 1970s and explained why system identification for MIMO LDSs is a challenging task [11, 15, 36].

On the other hand, one can verify that $(\widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}, \mathcal{SL}^{\text{a,mp,tC}}_{m,n,p})$ is a trivial bundle. Therefore, for such systems global canonical forms exist and they can be used to define distances, i.e., if $s : \mathcal{SL}^{\text{a,mp,tC}}_{m,n,p} \to \widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}$ is such a canonical form then $d_{\mathcal{SL}^{\text{a,mp,tC}}_{m,n,p}}(M_1, M_2) = \tilde{d}_{\widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}}(s(M_1), s(M_2))$ defines a distance on $\mathcal{SL}^{\text{a,mp,tC}}_{m,n,p}$ for any distance $\tilde{d}_{\widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}}$ on the realization space. In general, unless one has some specific knowledge there is no preferred choice for a section or canonical form. If one has a *group-invariant* distance on the realization space, then the distance induced from using a cross section might be inferior to the *group action induced distance*, in the sense it may result in an artificially larger distance. In the next section we review the basic idea behind group action induced distances in our application.

8.4.1 Group Action Induced Distances

Figure 8.1a schematically shows a realization bundle $\widetilde{\Sigma}$ and its base LDS space Σ. Systems $M_1, M_2 \in \Sigma$ have realizations R_1 and R_2 in $\widetilde{\Sigma}$, respectively. Let us assume that a $G = GL(n) \times O(n)$-invariant distance $\tilde{d}_G$ on the realization bundle is given. The realizations, R_1 and R_2, in general, are not aligned with each other, i.e., $\tilde{d}_G(R_1, R_2)$ can be still reduced by sliding one realization along its fiber as depicted in Fig. 8.1b. This leads to the definition of the group action induced distance:[13]

$$d_\Sigma(M_1, M_2) = \inf_{(P,\Theta)\in G} \tilde{d}_{\tilde{\Sigma}}((P, \Theta) \bullet R_1, R_2). \tag{8.13}$$

In fact, one can show that $d_\Sigma(\cdot, \cdot)$ is a true distance on Σ, i.e., it is symmetric and positive definite and obeys the triangle inequality (see e.g., [66]).[14]

The main challenge in the above approach is the fact that, due to non-compactness of $GL(n)$, constructing a $GL(n) \times O(n)$-invariant distance is computationally dif-

[13] We may call this an alignment distance. However, based on the same principle in Sect. 8.5 we define another group action induced distance, which we explicitly call the alignment distance. Since our main object of interest is that distance, we prefer not to call the distance in (8.13) an alignment distance.

[14] It is interesting to note that some of the good properties of the k-nearest neighborhood algorithms on a general metric space depend on the triangle inequality [21].

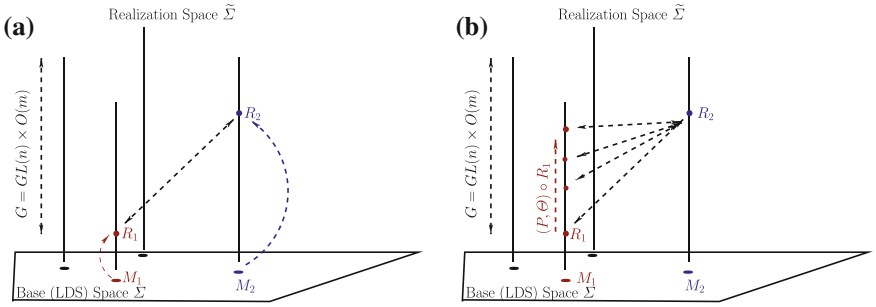


Fig. 8.1 Over each LDS in Σ sits a realization fiber. The fibers together form the realization space (bundle) $\tilde{\Sigma}$. If given a G-invariant distance on the realization bundle, then one can define a distance on the LDS space by aligning any realizations R_1, R_2 of the two LDSs M_1, M_2 as in (8.13)

ficult. The construction of such a distance can essentially be accomplished by defining a $GL(n) \times O(n)$-invariant Riemannian metric on the realization space and solving the corresponding geodesic equation, as well as searching for global minimizers.[15] Such a Riemannian metric for deterministic LDSs was proposed in [45, 46]. One could also start from (an already invariant) distance on a large ambient space such as $\mathcal{P}_p$ and specialize it to the desired submanifold Σ of LDSs to get a Riemannian manifold on Σ and then thereon solve geodesic equations, etc. to get an *intrinsic* distance (e.g., as reported in [33, 34]). Both of these approaches seem very complicated to implement for the case of very high-dimensional LDSs. Instead, our approach is to use *extrinsic* group action induced distances, which are induced from unitary-invariant distances on the realization space. For that we recall the notion of reduction of structure group on a principal fiber bundle.

8.4.2 Standardization: Reduction of the Structure Group

Next, we recall the notion of reducing a bundle with non-compact structure group to one with a compact structure group. This will be useful in our geometrization approach in the next section. Interestingly, bundle reduction also appears in *statistical analysis of shapes* under the name of *standardization* [43]. The basic fact is that any principal fiber G-bundle $(\tilde{\Sigma}, \Sigma)$ can be *reduced* to an OG-subbundle $\widetilde{\mathcal{O}\Sigma} \subset \tilde{\Sigma}$, where OG is the maximal compact subgroup of G [44]. This reduction means that Σ is *diffeomorphic* to $\widetilde{\mathcal{O}\Sigma}/OG$ (i.e., no *topological* information is lost by going to the subbundle and the subgroup). Therefore, in our cases of interest we can reduce a $GL(n) \times O(m)$-bundle to an $OG(n, m) = O(n) \times O(m)$-subbundle. We call

[15] This problem, in general, is difficult, among other things, because it is a non-convex (infinite-dimensional) variational problem. Recall that in Riemannian geometry the non-convexity of the arc length variational problem can be related to the non-trivial topology of the manifold (see e.g., [17]).

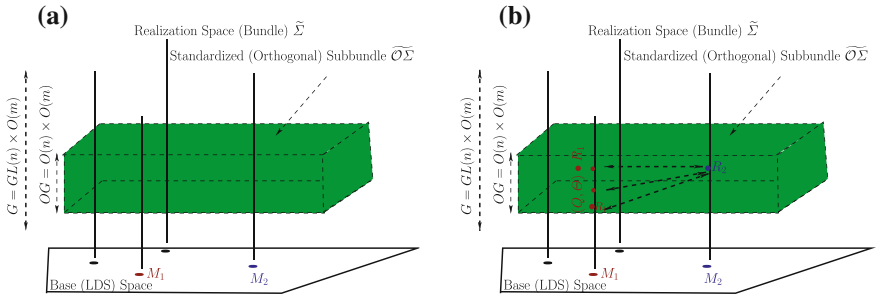


Fig. 8.2 A standardized subbundle $\widetilde{\mathcal{O}\Sigma}_{m,n,p}$ of $\widetilde{\Sigma}_{m,n,p}$ is a subbundle on which G acts via its *compact* subgroup OG. The quotient space $\widetilde{\mathcal{O}\Sigma}_{m,n,p}/OG$ still is diffeomorphic to the base space $\widetilde{\Sigma}_{m,n,p}$. One can define an alignment distance on the base space by aligning realizations $R_1, R_2 \in \widetilde{\mathcal{O}\Sigma}_{m,n,p}$ of $M_1, M_2 \in \Sigma_{m,n,p}$ as (8.15)

such a subbundle a *standardized* realization space or (sub)bundle. One can perform reduction to various standardized subbundles and there is *no* canonical reduction. However, in each application one can choose an *interesting* one. A reduction is in spirit similar to the Gram-Schmidt orthonormalization [44, Chap. 1]. Figure 8.2a shows a standardized subbundle $\widetilde{\mathcal{O}\Sigma}$ in the realization bundle $\widetilde{\Sigma}$.

8.4.3 Examples of Realization Standardization

As an example consider $R = (A, B, C, D) \in \widetilde{\mathcal{SL}}^{\mathrm{a,mp,tC}}_{m,n,p}$, and let $C = UP$ be an orthonormalization of C, where $U^\top U = I_n$ and $P \in GL(n)$. Now the new realization $\hat{R} = (P^{-1}, I_m) \bullet R$ belongs to the $O(n)$-subbundle $\widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC}}_{m,n,p} = \{R \in \widetilde{\mathcal{SL}}^{\mathrm{a,mp,tC}}_{m,n,p} | C^\top C = I_n\}$.

Other forms of bundle reduction, e.g., in the case of the nontrivial bundle $\widetilde{\mathcal{SL}}^{\mathrm{a,mp,min}}_{m,n,p}$ are possible. In particular, via a process known as *realization balancing* (see [2, 37]), we can construct a large family of standardized subbundles. For example, a more sophisticated one is in the case of $\widetilde{\mathcal{SL}}^{\mathrm{a,mp,min}}_{m,n,p}$ via the notion of (internal) *balancing*. Consider the symmetric $n \times n$ matrices $W_c = \mathcal{C}_\infty \mathcal{C}_\infty^\top$ and $W_o = \mathcal{O}_\infty^\top \mathcal{O}_\infty$, which are called controllability and observability Gramians, respectively, and where $\mathcal{C}_\infty$ and $\mathcal{O}_\infty$ are called extended controllability and observability matrices, respectively (see the definitions in Sect. 8.3.3.1 with $k = \infty$). Due to the minimality assumption, both W_o and W_c are positive definite. Notice that under the action $\bullet$, W_c transforms to $P^{-1} W_c P^{-\top}$ and W_o to $P^\top W_o P$. Consider the function $h : GL(n) \to \mathbb{R}$ defined as $h(P) = \mathrm{trace}(P^{-1} W_c P^{-\top} + P^\top W_o P)$. It is easy to see that h is constant on $O(n)$. More importantly, it can be shown that any critical point P_1 of h is global minimizer and if P_2 is any other minimizer then $P_1 = P_2 Q$

for some $Q \in O(n)$ [37]. Minimizing h is called balancing (in the sense of Helmke [37]). One can show that balancing is, in fact, a standardization in the sense that we defined (a proof of this fact will appear elsewhere). Note that a more specific form of balancing called diagonal balancing (due to Moore [52]) is more common in the control literature, however, that cannot be considered as a form of reduction of the structure group. The interesting intuitive reason is that it tries to reduce the structure group beyond the orthogonal group to the identity element, i.e., to get a canonical form (see also [55]). However, it fails in the sense that, as mentioned above, it cannot give a *smooth* canonical form, i.e., a section which is diffeomorphic to $\mathcal{SL}^{\mathrm{a,mp,min}}_{m,n,p}$.

8.5 Extrinsic Quotient Geometry and the Alignment Distance

In this section, we propose to use the large class of *extrinsic* unitary invariant distances on *a* standardized realization subbundle to build distances on *the* LDS base space. The main benefits are that such distances are abundant, the ambient space is *not* too large (e.g., not infinite dimensional), and calculating the distance in the base space boils down to a static optimization problem (albeit non-convex). Specifically, let $\tilde{d}_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}$ be a unitary invariant distance on a standardized realization subbundle $\widetilde{\mathcal{O}\Sigma}_{m,n,p}$ with the base $\Sigma_{m,n,p}$ (as in Theorem 1). One example of such a distance is

$$\tilde{d}^2_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}(R_1,R_2)=\lambda_A\|A_1-A_2\|_F^2+\lambda_B\|B_1-B_2\|_F^2+\lambda_C\|C_1-C_2\|_F^2+\lambda_D\|D_1-D_2\|_F^2, \tag{8.14}$$

where $\lambda_A, \lambda_B, \lambda_C, \lambda_D > 0$ are constants and $\|\cdot\|_F$ is the matrix Frobenius norm. A group action induced distance (called the *alignment* distance) between two LDSs $M_1, M_2 \in \Sigma_{m,n,p}$ with realizations $R_1, R_2 \in \widetilde{\mathcal{O}\Sigma}_{m,n,p}$ is found by solving the *realization alignment* problem (see Fig. 8.2b)

$$d^2_{\Sigma_{m,n,p}}(M_1, M_2) = \min_{(Q,\Theta)\in O(n)\times O(m)} \tilde{d}^2_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}\big((Q,\Theta)\bullet R_1, R_2\big). \tag{8.15}$$

In [39] a fast algorithm is developed which (with little modification) can be used to compute this distance.

Remark 4 We stress that, via the identification of a process with its canonical spectral factors (Proposition 1 and Theorem 1), $d_{\Sigma_{m,n,p}}(\cdot,\cdot)$ is (or induces) a distance on the space of processes generated by the LDSs in $\Sigma_{m,n,p}$. Therefore, in the sprit of distances studied in Sect. 8.2 we could have written $d_{\Sigma_{m,n,p}}(\boldsymbol{y}_1, \boldsymbol{y}_2)$ instead of $d_{\Sigma_{m,n,p}}(M_1, M_2)$, where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the processes generated by M_1 and M_2 when excited by the standard Gaussian process. However, the chosen notation seems more convenient.

Remark 5 Calling the static *global* minimization problem (8.15) "easy" in an absolute term is an oversimplification. However, even this *global* minimization

over orthogonal matrices is definitely simpler than solving the nonlinear geodesic ODEs and finding shortest geodesics globally (an infinite-dimensional dynamic programming problem). It is our ongoing research to develop fast and reliable algorithms to solve (8.15). Our experiments indicate that the Jacobi algorithm in [39] is quite effective in finding global minimizers.

In [1], this distance was first introduced on $\mathcal{SL}^{\mathrm{a,mp,tC}}_{m,n,p}$ with the standardized subbundle $\widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC}}_{m,n,p}$. The distance was used for efficient video sequence classification (using 1-nearest neighborhood and nearest mean methods) and clustering (e.g., via defining *averages* or a k-means like algorithm). However, it should be mentioned that in video applications (for reasons which are not completely understood) the comparison of LDSs based on the (A, C) part in (8.1) has proven quite effective (in fact, such distances are more commonly used than distances based on comparing the full model). Therefore, in [1], the alignment distance (8.15) with parameters $\lambda_B = \lambda_D = 0$ was used, see (8.14). An algorithm called the *align and average* is developed to do averaging on $\mathcal{SL}^{\mathrm{a,mp,tC}}_{m,n,p}$ (see also [2]). One defines the average $\bar{M}$ of LDSs $\{M_i\}_{i=1}^N \subset \mathcal{SL}^{\mathrm{a,mp,tC}}_{m,n,p}$ (the so-called Fréchet mean or average) as a minimizer of the sum of the squares of distances:

$$\bar{M} = \mathrm{argmin}_M \sum_{i=1}^{N} d^2_{\mathcal{SL}^{\mathrm{a,mp,tC}}_{m,n,p}}(M, M_i). \tag{8.16}$$

The align and average algorithm is essentially an alternating minimization algorithm to find a solution. As a result, in each step it *aligns* the realizations of the LDSs M_i to that of the current estimated average, then a Euclidean *average* of the aligned realizations is found and afterwards the found C matrix is orthonormalized, and the algorithm iterates these steps till convergence (see [1, 2] for more details). A nice feature of this algorithms is that (generically) the average LDS $\bar{M}$ by construction will be of order n and minimum phase (and under certain conditions stable). An interesting question is whether the average model found this way is asymptotically stable, by construction. The most likely answer is, in general, negative. However, in a special case it can be positive. Let $\|A\|_2$ denote the 2-norm (i.e., the largest singular value) of the matrix A. In the case the standardized realizations $R_i \in \widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC}}_{m,n,p}$, $(1 \leq i \leq N)$ are such that $\|A_i\|_2 < 1 (1 \leq i \leq N)$, then by construction the 2-norm of the A matrix of the average LDS will also be less than 1. Hence, the average LDS will be asymptotically stable. Moreover, as mentioned in Sect. 8.4.3, in the case of $\mathcal{SL}^{\mathrm{a,mp,min}}_{m,n,p}$ we may employ the subbundle of balanced realizations as the standardized subbundle. It turns out that in this case preserving stability (by construction) can be easier, but the averaging algorithm gets more involved (see [2] for some more details).

Obviously, the above alignment distance based on (8.14) is only an example. In a pattern recognition application, a large class of such distances can be constructed and among them a suitable one can be chosen or they can be combined in a machine learning framework (such distances may even correspond to different standardizations).

8.5.1 Extensions

Now, we briefly point to some possible directions along which this basic idea can be extended (see also [2]). First, note that the Frobenius norm in (8.14) can be replaced by any other unitary invariant matrix norm (e.g., the nuclear norm). A less trivial extension is to get rid of $O(m)$ in (8.15) by passing to covariance matrices. For example, in the case of $\widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC}}_{m,n,p}$ it is easy to verify that $\mathcal{SL}^{\mathrm{a,mp,tC}}_{m,n,p} = \widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC,cv}}_{m,n,p}/(O(n) \times I_m)$, where $\widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC,cv}}_{m,n,p} = \{(A, Z, C, S) | (A, B, C, D) \in \widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC}}_{m,n,p}, Z = BB^\top, S = DD^\top\}$. On this standardized subspace one only has the action of $O(n)$ which we denote as $Q \star (A, Z, C, S) = (Q^\top AQ, Q^\top ZQ, CQ, S)$. One can use the same ambient distance on this space as in (8.14) and get

$$d^2_{\Sigma_{m,n,p}}(M_1, M_2) = \min_{Q \in O(n)} \tilde{d}^2_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}\big(Q \star R_1, R_2\big), \tag{8.17}$$

for realizations $R_1, R_2 \in \widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC,cv}}_{m,n,p}$. One could also replace the $\|\cdot\|_F$ in the terms associated with B and D in (8.14) with some known distances in the spaces of positive definite matrices or positive-semi-definite matrices of fixed rank (see e.g., [14, 63]). Another possible extension is, e.g., to consider other submanifolds of $\widetilde{\mathcal{OSL}}^{\mathrm{a,mp,tC}}_{m,n,p}$, e.g., a submanifold where $\|C\|_F = \|B\|_F = 1$. In this case the corresponding alignment distance is essentially a scale invariant distance, i.e., two processes which are scaled version of one another will have zero distance. A more significant and subtle extension is to extend the underlying space of LDSs of fixed size and order n to that of fixed size but (minimal) order not larger than n. The details of this approach will appear elsewhere.

8.6 Conclusion

In this paper our focus was the geometrization of spaces of stochastic processes generated by LDSs of fixed size and order, for use in pattern recognition of high-dimensional time-series data (e.g., in the prototype Problem 1). We reviewed some of the existing approaches. We then studied the newly developed class of group action induced distances called the alignment distances. The approach is a general and flexible geometrization framework, based on the quotient structure of the space of such LDSs, which leads to a large class of extrinsic distances. The theory of alignment distances and their properties is still in early stages of development and we are hopeful to be able to tackle some interesting problems in control theory as well as pattern recognition in time-series data.

Acknowledgments The authors are thankful to the anonymous reviewers for their insightful comments and suggestions, which helped to improve the quality of this paper. The authors also thank the organizers of the GSI 2013 conference and the editor of this book Prof. Frank Nielsen. This work was supported by the Sloan Foundation and by grants ONR N00014-09-10084, NSF 0941362, NSF 0941463, NSF 0931805, and NSF 1335035.

References

1. Afsari, B., Chaudhry, R., Ravichandran, A., Vidal, R.: Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic visual scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
2. Afsari, B., Vidal, R.: The alignment distance on spaces of linear dynamical systems. In: IEEE Conference on Decision and Control (2013)
3. Afsari, B., Vidal, R.: Group action induced distances on spaces of high-dimensional linear stochastic processes. In: Geometric Science of Information, LNCS, vol. 8085, pp. 425–432 (2013)
4. Amari, S.I.: Differential geometry of a parametric family of invertible linear systems-Riemannian metric, dual affine connections, and divergence. Math. Syst. Theory **20**, 53–82 (1987)
5. Amari, S.I., Nagaoka, H.: Methods of information geometry. In: Translations of Mathematical Monographs, vol. 191. American Mathematical Society, Providence (2000)
6. Anderson, B.D., Deistler, M.: Properties of zero-free spectral matrices. IEEE Trans. Autom. Control **54**(10), 2365–5 (2009)
7. Aoki, M.: State Space Modeling of Time Series. Springer, Berlin (1987)
8. Barbaresco, F.: Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Frechet median. In: Matrix Information Geometry, pp. 199–255. Springer, Berlin (2013)
9. Basseville, M.: Distance measures for signal processing and pattern recognition. Sig. Process. **18**, 349–9 (1989)
10. Basseville, M.: Divergence measures for statistical data processingan annotated bibliography. Sig. Process. **93**(4), 621–33 (2013)
11. Bauer, D., Deistler, M.: Balanced canonical forms for system identification. IEEE Trans. Autom. Control **44**(6), 1118–1131 (1999)
12. Béjar, B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: Medical Image Computing and Computer Assisted Intervention, pp. 34–41 (2012)
13. Boets, J., Cock, K.D., Moor, B.D.: A mutual information based distance for multivariate Gaussian processes. In: Modeling, Estimation and Control, Festschrift in Honor of Giorgio Picci on the Occasion of his Sixty-Fifth Birthday, Lecture Notes in Control and Information Sciences, vol. 364, pp. 15–33. Springer, Berlin (2007)
14. Bonnabel, S., Collard, A., Sepulchre, R.: Rank-preserving geometric means of positive semi-definite matrices. Linear Algebra. Its Appl. **438**, 3202–16 (2013)
15. Byrnes, C.I., Hurt, N.: On the moduli of linear dynamical systems. In: Advances in Mathematical Studies in Analysis, vol. 4, pp. 83–122. Academic Press, New York (1979)
16. Chaudhry, R., Vidal, R.: Recognition of visual dynamical processes: Theory, kernels and experimental evaluation. Technical Report 09–01. Department of Computer Science, Johns Hopkins University (2009)
17. Chavel, I.: Riemannian Geometry: A Modern Introduction, vol. 98, 2nd edn. Cambridge University Press, Cambridge (2006)
18. Cock, K.D., Moor, B.D.: Subspace angles and distances between ARMA models. Syst. Control Lett. **46**(4), 265–70 (2002)

19. Corduas, M., Piccolo, D.: Time series clustering and classification by the autoregressive metric. Comput. Stat. Data Anal. **52**(4), 1860–72 (2008)
20. Deistler, M., Anderson, B.O., Filler, A., Zinner, C., Chen, W.: Generalized linear dynamic factor models: an approach via singular autoregressions. Eur. J. Control **3**, 211–24 (2010)
21. Devroye, L.: A probabilistic Theory of Pattern Recognition, vol. 31. Springer, Berlin (1996)
22. Doretto, G., Chiuso, A., Wu, Y., Soatto, S.: Dynamic textures. Int. J. Comput. Vision **51**(2), 91–109 (2003)
23. Ferrante, A., Pavon, M., Ramponi, F.: Hellinger versus Kullback-Leibler multivariable spectrum approximation. IEEE Trans. Autom. Control **53**(4), 954–67 (2008)
24. Forni, M., Hallin, M., Lippi, M., Reichlin, L.: The generalized dynamic-factor model: Identification and estimation. Rev. Econ. Stat. **82**(4), 540–54 (2000)
25. Georgiou, T.T., Karlsson, J., Takyar, M.S.: Metrics for power spectra: an axiomatic approach. IEEE Trans. Signal Process. **57**(3), 859–67 (2009)
26. Gray, R., Buzo, A., Gray Jr, A., Matsuyama, Y.: Distortion measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 367–76 (1980)
27. Gray, R.M.: Probability, Random Processes, and Ergodic Properties. Springer, Berlin (2009)
28. Gray, R.M., Neuhoff, D.L., Shields, P.C.: A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. The Ann. Probab. **3**, 315–328 (1975)
29. Gray Jr, A., Markel, J.: Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. **24**(5), 380–91 (1976)
30. Grenander, U.: Abstract Inference. Wiley, New York (1981)
31. Hannan, E.J.: Multiple Time Series, vol. 38. Wiley, New York (1970)
32. Hannan, E.J., Deistler, M.: The Statistical Theory of Linear Systems. Wiley, New York (1987)
33. Hanzon, B.: Identifiability, Recursive Identification and Spaces of Linear Dynamical Systems, vol. 63–64. Centrum voor Wiskunde en Informatica (CWI), Amsterdam (1989)
34. Hanzon, B., Marcus, S.I.: Riemannian metrics on spaces of stable linear systems, with applications to identification. In: IEEE Conference on Decision & Control, pp. 1119–1124 (1982)
35. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, New York (2003)
36. Hazewinkel, M.: Moduli and canonical forms for linear dynamical systems II: the topological case. Math. Syst. Theory **10**, 363–85 (1977)
37. Helmke, U.: Balanced realizations for linear systems: a variational approach. SIAM J. Control Optim. **31**(1), 1–15 (1993)
38. Jiang, X., Ning, L., Georgiou, T.T.: Distances and Riemannian metrics for multivariate spectral densities. IEEE Trans. Autom. Control **57**(7), 1723–35 (2012)
39. Jimenez, N.D., Afsari, B., Vidal, R.: Fast Jacobi-type algorithm for computing distances between linear dynamical systems. In: European Control Conference (2013)
40. Kailath, T.: Linear Systems. Prentice Hall, NJ (1980)
41. Katayama, T.: Subspace Methods for System Identification. Springer, Berlin (2005)
42. Kazakos, D., Papantoni-Kazakos, P.: Spectral distance measures between Gaussian processes. IEEE Trans. Autom. Control **25**(5), 950–9 (1980)
43. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley Series In Probability And Statistics. Wiley, New York (1999)
44. Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry Volume I. Wiley Classics Library Edition. Wiley, New York (1963)
45. Krishnaprasad, P.S.: Geometry of Minimal Systems and the Identification Problem. PhD thesis, Harvard University (1977)
46. Krishnaprasad, P.S., Martin, C.F.: On families of systems and deformations. Int. J. Control **38**(5), 1055–79 (1983)
47. Lee, J.M.: Introduction to Smooth Manifolds. Springer, Graduate Texts in Mathematics (2002)
48. Liao, T.W.: Clustering time series data—a survey. Pattern Recogn. **38**, 1857–74 (2005)
49. Makhoul, J.: Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–80 (1975)
50. Martin, A.: A metric for ARMA processes. IEEE Trans. Signal Process. **48**(4), 1164–70 (2000)

51. Moor, B.D., Overschee, P.V., Suykens, J.: Subspace algorithms for system identification and stochastic realization. Technical Report ESAT-SISTA Report 1990–28, Katholieke Universiteit Leuven (1990)
52. Moore, B.C.: Principal component analysis in linear systems: Controllability, observability, and model reduction. IEEE Trans. Autom. Control **26**, 17–32 (1981)
53. Ning, L., Georgiou, T.T., Tannenbaum, A.: Matrix-valued Monge-Kantorovich optimal mass transport. arXiv, preprint arXiv:1304.3931 (2013)
54. Nocerino, N., Soong, F.K., Rabiner, L.R., Klatt, D.H.: Comparative study of several distortion measures for speech recognition. Speech Commun. **4**(4), 317–31 (1985)
55. Ober, R.J.: Balanced realizations: canonical form, parametrization, model reduction. Int. J. Control **46**(2), 643–70 (1987)
56. Papoulis, A., Pillai, S.U.: Probability, random variables and stochastic processes with errata sheet. McGraw-Hill Education, New York (2002)
57. Piccolo, D.: A distance measure for classifying ARIMA models. J. Time Ser. Anal. **11**(2), 153–64 (1990)
58. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall International, NJ (1993)
59. Rao, M.M.: Stochastic Processes: Inference Theory, vol. 508. Springer, New York (2000)
60. Ravichandran, A., Vidal, R.: Video registration using dynamic textures. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 158–171 (2011)
61. Ravishanker, N., Melnick, E.L., Tsai, C.-L.: Differential geometry of ARMA models. J. Time Ser. Anal. **11**(3), 259–274 (1990)
62. Rozanov, Y.A.: Stationary Random Processes. Holden-Day, San Francisco (1967)
63. Vandereycken, B., Absil, P.-A., Vandewalle, S.: A Riemannian geometry with complete geodesics for the set of positive semi-definite matrices of fixed rank. Technical Report Report TW572, Katholieke Universiteit Leuven (2010)
64. Vishwanathan, S., Smola, A., Vidal, R.: Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. Int. J. Comput. Vision **73**(1), 95–119 (2007)
65. Youla, D.: On the factorization of rational matrices. IRE Trans. Inf. Theory **7**(3), 172–189 (1961)
66. Younes, L.: Shapes and Diffeomorphisms. In: Applied Mathematical Sciences, vol. 171. Springer, New York (2010)

Chapter 9
Discrete Ladders for Parallel Transport in Transformation Groups with an Affine Connection Structure

Marco Lorenzi and Xavier Pennec

9.1 Introduction

The analysis of complex information in medical imaging and in computer vision often requires to represent data in suitable manifolds in which we need to compute trajectories, distances, means and statistical modes. An important example is found in computational anatomy, which aims at developing statistical models of the anatomical variability of organs and tissues. Following Thompson [14], we can assume that the variability of a given observation of the population is encoded by a spatial deformation of a template shape or image (called an atlas in computational anatomy). The analysis of deformations thus enables the understanding of phenotypic variation and morphological traits in populations.

Deformations of images and shapes are usually estimated by image registration. Among the numerous methods used for registering medical images, the rich mathematical setting of *diffeomorphic* non-linear registration is particularly appealing since it provides elegant and grounded methods for atlas building [21], group-wise [9], and longitudinal statistical analysis of deformations [7, 15, 17]. In particular, temporal evolutions of anatomies can be modeled by transformat trajectories in the space of diffeomorphisms. However, developing population-based models requires reliable methods for comparing different trajectories.

Among the different techniques proposed so far [10, 17, 37], *parallel transport* represents a promising method which relies on a solid mathematical background. At the infinitesimal level, a trajectory is a tangent vector to a transformation. Parallel transport consists in transporting the infinitesimal deformation vector across the

M. Lorenzi (✉) · X. Pennec
Asclepios Research Project, INRIA Sophia Antipolis, 2004 route des Lucioles BP 93,
06902 Sophia Antipolis, France
e-mail: marco.lorenzi@inria.fr

X. Pennec
e-mail: xavier.pennec@inria.fr

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_9,

manifold by preserving its properties with respect to the space geometry. It is one of the fundamental operations of differential geometry which enables to compare tangent vectors, and thus the underlying trajectories, across the whole manifold.

For this reason, parallel transport in transformation groups is currently an important field of research with applications in medical imaging to the development of spatio-temporal atlases for brain images [26], the study of hippocampal shapes [35], or cardiac motion [1]. In computer vision one also finds applications to motion tracking and more generally to statistical analysis [12, 42, 45, 47].

Even though the notion of parallel transport comes in some way intuitively, its practical implementation requires the precise knowledge of the space geometry, in particular of the underlying connection. This is not always easy, especially in infinite dimensions such as in the setting of diffeomorphic image registration. Moreover, parallel transport is a continuous operation involving the computation of (covariant) derivatives which, from the practical point of view, might lead to issues concerning numerical stability and robustness. These issues are related to the unavoidable approximations arising when continuous energy functionals and operators are discretized on grids, especially concerning the evaluation of derivatives through finite difference schemes.

The complexity and limitations deriving from the direct computation of continuous parallel transport methods can be alleviated by considering discrete approximations. In the '70s of the past century [31] proposed a scheme for performing the parallel transport with a very simple geometrical constructions. This scheme was called *Schild's Ladder* since it was in the spirit of the work of the theoretical physicist Alfred Schild's. The computational interest of Schild's ladder resides in its generality, since it enables the transport of vectors in manifolds by computing geodesics only. This way, the implementation of covariant derivatives is not required anymore, and we can concentrate the implementation effort on the geodesics only.

We recently showed that numerical schemes derived from the Schild's ladder can be effectively applied in the setting of diffeomorphic image registration, by appropriately taking advantage of the underlying geometrical setting [28, 29]. Based on this experience, we believe that discrete transport methods represent promising and powerful techniques for the analysis of transformations due to their simplicity and generality. Bearing in mind the applicative context of the development of transport techniques, this chapter aims to illustrate the principles of discrete schemes for parallel transport in smooth groups equipped with affine connection or Riemannian structures.

The chapter is structured as follows. In Sect. 9.2 we provide fundamental notions of finite-dimensional Lie Groups and Riemannian geometry concerning affine connection and covariant derivative. These notions are the basis for the continuous parallel transport methods defined in Sect. 9.3. In Sect. 9.4 we introduce the Schild's ladder. After detailing its construction and mathematical properties, we derive from it the more efficient Pole ladder construction in Sect. 9.4.2. These theoretical concepts are then contextualized and discussed in the applicative setting of diffeomorphic image registration, in which some limitations arise when considering infinite dimensional Lie groups (Sect. 9.5). Finally, after illustrating numerical implementations of the

aforementioned discrete transport methods, we show in Sect. 9.6 their effectiveness when applied to the very practical problem of statistical analysis of the group-wise longitudinal brain changes in Alzheimer's disease.

This chapter summarizes and contextualizes a series of contributions concerning the theoretical foundations of diffeomorphic registration parametrized by stationary velocity fields (SVFs) and discrete transport methods. The application of Schild's ladder in the context of image registration was first presented in [29]. In [28], we highlighted the geometrical basis of the SVFs registration setting by investigating its theoretical foundations and its connections with Lie group theory and affine geometry. These insights allowed to define a novel discrete transport scheme, the *pole ladder*, which optimizes the Schild ladder by taking advantage of the geometrical properties of the SVF setting [27].

9.2 Basics of Lie Groups

We recall here the theoretical notions of Lie group theory and affine geometry, that will be extensively used in the following sections.

A Lie group $\mathbb{G}$ is a smooth manifold provided with an identity element id, a smooth associative composition rule $(g, h) \in \mathbb{G} \times \mathbb{G} \mapsto gh \in \mathbb{G}$ and a smooth inversion rule $g \mapsto g^{-1}$ which are both compatible with the differential manifold structure. As such, we have a tangent space $T_g\mathbb{G}$ at each point $g \in \mathbb{G}$. A vector field $\mathbf{X}$ is a smooth function that maps a tangent vector $\mathbf{X}|_g$ to each point g of the manifold. The set of vector fields (the tangent bundle) is denoted $T\mathbb{G}$. Vector fields can be viewed as the directional (or Lie) derivative of a scalar function ϕ along the vector field at each point: $\partial_{\mathbf{X}}\phi|_g = \frac{d}{dt}\phi(\varphi_t)|_{t=0}$, where φ_t is the flow of $\mathbf{X}$ and $\varphi_0 = g$. Composing directional derivatives $\partial_{\mathbf{X}}\partial_{\mathbf{Y}}\phi$ leads in general to a second order derivation. However, we can remove the second order terms by subtracting $\partial_{\mathbf{Y}}\partial_{\mathbf{X}}\phi$ (this can be checked by writing these expression in a local coordinate system). We obtain the Lie bracket that acts as an internal multiplication in the algebra of vector fields:

$$[\mathbf{X}, \mathbf{Y}](\phi) = \partial_{\mathbf{X}}\partial_{\mathbf{Y}}\phi - \partial_{\mathbf{Y}}\partial_{\mathbf{X}}\phi.$$

Given a group element $a \in \mathbb{G}$, we call left translation L_a the composition with the fixed element a on the left: $La : g \in \mathbb{G} \mapsto ag \in \mathbb{G}$. The differential DL_a of the left translation maps the tangent space $T_g\mathbb{G}$ to the tangent space $T_{ag}\mathbb{G}$. We say that a vector field $\mathbf{X} \in T(\mathbb{G})$ is left invariant if it remains unchanged under the action of the left translation: $DL_a\mathbf{X}|g = \mathbf{X}|_{ag}$. The sub-algebra of left-invariant vector fields is closed under the Lie bracket and is called the Lie algebra $\mathfrak{g}$ of the Lie group. Since a left-invariant vector field is uniquely determined by its value at identity through the one-to-one map $\tilde{\mathbf{X}}|_g = DL_gX$, the Lie algebra can be identified to the tangent space at the identity $T_{id}\mathbb{G}$. One should notice that any smooth vector field can be written as a linear combination of left-invariant vector fields with smooth *functional* coefficients.

Left-invariant vector fields are *complete* in the sense that their flow φ_t is defined for all time. Moreover, this flow is such that $\varphi_t(g) = g\varphi_t(id)$ by left invariance. The map $X \mapsto \varphi_1(id)$ of $\mathfrak{g}$ into $\mathbb{G}$ is called *Lie group exponential* and denoted by exp. In particular, the group exponential defines the *one-parameter subgroup* associated to the vector X and has the following properties:

- $\varphi_t(id) = \exp(tX)$, for each $t \in \mathbb{R}$;
- $\exp((t+s)X) = \exp(tX)\exp(sX)$, for each $t, s \in \mathbb{R}$.

In finite dimension, it can be shown that the Lie group exponential is a diffeomorphism from a neighborhood of 0 in $\mathfrak{g}$ to a neighborhood of id in $\mathbf{G}$.

For each tangent vector $X \in \mathfrak{g}$, the one parameter subgroup $\exp(tX)$ is a curve that starts from identity with this tangent vector. One could question if this curve could be seen as a geodesic like in Riemannian manifolds. To answer this question, we first need to define what are geodesics. In a Euclidean space, straight lines are curves which have the same tangent vector at all times. In a manifold, tangent vectors at different times belong to different tangent spaces. When one wants to compare tangent vectors at different points, one needs to define a specific mapping between their tangent spaces: this is the notion of parallel transport. There is generally no way to define globally a linear operator $\Pi_g^h : T_g\mathbb{G} \to T_h\mathbb{G}$ which is consistent with composition (i.e. $\Pi_g^h \circ \Pi_f^g = \Pi_f^h$). However, specifying the parallel transport for infinitesimal displacements allows integrating along a path, thus resulting in a parallel transport that depend on the path. This specification of the parallel transport for infinitesimal displacements is called the (affine) connection.

9.2.1 Affine Connection Spaces

An *affine connection* on $\mathbb{G}$ is an operator which assigns to each $\mathbf{X} \in T(\mathbb{G})$ a linear mapping $\nabla_{\mathbf{X}} : T(\mathbb{G}) \to T(\mathbb{G})$ such that, for each vector field $\mathbf{X}, \mathbf{Y} \in T(\mathbb{G})$, and smooth function $f, g \in C^\infty(\mathbb{G}, \mathbb{R})$

$$\nabla_{f\mathbf{X}+g\mathbf{Y}} = f\nabla_{\mathbf{X}} + g\nabla_{\mathbf{Y}} \quad \text{(Linearity)}; \tag{9.1}$$

$$\nabla_{\mathbf{X}}(f\mathbf{Y}) = f\nabla_{\mathbf{X}}(\mathbf{Y}) + (\mathbf{X}f)\mathbf{Y} \quad \text{(Leibniz rule)}. \tag{9.2}$$

The *affine connection* is therefore a derivation on the tangent space which infinitesimally maps them from one tangent plane to another.

The connection give rise to two very important geometrical objects: the *torsion* and *curvature* tensors. The torsion quantifies the failure to close infinitesimal geodesic parallelograms:

$$T(\mathbf{X}, \mathbf{Y}) = \nabla_{\mathbf{X}}\mathbf{Y} - \nabla_{\mathbf{Y}}\mathbf{X} - [\mathbf{X}, \mathbf{Y}],$$

while the curvature measures the local deviation of the space from being flat, and is defined as

$$R(\mathbf{X}, \mathbf{Y})\mathbf{Z} = \nabla_{\mathbf{X}}\nabla_{\mathbf{Y}}\mathbf{Z} - \nabla_{\mathbf{Y}}\nabla_{\mathbf{X}}\mathbf{Z} - \nabla_{[\mathbf{X},\mathbf{Y}]}\mathbf{Z}.$$

Once the manifold is provided with a connection, it is possible to generalize the notion of "straight line": a vector field $\mathbf{X}$ is *parallel* along a curve $\gamma(t)$ if $\nabla_{\dot{\gamma}(t)}\mathbf{X} = 0$ for each t. A path $\gamma(t)$ on $\mathbf{G}$ is said to be straight or *geodesic* if $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$, i.e. if its tangent vector remains parallel to itself along the path.

In a local coordinate system, the geodesic equation is a second order differential equation. Thus, given a point $p \in \mathbb{G}$ and a vector $X \in T_p\mathbb{G}$, there exist a unique geodesic $\gamma(t, p, X)$ that passes through p with velocity X at the instant $t = 0$ [34]. We define therefore the *Affine exponential* as the application $\exp : \mathbb{G} \times T(\mathbb{G}) \to \mathbb{G}$ given by $\exp_p(X) = \gamma(1, p, X)$.

If, as in the Euclidean case, we want to associate to the straight lines the property of minimizing the distance between points, we need to provide the group $\mathbb{G}$ with a Riemannian manifold structure, i.e. with a metric operator g on the tangent space. In this case there is a unique connection, called the *Levi Civita connection*, which, for each $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in T(\mathbb{G})$:

- Preserves the metric, i.e. the parallel transport along any curve connecting f to g is an isometry:

$$g(\mathbf{X}, \mathbf{Y})_g = g(\Pi_g^f\mathbf{X}, \Pi_g^f\mathbf{Y})_f.$$

- Is torsion free:

$$\nabla_{\mathbf{X}}\mathbf{Y} - \nabla_{\mathbf{Y}}\mathbf{X} = [\mathbf{X}, \mathbf{Y}],$$

  thus the parallel transport is symmetric with respect to the Lie bracket.

By choosing the Levi Civita connection of a given Riemannian metric, the affine geodesics are the length minimizing paths (i.e. classical Riemannian geodesics). However, given a general affine connection, there may not exist any Riemannian metric for which affine geodesics are length minimizing.

9.2.2 Cartan-Schouten Connections

Given an affine connection ∇ and a vector X on $T_{id}\mathbb{G}$, we can therefore define two curves on $\mathbb{G}$ passing through id and having X as tangent vector, one given by the Lie group exponential exp and the other given by the affine exponential $\exp_{id}$. When do they coincide?

The connection ∇ on $\mathbb{G}$ is *left-invariant* if, for each left translation L_a ($a \in \mathbb{G}$) and any vector fields $\mathbf{X}$ and $\mathbf{Y}$, we have $\nabla_{DL_a\mathbf{X}}(DL_a\mathbf{Y}) = DL_a\nabla_{\mathbf{X}}(\mathbf{Y})$. Using two left invariant vector fields $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \in \mathfrak{g}$ generated by the tangent vectors $X, Y \in T_{id}\mathbb{G}$, we see that $\nabla_{\tilde{\mathbf{X}}}\tilde{\mathbf{Y}}$ is itself a left-invariant vector field generated by its value at identity. Since a connection is completely determined by its action on the left-invariant vector fields

(we can recover the connection on arbitrary vector fields using Eqs. (9.1 and 9.2) from their decomposition on the Lie Algebra), we conclude that each left-invariant connection ∇ is uniquely determined by a product α (symmetric bilinear operator) on $T_{id}\mathbb{G}$ through

$$\alpha(X, Y) = \nabla_{\tilde{\mathbf{X}}} \tilde{\mathbf{Y}}\Big|_{id}.$$

Notice that such a product can be uniquely decomposed into a commutative part $\alpha' = \frac{1}{2}(\alpha(X, Y) + \alpha(Y, X))$ and a skew symmetric part $\alpha'' = \frac{1}{2}(\alpha(X, Y) - \alpha(Y, X))$. The symmetric part specifies the geodesics (i.e. the parallel transport of a vector along its own direction) while the skew-symmetric part specifies the torsion which governs the parallel transport of a vector along a transverse direction (the rotation around the direction of the curve if we have a metric connection with torsion).

Following [34], a left-invariant connection ∇ on a Lie group $\mathbb{G}$ is a *Cartan-Schouten connection* if, for any tangent vector X at the identity, the one-parameter subgroups and the affine geodesics coincide, i.e. $\exp(tX) = \gamma(t, id, X)$. We can see that a Cartan connection satisfies $\alpha(X, X) = 0$ or, equivalently, is purely skew-symmetric.

The one-dimensional family of connections generated by $\alpha(X, Y) = \lambda[X, Y]$ obviously satisfy this skew-symmetry condition. Moreover, the connections of this family are also invariant by right translation [33], thus invariant by inversion also since they are already left invariant. This make them particularly interesting since they are fully compatible with all the group operations.

In this family, three connections have special curvature or symmetric properties and are called the canonical Cartan-Schouten connections [11]. The zero curvature connections given by $\lambda = 0, 1$ (with torsion $T = -[\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}]$ and $T = [\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}]$ respectively on left invariant vector fields) are called *left* and *right* Cartan connections. The choice of $\lambda = 1/2$ leads to average the left and right Cartan connections. It is called the *symmetric* (or *mean*) Cartan connection. It is torsion-free, but has curvature $R(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})\tilde{\mathbf{Z}} = -\frac{1}{4}\left[[\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}], \tilde{\mathbf{Z}}\right]$.

As a summary, the three canonical Cartan connections of a Lie group are (for two left-invariant vector fields):

$$\begin{aligned}
&\nabla_{\tilde{\mathbf{X}}} \tilde{\mathbf{Y}} = 0 && \text{Left (Torsion, Flat);}\\
&\nabla_{\tilde{\mathbf{X}}} \tilde{\mathbf{Y}} = \frac{1}{2}\left[\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right] && \text{Symmetric (Torsion-Free, Curved);}\\
&\nabla_{\tilde{\mathbf{X}}} \tilde{\mathbf{Y}} = \left[\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right] && \text{Right (Torsion, Flat).}
\end{aligned}$$

Since the three canonical Cartan connections only differ by torsion, they share the same affine geodesics which are the left and right translations of one parameter subgroups. In the following, we call them *group geodesics*. However, the parallel transport of general vectors along these group geodesics is specific to each connection as we will see below.

9.2.3 Riemannian Setting: Levi Civita Connection

Given a metric $<X, Y>$ on the tangent space at identity of a group, one can propagate this metric to all tangent spaces using left (resp. right) translation to obtain a left- (resp. right-) invariant Riemannian metric on the group. In the left-invariant case we have $<DL_aX, DL_aY>_a = <X, Y>$ and one can show [24] that the Levi Civita connection is the left-invariant connection generated by the product

$$\alpha(X, Y) = \frac{1}{2}[X, Y] - \frac{1}{2}(ad^*(X, Y) + ad^*(Y, X)),$$

where the operator ad^* is defined by $<ad^*(Y, X), Z> = <[X, Z], Y>$ for all $X, Y, Z \in \mathfrak{g}$. A similar formula can be established for right-invariant metrics using the algebra of right-invariant vector fields .

We clearly see that this left-invariant Levi Civita connection has a symmetric part which make it differ from the Cartan symmetric connection $\alpha(X, Y) = \frac{1}{2}[X, Y]$. In fact, the quantity $ad^*(X, X)$ specifies the rate at which a left invariant geodesic and a one parameter subgroup starting from the identity with the same tangent vector X deviates from each-other. More generally, the condition $ad^*(X, X) = 0$ for all $X \in \mathfrak{g}$ turns out to be a necessary and sufficient condition to have a bi-invariant metric [34]. It is important to notice that geodesics of the left- and right-invariant metrics differ in general as there do not exists bi-invariant metrics even for simple groups like the Euclidean motions [33]. However, right invariant geodesics can be easily obtained from the left invariant one through inversion: if $\phi(t)$ is a left invariant geodesic joining identity to the transformation ϕ_1, then $\phi^{-1}(t)$ is a right-invariant geodesic joining identity to ϕ_1^{-1}.

9.3 Continuous Methods for Parallel Transport

After having introduced the theoretical bases of affine connection spaces, in this section we detail the theoretical relationship between parallel transport of tangent vectors and respectively Cartan-Schouten and Riemannian (Levi Civita) connections.

9.3.1 Cartan-Schouten Parallel Transport

For the left Cartan connection, the unique fields that are covariantly constant are the *left-invariant* vector fields, and the parallel transport is induced by the differential of the left translation [34], i.e. $\Pi^L : T_p\mathbb{G} \to T_q\mathbb{G}$ is defined as

$$\Pi^L(X) = DL_{qp^{-1}}X. \tag{9.3}$$

One can see that the parallel transport is actually independent of the path, which is due to the fact that the curvature is null: we are in a space with absolute parallelism. Similarly, the *right-invariant* vector fields are covariantly constant with respect to the right invariant connection only. As above, the parallel transport is given by the differential of the right translation

$$\Pi^R(X) = DR_{p^{-1}q}X, \tag{9.4}$$

and we have an absolute parallelism as well.

Finally, the parallel transport for the symmetric Cartan connection is given by the infinitesimal alternation of the left and right transports. However, as there is curvature, it depends on the path: it can be shown [18] that the parallel transport of X along the geodesic $\exp(tY)$ is:

$$\Pi^S(X) = DL_{\exp(\frac{1}{2}Y)}DR_{\exp(\frac{1}{2}Y)}X. \tag{9.5}$$

9.3.2 Riemannian Parallel Transport

In the Riemannian setting the parallel transport with the Levi Civita connection can be computed by solving a system of PDEs which locally depend on the associated metric (the interested reader can refer to [16] for a more comprehensive description of the parallel transport in Riemannian geometry). Let x_i be a local coordinate chart x_i with $\partial_i = \frac{\partial}{\partial x_i}$ a local basis of the tangent space. The tangent vector to the curve γ is $\dot{\gamma} = \sum_i v^i \partial_i$. It can be easily shown that a vector $Y = \sum_i y^i \partial_i$ is parallel transported along γ with respect to the affine connection ∇ iff

$$\nabla_{\dot{\gamma}} Y = \sum_k \left(\sum_{ij} \Gamma_{ij}^k v^j y^k + \dot{\gamma}(y^k) \right) \partial_k = 0, \tag{9.6}$$

with the Christoffel symbols of the connection being defined from the covariant derivative $\nabla_{\partial_i}\partial_j = \sum_k \Gamma_{ij}^k \partial_k$.

Let us consider the local expression for the metric tensor $g_{lk} = <\frac{\partial}{\partial x_l}, \frac{\partial}{\partial x_k}>$. Thanks to the compatibility condition, the Christoffel symbols of the Levi Civita connection can be locally expressed via the metric tensor leading to

$$\Gamma_{ij}^k = \frac{1}{2} \sum_l \left\{ \frac{\partial}{\partial x_i} g_{jl} + \frac{\partial}{\partial x_j} g_{li} + \frac{\partial}{\partial x_l} g_{ij} \right\} g^{lk},$$

with $(g^{lk}) = (g_{lk})^{-1}$.

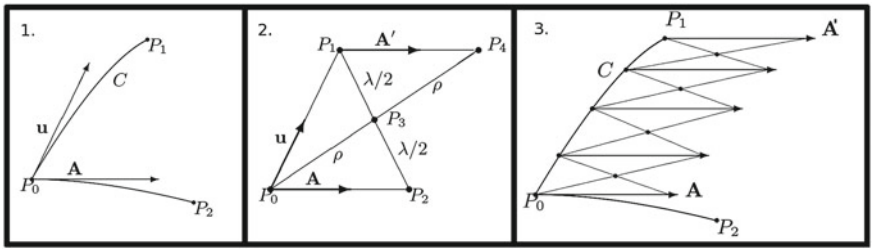


Fig. 9.1 (*1*) The transport of the vector A along the curve C is performed by the Schild's ladder by (*2*) the construction of geodesic parallelograms in a sufficiently small neighborhood. (*3*) The construction is iterated for a sufficient number of neighborhoods

Thus, in the Riemannian setting the covariant derivative is uniquely defined by the metric, and the parallel transport thus depends from the path γ and from the local expression of the metric tensor g_{ij} (formula 9.6).

9.4 Discrete Methods for Parallel Transport

In Sect. 9.3 we showed that the parallel transport closely depends on the underlying connection, and that thus it assumes very specific formulations depending on the underlying geometry. In this section we introduce discrete methods for the computation of the parallel transport which do not explicitly depend on the connection and only make use of geodesics. Such techniques could be applied more generally when working on arbitrary geodesic spaces.

9.4.1 Schild's Ladder

Schild's ladder is a general method for the parallel transport, introduced in the theory of gravitation in [31] after Schild's similar constructions [39]. The method infinitesimally transports a vector along a given curve through the construction of geodesic parallelograms (Fig. 9.1). The Schild's ladder provides a straightforward method to compute a second order approximation of the parallel transport of a vector along a curve using geodesics only.

Let M a manifold and C a curve parametrized by the parameter τ with $\frac{\partial C}{\partial \tau}|_{T_0} = u$, and $A \in T_{P_0}M$, a tangent vector on the curve at the point $P_0 = C(0)$. Let P_1 be a point on the curve relatively close to P_0, i.e. separated by a sufficiently small parameter value τ.

The Schild's ladder computes the parallel transport of A along the curve C as follows:

1. Define a *curve* on the manifold parametrized by a parameter σ passing through the point P_0 with tangent vector $\frac{\partial}{\partial\sigma}|_{P_0} = A$. Chose a point P_2 on the curve separated by P_0 by the value of the parameters σ. The values of the parameters σ and τ should be chosen in order to construct this step of the ladder within a single coordinate neighborhood.
2. Let l be the *geodesic* connecting $P_2 = l(0)$ and $P_1 = l(\lambda)$, we choose the "middle point" $P_3 = l(\lambda/2)$. Now, let us define the *geodesic* r connecting the starting point P_0 and P_3 parametrized by ρ such that $P_3 = r(\rho)$. Extending the geodesic at the parameter 2ρ we reach the point P_4. We can now compute the geodesic *curve* connecting P_1 and P_4. The vector A' tangent to the curve at the point $P1$ is the parallel translation of A along C.
3. If the distance between the points P_0 and P_1 is large, the above construction can be iterated for a sufficient number of steps.

The algorithmic interest of the Schild's ladder is that it only relies on the computation of geodesics. Although the geodesics on the manifold are not sufficient to recover all the information about the space properties, such as the torsion of the connection, it has been shown that the Schild's ladder implements the parallel transport with respect to the symmetric part of the connection of the space [23]. An intuitive view of that point is that the construction of the above diagram is commutative and can be symmetrized with respect to the points P_1 and P_2. If the original connection is symmetric, then this procedure provides a correct linear approximation of the parallel transport of vectors.

9.4.2 Pole Ladder

We proposed in [27] a different construction for the parallel transport of vectors based on geodesics parallelograms. If the curve C is geodesic, then it can be itself one of the diagonals and the Schild's ladder can therefore be adapted by requiring the computation of only one new diagonal of the parallelogram. We define in this way a different ladder scheme, that we name "pole ladder" since its geometrical construction recalls the type of ladders with alternating or symmetric steps with respect to a central axis.

We now prove that the pole ladder is correctly implementing the parallel transport. In the diagram of Fig. 9.2, the parallel transport of the tangent vector $v = \dot{C}$ along the geodesic C is specified by the geodesic equation $\dot{\mathbf{v}} + \Gamma^k_{ij}\mathbf{v}^i\mathbf{v}^j = 0$ using the Christoffel symbols $\Gamma^k_{ij}(x)$. In a sufficiently small neighborhood the relationships can be linearized to give

$$\mathbf{v}^k(t) = \mathbf{v}^k(0) - t\Gamma^k_{ij}(x(0))\mathbf{v}^i(0)\mathbf{v}^j(0) + O(t^2),$$

and by integrating:

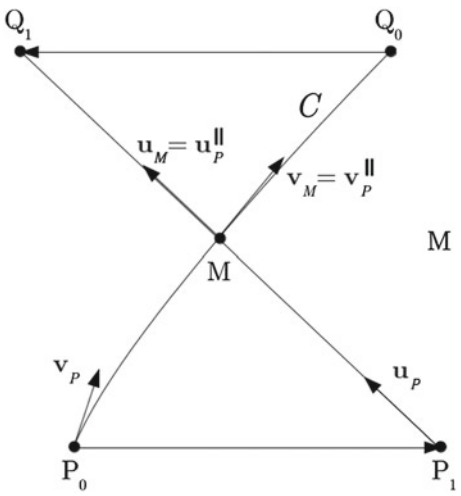


Fig. 9.2 The pole ladder parallel transports the vector A along the geodesic C. Contrarily to the Schild's ladder, it only requires to compute one diagonal geodesic

$$x^k(t) = x^k(0) + t\mathbf{v}^k(0) - \frac{t^2}{2}\Gamma^k_{ij}(x(0))\mathbf{v}^i(0)\mathbf{v}^j(0) + O(t^3).$$

By renormalizing the length of the vector v so that $C(-1) = P_0$, $C(0) = M$ and $C(1) = Q_0$ (and denoting $\Gamma^k_{ij} = \Gamma^k_{ij}(M)$), we obtain the relations:

$$P_0{}^k = M^k - \mathbf{v}^k_M - \frac{1}{2}\Gamma^k_{ij}\mathbf{v}^i_M\mathbf{v}^j_M + O(\|\mathbf{v}\|^3),$$
$$Q_0{}^k = M^k + \mathbf{v}^k_M - \frac{1}{2}\Gamma^k_{ij}\mathbf{v}^i_M\mathbf{v}^j_M + O(\|\mathbf{v}\|^3).$$

Similarly, we have along the second geodesic:

$$P_1{}^k = M^k - \mathbf{u}^k_M - \frac{1}{2}\Gamma^k_{ij}\mathbf{u}^i_M\mathbf{u}^j_M + O(\|\mathbf{u}\|^3),$$
$$Q_1{}^k = M^k + \mathbf{u}^k_M - \frac{1}{2}\Gamma^k_{ij}\mathbf{u}^i_M\mathbf{u}^j_M + O(\|\mathbf{u}\|^3).$$

Now, to compute the geodesics joining P_0 to P_1 and Q_0 to Q_1, we have to use a Taylor expansion of the Christoffel symbols Γ^k_{ij} around the point M. In the following, we indicate the coordinate according to which the quantity is derived by the index after a comma: $\Gamma^k_{ij,a} = \partial_a\Gamma^k_{ij}$:

$$\Gamma^k_{ij}(P_0) = \Gamma^k_{ij} + \Gamma^k_{ij,a}(-\mathbf{v}^k_M - \frac{1}{2}\Gamma^k_{ij}\mathbf{v}^i_M\mathbf{v}^j_M) + \frac{1}{2}\Gamma^k_{ij,ab}\mathbf{v}^a_M\mathbf{v}^b_M + O(\|\mathbf{v}\|^3).$$

However, the Christoffel symbols are multiplied by a term of order $O(\|A\|^2)$, so that only the first term will be quadratic and all others will be of order 3 with respect to A and $\mathbf{v}_M$. Thus, the geodesics joining P_0 to P_1 and Q_0 to Q_1 have equations:

$$P_1^k = P_0^k + A^k - \frac{1}{2}\Gamma_{ij}^k A^i A^j + O((\|A\| + \|\mathbf{v}_M\|)^3),$$
$$Q_1^k = Q_0^k + B^k - \frac{1}{2}\Gamma_{ij}^k B^i B^j + O((\|B\| + \|\mathbf{v}_M\|)^3).$$

Equating P_1^k in the previous equations gives

$$\mathbf{u}_M^k + \frac{1}{2}\Gamma_{ij}^k \mathbf{u}_M^i \mathbf{u}_M^j = \mathbf{v}_M^k - A^k + \frac{1}{2}\Gamma_{ij}^k (\mathbf{v}_M^i \mathbf{v}_M^j + A^i A^j) + O((\|B\| + \|\mathbf{v}_M\|)^3).$$

Solving for $\mathbf{u}$ as a second order polynomial in v_M and A gives

$$\mathbf{u}^k = \mathbf{v}_M^k - A^k + \frac{1}{2}(\Gamma_{ij}^k + \Gamma_{ji}^k) A^i \mathbf{v}_M^j + O((\|A\| + \|\mathbf{v}_M\|)^3).$$

Now equating Q_1^k in the previous equations gives

$$B^k - \frac{1}{2}\Gamma_{ij}^k B^i B^j = -A^k + (\Gamma_{ij}^k + \Gamma_{ji}^k) A^i \mathbf{v}_M^j + \frac{1}{2}\Gamma_{ij}^k A^i A^j + O((\|A\| + \|\mathbf{v}_M\|)^3).$$

Solving for B^k as a second order polynomial in v_M and A gives:

$$B^k = -A^k + (\Gamma_{ij}^k + \Gamma_{ji}^k) A^i \mathbf{v}^j + O((\|A\| + \|\mathbf{v}_M\|)^3). \tag{9.7}$$

To verify that this is the correct formula for the parallel transport of A, let us observe that the field $A(x)$ is parallel in the direction of $\mathbf{v}^j$ if $\nabla_V A = 0$, i.e. if $\partial_v A^k + \Gamma_{ij}^k A^i \mathbf{v}^j = 0$, which means that $A^k(x + \epsilon v) = A^k - \epsilon \Gamma_{ij}^k A^i \mathbf{v}^j + O(\epsilon^2)$. If the connection is symmetric, i.e. if $\Gamma_{ij}^k = \Gamma_{ji}^k$, Eq. (9.7) shows that the pole ladder leads to $B^k \simeq -A^k + 2\Gamma_{ij}^k A^i \mathbf{v}^j$. Thus the pole ladder is realizing the parallel transport for a length $\epsilon = 2$ (remember that our initial geodesic was defined from -1 to 1).

We have thus demonstrated that the vector $-B$ of Fig. 9.2 is a second order approximation of the transport of A. In order to optimize the number of time steps we should evaluate the error in Eq. (9.7) at high orders on $\|A\|$ and $\|\mathbf{v}_M\|$. The computation is not straightforward and involves a large number of terms, thus preventing the possibility to synthesize a useful result. However, we believe that the dependency on $\|A\|$ is more important that the one on $\|\mathbf{v}_M\|$, and that we could obtain larger time steps provided that $\|A\|$ is sufficiently small.

9.5 Diffeomorphic Medical Image Registration

We now describe a practical context in which the previous theoretical insights find useful application. For this purpose, we describe here the link between the theory

described in Sect. 9.2 and the context of computational anatomy, in particular through the diffeomorphic non-linear registration of time series of images.

9.5.1 Longitudinal and Cross-Sectional Registration Settings

Modeling the temporal evolution of the tissues of the body is an important goal of medical image analysis for understanding the structural changes of organs affected by a pathology, or for studying the physiological growth during the life span. For such purposes we need to analyze and compare the observed anatomical differences between follow-up sequences of anatomical images of different subjects. Non-rigid registration is one of the main instruments for modeling anatomical differences from images. The aim of non-linear registration is to encode the observed structural changes as deformation fields densely represented in the image space, which represent the warping required to match the observed differences. This way, the anatomical changes can be modeled and quantified by analyzing the associated deformations.

We can identify two distinct settings for the application of non-linear registration: longitudinal and cross-sectional. In the former, non-linear registration estimates the deformation field which explains the longitudinal anatomical (intra-subject) changes that usually reflect biological phenomena of interest, like atrophy or growth. In the latter, the deformation field accounts for the anatomical differences between different subjects (inter-subject), in order to match homologous anatomical regions. These two settings are profoundly different: the cross-sectional setting does not involve any physical or mechanical deformations and we might wish to compare different anatomies with different topologies. Moreover, inter-subject deformations are often a scale of magnitude higher than the ones characterizing the usually subtle variations of the longitudinal setting.

In case of group-wise analysis of longitudinal deformations, longitudinal and cross-sectional settings must be integrated in a consistent manner. In fact, the comparison of longitudinal deformations is usually performed after normalizing them in a common reference frame through the inter-subject registration, and the choice of the normalization method might have a deep impact on the following analysis. In order to accurately identify longitudinal deformations in a common reference frame space, a rigorous and reliable normalization procedure need thus to be defined.

Normalization of longitudinal deformations can be done in different ways, depending on the analyzed feature. For instance, the scalar Jacobian determinant of longitudinal deformations represents the associated local volume change, and can be compared by scalar resampling in a common reference frame via inter-subject registration. This simple transport of scalar quantities is the basis of the classical deformation/tensor based morphometry techniques [5, 38]. However, transporting the Jacobian determinant is not sufficient to reconstruct a deformation in the Template space.

If we consider vector-values characteristics of deformations instead of scalar quantities, the transport is not uniquely defined anymore. For instance, a simple method

of transport consists in *reorienting* the longitudinal intra-subject displacement vector field by the Jacobian matrix of the subject-to-reference deformation. Another intuitive method was proposed by [37] and uses the *transformation conjugation* (change of coordinate system) in order to compose the longitudinal intra-subject deformation with the subject-to-reference one. As pointed out in [10], this practice could potentially introduce variations in the transported deformation and relies on the inverse consistency of the estimated deformations, which can raise numerical problems for large deformations.

In the geometric setting, when we have a Riemannian or affine manifold structure for the space our deformations, one would like to use a normalization which is consistent with the manifold structure. This requirement naturally raise parallel transport as the natural tool for normalizing measurements at different points. In order to elaborate along this idea, we first need to describe the geometric structures on diffeomorphic deformations.

A first formulation of diffeomorphic registration was proposed with the "Large Deformation Diffeomorphic Metric Mapping (LDDMM)" setting [8, 44]. In this framework the images are registered by minimizing the length of the trajectory of transformations in the space of diffeomorphism, once specified an opportune right invariant metric. The solution is the endpoint of the flow of a *time-varying* velocity field, which is a geodesic parametrized through the Riemannian exponential. The LDDMM deformations are thus Riemannian (metric) geodesics, which are also geodesics of the Levi Civita connection.

Since LDDMM is generally computationally intensive, a different diffeomorphic registration method was later proposed with the stationary velocity field (SVF) setting [3]. In this case the diffeomorphisms are parametrized by *stationary* velocity fields, in opposition to the time varying velocity fields of the LDDMM framework, through the Lie group exponential. The restriction to stationary velocity fields simplifies the registration problem and provides efficient numerical schemes for the computation of deformations. This time the flow associated to SVFs is a one-parameter subgroup, which is a geodesic with respect to the Cartan-Schouten connections. One-parameter subgroups are generally not metric geodesics, since there do not exist any left and right invariant metric on non-compact and non-commutative groups.

In both the LDDMM and SVF settings, the longitudinal deformation is encoded by the initial tangent velocity field. The transport of longitudinal deformations can be then naturally formulated as the parallel transport of tangent vectors along geodesics according to the underlying connection, i.e. the Levi Civita connection in LDDMM, and the canonical symmetric Cartan-Schouten connection in the SVF setting.

9.5.2 A Glimpse of Lie Group Theory in Infinite Dimension

In Sect. 9.2.2, we derived the equivalence of one-parameter subgroups and the affine geodesics of the canonical Cartan connections in a finite dimensional Lie group. In order to use such a framework for diffeomorphisms, we have to generalize the theory

to infinite dimensions. However, defining infinite dimensional Lie groups is raising much more difficulties. This is in fact the reason why Lie himself restricted to finite dimensions. The theory was developed since the '70s and is now an active field of research. We refer the reader to the recent books [22, 48] for more details on this theory and to [40] for a good overview of the problems and applications.

The basic construction scheme is to consider an infinite dimensional manifold endowed with smooth group operations. Such a Lie group is locally diffeomorphic to an infinite-dimensional vector space which can be a Fréchet space (a locally convex space which is complete with respect to a translation invariant distance), a Banach space (where the distance comes from a norm) or a Hilbert space (where the norm is derived from a scalar product). We talk about Fréchet, Banach or Hilbert Lie groups, respectively. Extending differential calculus from $\mathbb{R}^n$ to Banach and Hilbert spaces is straightforward, but this is not so simple for Fréchet spaces. In particular, the dual of a Fréchet space need not be Fréchet, which means that some extra care must be taken when defining differential forms. Moreover, some important theorems such as the inverse function theorem hold for Banach spaces but not necessarily for Fréchet spaces.

For instance, the set $\mathrm{Diff}^k(\mathcal{M})$ of C^k diffeomorphisms of a compact manifold $\mathcal{M}$ is a Banach manifold and the set of Sobolev H^s diffeomorphisms $\mathrm{Diff}^s(\mathcal{M})$ is a Hilbert manifold (if $s > dim\mathcal{M}/2$). However, these are no-classical "Lie groups" since one loses derivatives when differentiating the composition and inversion maps. To obtain the complete smoothness of the composition and inversion maps, one has to go to infinity, but the Banach structure is lost in the process [40, p.12] and we are left with $\mathrm{Diff}^\infty(\mathcal{M})$ being only a Fréchet Lie group. Some additional structure can be obtained by considering the sequence of $\mathrm{Diff}^k(\mathcal{M})$ spaces as a succession of dense inclusions as k goes to infinity: this the Inverse Limit of Banach (ILB)-Lie group setting. Likewise, the succession of dense inclusions of Sobolev H^s diffeomorphisms give rise to the Inverse Limit of Hilbert (ILH)-Lie group setting.

As the diffeomorphisms groups considered are Fréchet but not Banach, the usual setting of infinite dimensional Lie groups is the general framework of Fréchet manifolds. This implies that many of the important properties which are true in finite dimension do not hold any more for general infinite dimensional Lie groups [41].

First, there is no implicit or inverse function theorem (except Nash-Moser type theorems.) This implies for instance that the log-map (the inverse of the exponential map) may not be smooth even if the differential of the exponential map is the identity.

Second, the exponential map is not in general a diffeomorphism from a neighborhood of zero in the Lie algebra onto a neighborhood of the identity in the group. This means that it cannot be used as a local chart to work on the manifold. For instance in $\mathrm{Diff}^s(\mathcal{M})$, in every neighborhood of the identity there may exist diffeomorphisms which are not the exponential of an H^s vector field. A classical example of the non-surjectivity of the exponential map is the following function in $\mathrm{Diff}(\mathbb{S}^1)$ [30]:

$$f_{n,\epsilon}(\theta) = \theta + \pi/n + \epsilon \sin^2(n\theta). \tag{9.8}$$

This function can be chosen as close as we want to the identity by opportunely dimensioning ϵ and θ. However, it can be shown that it cannot be reached by any one-parameter subgroup, and therefore the Lie group exponential is not a local diffeomorphisms of $\mathrm{Diff}(\mathbb{S}^1)$.

This example is quite instructive and shows that this theoretical problem might actually be a very practical advantage: the norm of the k-th derivative of $f_{n,\epsilon}$ is exploding when k is going to infinity, which shows that we would rather want to exclude this type of diffeomorphisms from the group under consideration.

9.5.3 Riemannian Structure and Stationary Velocity Fields

In the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [48], a different construction is leading to a more restricted subgroup of diffeomorphisms which is more rational from the computational point of view. One first chooses a Hilbert norm on the Lie Algebra which turn it into an admissible Hilbert space. Admissible means that it can be embedded into the space of vector fields which are bounded and vanishing at infinity, as well as all the first order derivatives. Typically, this is a Sobolev norm of a sufficiently high order. Then, one restricts to the subgroup of diffeomorphisms generated by the flow of integrable sequences of such vector fields for a finite time. To provide this group with a Riemannian structure, a right invariant metric is chosen. A first reason for choosing right translation is that it is simply a composition which does not involve a differential operator as for the left translation. A second reason is that the resulting metric on the group is generating an invariant metric on the object space with a right action. One can show that the group provided with this right-invariant metric is a complete metric space [44, 48]: the choice of the norm on the Lie algebra is specifying the subgroup of diffeomorphisms which are reachable, i.e. which are at a finite distance.

In the SVF setting, the fact that the flow in an autonomous ODE allows us to generalize efficient algorithms such as the scaling and squaring algorithm: given an approximation $\exp(\delta Y) = id + \delta Y$ for small vector fields δY, the exponential of a SVF Y can be efficiently and simply computed by recursive compositions:

$$\exp(Y) = \exp\left(\frac{Y}{2}\right) \circ \exp\left(\frac{Y}{2}\right) = \left(\exp\left(\frac{Y}{2^n}\right)\right)^{2^n}.$$

A second algorithm is at the heart of the efficiency of the optimization algorithms with SVFs: the Baker-Campbell-Hausdorff (BCH) formula [9] tells us how to approximate the log of the composition:

$$\begin{aligned} BCH(X, \delta Y) &= \log(\exp(X) \circ \exp(\delta Y)) \\ &= X + \delta Y + \frac{1}{2}[X, \delta Y] + \frac{1}{12}[X, [X, \delta Y]] + \ldots. \end{aligned}$$

In order to have a well-posed space of deformations, we need to specify on which space is modeled the Lie algebra, as previously. This is the role of the regularization term of the SVF registration algorithms [19, 46] or of the spline parametrization of the SVF in [6, 32]: this restricts the Lie algebra to the sub-algebra of sufficiently regular velocity fields. The subgroup of diffeomorphisms considered is then generated by the flow of these stationary velocity fields and their finite composition. So far, the theoretical framework is very similar to the LDDMM setting and we can see that the diffeomorphisms generated by the one-parameter subgroups (the exponential of SVFs) all belong to the group considered in the LDDMM setting, provided that we model the Lie algebra on the same admissible Hilbert space. As in finite dimension, the affine geodesics of the Cartan connections (group geodesics) are metric-free (the Hilbert metric is only used to specify the space on which is modeled the Lie Algebra) and generally differ from the Riemannian geodesics of LDDMM.

It is well known that the subgroup of diffeomorphisms generated by this Lie algebra is significantly larger than what is covered by the group exponential. Indeed, although our affine connection space is geodesically complete (all geodesics can be continued for all time without hitting a boundary), there is no Hopf-Rinow theorem which state that any two points can be joined by a geodesic (metric completeness). Thus, in general, not all the elements of the group $\mathbb{G}$ may be reached by the one-parameter subgroups. An example in finite dimension is given by $SL(2)$.

However, this might not necessarily results into a problem in the image registration context since we are not interested in recovering "all" the possible diffeomorphisms, but only those which lead to admissible anatomical transformations. For instance, the diffeomorphism on the circle defined above at Eq. (9.8) cannot be reached by any one-parameter subgroup of $\mathbb{S}^1$. However, since

$$\lim_{k\to\infty} \|f_{n,\epsilon}\|_{H^k} \to \infty,$$

this function is not well behaved from the regularity point of view, which is a critical feature when dealing with image registration.

In practice, we have a spatial discretization of the SVF (and of the deformations) on a grid, and the temporal discretization of the time varying velocity fields by a fixed number of time steps. This intrinsically limits the frequency of the deformation below a kind of "Nyquist" threshold, which prevents these diffeomorphisms to be reached anyway both by the SVF and by the "discrete" LDDMM frameworks. Therefore, it seems more importance to understand the impact of using stationary velocity fields in registration from the *practical* point of view, than from the theoretical point of view, because we will have necessarily to deal with the unavoidable numerical implementation and relative approximation issues.

9.6 Parallel Transport in Diffeomorphic Registration

Continuous and discrete methods for the parallel transport provided in Sects. 9.3 and 9.4 can be applied in the diffeomorphic registration setting, once provided the appropriate geometrical context (Sect. 9.5). In this section we discuss and illustrate practical implementations of the parallel transport in diffeomorphic registration, with special focus on the application of the ladder schemes exposed in Sect. 9.4.

9.6.1 Continuous Versus Discrete Transport Methods

As illustrated in abstract form in [2], parallel transport can be approximated infinitesimally by Jacobi fields. Following this intuition, a computational method for the parallel transport along geodesics of diffeomorphisms provided with a right invariant metric was proposed in the LDDMM context by [49]. This framework enables to transport diffeomorphic deformations of point supported and image data, and it was applied to study the hippocampal shape changes in Alzheimer's disease [35, 36]. Although it represents a rigorous implementation of the parallel transport, it comes to the price of the computationally intense scheme. More importantly, it is limited to the transport along geodesics of the considered right invariant metric, and does not allow to specify different metrics for longitudinal and inter-subject registrations. While from the theoretical point of view parallel transporting along a generic curve can be approximated by the parallel transport on piecewise geodesics, the effectiveness of the above methods was shown only on LDDMM geodesics, and no general computational schemes were provided.

The parallel transport in the SVF setting was investigated in [28], in which explicit formula for the parallel transport with respect to the standard Cartan-Schouten connections (left, right and symmetric) in the case of *finite dimensional Lie groups* were derived. Then it was proposed to seamlessly apply these formulas in the *infinite dimensional* case of the diffeomorphic registration of images. Although further investigations would be needed to better understand the impact of generalizing to infinite dimensions the concepts defined for the Lie Group theory in finite dimension, practical examples of parallel transport of longitudinal diffeomorphisms in synthetic and real images with respect to the Cartan-Schouten connections showed to be an effective and simple way to transport tangent SVFs. In particular the parallel transport of the left, right and symmetric Cartan connection defined in Eqs. (9.3), (9.4), and (9.5) was directly applied to longitudinal deformations and compared against the discrete transport implemented with the Schild's ladder. These experiments highlighted the central role of the numerical implementation on stability and accuracy of the methods. For instance, the left and symmetric Cartan transports were much less stable than the right one because they involve the computation of the Jacobian matrix, computed here with standard finite differences. More robust numerical schemes to

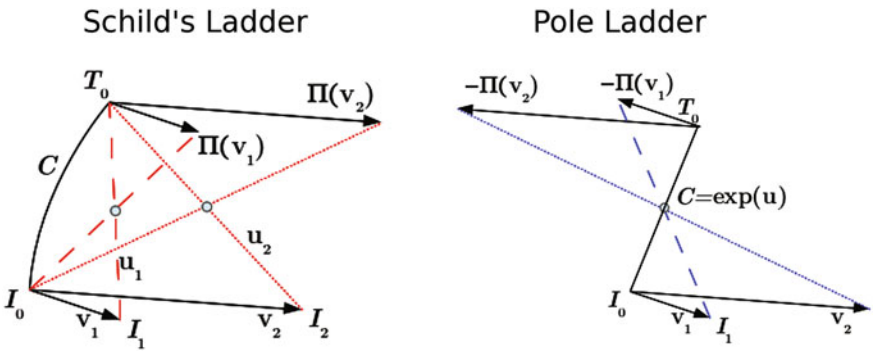


Fig. 9.3 Geometrical schemes in the Schild's ladder and in the pole ladder. By using the curve C as diagonal, the pole ladder requires the computation of half times of the geodesics (*blue*) required by the Schild's ladder (*red*)

compute differential operators on discrete image grids are definitely required to compare them on a fair basis.

9.6.2 Discrete Ladders: Application to Image Sequences

Let I_i $(i = 1 \dots n)$ be a time series of images with the baseline I_0 as reference. Consider a template image T_0, the aim of the procedure is to compute the image T_i in order to define the transport of the sequence $I_0, \dots, I_i$ in the reference of T_0. In the sequel, we focus on the transport of a single image I_1.

Algorithm 1: Schild's ladder for the transport of a longitudinal deformation.

Let I_0 and I_1 be a series of images, and T_0 a reference frame.
1. Compute the geodesic $l(\lambda)$ in the space $\mathbb{I}$ connecting I_1 and T_0 such that $l(0) = I_1$, and $l(1) = T_0$.
2. Define the half-space image $l(1/2) = I_{\frac{1}{2}}$.
3. Compute the geodesic $r(\rho)$ connecting I_0 and $I_{\frac{1}{2}}$ such that $r(0) = I_0$ and $r(1) = I_{\frac{1}{2}}$.
4. Define the transported follow-up image as $T_1 = r(2)$.
5. The transported deformation is given by registering the images T_0 and T_1.

We assume that a well posed Riemannian metric is given on the space of images. This could be L_2, H_k or the metric induced on the space of images by the action of diffeomorphisms of a well chosen right-invariant metric (LDDMM).

Schild's ladder can be naturally translated in the image context (Algorithm 1), by requiring the computation of two diagonal geodesics.

The pole ladder is similar to the Schild's one, with the difference of explicitly using as a diagonal the geodesic C which connects I_0 and T_0 (Algorithm 2). This is an interesting property since, given C, it requires the computation of only one additional geodesic, thus the transport of time series of several images is based on the same baseline-to-reference curve C (Fig. 9.3).

Algorithm 2: Pole ladder for the transport of a longitudinal deformation.

Let I_0 and I_1 be a series of images, and T_0 a reference frame.
1. Compute the geodesic $C(\mu)$ in the space $\mathbb{I}$ connecting I_0 and T_0 such that $C(0) = I_0$ and $C(1) = T_0$.
2. Define the half-space image $C(1/2) = I_{\frac{1}{2}}$.
3. Compute the geodesic $g(\eta)$ connecting I_1 and $I_{\frac{1}{2}}$ such that $g(0) = I_1$ and $g(1) = I_{\frac{1}{2}}$.
4. Define the transported image as $T_1' = g(2)$
5. Compute the path $p(t)$ such that $p(0) = T_0$ and $p(1) = T_1'$.
   The transported deformation is the inverse of the registration of $p(0) = T_0$ to $p(1) = T_1'$.

9.6.2.1 Lifting the Transport to Diffeomorphisms

Despite the straightforward formulation, algorithms (1) and (2) require multiple evaluations of image geodesics, and a consequent a high cost in terms of computation time and resources if we compute them with registration. Moreover, since we look for regular transformations of the space, the registration is usually constrained to be smooth and the perfect match of correspondent intensities in the registered images is not possible. For instance, the definition of $I_{\frac{1}{2}}$ using the forward deformation on I_1 or the backward from T_0 would lead to different results. Since we work in computational anatomy with deformations, it seems more natural to perform the parallel transport directly in the group of diffeomorphisms

9.6.3 Effective Ladders Within the SVF Setting

Given a pair of images I_i, $i \in \{0, 1\}$, the SVF framework parametrizes the diffeomorphism φ required to match the reference I_0 to the moving image I_1 by a SVF u. The velocity field u is an element of the Lie Algebra $\mathfrak{g}$ of the Lie group of diffeomorphisms G, i.e. an element of the tangent space at the identity $T_{id}G$. The diffeomorphism φ belongs to the one parameter subgroup $\varphi = \exp(t\, u)$ generated by the flow of u.

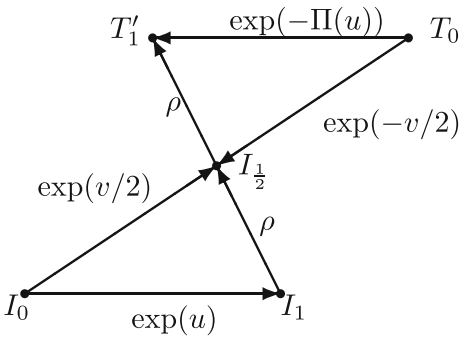


Fig. 9.4 Ladder with the one parameter subgroups. The transport $\exp(\Pi(u))$ is the deformation $\exp(v/2) \circ \exp(u) \circ \exp(-v/2)$

We can therefore define the paths in the space of the diffeomorphisms from the one parameter subgroup parametrization $l(\lambda) = \exp(\lambda \cdot u)$.

Figure 9.4 illustrate how we can take advantage of the stationarity properties of the one-parameter subgroup in order to define the following robust scheme:

1. Let $I_1 = \exp(u) * I_0$.
2. Compute $v = argmin_{v \in \mathbb{G}} E\,(T_0 \circ \exp(-v/2), I_0 \circ \exp(v/2))$, where E is a generic registration energy functional to be minimized.
   The half space image $I_{\frac{1}{2}}$ can be defined in terms of $v/2$ as $\exp(-v/2) * T_0$ or $\exp(v/2) * I_0$. While from the theoretical point of view the two images are identical, the choice of one of them, or even their mean, introduces a bias in the construction. The definition of the half step image can be bypassed by relying on the symmetric construction of the parallelogram.
3. The transformation from I_1 to $I_{\frac{1}{2}}$ is $\rho = \exp(v/2) \circ \exp(-u)$ and the symmetry leads to $\exp(\Pi(u)) = \exp(v/2) \circ \rho^{-1} = \exp(v/2) \circ \exp(u) \circ \exp(-v/2)$.

The transport of the deformation $\varphi = \exp(u)$ can be therefore obtained through the conjugate action operated by the deformation parametrized by $v/2$.

Since the direct computation of the conjugation by composition is potentially biased by the spatial discretization, we propose a numerical scheme to more robustly evaluate the transport directly in the Lie Algebra.

9.6.3.1 BCH Formula for the Conjugate Action

The Baker Campbell Hausdorff (BCH) formula was introduced in the SVF diffeomorphic registration in [9] and provides an explicit way to compose diffeomorphisms parametrized by SVFs by operating in the associated Lie Algebra. More specifically, if v, u are SVFs, then $\exp(v) \circ \exp(u) = \exp(w)$ with $w = BCH(v, u) = v + u + \frac{1}{2}[v, u] + \frac{1}{12}[v, [v, u]] - \frac{1}{12}[u, [v, u]] + \ldots$. In particular, for small u, the computation can be truncated to any order to obtain a valid approximation for the composition of diffeomorphisms. Applying the truncate BCH to the conjugate action

leads to

$$\Pi_{BCH}(u) \simeq u + [v/2, u] + \frac{1}{2}[v/2, [v/2, u]]. \tag{9.9}$$

To establish this formula, let consider the following second order truncation of the BCH formula

$$BCH((v/2, u) \simeq v/2 + u + \frac{1}{2}[v/2, u] + \frac{1}{12}[v/2, [v/2, u]] - \frac{1}{12}[u, [v/2, u]].$$

The composition

$$\Pi_{BCH}(u) = BCH\,(v/2, BCH(u, -v/2))$$

is

$$\Pi(u)^v = \underbrace{v/2 + BCH(u, -v/2)}_{A} + \underbrace{\frac{1}{2}[v/2, BCH(u, -v/2)]}_{B} + \underbrace{\frac{1}{12}[v/2, [v/2, BCH(u, -v/2)]}_{C} - \underbrace{\frac{1}{12}[BCH(u, -v/2), [v/2, BCH(u, -v/2)]]}_{D}.$$

The second order truncation of the four terms is:

$$\begin{aligned}
A &\simeq u + \frac{1}{2}[u, -v/2] + \frac{1}{12}[u, [u, -v/2]] - \frac{1}{12}[-v/2, [u, -v/2]], \\
B &\simeq \frac{1}{2}[v/2, u] + \frac{1}{4}[v/2, [u, -v/2]], \\
C &\simeq \frac{1}{12}[v/2, [v/2, u]], \\
D &\simeq -\frac{1}{12}[u, [v/2, u]] + \frac{1}{12}[v/2, [v/2, u]].
\end{aligned}$$

From the additive and anticommutative properties of the Lie bracket, adding the four terms leads to (9.9).

9.6.3.2 Iterative Computation of the Ladder

Once defined the formula for the computation of the ladder, we need a consistent scheme for the iterative construction along trajectories. We recall that the transport by geodesic parallelograms holds only if both sides of the parallelogram are sufficiently small, which in our case means that both longitudinal and inter-subject vectors must be small. This is not the case in practice, since the inter-subject deformation is usually very large. By definition, the ladder requires to scale down vectors to a

sufficiently small neighborhood, in order to correctly approximate the transport by parallelograms.

From the theoretical point of view, the degree of approximation of the ladder is approximately proportional to the curvature of the space of deformations. This can be seen by the higher order terms that we dropped off in the proof of Sect. 9.4.2, which are all derivatives of the Christoffel symbols. While on a linear space the ladder is the exact parallel transport, when working on curved spaces the error resulting from the non-infinitesimal geodesic parallelogram is proportional to the distance between the points.

From the numerical point of view, we notice that Formula (9.9) requires the computation of the Lie brackets of the velocity fields. Lie brackets involve the differentiation of the vector which is usually computed on images by finite differences, and which are know to be very sensitive to noise and to be unstable in case of large deformations.

For all these reasons we propose the following iterative scheme based on the properties of SVFs. To provide a sufficiently small vector for the computation of the conjugate we observe that

$$\begin{aligned} &\exp(v) \circ \exp(u) \circ \exp(-v) \\ &= \exp\left(\frac{v}{n}\right) \circ \ldots \circ \exp\left(\frac{v}{n}\right) \circ \exp(u) \circ \exp\left(-\frac{v}{n}\right) \circ \ldots \circ \exp\left(-\frac{v}{n}\right) \end{aligned}$$

The conjugation can then be recursively computed in the following way:

1. Scaling step. Find n such that v/n is small.
2. Ladder step. Compute $w = u + \left[\frac{v}{n}, u\right] + \frac{1}{2}\left[\frac{v}{n}, \left[\frac{v}{n}, u\right]\right]$.
3. Let $u = w$.
4. Iterate the steps 2 and 3 n times.

The BCH formula allows to perform the transport directly in the Lie algebra and avoids multiple exponentiation and interpolations, thus reducing the bias introduced by the numerical approximations. Moreover, this method preserves the original "ladder" formulation, operated along the inter-subject geodesic $\exp(tv)$. In fact it iterates the construction of the ladder along the path $\exp(tv)$ over small steps of size $\exp(\frac{v}{n})$.

The stability of the proposed method critically depends from the initial scaling step n, which determines the step-size of the numerical scheme. Ideally the step-size should depend on the curvature, and should be therefore small enough in order to minimize the error in case of highly curved space. For this purpose, given the image domain Ω, we define a global scaling factor n in order to guarantee that the given SVF stays sufficiently close to 0, i.e. in order to satisfy the global condition $max_{x\in\Omega}\|v(x)\|/n < \delta$, with $\delta = 0.5 * voxel\_size$. This condition ensures reasonably small SVFs, and thus enables the iterative construction of the parallelogram in small neighborhoods.

9.6.4 Pole Ladder for Estimating Longitudinal Changes in Alzheimer's Disease

We provide here an application of the pole ladder for the estimation of a group-wise model of the longitudinal changes in a group of patients affected by Alzheimer's disease (AD). AD is a neurodegenerative pathology of the brain, characterized by the co-occurrence of different phenomena, starting from the deposition of amyloid plaques and neurofibrillary tangles, to the development of functional loss and finally to cell deaths [20]. In particular brain atrophy detectable from magnetic resonance imaging (MRI) is currently considered as a potential outcome measure for the monitoring of the disease progression. Structural atrophy was shown to strongly correlate with cognitive performance and neuropsychological scores, and characterizes the progression from pre-clinical to pathological stages [20]. For this reason, the development of reliable atlases of the pathological longitudinal evolution of the brain is of paramount importance for improving the understanding of the pathology.

A preliminary approach to the group-wise analysis of longitudinal morphological changes in AD consists in performing the longitudinal analysis after the subject-to-template normalization [13, 43]. A key issue here is the different nature of the changes occurring at the intra-subject level, which reflects the biological phenomena of interest, and the changes across different subjects, which are usually large and not related to any biological process. In fact, the inter-subject variability is a scale of magnitude higher than the more subtle longitudinal subject-specific variations. To provide a more sensitive quantification of the longitudinal dynamics, the intra-subject changes should be modeled independently from the subject-to-template normalization, and only transported in the common reference for statistical analysis afterward. Thus, novel techniques such as the parallel transport of longitudinal deformations might lead to better accuracy and precision for the modeling and quantification of longitudinal pathological brain changes.

9.6.4.1 Data Analysis and Results

Images corresponding to the baseline I_0 and the one-year follow-up I_1 scans were selected for 135 subjects affected by Alzheimer's disease. For each subject i, the pairs of scans were rigidly aligned. The baseline was linearly registered to a reference template and the parameters of the transformation were applied to I_1^i. Finally, for each subject, the longitudinal changes were measured by non-linear registration using the LCC-Demons algorithm [25].

The resulting deformation fields $\varphi_i = \exp(v_i)$ were transported with the pole ladder (BCH scheme) in the template reference along the subject-to-template deformation. The group-wise longitudinal progression was modeled as the mean of the transported SVFs v_i. The areas of significant longitudinal changes were investigated by one-sample t-test on the group of log-Jacobian scalar maps corresponding

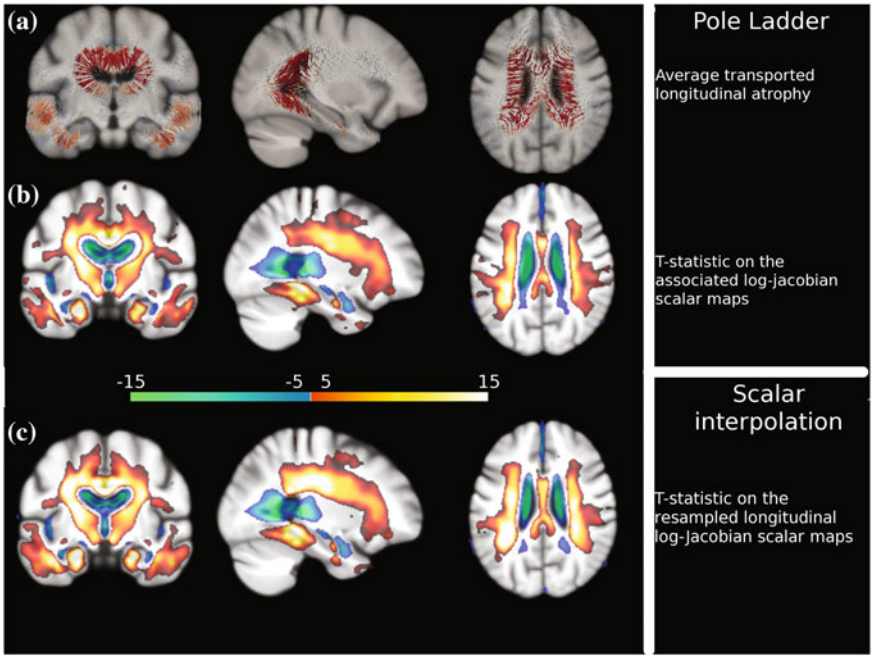


Fig. 9.5 One year structural changes for 135 Alzheimer's patients. **a** Mean of the longitudinal SVFs transported in the template space with the pole ladder. We notice the lateral expansion of the ventricles and the contraction in the temporal areas. **b** T-statistic for the correspondent log-Jacobian values significantly different from 0 ($p < 0.001$ FDR corrected). **c** T-statistic for longitudinal log-Jacobian scalar maps resampled from the subject to the template space. *Blue color* significant expansion, *Red color* significant contraction. The figure is reproduced from [27]

to the transported deformations, in order to detect the areas of measured expansion/contraction significantly different from zero.

For the sake of comparison, the one sample t-statistic was tested on the subject specific longitudinal log-Jacobian scalar maps warped into the template space along the subject-to-template deformation. This is the classical transport used in tensor's based morphometry studies [4].

Figure 9.5 shows a detail from the mean SVF from the transported one-year longitudinal trajectories. The field flows outward from the ventricles indicates a pronounced enlargement. Moreover, we notice an expansion in the temporal horns of the ventricles as well as a consistent contracting flow in the temporal areas. The same effect can be statistically quantified by evaluating the areas where the log-Jacobian maps are statistically different from zero. The areas of significant expansion are located around the ventricles and spread in the CSF areas, while a significant contraction is appreciable in the temporal lobes, hippocampi, parahippocampal gyrus and in the posterior cingulate. The statistical result is in agreement with the one provided by the simple scalar interpolation of the longitudinal subject specific log-

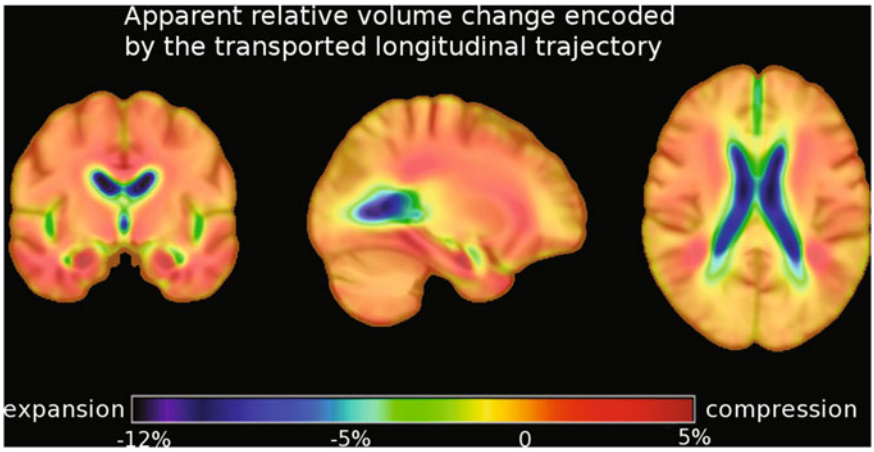


Fig. 9.6 Apparent relative volume changes encoded by the average longitudinal trajectory computed with the pole ladder (Fig. 9.5). The trajectory describes a pattern of apparent volume gain in the CSF areas, and of apparent volume loss in temporal areas and around the ventricles

Jacobian maps. In fact we do not experience any substantial loss of localization power by transporting SVFs instead of scalar log-Jacobian maps. However by parallel transporting we preserve also the multidimensional information of the SVFs that, as experienced in [26], potentially leads to more powerful voxel-by-voxel comparisons than the ones obtained with univariate tests on scalars.

Finally, Fig. 9.6 shows that the apparent volume changes associated to the average trajectory computed with the pole ladder describe biologically plausible dynamics of longitudinal atrophy. We notice that the estimated one-year longitudinal trajectory is associated to local volume changes ranging from $+12\,\%$ for the expansion of the ventricles, to $5\,\%$ for the volume loss of the hippocampi.

We recall that the apparent expansion of the CSF areas is detectable thanks to the diffeomorphic registration constraint. In fact, since the deformation is spatially smooth, the apparent volume loss (e.g. brain atrophy) detectable in the image is modeled as voxels shrinkage, which is associated to the enlargement of the surrounding areas (e.g. ventricles).

9.7 Conclusions

This chapter illustrates the principles of parallel transporting in transformation groups, with particular focus on discrete transport methods. The use of discrete transport is motivated from both theoretical and practical point of view. In fact, discrete methods such as the pole ladder are based only on the computation of geodesics, and thus they do not require the explicit knowledge of the connection of the space.

This is a rather interesting characteristic that enables to employ the ladder without requiring the design of any additional tool outside geodesics.

From the practical point of view, discrete methods can alleviate the numerical problems arising from the discretization of continuous functional on finite grids, and thus provide feasible and numerically stable alternatives to continuous transport approaches. The application shown in Sect. 9.6.4 is a promising example of the potential of such approaches when applied to challenging problems such as the estimation of longitudinal atlases in diffeomorphic registration.

As shown in Sect. 9.4 the construction of the ladder holds in sufficiently small neighborhoods. From the practical point of view this is related to the choice of an appropriate step size for the iterative scheme proposed in Sect. 9.6, and future studies are required in order to investigate the impact of the step size from the numerical point of view.

Finally, future studies aimed to directly compare discrete versus continuous approaches might shed more light on the theoretical and numerical properties of different methods of transport.

References

1. Ardekani, S., Weiss, R.G., Lardo, A.C., George, R.T., Lima, J.A.C., Wu, K.C., Miller, M.I., Winslow, R.L., Younes, L.: Cardiac motion analysis in ischemic and non-ischemic cardiomyopathy using parallel transport. In: Proceedings of the Sixth IEEE International Conference on Symposium on Biomedical Imaging: From Nano to Macro, ISBI'09, pp. 899–902. IEEE Press, Piscataway (2009)
2. Arnold, V.I.: Mathematical Methods of Classical Mechanics, vol. 60. Springer, New York (1989)
3. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Proceedings of Medical Image Computing and Computer-Assisted Intervention - MICCAI, vol. 9, pp. 924–931. Springer, Heidelberg (2006)
4. Ashburner, J., Ridgway, G.R.: Symmetric diffeomorphic modeling of longitudinal structural MRI. Front. Neurosci. 6, (2012)
5. Ashburner, J., Friston, K.J.: Voxel-based morphometry—the methods. NeuroImage **11**, 805–21 (2000)
6. Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**(1), 95–113 (2007)
7. Avants, B., Anderson, C., Grossman, M., Gee, J.: Spatiotemporal normalization for longitudinal analysis of gray matter atrophy in frontotemporal dementia. In Ayache, N., Ourselin, S., Maeder, A. (eds.) Medical Image Computing and Computer-Assisted Intervention, MICCAI, pp. 303–310. Springer, Heidelberg (2007)
8. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
9. Bossa, M., Hernandez, M., Olmos, S.: Contributions to 3d diffeomorphic atlas estimation: application to brain images. In: Proceedings of Medical Image Computing and Computer-Assisted Intervention- MICCAI, vol. 10, pp. 667–74 (2007)
10. Bossa, M.N., Zacur, E., Olmos, S.: On changing coordinate systems for longitudinal tensor-based morphometry. In: Proceedings of Spatio Temporal Image Analysis Workshop (STIA), (2010)

11. Cartan, E., Schouten, J.A.: On the geometry of the group-manifold of simple and semi-simple groups. Proc. Akad. Wekensch (Amsterdam) **29**, 803–815 (1926)
12. Charpiat, G.: Learning shape metrics based on deformations and transport. In: Second Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment, Kyoto, Japon (2009)
13. Chetelat, G., Landeau, B., Eustache, F., Mezenge, F., Viader, F., de la Sayette, V., Desgranges, B., Baron, J.-C.: Using voxel-based morphometry to map the structural changes associated with rapid conversion to mci. NeuroImage **27**, 934–46 (2005)
14. Thompson, D.W.: On growth and form by D'Arcy Wentworth Thompson. University Press, Cambridge (1945)
15. Davis, B.C., Fletcher, P.T., Bullit, E., Joshi, S.: Population shape regression from random design data. In: ICCV vol.4, pp. 375–405 (2007)
16. do Carmo, M.P.: Riemannian Geometry. Mathematics. Birkhäuser, Boston, Basel, Berlin (1992)
17. Durrleman, S., Pennec, X., Trouvé, A., Gerig, G., Ayache, N.: Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI, vol. 12, pp. 297–304 (2009)
18. Helgason, S.: Differential Geometry, Lie groups, and Symmetric Spaces. Academic Press, New York (1978)
19. Hernandez, M., Bossa, M., Olmos, S.: Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows. Int. J. Comput. Vis. **85**, 291–306 (2009)
20. Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.: Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. Lancet Neurol. **9**, 119–28 (2010)
21. Joshi, S., Miller, M.I.: Landmark matching via large deformation diffeomorphisms. IEEE Trans. Image Process. **9**(8), 1357–70 (2000)
22. Khesin, B.A., Wendt, R.: *The Geometry of Infinite Dimensional Lie groups*, volume 51 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer (2009)
23. Kheyfets, A., Miller, W., Newton, G.: Schild's ladder parallel transport for an arbitrary connection. Int. J. Theoret. Phys. **39**(12), 41–56 (2000)
24. Kolev, B.: Groupes de Lie et mécanique. http://www.cmi.univ-mrs.fr/kolev/. Notes of a Master course in 2006–2007 at Université de Provence (2007)
25. Lorenzi, M., Ayache, N., Frisoni, G.B., Pennec, X.: Lcc-demons: a robust and accurate symmetric diffeomorphic registration algorithm. NeuroImage **1**(81), 470–83 (2013)
26. Lorenzi, M., Ayache, N., Frisoni, G.B., Pennec, X.: Mapping the effects of $A\beta_{1-42}$ levels on the longitudinal changes in healthy aging: hierarchical modeling based on stationary velocity fields. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI, pp. 663–670, (2011)
27. Lorenzi, M., Pennec, X.: Efficient parallel transport of deformations in time series of images: from Schild's to pole ladder. J. Math. Imaging Vis. (2013) (Published online)
28. Lorenzi, M., Pennec, X.: Geodesics, parallel transport and one-parameter subgroups for diffeomorphic image registration. Int. J. Comput. Vis.—IJCV **105**(2), 111–127 (2012)
29. Lorenzi, M., Ayache, N., Pennec, X.: Schild's ladder for the parallel transport of deformations in time series of images. Inf. Process. Med. Imaging—IPMI **22**, 463–74 (2011)
30. Milnor, J.: Remarks on infinite-dimensional Lie groups. In: Relativity, Groups and Topology, pp. 1009–1057. Elsevier Science Publishers, Les Houches (1984)
31. Misner, C.W., Thorne, K.S., Wheeler, J.A.: Gravitation. W.H. Freeman and Compagny , San Francisco, California (1973)
32. Modat, M., Ridgway, G.R., Daga, P., Cardoso, M.J., Hawkes, D.J., Ashburner, J., Ourselin, S.: Log-Euclidean free-form deformation. In: Proceedings of SPIE Medical Imaging 2011. SPIE, (2011)
33. Pennec, X., Arsigny, V.: Exponential barycenters of the canonical cartan connection and invariant means on Lie groups. In: Barbaresco, F., Mishra, A., Nielsen, F. (eds.) Matrix Information Geometry. Springer, Heidelberg (2012)

34. Postnikov, M.M.: Geometry VI: Riemannian Geometry. Encyclopedia of mathematical science. Springer, Berlin (2001)
35. Qiu, A., Younes, L., Miller, M., Csernansky, J.G.: Parallel transport in diffeomorphisms distinguish the time-dependent pattern of hippocampal surface deformation due to healthy aging and dementia of the Alzheimer's type. NeuroImage, **40**(1):68–76 (2008)
36. Qiu, A., Albert, M., Younes, L., Miller, M.: Time sequence diffeomorphic metric mapping and parallel transport track time-dependent shape changes. NeuroImage **45**(1), S51–60 (2009)
37. Rao, A., Chandrashekara, R., Sanchez-Hortiz, G., Mohiaddin, R., Aljabar, P., Hajnal, J., Puri, B., Rueckert, D.: Spatial trasformation of motion and deformation fields using nonrigid registration. IEEE Trans. Med. Imaging **23**(9), 1065–76 (2004)
38. Riddle, W.R., Li, R., Fitzpatrick, J.M., DonLevy, S.C., Dawant, B.M., Price, R.R.: Characterizing changes in mr images with color-coded jacobians. Magn. Reson. Imaging **22**(6), 769–77 (2004)
39. Schild, A.: Tearing geometry to pieces: More on conformal geometry. unpublished lecture at Jan 19 1970 Princeton University relativity seminar (1970)
40. Schmid, R.: Infinite dimensional lie groups with applications to mathematical physics. J. Geom. Symmetry Phys. **1**, 1–67 (2004)
41. Schmid, R.: Infinite-dimensional lie groups and algebras in mathematical physics. Adv. Math. Phys. **2010**, 1–36 (2010)
42. Subbarao, R.: Robust Statistics Over Riemannian Manifolds for Computer Vision. Graduate School New Brunswick, Rutgers The State University of New Jersey, New Brunswick, (2008)
43. Thompson, P., Ayashi, K.M., Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman, D., Hong, M.S., Dittmer, S.S., Dodrell, D.M., Toga, A.W.: Dynamics of gray matter loss in alzheimer's disease. J. Neurosci. **23**(3), 994–1005 (2003)
44. Trouvé, A.: Diffeomorphisms groups and pattern matching in image analysis. Int. J. Comput. Vis. **28**(3), 213–21 (1998)
45. Twining, C., Marsland, S., Taylor, C.: Metrics, connections, and correspondence: the setting for groupwise shape analysis. In: Proceedings of the 8th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, EMMCVPR'11, pp. 399–412. Springer, Berlin, Heidelberg (2011)
46. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Symmetric log-domain diffeomorphic registration: a demons-based approach. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI. Lecture Notes in Computer Science, vol. 5241, pp. 754–761. Springer, Heidelberg (2008)
47. Wei, D., Lin, D., Fisher, J.: Learning deformations with parallel transport. In: ECCV, pp. 287–300 (2012)
48. Younes, L.: Shapes and diffeomorphisms. Number 171 in Applied Mathematical Sciences. Springer, Berlin (2010)
49. Younes L.: Jacobi fields in groups of diffeomorphisms and applications. Q. Appl. Math. pp. 113–134 (2007)

Chapter 10
Diffeomorphic Iterative Centroid Methods for Template Estimation on Large Datasets

Claire Cury, Joan Alexis Glaunès and Olivier Colliot

Abstract A common approach for analysis of anatomical variability relies on the estimation of a template representative of the population. The Large Deformation Diffeomorphic Metric Mapping is an attractive framework for that purpose. However, template estimation using LDDMM is computationally expensive, which is a limitation for the study of large datasets. This chapter presents an iterative method which quickly provides a centroid of the population in the shape space. This centroid can be used as a rough template estimate or as initialization of a template estimation method. The approach is evaluated on datasets of real and synthetic hippocampi segmented from brain MRI. The results show that the centroid is correctly centered within the population and is stable for different orderings of subjects. When used as an initialization, the approach allows to substantially reduce the computation time of template estimation.

C. Cury (✉) · O. Colliot
UPMC Univ Paris 06, UM 75, Sorbonne Universités, ICM, F-75013 Paris, France
e-mail: claire.cury.pro@gmail.com

C. Cury · O. Colliot
UMR 7225, CNRS, ICM, F-75013, Paris, France

C. Cury · O. Colliot
U1127, Inserm, ICM, F-75013, Paris, France

C. Cury · O. Colliot
ICM, Institut du Cerveau et de la Moëlle épinière, 47 bd de l'hôpital, Paris, France

C. Cury · O. Colliot
Aramis project-team, Inria Paris-Rocquencourt, Paris, France

J. A. Glaunès
MAP5, Université Paris Descartes, Sorbonne Paris Cité, France

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_10,

10.1 Introduction

Large imaging datasets are being increasingly used in neuroscience, thanks to the wider availability of neuroimaging facilities, the development of computing infrastructures and the emergence of large-scale multi-center studies. Such large-scale datasets offer increased statistical power which is crucial for addressing questions such as the relationship between anatomy and genetics or the discovery of new biomarkers using machine learning techniques for instance.

Computational anatomy aims at developing tools for the quantitative analysis of variability of anatomical structures, its variation in healthy and pathological cases and relations between functions and structures [1]. A common approach in computational anatomy is template-based analysis, where the idea is to compare anatomical objects by analyzing their variations relatively to a common template. These variations are analyzed using the ambient space deformations that match each individual structure to the template.

A common requirement is that transformations must be diffeomorphic in order to preserve the topology and to consistently transform coordinates. The Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [2, 3] has been widely used for the study of the geometric variation of human anatomy, of intra-population variability and inter-population differences. It focuses the study on the spatial transformations which can match subject's anatomies one to another, or one to a template structure which needs to be estimated. These transformations not only provide a diffeomorphic correspondence between shapes, but also define a metric distance in shape space.

Several methods have been proposed to estimate templates in the LDDMM framework [4–8]. Vaillant et al. proposed in 2004 [4] a method based on geodesic shooting which iteratively updates a shape by shooting towards the mean of directions of deformations from this shape to all shapes of the population. The method proposed by Glaunès and Joshi [5] starts from the whole population and estimates a template by co-registering all subjects. The method uses a backward scheme: deformations are defined from the subjects to the template. The method optimizes at the same time the deformations between subjects and the template, and the template itself. The template is composed, in the space of currents (more details on Sect. 10.2.2), by all surfaces of the population. A different approach was proposed by Durrleman et al. [6, 7]. The method initializes the template with a standard shape, in practice it is often an ellipsoid. The method uses a forward scheme: deformations are defined from the template to the subjects. Again, it optimizes at the same time the deformations and the template. The template is composed by one surface which presents the same configuration as the initial ellipsoid. The method presented by Ma et al. [8] introduces an hyper template which is an extra fixed shape (which can be a subject of the population). The method aims at optimizing at the same time deformations from the hyper template to the template and deformations from the template to subjects of the population. The template is optimized via the deformation of the hyper template, not directly.

A common point of all these methods is that they need a surface matching algorithm, which is very expensive in terms of computation time in the LDDMM framework. When no specific optimization is used, computing only one matching between two surfaces, each composed of 3000 vertices, takes approximately 30–40 min. Then, computing a template composed of one hundred such surfaces until convergence can take a few days or some weeks. This is a limitation for the study of large databases. Different strategies can be used to reduce computation time. GPU implementation can substantially speed up the computation of convolutions that are heavily used in LDDMM deformations. Matching pursuit on current can also be used to reduce the computation time [9]. Sparse representations of deformations allow to reduce the number of optimized parameters of the deformations [7].

Here, we propose a new approach to reduce the computation time called diffeomorphic iterative centroid using currents. The method provides in $N-1$ steps (with N the number of shapes of the population) a centroid already correctly centered within the population of shapes. It increases the convergence speed of the template estimation by providing an initialization that is closer to the target.

Our method has some close connections with more general iterative methods to compute means on Riemannian manifolds. For example Arnaudon et al. [10] defined a stochastic iterative method which converges to the Fréchet mean of the set of points. Ando et al. [11] gave a recursive definition of the mean of positive definite matrices which verifies important properties of geometric means. However these methods require a large number of iterations (much larger than the number of points of the dataset), while in our case, due to the high computational cost of matchings, we aim at limiting as much as possible the number of iterations.

The chapter is organized as follows. First, we present the mathematical framework of LDDMM and currents (Sect. 10.2). Section 10.3.1 then introduces the template estimation and the iterative centroid method. In Sect. 10.4.1, we evaluate the approach on datasets of real and synthetic hippocampi extracted from brain magnetic resonance images (MRI).

10.2 Notation and Mathematical Setup

For the Diffeomorphic Iterative Centroid method, we use the LDDMM framework (Sect. 10.2.1) to quantify the difference between shapes. To model surfaces of the population, we use the framework of currents (Sect. 10.2.2) which does not assume point-to-point correspondences.

10.2.1 Large Diffeomorphic Deformations

The Large Diffeomorphic Deformation Metric Mapping framework allows quantifying the difference between shapes and provides a shape space representation: shapes of the population are seen as points in an infinite dimensional smooth manifold,

providing a continuum between shapes of the population. In this framework a diffeomorphism deforms the whole space, not only a shape.

Diffeomorphisms as flows of vector fields In the LDDMM framework, deformation maps $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ are generated by integration of time-dependent vector fields $v(x, .)$ in an Hilbert space V, with $x \in \mathbb{R}^3$ and $t \in [0, 1]$. If $v(x, t)$ is regular enough, i.e. if we consider the vector fields $(v(\cdot, t))_{t\in[0,1]}$ in $L^2([0, 1], V)$, where V is a Reproducing Kernel Hilbert Space (R. K. H. S.) embedded in the space of $C^1(\mathbb{R}^3, \mathbb{R}^3)$ vector fields vanishing at infinity, then the transport equation:

$$\begin{cases} \frac{d\phi_v}{dt}(x, t) = v(\phi_v(x, t), t) & \forall t \in [0, 1] \\ \phi_v(x, 0) = x & \forall x \in \mathbb{R}^3 \end{cases} \tag{10.1}$$

has a unique solution, and one sets $\varphi_v = \phi_v(\cdot, 1)$ the diffeomorphism induced by $v(x, t)$. The induced set of diffeomorphisms $\mathcal{A}_V$ is a subgroup of the group of C^1 diffeomorphisms. To enforce velocity fields to stay in this space, one must control the energy

$$E(v) := \int_0^1 \|v(\cdot, t)\|_V^2 \; dt. \tag{10.2}$$

Metric structure on the diffeomorphisms group The induced subgroup of diffeomorphisms $\mathcal{A}_V$ is equipped with a right-invariant metric defined by the rules: $\forall \varphi, \psi \in \mathcal{A}_V$,

$$\begin{cases} D(\varphi, \psi) = D(Id, \varphi^{-1} \circ \psi) \\ D(Id, \varphi) = \inf \left\{ \int_0^1 \|v(\cdot, t)\|_V \; dt \; ; \; v \in L^2([0, 1], V), \varphi_v = \varphi \right\} \end{cases} \tag{10.3}$$

$D(\varphi, \psi)$ represents the shortest length of paths connecting φ to ψ in the diffeomorphisms group. Moreover, as in the classical Riemannian theory, minimizing the length of paths is equivalent to minimizing their energy, and one has also:

$$D(Id, \varphi)) = \inf \left\{ E(v) \; ; \; v \in L^2([0, 1], V), \varphi_v = \varphi \right\} \tag{10.4}$$

Discrete matching functionals Considering two surfaces S and T, the optimal matching between them is defined in an ideal setting, as the map φ_v minimizing $E(v)$ under the constraint $\varphi_v(S) = T$. In practice such an exact matching is often not feasible and one writes inexact unconstrained matching functionals which minimize both $E(v)$ and a matching criterion which evaluates the spatial proximity between $\varphi_v(S)$ and T, as we will see in the next section.

In a discrete setting, when the matching criterion depends only on φ_v via the images $\varphi_v(x_i)$ of a finite number of points x_i (such as the vertices of the mesh S) one can show that the vector fields $v(x, t)$ which induce the optimal deformation map

can be written via a convolution formula over the surface involving the reproducing kernel K_V of the R.K.H.S. V. This is due to the reproducing property of V; indeed V is the closed span of vectors fields of the form $K_V(x, .)\alpha$, and therefore $v(x, t)$ writes

$$v(x,t) = \sum_{i=1}^{n} K_V(x, x_i(t))\alpha_i(t), \tag{10.5}$$

where $x_i(t) = \phi_v(x_i, t)$ are the trajectories of points x_i, and $\alpha_i(t) \in \mathbb{R}^3$ are time-dependent vectors called momentum vectors, which parametrize completely the deformation. Trajectories $x_i(t)$ depend only on these vectors as solutions of the following system of ordinary differential equations:

$$\frac{dx_j(t)}{dt} = \sum_{i=1}^{n} K_V(x_j(t), x_i(t))\alpha_i(t), \tag{10.6}$$

for $1 \leq j \leq n$. This is obtained by plugging formula 10.5 for the optimal velocity fields into the flow equation 10.1 taken at $x = x_j$. Moreover, the energy $E(v)$ takes an explicit form as expressed in terms of trajectories and momentum vectors:

$$E(v) = \int_0^1 \sum_{i,j=1}^{n} \alpha_i(t)^T K_V(x_i(t), x_j(t))\alpha_j(t)\, dt. \tag{10.7}$$

These equations reformulate the problems in a finite dimensional Riemannian setting. Indeed $E(v)$ appears as the energy of the path $t \mapsto (x_i(t))_{1\leq i\leq n}$ in the space of landmarks $\mathcal{L}^n = \{\mathbf{x} = (x_i)_{1\leq i\leq n}, x_i \neq x_j \quad \forall i, j\}$ equipped with local metric $g(\mathbf{x}) = K(\mathbf{x})^{-1}$, where $K_V(\mathbf{x})$ is the $3n \times 3n$ matrix with block entries $K_V(x_i, x_j)$, $1 \leq i, j \leq n$.

Geodesic equations and local encoding As introduced previously, the minimization of the energy $E(v)$ in matching problems can be interpreted as the estimation of a length-minimizing path in the group of diffeomorphisms $\mathcal{A}_V$, and also additionally as a length-minimizing path in the space of landmarks when considering discrete problems. Such length-minimizing paths obey some geodesic equations [4] (distances are define as in 10.3), which write as follows in the case of landmarks (using matrix notations):

$$\begin{cases} \frac{d\mathbf{x}(t)}{dt} = K_V(\mathbf{x}(t))\boldsymbol{\alpha}(t) \\ \frac{d\boldsymbol{\alpha}(t)}{dt} = -\frac{1}{2}\nabla_{\mathbf{x}(t)}\left[\boldsymbol{\alpha}(t)^T K_V(\mathbf{x}(t))\boldsymbol{\alpha}(t)\right], \end{cases} \tag{10.8}$$

Note that the first equation is nothing more than Eq. 10.6 which allows to compute trajectories $x_i(t)$ from any time-dependant momentum vectors $\alpha_i(t)$, while the second equation gives the evolution of the momentum vectors themselves. This new

set of ODEs can be solved from any initial conditions $x_i(0)$, $\alpha_i(0)$, which means that the initial momentum $\alpha_i(0)$ fully determine the subsequent time evolution of the system (since the $x_i(0)$ are fixed points). As a consequence, these initial momentum vectors encode all information of the optimal diffeomorphism. This is a very important point for applications, specifically for group studies, since it allows to analyse the set of deformation maps from a given template to the observed shapes by performing statistics on the initial momentum vectors located on the template shape. We also can use geodesic shooting from initial conditions $(x_i(0), \alpha_i(0))$ in order to generate any arbitrary deformation of a shape in the shape space. We will use this tool for the construction of our synthetic dataset Data1 (see Sect. 10.4.1).

10.2.2 Currents

The idea of the mathematical object named "currents" is related to the theory of distributions as presented by Schwartz in 1952 [12], in which distributions are characterized by their action on any smooth functions with compact support. In 1955, De Rham [13] generalized distributions to differential forms to represent submanifolds, and called this representation currents. This mathematical object serves to model geometrical objects using a non parametric representation.

The use of currents in computational anatomy was introduced by J. Glaunés and M. Vaillant in 2005 [14, 15] and subsequently developed by Durrleman [16] in order to provide a dissimilarity measure between meshes which does not assume point-to-point correspondence between anatomical structures. The approach proposed by Vaillant and Glaunès is to represent meshes as objects in a linear space and supply it with a computable norm. Using currents to represent surfaces has some benefits. First it avoids the point correspondence issue: one does not need to define pairs of corresponding points between two surfaces to evaluate their spatial proximity. Moreover, metrics on currents are robust to different samplings and topologies and take into account not only the global shapes but also their local orientations. Another important benefit is that the space of currents is a vector space, which allows to consider linear combinations such as means of shapes in the space of currents. This property will be used in the centroid and template methods that we introduce in the following.

We limit the framework to surfaces embedded in $\mathbb{R}^3$. Let S be an oriented compact surface, possibly with boundary. Any smooth and compactly supported differential 2-form ω of $\mathbb{R}^3$—i.e. a mapping $x \mapsto \omega(x)$ such that for any $x \in \mathbb{R}^3$, $\omega(x)$ is a 2-form, an alternated bilinear mapping from $\mathbb{R}^3 \times \mathbb{R}^3$ to $\mathbb{R}$—can be integrated over S

$$\int_S \omega = \int_S \omega(x)(u_1(x), u_2(x))d\sigma(x). \tag{10.9}$$

where $(u_1(x), u_2(x))$ is an orthonormal basis of the tangent plane at point x, and $d\sigma$ the Lebesgue measure on the surface S. Hence one can define a linear form $[S]$ over

the space of 2-forms via the rule $[S](\omega) := \int_S \omega$. If one defines a Hilbert metric on the space of 2-forms such that the corresponding space is continuously embedded in the space of continuous bounded 2-forms, this mapping will be continuous [14], which will make $[S]$ an element of the space of 2-currents, the dual space to the space of 2-forms.

Note that since we are working with 2-forms on $\mathbb{R}^3$, we can use a vectorial representation via the cross product: for every 2-form ω and $x \in \mathbb{R}^3$ there exists a vector $\overline{\omega}(x) \in \mathbb{R}^3$ such that for every $\alpha, \beta \in \mathbb{R}^3$,

$$\omega(x)(\alpha, \beta) = \langle \overline{\omega}(x) \ , \ \alpha \times \beta \rangle = det(\alpha, \beta, \overline{\omega}(x)), \tag{10.10}$$

Therefore we can work with vector fields $\overline{\omega}$ instead of 2-forms ω. In the following, with a slight abuse of notation, we will use $\omega(x)$ to represent both the bilinear alternated form and its vectorial representative. Hence the current of a surface S can be re-written from Eq. 10.9 as follows:

$$[S](\omega) = \int_S \langle \omega(x) \ , \ n(x) \rangle \ d\sigma(x) \tag{10.11}$$

with $n(x)$ the unit normal vector to the surface: $n(x) := u_1(x) \times u_2(x)$.

We define a Hilbert metric $\langle \cdot \ , \ \cdot \rangle_W$ on the space of vector fields of $\mathbb{R}^3$, and require the space W to be continuously embedded in $C_0^1(\mathbb{R}^3, \mathbb{R}^3)$. The space of currents we consider is the space of continuous linear forms on W, i.e. the dual space W^*, and the required embedding property ensures that for a large class of oriented surfaces S in $\mathbb{R}^3$, comprising smooth surfaces and also triangulated meshes, the associated linear mapping $[S]$ is indeed a current, i.e. it belongs to W^*.

The central object from the computational point of view is the reproducing kernel of space W, which we introduce here. For any point $x \in \mathbb{R}^3$ and vector $\alpha \in \mathbb{R}^3$ one can consider the Dirac functional $\delta_x^\alpha : \omega \mapsto \langle \omega(x) \ , \ \alpha \rangle$ which is an element of W^*. The Riesz representation theorem then states that there exists a unique $u \in W$ such that for all $\omega \in W$, $\langle u \ , \ \omega \rangle_W = \delta_x^\alpha(\omega) = \langle \omega(x) \ , \ \alpha \rangle$. u is thus a vector field which depends on x and linearly on α, and we write it $u = K_W(\cdot, x)\alpha$. Thus we have the rule

$$\langle K_W(\cdot, x)\alpha, \omega \rangle_W = \langle \omega(x) \ , \ \alpha \rangle \, . \tag{10.12}$$

Moreover, applying this formula to $\omega = K_W(\cdot, y)\beta$ for any other point $y \in \mathbb{R}^3$ and vector $\beta \in \mathbb{R}^3$, we get

$$\begin{aligned} \langle K_W(\cdot, x)\alpha, K_W(\cdot, y)\beta \rangle_W &= \left\langle K_W(x, y)\beta \ , \ \alpha \right\rangle \\ &= \alpha^T K_W(x, y)\beta = \left\langle \delta_x^\alpha \ , \ \delta_y^\beta \right\rangle_{W^*} . \end{aligned} \tag{10.13}$$

$K_W(x, y)$ is a 3×3 matrix, and the mapping $K_W : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ is called the reproducing kernel of the space W. Now, note that we can rewrite Eq. 10.11 as

$$[S](\omega) = \int_S \delta_x^{n(x)}(\omega)\, d\sigma(x) \tag{10.14}$$

Thus using Eq. 10.13, one can prove that for two surfaces S and T,

$$\langle [S]\ ,\ [T] \rangle_{W^*}^2 = \int_S \int_T \left\langle n_S(x)\ ,\ K_W(x, y) n_T(y) \right\rangle\, ds(x) ds(y) \tag{10.15}$$

This formula defines the metric we use for evaluating spatial proximity between shapes. It is clear that the type of kernel one uses fully determines the metric and therefore will have a direct impact on the behaviour of the algorithms. We use scalar invariant kernels of the form $K_W(x, y) = h(\|x - y\|^2/\sigma_W^2) I_3$, where h is a real function such as $h(r) = e^{-r}$ (gaussian kernel) or $h(r) = 1/(1 + r)$ (Cauchy kernel), and σ_W a scale factor. In practice this scale parameter has a strong influence on the results; we will go back to this point later.

We can now define the optimal match between two currents $[S]$ and $[T]$, which is the diffeomorphism minimizing the functional

$$J_{S,T}(v) = \gamma E(v) + \|[\varphi_v(S)] - [T]\|_{W^*}^2 \tag{10.16}$$

This functional is non convex and in practice we use a gradient descent algorithm to perform the optimization, which cannot guarantee to reach a global minimum. We observed empirically that local minima can be avoided by using a multi-scale approach in which several optimization steps are performed with decreasing values of the width σ_W of the kernel K_W (each step provides an initial guess for the next one).

In practice, surfaces are given as triangulated meshes, which we discretize in the space of currents W^* by combinations of Dirac functionals: $[S] \simeq \sum_{f\in S} \delta_{c_f}^{n_f}$, where the sum is taken over all triangles $f = (f^1, f^2, f^3)$ of the mesh S, and $c_f = \frac{1}{2}(f^1 + f^2 + f^3)$, $n_f = \frac{1}{2}(f^2 - f^1) \times (f^3 - f^1)$ denote respectively the center and normal vector of the triangle. Given a deformation map φ and a triangulated surface S, we also approximate its image $\varphi(S)$ by the triangulated mesh obtained by letting ϕ act only on the vertices of S. This leads us to the following discrete formulation of the matching problem:

$$\begin{aligned} J_{S,T}^d(\alpha) = &\gamma \int_0^1 \sum_{i=1}^n \alpha_i(t)^T K_V(x_i(t), x_j(t)) \alpha_j(t)\, dt \\ &+ \sum_{f, f' \in S} n_{\varphi(f)}^T K_W(c_{\varphi(f)}, c_{\varphi(f')}) n_{\varphi(f')} \end{aligned}$$

$$+ \sum_{g,g' \in T} n_g K_W(c_g, c_{g'}) n_{g'} - 2 \sum_{f \in S, g \in T} n^T_{\varphi(f)} K_W(c_{\varphi(f)}, c_g) n_g \quad (10.17)$$

where φ denotes the diffeomorphism associated to momentum vectors $\alpha_i(t)$ and trajectories $x_i(t)$, $x_i = x_i(0)$ being the vertices of mesh S, and where we have noted for any face f, $\varphi(f) = (\varphi(f^1), \varphi(f^2), \varphi(f^3))$. We note 3 important parameters, γ which controls the regularity of the map, σ_V which controls the scale in the space of deformations and σ_W which controls the scale in the space of currents.

10.3 A Template Estimation for Large Database via LDDMM

10.3.1 Why Building a Template?

A central notion in computational anatomy is the generation of registration maps, mapping a large set of anatomical data to a common coordinate system to study intra-population variability and inter-population differences. In this chapter, we use the method introduced by Glaunès et al. [5] which estimates a template given a collection of unlabelled points sets or surfaces in the framework of scalar measures and currents. In our case we use the framework of currents. This method is posed as a minimum mean squared error estimation problem and uses the metric on the space of diffeomorphisms. Let S_i be N surfaces in $\mathbb{R}^3$ (i.e. the whole surface population). Let $[S_i]$ be the corresponding current of S_i, or its approximation by a finite sum of vectorial Diracs. The problem is formulated as follows:

$$\left\{\hat{v}_i, \hat{\mathcal{T}}\right\} = \arg\min_{v_i, \mathcal{T}} \sum_{i=1}^{N} \left\{ \| \mathcal{T} - \left[\varphi_{v_i}(S_i)\right] \|^2_{W^*} + \gamma E(v_i) \right\}, \quad (10.18)$$

where the minimization is performed over the spaces $L^2([0, 1], V)$ for the velocity fields v_i and over the space of currents W^* for $\mathcal{T}$. The method uses an alternated optimization i.e. surfaces are successively matched to the template, then the template is updated and this sequence is iterated until convergence. One can observe that when φ_i is fixed, the functional is minimized when $\mathcal{T}$ is the average of $[\varphi_i(S_i)]$ in space W^*:

$$\mathcal{T} = \frac{1}{N} \sum_{i=1}^{N} \left[\varphi_{v_i}(S_i)\right], \quad (10.19)$$

which makes the optimization with respect to $\mathcal{T}$ straightforward. This optimal current is not a surface itself; in practice it is constituted by the union of all surfaces $\varphi_{v_i}(S_i)$, and the $\frac{1}{N}$ factor acts as if all normal vectors to these surfaces were weighted by $\frac{1}{N}$.

At the end of the optimization process however, all surfaces being co-registered, the $\hat{\varphi}_{v_i}(S_i)$ are close to each other, which makes the optimal template $\hat{\mathcal{T}}$ close to being a true surface.

In practice, we stop the template estimation method after P loops, and with the datasets we use, $P = 7$ seems to be sufficient to obtain an adequate template.

As detailed in Sect. 10.2, obtaining a template allows to perform statistical analysis of the deformation maps via the initial momentum representation to characterize the population. One can run analysis on momentum vectors such as Principal Component Analysis (PCA), or estimate an approximation of pairwise diffeomorphic distances between subjects using the estimated template [17] in order to use manifold learning methods like Isomap [18].

In the present case, the optimal template for the population is not a true surface but is defined, in the space of currents, by the mean $\hat{\mathcal{T}} = \frac{1}{N}\sum_{j=1}^{N} \hat{\varphi}_{v_j}\left[S_j\right]$. However this makes no difference from the point of view of statistical analysis, because this template can be used in the LDDMM framework exactly as if it was a true surface.

One may speed up the estimation process and avoid local minima issues by defining a good initialization of the optimization process. Standard initialization consists in setting $\mathcal{T} = \frac{1}{N}\sum_{i=1}^{N}[S_i]$, which means that the initial template is defined as the combination of all unregistered shapes in the population. Alternatively, if one is given a good initial guess $\mathcal{T}$, the convergence speed of the method can be improved. This is the primary motivation for the introduction of the iterative centroid method which we present in the next section.

10.3.2 The Iterative Centroid Method

As presented in the introduction, computing a template in the LDDMM framework can be highly time consuming, taking a few days or some weeks for large real-world databases. To increase the speed of the method, one of the key points may be to start with a good initial template, already correctly centred among shapes in the population. Of course the computation time of such an initialization method must be substantially lower than the template estimation itself. The Iterative Centroid method presented here performs such an initialization with $N - 1$ pairwise matchings only.

The LDDMM framework, in an ideal setting (exact matching between shapes), sets the template estimation problem as the computation of a centroid on a Riemannian manifold, which is of finite dimension in the discrete case (we limit our analysis to this finite dimensional setting in what follows). The Fréchet mean is the standard way for defining such a centroid and provides the basic inspiration of all LDDMM template estimation methods. Since our Iterated Centroid method is also inspired by considerations about computation of centroids in Euclidean space and their analogues on Riemannian manifolds, we will briefly discuss these ideas in the following.

Centroid computation on Euclidean and Riemannian spaces If $x_i, 1 \leq i \leq N$ are points in $\mathbb{R}^d$, then their centroid is defined as

$$b_N = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{10.20}$$

It satisfies also the following:

$$b_N = \arg\min_{y \in \mathbb{R}^d} \sum_{1 \leq i \leq N} \|y - x_i\|^2. \tag{10.21}$$

Now, when considering points x_i living on a Riemannian manifold M (we assume M is path-connected and geodesically complete), the definition of b_N cannot be used because M is not a vector space. However the variational characterization of b_N has an analogue, which leads to the definition of the Fréchet mean, also called 2-mean, which is uniquely defined under some constraints (see [10]) on the relative locations of points x_i in the manifold:

$$b_N = \arg\min_{y \in M} \sum_{1 \leq i \leq N} d_M(y, x_i)^2. \tag{10.22}$$

Many mathematical studies (as for example Kendall [19], Karcher [20] Le [21], Afsari [22, 23]), have focused on proving the existence and uniqueness of the mean, as well as proposing algorithms to compute it. The more general notion of p-mean of a probability measure μ on a Riemannian manifold M is defined by:

$$b = \arg\min_{x \in M} F_p(x), \quad F_p(x) = \int_M d_M(x, y)^p \mu(dy). \tag{10.23}$$

Arnaudon et al. [10] published in 2012 for $p \geq 1$ a stochastic algorithm which converges almost surely to the p-mean of the probability measure μ. This algorithm does not require to compute the gradient of the functional F_p to minimize. The authors construct a time inhomogeneous Markov chain by choosing at each step a random point P with distribution μ and moving the current point X to a new position along the geodesic connecting X to P. As it will be obvious in the following, our method shares similarities with this method for the case $p = 2$, in that it also uses an iterative process which at each step moves the current position to a new position along a geodesic. However our method is not stochastic and does not compute the 2-mean of the points. Moreover, our approach stops after $N - 1$ iterations, while on the contrary the stochastic method does not ensure to have considered all subjects of the population after N iterations.

Other definitions of centroids in the Riemannian setting can be proposed. The following ideas are more directly connected to our method. Going back to the Euclidean case, one can observe that b_N satisfies the following iterative relation:

$$\begin{cases} b_1 = x_1 \\ b_{k+1} = \frac{k}{k+1} b_k + \frac{1}{k+1} x_{k+1}, \quad 1 \leq k \leq N-1, \end{cases} \tag{10.24}$$

which has the side benefit that at each step b_k is the centroid of the x_i, $1 \leq i \leq k$. This iterative process has an analogue in the Riemannian case, because one can interpret the convex combination $\frac{k}{k+1} b_k + \frac{1}{k+1} x_{k+1}$ as the point located along the geodesic linking b_k to x_{k+1}, at a distance equal to $\frac{1}{k+1}$ of the total length of the geodesic, which we can write $geod(b_k, x_{k+1}, \frac{1}{k+1})$. This leads to the following definition in the Riemannian case:

$$\begin{cases} \tilde{b}_1 = x_1 \\ \tilde{b}_{k+1} = \text{geod}(\tilde{b}_k, x_{k+1}, \frac{1}{k+1}), \quad 1 \leq k \leq N-1, \end{cases} \tag{10.25}$$

Of course this new definition of centroid does not coincide with the Fréchet mean when the metric is not Euclidean, and furthermore it has the drawback to depend on the ordering of points x_i. Moreover one may consider other iterative procedures such as computing midpoints between arbitrary pairs of points x_i, and then midpoints of the midpoints, etc. In other words, all procedures that are based on decomposing the Euclidean equality $b_N = \frac{1}{N} \sum_{i=1}^{N} x_i$ as a sequence of pairwise convex combinations lead to possible alternative definitions of centroid in a Riemannian setting. Based on these remarks, Emery and Mokobodzki [24] proposed to define the centroid not as a unique point but as the set B_N of points $x \in M$ satisfying

$$f(x) \leq \frac{1}{N} \sum_{i=1}^{N} f(x_i), \tag{10.26}$$

for any convex function f on M (a convex function f on M being defined by the property that its restriction to all geodesics is convex). This set B_N takes into account all centroids obtained by bringing together points x_i by all possible means, i.e. recursively by pairs, or by iteratively adding a new point, as explained above (see Fig. 10.2).

Outline of the method The Iterated Centroid method consists roughly in applying the following procedure: given a collection of N shapes S_i, we successively update the centroid by matching it to the next shape and moving along the geodesic flow. Figure 10.1 illustrates the general idea. We propose two alternative ways for the update step (algorithms 1 and 2 below).

Direct iterative centroid: IC1 The first version of the method computes a centroid between two objects O_1 and O_2 by transporting a first object O_1 along the geodesic going from this object to O_2. The transport is stopped depending of the weights of

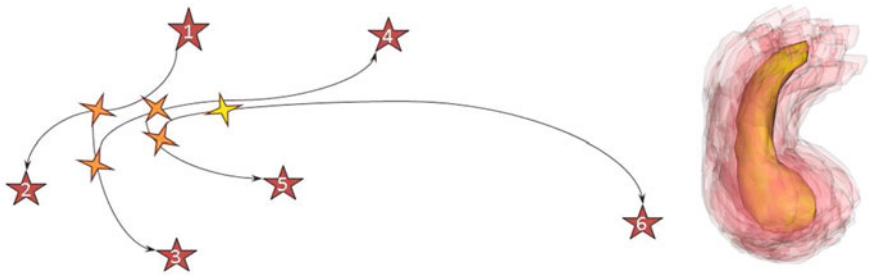


Fig. 10.1 Illustration of the method. *Left image red stars* are subjects of the population, the *yellow star* is the final Centroid, and *orange stars* are iterations of the centroid. *Right image* Final centroid with the hippocampus population from Data1 (*red*). See Sect. 10.4.1 for more details about datasets

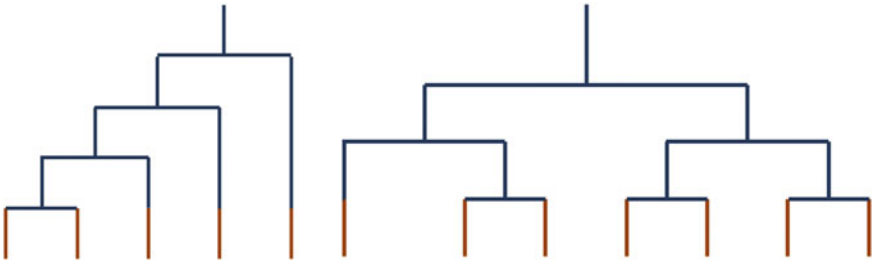


Fig. 10.2 Diagrams of the iterative processes which lead to the centroid computation. The *tops* of the diagrams represent the final centroid. The diagram on the *left* corresponds to the iterative centroid algorithms (IC1 and IC2). The diagram on the *right* corresponds to the pairwise algorithm (PW)

the objects. If the weight of O_1 is w_1, and the weight of O_2 is w_2 with $w_1 + w_2 = 1$, we stop the deformation of O_1 at time $t = w_2$. Since the method is iterative, the first two objects are two subjects of the population, for the next step we have as a first object the previous centroid and as a second object a new subject of the population. The algorithm proceeds as presented in the Algorithm 1.

Data: N surfaces S_i
Result: 1 surface B_N representing the centroid of the population
$B_1 = S_1$;
for *i from* 1 *to* $N - 1$ **do**
  B_i is matched using the Eq. (10.16) to S_{i+1} which results in a deformation map $\phi_{v_i}(x, t)$;
  Set $B_{i+1} = \phi_{v_i}(B_i, \frac{1}{i+1})$ which means we transport B_i along the geodesic and stop at time $t = \frac{1}{i+1}$;
end

Algorithm 1: Iterative Centroid 1 (IC1)

Iterative centroid with averaging in the space of currents: IC2 Because matchings are inaccurate, the centroid computed with the method presented above accumulates small errors which can have an impact on the final centroid. Furthermore, the centroid computed with algorithm 1 is in fact a deformation of the first shape S_1, which makes the procedure even more dependent on the ordering of subjects than it would be in an ideal exact matching setting. In this second algorithm, we modify the updating step by computing a mean in the space of currents between the deformation of the current centroid and the backward flow of the curent shape being matched. Hence the computed centroid is not a true surface but a current, i.e. combination of surfaces, as in the template estimation method. The weights chosen in the averaging reflects the relative importance of the new shape, so that at the end of the procedure, all shapes forming the centroid have equal weight $\frac{1}{N}$. The algorithm proceeds as presented in Algorithm 2.

Data: N surfaces S_i
Result: 1 current $\mathcal{B}_N$ representing the centroid of the population
$\mathcal{B}_1 = [S_1]$;
for *i from* 1 *to* $N-1$ **do**
  $\mathcal{B}_i$ is matched using the Eq. (10.16) to S_{i+1} which results in a deformation map $\phi_{v_i}(x,t)$;
  Set $\mathcal{B}_{i+1} = \frac{i}{i+1} * \phi_{v_i}(\mathcal{B}_i, \frac{1}{i+1}) + \frac{1}{i+1}[\phi_{u_i}(S_{i+1}, \frac{i}{i+1})]$ which means we transport $\mathcal{B}_i$ along the geodesic and stop at time $t = \frac{1}{i+1}$;
  where $u_i(x,t) = -v_i(x, 1-t)$, i.e. ϕ_{u_i} is the reverse flow map.
end

Algorithm 2: Iterative Centroid 2 (IC2)

Note that we have used the notation $\phi_{v_i}(\mathcal{B}_i, \frac{1}{i+1})$ to denote the transport (push-forward) of the current $\mathcal{B}_i$ by the diffeomorphism. Here $\mathcal{B}_i$ is a linear combination of currents associated to surfaces, and the transported current is the linear combination (keeping the weights unchanged) of the currents associated to the transported surfaces.

An alternative method: Pairwise Centroid (*PW*) Another possibility is to group objects by pairs, compute centroids (middle points) for each pair, and then recursively apply the same procedure to the set of centroids, until having only one centroid (see Fig. 10.2). This pairwise method also depends on the ordering of subjects, and also provides a centroid which satisfies the definition of Emery and Mokobodzki (disregarding the inaccuracy of matchings).

When the population is composed of more than 3 subjects, we split the population in two parts and recursively apply the same splitting until having two or three objects in each group. We then apply algorithm 1 to obtain the corresponding centroid before going back up along the dyadic tree, and keeping attention to the weight of each object. This recursive algorithm is described in algorithm 3.

Data: N surfaces S_i
Result: 1 surface B representing the centroid of the population
if $N \geq 2$ **then**
  B_{left} = Pairwise Centroid (S_1, ..., $S_{[N/2]}$);
  B_{right} = Pairwise Centroid ($S_{[N/2]+1}$, ..., S_N);
  B_{left} is matched to B_{right} which results in a deformation map $\phi_v(x, t)$;
  Set $B = \phi_v(B_{left}, \frac{[N/2]+1}{N})$ which means we transport B_{left} along the geodesic and stop at time $t = \frac{[N/2]+1}{N}$;
end
else
  $B = S_1$
end

Algorithm 3: Pairwise Centroid

10.3.3 Implementation

The methods presented just before need some parameters. Indeed, in each algorithm we have to compute the matching from one surface to another. For each matching we minimize the corresponding functional (see Eq. 10.17 at the end of Sect. 10.2.2) which estimates the news momentum vectors α, which then are used to update the positions of points x_i of the surface. A gradient descent with adaptive step size is used for the minimization of the functional J. Evaluation of the functional and its gradient require numerical integrations of high-dimensional ordinary differential equations, which is done using Euler trapezoidal rule.

The main parameters for computing J are $maxiter$ which is the maximum number of iterations for the adaptive step size gradient descent algorithm, γ for the regularity of the matching, and σ_W and σ_V the sizes of the kernels which control the metric of the spaces W and V.

We selected parameters in order to have relatively good matchings in a short time. We chose γ close enough to zero to enforce the matching to bring the first object to the second one. Nevertheless, we must be prudent: choosing a γ too small could be hazardous because the regularity of the deformation could not be preserved. For each pairwise matching, we use the multi-scale approach described in Sect. 10.2.2 page 5, performing four consecutive optimization processes with decreasing values by a constant factor of the σ_W parameter which is the size of the R. K. H. S. W, to increase the precision of the matching. At the beginning, we fix this σ_W parameter with a sufficient large value in order the capture the possible important variations or differences between shapes. This is for this reason that for the two first minimizations of the functional, we use a small $maxiter$ parameter. For the results presented after, we used very small values for the parameter $maxiter = [50, 50, 100, 300]$, to increase the velocity of the method. Results can be less accurate than in our previous study [25] which used different values for $maxiter$: [40, 40, 100, 1000], which take twice as much time to compute. For the kernel size σ_V of the deformation space, we fix this parameter at the beginning and have to adapt it to the size of the data.

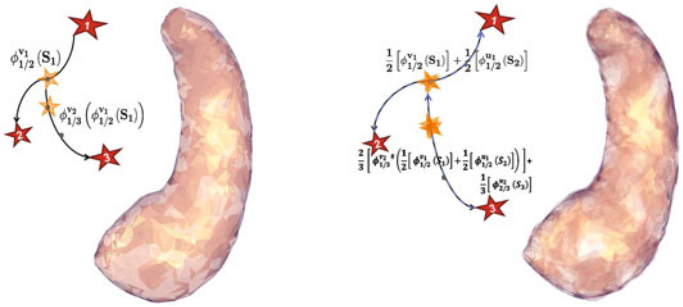


Fig. 10.3 On the *left*, an iterative centroid of the dataset data2 (see Sect. 10.4.1 for more details about datasets) computed using the IC1 algorithm, and on the *right* the IC2 algorithm

The first method starts from N surfaces, and gives a centroid composed by only one surface, which is a deformation of the surface used at the initialization step. An example is shown in Fig. 10.3. This method is rather fast, because at each step we have to match only one mesh composed by n_1 vertices to another, where n_1 is the number of vertices of the first mesh of the iterative procedure.

The second method starts from N surfaces and gives a centroid composed of deformations of all surfaces of the population. At each step it forms a combination in the space of currents between the current centroid and a backward flow of the new surface being matched. In practice this implies that the centroid grows in complexity; at step i its number of vertices is $\sum_{j=1}^{i} j * n_j$. Hence this algorithm is slower than the first one, but the mesh structure of the final centroid does not depend on the mesh of only one subject of the population, and the combination compensates the bias introduced by the inaccuracy of matchings.

The results of the Iterative Centroid algorithms depend on the ordering of subjects. We will study this dependence in the experimental part, and also study the effect of stopping the I. C. before it completes all iterations.

10.4 Experiments and Results

10.4.1 Data

To evaluate our approach, we used data from 95 young (14–16 years old) subjects from the European database IMAGEN. The anatomical structure that we considered was the hippocampus, which is a small bilateral structure of the temporal lobe of the brain involved in memory processes. The hippocampus is one of the first structures to be damaged in Alzheimer's disease; it is also implicated in temporal lobe epilepsy, and is altered in stress and depression. Ninety five left hippocampi were segmented from T1-weighted Magnetic Resonance Images (MRI) of this database (see Fig. 10.4)

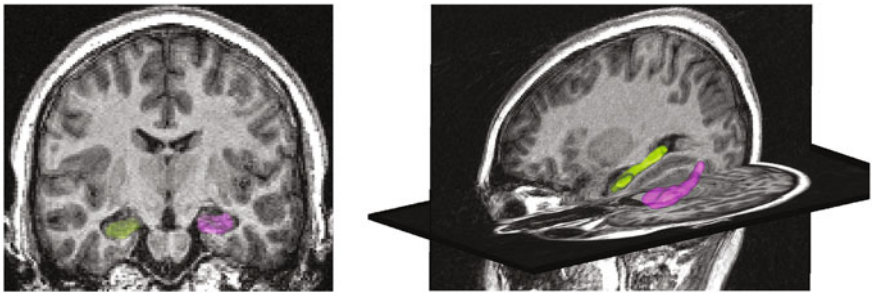


Fig. 10.4 *Left panel* coronal view of the MRI with the meshes of hippocampi segmented by the SACHA software [26], the *right* hippocampus is in *green* and the *left* one in *pink*. *Right panel* 3D view of the hippocampi

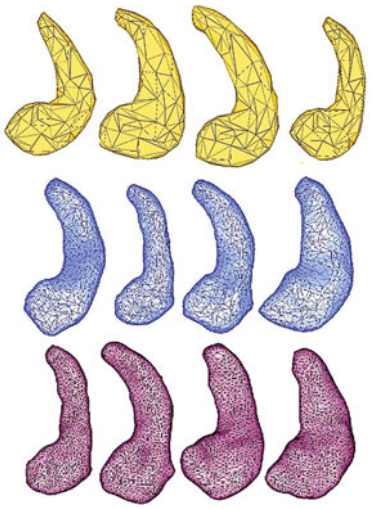


Fig. 10.5 *Top* to *bottom* meshes from Data1 (n = 500), Data2 (n = 95) and RealData (n = 95)

with the software SACHA [26], before computing meshes from the binary masks using BrainVISA software.[1]

We denote as RealData the dataset composed of all 95 hippocampi meshes. We rigidly aligned all hippocampi to one subject of the population. For this rigid registration, we used a similarity term based on measures (as in [27]) rather than currents.

We also built two synthetic populations of hippocampi meshes, denoted as Data1 and Data2. Data1 is composed of a large number of subjects, in order to test our algorithms on a large dataset. In order to study separately the effect of the population size, meshes of this population are simple. Data2 is a synthetic population close to the real one, with the difference that all subjects have the same mesh structure.

[1] http://www.brainvisa.info

This allows to test our algorithms in a population with a single mesh structure, thus disregarding the effects of different mesh structures. These two datasets are defined as follows (examples of subjects from these datasets are shown on Fig. 10.5):

- *Data*1 We chose one subject S_0 that we decimated (down to 135 vertices) and deformed using geodesic shooting in 500 random directions with a sufficiently large kernel and a reasonable momentum vector norm in order to preserve the overall hippocampal shape, resulting in 500 deformed objects. Each deformed object was then further transformed by a translation and a rotation of small magnitude. This resulted in the 500 different shapes of Data1. All shapes in Data1 have the same mesh structure. Data1 thus provides a large dataset with simple meshes and mainly global deformations.
- *Data*2 We chose the same initial subject S_0 that we decimated to 1001 vertices. We matched this mesh to each subject of the dataset RealData ($n = 95$), using diffeomorphic deformation, resulting in 95 meshes with 1001 vertices. Data2 has more local variability than Data1, and is closer to the anatomical truth.

10.4.2 Effect of Subject Ordering

Each of the 3 proposed algorithms theoretically depends on the ordering of subjects. Here, we aim to assess the influence of the ordering of subjects on the final centroid for each algorithm.

For that purpose, we compared several centroids computed with different orderings. For each dataset and each algorithm, we computed 10 different centroids. We computed the mean $m1$ and maximal distance between all pairs of centroids. The three datasets have different variabilities. In order to relate the previous mean distance to the variability, we also computed the mean distance $m2$ between each centroid and all subjects of a given dataset. We finally computed the ratio between these two mean distances $m1/m2$. Distances between surfaces were computed in the space of currents, i.e. to compare two surfaces S and T, we computed the squared norm $\| [S] - [T] \|^2_{W^*}$. Results are presented in Table 10.1. Additionnaly, we computed the mean of distances between centroids computed using the different methods. Results are presented in Table 10.2.

For each dataset and for each type of centroid, the mean of distances between all 10 centroids is small compared to the mean of distances between the centroid and the subjects. However, the three algorithms IC1, IC2 and PW were not equally influenced by the ordering. IC2 seems to be the most stable: the different centroids are very close one to each other, this being true for all datasets. This was expected since we reduce the matching error by combining in the space of currents the actual centroid with the deformation of the new subject along the reverse flow. For IC1, the distance was larger for Data2 and RealData, which have anatomically more realistic deformations, than for Data1, which has rather simplistic shapes. This suggests that, for real datasets, IC1 is more dependent on the ordering than IC2. This is due to the

Table 10.1 Distances between centroids computed with different subjects orderings, for each dataset and each of the 3 algorithms

		From different order			To the dataset	
		mean ($m1$)	max	std	mean ($m2$)	$m1/m2$
Data1	IC1	0.8682	1.3241	0.0526	91.25	0.0095
	IC2	0.5989	0.9696	0.0527	82.66	0.0072
	PW	3.5861	7.1663	0.1480	82.89	0.0433
Data2	IC1	2.4951	3.9516	0.2205	16.29	0.1531
	IC2	0.2875	0.4529	0.0164	15.95	0.0181
	PW	3.8447	5.3172	0.1919	17.61	0.2184
RealData	IC1	4.7120	6.1181	0.0944	18.54	0.2540
	IC2	0.5583	0.7867	0.0159	17.11	0.0326
	PW	5.3443	6.1334	0.1253	19.73	0.2708

The three first columns present the mean, standard deviation and the maximum of distances between all pairs of centroids computed with different orderings. The fourth column displays the mean of distances between each centroid algorithm and all subjects of the datasets. Distances are computed in the space of currents

Table 10.2 In columns, average distances between centroids computed using the different algorithms

	IC1 versus IC2	IC1 versus PW	IC2 versus PW
Data1	1.57	5.72	6.31
Data2	1.89	3.60	3.42
RealData	3.51	5.31	4.96

fact that IC1 provides a less precise estimate of the centroid between two shapes since it does not incorporate the reverse flow. For all datasets, distances for PW were larger than those for IC1 and IC2, suggesting that the PW algorithm is the most dependent on the subjects ordering. Furthermore, centroids computed with PW are also farther from those computed using IC1 or IC2. Furthermore, we speculate that the increased sensitivity of PW over IC1 may be due to the fact that, in IC1, $n-1$ levels of averaging are performed (and only $\log_2 n$ for PW) leading to a reduction of matching errors.

Finally, in order to provide a visualization of the differences, we present matchings between 3 centroids computed with the IC1 algorithm, in the case of RealData. Figure 10.6 shows that shape differences are local and residual. Visually, the 3 centroids are almost similar, and the amplitudes of momentum vectors, which bring one centroid to another, are small and local.

10.4.3 Position of the Centroids Within the Population

We also assessed whether the centroids are close to the center of the population. To that purpose, we calculated the ratio

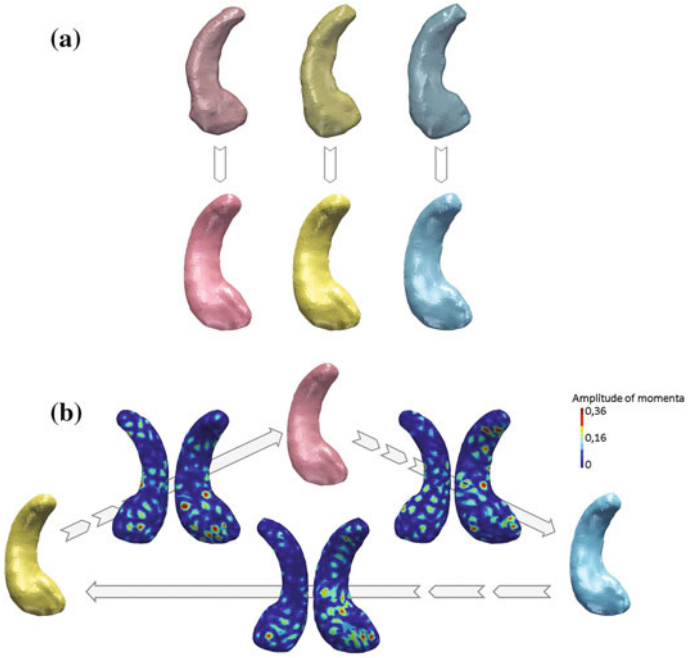


Fig. 10.6 **a** *First row* 3 initial subjects used for 3 different centroid computations with IC1 (mean distance between such centroids, in the space of currents, is 4.71) on RealData. *Second row* the 3 centroids computed using the 3 subjects from the first row as initialization. **b** Maps of the amplitude of the momentum vectors that map each centroid to another. *Top* and *bottom* views of the maps are displayed. One can note that the differences are small and local

$$R = \frac{\| \frac{1}{N} \sum_{i=1}^{N} v_0(S_i) \|_V}{\frac{1}{N} \sum_{i=1}^{N} \| v_0(S_i) \|_V}, \tag{10.27}$$

with $v_0(S_i)$ the vector field corresponding to the initial momentum vector of the deformation from the template or the centroid to the subject i. This ratio gives some indication about the centering of the centroid, because in a pure Riemannian setting (i.e. disregarding the inaccuracies of matchings), a zero ratio would mean that we are at a critical point of the Fréchet functional, and under some reasonable assumptions on the curvature of the shape space in the neighbourhood of the dataset (which we cannot check however), it would mean that we are at the Fréchet mean. To compute R, we need to match the centroid to all subjects of the population. We computed this ratio on the best (i.e. the centroid which is the closest to all other centroids) centroid for each algorithm and for each dataset.

Results are presented in Table 10.3. We can observe that the centroids obtained with the three different algorithms are reasonably centered for all datasets. Centroids for Data1 are particularly well centered, which was expected given the nature of this population. Centroids for Data2 and RealData are slightly less well centered but they

Table 10.3 Ratio values for assessing the position of the representative centroid within the population, computed using Eq. 10.27 (for each algorithm and for each dataset)

R	IC1	IC2	PW
Data1	0.046	0.038	0.085
Data2	0.106	0.102	0.107
RealData	0.106	0.107	0.108

are still close to the Fréchet mean. It is likely that using more accurate matchings (and thus increasing the computation time of the algorithms) we could reduce this ratio for RealData and Data2. Besides, one can note that ratios for Data2 and RealData are very similar; this indicates that the centering of the centroid is not altered by the variability of mesh structures in the population.

10.4.4 Effects of Initialization on Estimated Template

The initial idea was to have a method which provides a good initialization for template estimation methods for large databases. We just saw that IC1 and IC2 centroids are both reasonably centered and do not depend on the subjects ordering. Despite the fact that IC2 has the smallest sensitivity to the subjects ordering, the method is slower and provides a centroid composed of N meshes. Because we want to decrease the computation time for the template estimation of a large database, it is natural to choose as initialization a centroid composed by only one mesh (time saved in kernel convolution) in a short time. We advocate to choose IC1 over PW because we can stop the IC1 algorithm at any step to get a centroid of the sub-population used so far. Furthermore, PW seems to be more sensitive to subjects ordering.

Now, we study the impact of the use of a centroid, computed with the IC1 algorithm, as initialization for the template estimation method presented in Sect. 10.3.1. To that purpose, we compared the template obtained using a standard initialization, denoted as $T(StdInit)$, to the template initialized with IC1 centroid, denoted as $T(IC1)$. We chose to stop the template estimation method after 7 iterations of the optimization process. We arbitrarily chose this number of iterations, it is large enough to have a good convergence for $T(IC1)$ and to have an acceptable convergence for $T(StdInit)$. We did not use a stopping criterion based on the W^* metric because it is highly dependent on the data and is difficult to establish when using a multiscale approach. In addition to comparing $T(IC1)$ to $T(StdInit)$, we also compared the templates corresponding to two different IC1 intialization based on two different orderings. We compared the different templates in the space of currents. Results are presented in Table 10.4. We also computed the same ratios R as in equation 10.27. Results are presented in Table 10.5.

One can note that the differences between $T(IC1)$ for different orderings are small for Data1 and Data2 and larger for RealData, suggesting that these are due to the mesh used for the initialization step. We can also observe that templates initialized

Table 10.4 Distances between templates initialized via differents IC1 ($T(IC1)$) for each datasets, and the distance between template initialized via the standard initialization ($T(StdInit)$) and templates initialized via IC1

	T(IC1) versus T(IC1)	T(IC1) versus T(StdInit)
Data1	0.9833	40.9333
Data2	0.6800	20.4666
RealData	4.0433	26.8667

Table 10.5 Ratios R for templates ($T(IC1)$) and for the template with its usual initialization $T(StdInit)$, for each datasets

R	$T(IC1)$	T(StdInit)
Data1	0.0057	0.0062
Data2	0.0073	0.0077
RealData	0.0073	0.0074

Fig. 10.7 Estimated template from RealData. On the *left*, initialized via the standard initialization which is the whole population. On the *right*, estimated template initialized via a IC1 centroid

via IC1 are far, in terms of distances in the space of currents, from the template initialized by the standard initialization. These results could be alarming, but the results of ratios (see Table 10.5) prove that templates are all very close to the Fréchet mean, and that the differences are not due to a bad template estimation. Moreover, both templates are visually similar as seen on Fig. 10.7.

10.4.5 Effect of the Number of Iterations for Iterative Centroids

Since it is possible to stop the Iterative Centroid methods IC1 and IC2 at any step, we wanted to assess the influence of computing only a fraction of the N iterations on the estimated template. Indeed one may wonder if computing an I.C. at e.g. 40 % (then saving 60 % of computation time for the IC method) could be enough

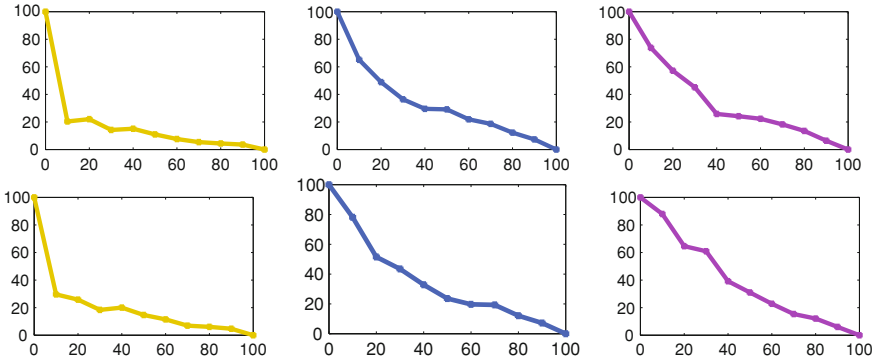


Fig. 10.8 *First row* Graphs of average W^*-distances between the IC1 at $x\%$ and the final one. The *second row* present the same results with IC2

Table 10.6 Results of initialization of template estimation method by a IC1 at 40 %

	Data1	Data2	RealData
T(IC1 at 40 %) versus T(StdInit)	41.41	24.41	24.82
T(IC1 at 40 %) versus T(IC1 at 100 %)	9.36	9.56	6.18
R value for T(IC1 at 40 %)	0.040	0.106	0.105

to initialize a template estimation. Moreover, for large datasets, the last subject will have a very small influence: for a database composed of 1000 subjects, the weight of the last subject is $1/1000$. We performed this experiment in the case of IC1. In the following, we call "IC1 at $x\%$" an IC1 computed using $x \times N/100$ subjects of the population.

We computed the distance in the space of currents between "IC1 at $x\%$" and IC1. Results are presented on Fig. 10.8. These distances are averaged over the 10 centroids computed for each datasets. We can note that after processing 40 % of the population, the IC1 covers more than 75 % of the distance to the final centroid for all datasets.

We also compared T(IC1 at 40 %) to T(IC1) and to T(StdInit), using distances in the space of currents, as well as the R ratio defined in Eq. 10.27. Results are shown on Table 10.6). They show that using 40 % of subjects lowers substantially the quality of the resulting template. Indeed the estimated template seems trapped in the local minimum found by the IC1 at 40 %. We certainly have to take into account the size of the dataset. Nevertheless, we believe that if the dataset is very large and sufficiently homogeneous we could stop the Iterative Centroid method before the end.

10.4.6 Computation Time

To speed up the matchings, we use a GPU implementation for the computation of kernel convolutions, which constitutes the most time-consuming part of LDDMM

Table 10.7 Computation time (in h) for Iterative Centroids and for template estimation initialised by IC1 ($T(IC1)$), the standard initialization (T(StdInit)) and by IC1 at 40 % (T(IC1 at 40 %))

Computation time (h)	Data1	Data2	RealData
IC1	1.7	0.7	1.2
IC2	5.2	2.4	7.5
PW	1.4	0.7	1.2
$T(IC1)$	$21.1(= 1.7 + 19.4)$	$13.3(= 0.7 + 12.6)$	$27.9(= 1.2 + 26.7)$
T(StdInit)	96.1	20.6	99
T(IC1 at 40 %)	$24.4(= 0.7 + 23.7)$	$10.4(= 0.3 + 10.1)$	$40.7(= 0.5 + 40.2)$

For $T(IC1)$, we give the complete time for the whole process i.e. the time for the IC1 computation plus the time for $T(IC1)$ computation itself

methods. Computations were performed on a Nvidia Tesla C1060 card. Computation times are displayed in Table 10.7.

We can note that the computation time of IC1 is equal to the one of PW and that these algorithms are faster than the IC2 algorithm, as expected. The computation time for any IC method (even for IC2) is much lower (by a factor from 10 to 80 %) than the computation time of the template estimation method. Morevoer, initializing the template estimation with IC1 can save up to 70 % of computation time over the standard initialization. On the other hand, using T(IC1 at 40 %) does not reduce computation time compared to using T(IC1).

It could be interesting to evaluate the parameters which would lead to a more precise centroid estimate in a time that would still be inferior to that needed for the template estimation. We should also mention that one could speed up computations by adding a Matching Pursuit on currents as described in [9].

10.5 Conclusion

We have proposed a new approach for the initialization of template estimation methods. The aim was to reduce computation time by providing a rough initial estimation, making more feasible the application of template estimation on large databases.

To that purpose, we proposed to iteratively compute a centroid which is correctly centered within the population. We proposed three different algorithms to compute this centroid: the first two algorithms are iterative (IC1 and IC2) and the third one is recursive (PW). We have evaluated the different approaches on one real and two synthetic datasets of brain anatomical structures. Overall, the centroids computed with all three approaches are close to the Fréchet mean of the population, thus providing a reasonable centroid or initialization for template estimation method. Furthermore, for all methods, centroids computed using different orderings are similar. It can be noted that IC2 seems to be more robust to the ordering than IC1 which in turns seems more robust than PW. Nevertheless, in general, all methods appear relatively robust with respect to the ordering.

The advantage of iterative methods, like IC1 and IC2, is that we can stop the deformation at any step, resulting in a centroid built with part of the population. Thus, for large databases (composed for instance of 1000 subjects), it may not be necessary to include all subjects in the computation since the weight of these subjects will be very small. The iterative nature of IC1 and IC2 provides another interesting advantage which is the possible online refinement of the centroid estimation as new subjects are entered in the dataset. This leads to an increased possibility of interaction with the image analysis process. On the other hand, the recursive PW method has the advantage that it can be parallelized (still using GPU implementation), although we did not implement this specific feature in the present work.

Using the centroid as initialization of the template estimation can substantially speed up the convergence. For instance, using IC1 (which is the fastest one) as initialization saved up 70 % of computation time. Moreover, this method could certainly be used to initialize other template estimation methods, such as the method proposed by Durrleman et al. [6].

As we observed, the centroids, obtained with rough parameters, are close to the Fréchet mean of the population, thus we believe that computing IC with more precise parameters (but still reasonable in terms of computation time), we could obtain centroids closer to the center. This accurate centroid could be seen as a cheap alternative to true template estimation methods, particularly if computing a precise mean of the population of shapes is not required. Indeed, in the LDDMM framework, template-based shape analysis gives only a first-order, linearized approximation of the geometry in shape space. In a future work, we will study the impact of using IC as a cheap template on results of population analysis based for instance on kernel principal component analysis. Finally, the present work deals with surfaces for which the metric based on currents seems to be well-adapted. Nevertheless, the proposed algorithms for centroid computation are general and could be applied to images, provided that an adapted metric is used.

Acknowledgments The authors are grateful to Vincent Frouin, Jean-Baptiste Poline, Roberto Toro and Edouard Duschenay for providing a sample of subjects from the IMAGEN dataset and to Marie Chupin for the use of the SACHA software.The authors thank Professor Thomas Hotz for his suggestion on the pairwise centroid, during the discussion of the GSI'13 conference. The research leading to these results has received funding from ANR (project HM-TC, grant number ANR-09-EMER-006, and project KaraMetria, grant number ANR-09-BLAN-0332) and from the program Investissements d' avenir ANR-10-IAIHU-06.

References

1. Grenander, U., Miller, M.I.: Computational anatomy: an emerging discipline. Q. Appl. Math. **56**(4), 617–694 (1998)
2. Christensen, G.E., Rabbitt, R.D., Miller, M.I.: Deformable templates using large deformation kinematics. IEEE Trans. Image Process. **5**(10), 1435–1447 (1996)

3. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision **61**(2), 139–157 (2005)
4. Vaillant, M., Miller, M.I., Younes, L., Trouvé, A.: Statistics on diffeomorphisms via tangent space representations. Neuroimage **23**, S161–S169 (2004)
5. Glaunès, J., Joshi, S.: Template estimation from unlabeled point set data and surfaces for computational anatomy. In: Pennec, X., Joshi, S., (eds.) Proceedings of the International Workshop on the Mathematical Foundations of Computational Anatomy (MFCA-2006), pp. 29–39, 1 Oct 2006
6. Durrleman, S., Pennec, X., Trouvé, A., Ayache, N., et al.: A forward model to build unbiased atlases from curves and surfaces. In: 2nd Medical Image Computing and Computer Assisted Intervention. Workshop on Mathematical Foundations of Computational Anatomy, pp. 68–79 (2008)
7. Durrleman, S., Prastawa, M., Korenberg, J.R., Joshi, S., Trouvé, A., Gerig, G.: Topology preserving atlas construction from shape data without correspondence using sparse parameters. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012. Lecture Notes in Computer Science, vol. 7512, pp. 223–230. Springer, Berlin (2012)
8. Ma, J., Miller, M.I., Trouvé, A., Younes, L.: Bayesian template estimation in computational anatomy. Neuroimage **42**(1), 252–261 (2008)
9. Durrleman, S., Pennec, X., Trouvé, A., Ayache, N.: Statistical models of sets of curves and surfaces based on currents. Med. Image Anal. **13**(5), 793–808 (2009)
10. Arnaudon, M., Dombry, C., Phan, A., Yang, L.: Stochastic algorithms for computing means of probability measures. Stoch. Process. Appl. **122**(4), 1437–1455 (2012)
11. Ando, T., Li, C.K., Mathias, R.: Geometric means. Linear Algebra Appl. **385**, 305–334 (2004)
12. Schwartz, L.: Théorie des distributions. Bull. Amer. Math. Soc. **58**, 78–85 (1952) 0002–9904
13. de Rham, G.: Variétés différentiables. Formes, courants, formes harmoniques. Actualits Sci. Ind., no. 1222, Publ. Inst. Math. Univ. Nancago III. Hermann, Paris (1955)
14. Vaillant, M., Glaunes, J.: Surface matching via currents. In: Information Processing in Medical Imaging, pp. 381–392. Springer, Berlin (2005)
15. Glaunes, J.: Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l'anatomie numérique. PhD thesis, Université Paris 13 (2005)
16. Durrleman, S.: Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution. PhD thesis, University of Nice-Sophia Antipolis (2010)
17. Yang, X.F., Goh, A., Qiu, A.: Approximations of the diffeomorphic metric and their applications in shape learning. In: Information Processing in Medical Imaging (IPMI), pp. 257–270 (2011)
18. Tenenbaum, J., Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
19. Kendall, W.S.: Probability, convexity, and harmonic maps with small image i: uniqueness and fine existence. Proc. Lond. Math. Soc. **3**(2), 371–406 (1990)
20. Karcher, H.: Riemannian center of mass and mollifier smoothing. Commun. Pure Appl. Math. **30**(5), 509–541 (1977)
21. Le, H.: Estimation of riemannian barycentres. LMS J. Comput. Math. **7**, 193–200 (2004)
22. Afsari, B.: Riemannian L^p center of mass: existence, uniqueness, and convexity. Proc. Am. Math. Soc. **139**(2), 655–673 (2011)
23. Afsari, B., Tron, R., Vidal, R.: On the convergence of gradient descent for finding the riemannian center of mass. SIAM J. Control Optim. **51**(3), 2230–2260 (2013)
24. Emery, M., Mokobodzki, G.: Sur le barycentre d'une probabilité dans une variété. In: Séminaire de probabilités, vol. 25, pp. 220–233. Springer, Berlin (1991)
25. Cury, C., Glaunès, J.A., Colliot, O.: Template estimation for large database: a diffeomorphic iterative centroid method using currents. In: Nielsen, F., Barbaresco, F. (eds.) GSI. Lecture Notes in Computer Science, vol. 8085, pp. 103–111. Springer, Berlin (2013)

26. Chupin, M., Hammers, A., Liu, R.S.N., Colliot, O., Burdett, J., Bardinet, E., Duncan, J.S., Garnero, L., Lemieux, L.: Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. Neuroimage **46**(3), 749–761 (2009)
27. Glaunes, J., Trouvé, A., Younes, L.: Diffeomorphic matching of distributions: a new approach for unlabelled point-sets and sub-manifolds matching. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 712–718 (2004)

Chapter 11
Hartigan's Method for k-MLE: Mixture Modeling with Wishart Distributions and Its Application to Motion Retrieval

Christophe Saint-Jean and Frank Nielsen

Abstract We describe a novel algorithm called k-Maximum Likelihood Estimator (k-MLE) for learning finite statistical mixtures of exponential families relying on Hartigan's k-means swap clustering method. To illustrate this versatile Hartigan k-MLE technique, we consider the exponential family of Wishart distributions and show how to learn their mixtures. First, given a set of symmetric positive definite observation matrices, we provide an iterative algorithm to estimate the parameters of the underlying Wishart distribution which is guaranteed to converge to the MLE. Second, two initialization methods for k-MLE are proposed and compared. Finally, we propose to use the Cauchy-Schwartz statistical divergence as a dissimilarity measure between two Wishart mixture models and sketch a general methodology for building a motion retrieval system.

Keywords Mixture modeling · Wishart · k-MLE · Bregman divergences · Motion retrieval

11.1 Introduction and Prior Work

Mixture models are a powerful and flexible tool to model an unknown probability density function $f(x)$ as a weighted sum of parametric density functions $p_j(x;\theta_j)$:

C. Saint-Jean (✉)
Mathématiques, Image, Applications (MIA), Université de La Rochelle,
17000 La Rochelle, France
e-mail: christophe.saint-jean@univ-lr.fr

F. Nielsen
Sony Computer Science Laboratories, Inc., 3-14-13 Higashi Gotanda,141-0022 Shinagawa-Ku, Tokyo, Japan

F. Nielsen
Laboratoire d'Informatique (LIX), Ecole Polytechnique, Palaiseau Cedex, France

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_11,

$$f(x) = \sum_{j=1}^{K} w_j p_j(x; \theta_j), \quad \text{with} \quad w_j > 0 \quad \text{and} \quad \sum_{j=1}^{K} w_j = 1. \tag{11.1}$$

By far, the most common case are mixtures of Gaussians for which the Expectation-Maximization (EM) method is used for decades to estimate the parameters $\{(w_j, \theta_j)\}_j$ from the maximum likelihood principle. Many extensions aimed at overcoming its slowness and lack of robustness [1]. From the seminal work of Banerjee et al. [2], several methods have been generalized for the exponential families in connection with the Bregman divergences. In particular, the Bregman soft clustering provides a unifying and elegant framework for the EM algorithm with mixtures of exponential families. In a recent work [3], the k-Maximum Likelihood Estimator (k-MLE) has been proposed as a fast alternative to EM for learning any exponential family mixtures: k-MLE relies on the bijection of exponential families with Bregman divergences to transform the mixture learning problem into a geometric clustering problem. Thus we refer the reader to the review paper [4] for an introduction to clustering.

This paper proposes several variations around the initial k-MLE algorithm with a specific focus on mixtures of Wishart [5]. Such a mixture can model complex distributions over the set $\mathcal{S}_{++}^d$ of $d \times d$ symmetric positive definite matrices. Data of this kind comes naturally in some applications like diffusion tensor imaging, radar imaging but also artificially as signature for a multivariate dataset (region of interest in an multispectral image or a temporal sequence of measures for several sensors).

In the literature, the Wishart distribution is rarely used for modeling data but more often in bayesian approaches as a (conjugate) prior for the inverse covariance-matrix of a gaussian vector. This justifies that few works concern the estimation of the parameters of Wishart from a set of matrices. To the best of our knowledge, the only and most related work is the one of Tsai [6] concerning MLE and Restricted-MLE with ordering constraints. From the application viewpoint, one may cite polarimetric SAR imaging [7], bio-medical imaging [8]. Another example is a recent paper on people tracking [9] which applies Dirichlet process mixture model (infinite mixture model) to the clustering of covariance matrices.

The paper is organized as follows: Sect. 11.2 recalls the definition of an exponential family (EF), the principle of maximum likelihood estimation in EFs and how it is connected with Bregman divergences. From these definitions, the complete description of k-MLE technique is derived by following the formalism of the Expectation-Maximization algorithm in Sect. 11.3. In the same section, the Hartigan approach for k-MLE is proposed and discussed as well as how to initialize it properly. Section 11.4 concerns the learning of a mixture of Wishart with k-MLE. For this purpose, a iterative procedure that converges to the MLE when it exists. In Sect. 11.5, we describe an application scenario to motion retrieval before concluding in Sect. 11.6.

11.2 Preliminary Definitions and Notations

An exponential family is a set of probability distributions admitting the following canonical decomposition:

$$p_F(x;\theta) = \exp\{\langle t(x),\theta\rangle + k(x) - F(\theta)\}$$

with $t(x)$ the sufficient statistic, θ the natural parameter, k the carrier measure and F the log-normalizer [10]. Most of commonly used distributions such Bernoulli, Gaussian, Multinomial, Dirichlet, Poisson, Beta, Gamma, von Mises are indeed exponential families (see above reference for a complete list). Later on in the chapter, a canonical decomposition of the Wishart distribution as an exponential family will be detailed.

11.2.1 Maximum Likelihood Estimator

The framework of exponential families gives a direct solution for finding the maximum likelihood estimator $\hat{\theta}$ from a set of i.i.d observations $\chi = \{x_1,\ldots,x_N\}$. Denoting $\mathcal{L}$ the likelihood function

$$\mathcal{L}(\theta;\chi) = \prod_{i=1}^{N} p_F(x_i;\theta) = \prod_{i=1}^{N} \exp\{\langle t(x_i),\theta\rangle + k(x_i) - F(\theta)\} \tag{11.2}$$

and $\bar{l}$ the average log-likelihood function

$$\bar{l}(\theta;\chi) = \frac{1}{N}\sum_{i=1}^{N} \left(\langle t(x_i),\theta\rangle + k(x_i) - F(\theta)\right). \tag{11.3}$$

It follows that the MLE $\hat{\theta} = \arg\max_{\Theta} \bar{l}(\theta;\chi)$ for θ satisfies

$$\nabla F(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} t(x_i). \tag{11.4}$$

Recall that the functional reciprocal $(\nabla F)^{-1}$ of ∇F is also ∇F^* for F^* the convex conjugate of F [11]. It is a mapping from the expectation parameter space $\mathbf{H}$ to the natural parameter space Θ. Thus, the MLE is obtained by mapping $(\nabla F)^{-1}$ on the average of sufficient statistics:

$$\hat{\theta} = (\nabla F)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} t(x_i)\right). \tag{11.5}$$

Whereas determining $(\nabla F)^{-1}$ may be trivial for some univariate distributions like Bernoulli, Poisson, Gaussian, multivariate case is much challenging and lead to consider approximate methods to solve this variational problem [12].

11.2.2 MLE and Bregman Divergence

In this part, the link between MLE and Kullback-Leibler (KL) divergence is recalled. Banerjee et. al. [2] interpret the log-density of a (regular) exponential family as a (regular) Bregman divergence:

$$\log p_F(x;\theta) = -B_{F^*}(t(x):\eta) + F^*(t(x)) + k(x), \tag{11.6}$$

where F^* is the convex conjugate (Legendre transform) of F. Skipping a formal definition, a Bregman divergence for a strictly convex and differentiable function $\varphi : \Omega \mapsto \mathbb{R}$ is

$$B_\varphi(\omega_1 : \omega_2) = \varphi(\omega_1) - \varphi(\omega_2) - \langle \omega_1 - \omega_2, \nabla\varphi(\omega_2)\rangle. \tag{11.7}$$

From a geometric viewpoint, $B_\varphi(\omega_1 : \omega_2)$ is the difference between the value of φ at ω_1 and its first-order Taylor expansion around ω_2 evaluated at ω_1. Since φ is convex, B_φ is positive, zero if and only if (iff) $\omega_1 = \omega_2$ but not symmetric in general. The expression of F^* (and thus of B_{F^*}) follows from $(\nabla F)^{-1}$

$$F^*(\eta) = \langle (\nabla F)^{-1}(\eta), \eta\rangle - F((\nabla F)^{-1}(\eta)). \tag{11.8}$$

In Eq. 11.6, term $B_{F^*}(t(x) : \eta)$ says how much sufficient statistic $t(x)$ on observation x is dissimilar to $\eta \in \mathbf{H}$.

The Kullback-Leibler divergence on two members of the same exponential family is equivalent to the Bregman divergence of the associated log-normalizer on swapped natural parameters [10]:

$$\mathrm{KL}(p_F(.;\theta_1)||p_F(.;\theta_2)) = B_F(\theta_2 : \theta_1) = B_{F^*}(\eta_1 : \eta_2). \tag{11.9}$$

Let us remark that B_F is always known in a closed-form using the canonical decomposition of p_F whereas B_{F^*} requires the knowledge of F^*. Finding the maximizer of the log likelihood on Θ amounts to find the minimizer $\hat{\eta}$ of

$$\sum_{i=1}^{N} B_{F^*}(t(x_i) : \eta) = \sum_{i=1}^{N} \mathrm{KL}\left(p_{F^*}(.; t(x_i))||p_{F^*}(.;\eta)\right)$$

on $\mathbf{H}$ since the two last terms in Eq. (11.6) are constant with respect to η.

11.3 Learning Mixtures with k-MLE

This section presents how to fit a mixture of exponential families with k-MLE. This algorithm requires to have a MLE (see previous section) for each component distribution p_{F_j} of the considered mixture. As it shares many properties with the EM

algorithm for mixtures, this latter is recalled first. The heuristics Lloyd and Hartigan for k-MLE are completely described. Also, two methods for the initialization of k-MLE are proposed depending whether or not component distributions are known.

11.3.1 EM Algorithm

Mixture modeling is a convenient framework to address the problem of clustering defined as the partitioning of a set of i.i.d observations $\chi = \{x_i\}_{i=1,..,N}$ into "meaningful" groups regarding to some similarity. Consider a finite mixture model of exponential families (see Eq. 11.1)

$$f(x) = \sum_{j=1}^{K} w_j \, p_{F_j}(x; \theta_j), \tag{11.10}$$

where K is the number of components and w_j are the mixture weights which sum up to unity. Finding mixture parameters $\{(w_j, \theta_j)\}_j$ can be again addressed by maximizing the log likelihood of the mixture distribution

$$\mathcal{L}(\{(w_j, \theta_j)\}_j; \chi) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} w_j \, p_{F_j}(x_i; \theta_j). \tag{11.11}$$

For $K > 1$, a sum of terms appearing inside a logarithm makes optimization much more difficult than the one of Sect. 11.2.1 ($K = 1$). A classical solution, also well suitable for clustering purpose, is to augment model with indicatory hidden vector variables z_i where $z_{ij} = 1$ iff observation x_i is generated for jth component and 0 otherwise. Previous equation is now replaced by the complete log likelihood of the mixture distribution

$$\mathcal{L}_{\mathbf{c}}(\{(w_j, \theta_j)\}_j; \{(x_i, z_i)\}_i) = \sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij} \log \left(w_j \, p_{F_j}(x_i; \theta_j) \right). \tag{11.12}$$

This is typically the framework of the Expectation-Maximization (EM) algorithm [13] which optimizes this function by repeating two steps:

1. Compute $\mathcal{Q}(\{(w_j, \theta_j)\}_j, \{(w_j^{(t)}, \theta_j^{(t)})\}_j)$ the conditional expectation of $\mathcal{L}_{\mathbf{c}}$ w.r.t. the observed data χ given an estimate $\{(w_j^{(t)}, \theta_j^{(t)})\}_j$ for mixture parameters. This step amounts to compute $\hat{z}_i^{(t)} = \mathbb{E}_{\{(w_j^{(t)}, \theta_j^{(t)})\}_j}[z_i | x_i]$, the vector of responsibilities for each component to have generated x_i.

$$\hat{z}_{ij}^{(t)} = \frac{w_j^{(t)} p_{F_j}(x_i; \theta_j^{(t)})}{\sum_{j'} w_{j'}^{(t)} p_{F_{j'}}(x_i; \theta_{j'}^{(t)})}. \tag{11.13}$$

2. Update mixture parameters by maximizing $\mathcal{Q}$ (i.e. Eq. (11.12) where hidden values z_{ij} are replaced by $\hat{z}_{ij}^{(t)}$).

$$\hat{w}_j^{(t+1)} = \frac{\sum_{i=1}^N \hat{z}_{ij}^{(t)}}{N}, \quad \hat{\theta}_j^{(t+1)} = \arg\max_{\theta_j \in \Theta_j} \sum_{i=1}^N \hat{z}_{ij}^{(t)} \log\left(p_{F_j}(x_i; \theta_j)\right). \tag{11.14}$$

While $\hat{w}_j^{(t+1)}$ is always known in closed-form whatever F_j are, $\hat{\theta}_j^{(t+1)}$ are obtained by component-wise specific optimization involving all observations.

Many properties of this algorithm are known (e.g. maximization of $\mathcal{Q}$ implies maximization of $\mathcal{L}$, slow convergence to local maximum, etc...). In a clustering perspective, components are identified to clusters and values $\hat{z}_{ij}$ are interpreted as soft membership of x_i to cluster $\mathcal{C}_j$. In order to get a strict partition after the convergence, each x_i is assigned to the cluster $\mathcal{C}_j$ iff $\hat{z}_{ij}$ is maximum over $\hat{z}_{i1}, \hat{z}_{i2}, \ldots, \hat{z}_{iK}$.

11.3.2 k-MLE with Lloyd Method

A main reason for the slowness of EM is that all observations are taken into account for the update of parameters for each component since $\hat{z}_{ij}^{(t)} \in [0, 1]$. A natural idea is then to generate smaller sub-samples of χ from $\hat{z}_{ij}^{(t)}$ in a deterministic manner.[1] The simplest way to do this is to get a strict partition of χ with MAP assignment:

$$\tilde{z}_{ij}^{(t)} = \begin{cases} 1 & \text{if } \hat{z}_{ij}^{(t)} = \max_k \hat{z}_{ik}^{(t)} \\ 0 & \text{otherwise} \end{cases}.$$

When multiple maxima exist, the component with the smallest index is chosen. If this classification step is inserted between E-step and M-step, Classification EM (CEM) algorithm [14] is retrieved. Moreover, for isotropic gaussian components with fixed unit variance, CEM is shown to be equivalent to the Lloyd K-means algorithm [4]. More recently, CEM was reformulated in a close way under the name k-MLE [3] for the context of exponential families and Bregman divergences. In the following of the paper, we will refer only to this latter. Replacing $z_{ij}^{(t)}$ by $\tilde{z}_{ij}^{(t)}$ in Eq. (11.12), the criterion to be maximized in the M-step can be reformulated as

[1] Otherwise, convergence to a pointwise estimate of the parameters would be replaced by convergence in distribution of a Markov chain.

$$\tilde{\mathcal{L}}_{\mathbf{c}}(\{(w_j,\theta_j)\}_j;\{(x_i,\tilde{z}_i^{(t)})\}_i)=\sum_{j=1}^{K}\sum_{i=1}^{N}\tilde{z}_{ij}^{(t)}\log\left(w_j p_{F_j}(x_i;\theta_j)\right). \tag{11.15}$$

Following CEM terminology, this quantity is called the "classification maximum likelihood". Letting $C_j^{(t)}=\left\{x_i\in\chi|\tilde{z}_{ij}^{(t)}=1\right\}$, this equation can be conveniently rewritten as

$$\begin{aligned}\tilde{\mathcal{L}}_{\mathbf{c}}(\{(w_j,\theta_j)\}_j;\{\mathcal{C}_j^{(t)}\}_j)=&\sum_{x\in\mathcal{C}_1^{(t)}}\log\left(w_1 p_{F_1}(x;\theta_1)\right)\\&+\cdots+\sum_{x\in\mathcal{C}_K^{(t)}}\log\left(w_K p_{F_K}(x;\theta_K)\right).\end{aligned} \tag{11.16}$$

Each term leads to a separate optimization to get the parameters of the corresponding component:

$$\hat{w}_j^{(t+1)}=\frac{|\mathcal{C}_j^{(t)}|}{N},\quad \hat{\theta}_j^{(t+1)}=\arg\max_{\theta_j\in\Theta_j}\sum_{x\in\mathcal{C}_j^{(t)}}\log p_{F_j}(x;\theta_j). \tag{11.17}$$

Last equation is nothing but the equation of the MLE for the j-th component with a subset of χ. Algorithm 1 summarizes k-MLE with Lloyd method given an initial description of the mixture.

Algorithm 1: k-MLE (Lloyd method)

Input: A sample $\chi=\{x_1,x_2,...,x_N\}$, initial mixture parameters $\{\hat{w}_j^{(0)},\hat{\theta}_j^{(0)}\}_j$, $\{F_j\}_j$ log-normalizers of exponential families

Output: Ending values for $\{\hat{w}_j^{(t)},\hat{\theta}_j^{(t)}\}_j$ are estimates of mixture parameters, $\mathcal{C}_j^{(t)}$ a partition of χ

```
t = 0;
repeat
    repeat
        // Partition χ in K disjoint subsets with MAP assignment
        foreach x_i ∈ χ do z̃_i^(t) = arg max_j log ŵ_j^(t) p_{F_j}(x_i; θ̂_j^(t));
        C_j^(t) = {x_i ∈ χ | z̃_i^(t) = j};
        // Update parameters {θ_j}_j with MLE ({w_j}_j unchanged)
        foreach j ∈ 1, ..., K do θ̂_j^(t+1) = arg max_{θ_j∈Θ_j} Σ_{x∈C_j^(t)} log p_{F_j}(x; θ_j);
        t = t + 1;
    until Convergence of the classification maximum likelihood (Eq. 11.16);
    // Update mixture weights {w_j}_j
    foreach j ∈ 1, ..., K do ŵ_j^(t+1) = |C_j^(t)| / N;
until Further convergence of the classification maximum likelihood (Eq. 11.16);
```

Contrary to CEM algorithm, k-MLE algorithm updates mixture weights after the convergence of the $\tilde{\mathcal{L}}_{\mathbf{c}}$ (line 8) and not simultaneously with component parameters. Despite this difference, both algorithms can be proved to converge to a local maximum of $\tilde{\mathcal{L}}_{\mathbf{c}}$ with same kind of arguments (see [3, 14]). In practice, the local maxima (and also the mixture parameters) are not necessary equal for the two algorithms.

11.3.3 k-MLE with Hartigan Method

In this section, a different optimization of the classification maximum likelihood is presented. A drawback of previous methods is that they can produce empty clusters without any mean of control. It occurs especially when observations are in a high dimensional space. A mild solution is to discard empty clusters by setting their weights to zero and their parameters to $\emptyset$. A better approach, detailed in the following, is to rewrite the k-MLE following the same principle as Hartigan method for k-means [15]. Moreover, this heuristic is preferred to Lloyd's one since it generally provides better local maxima [16].

Hartigan method is generally summarized by the sentence "Pick an observation, say x_c in cluster $\mathcal{C}_c$, and optimally reassign it to another cluster." Let us first consider as "optimal" the assignment x_c to its most probable cluster, say $\mathcal{C}_{j^*}$:

$$j^* = \arg\max_j \log \hat{w}_j^{(t)} p_{F_j}(x_c; \hat{\theta}_j^{(t)}),$$

where $\hat{w}_j^{(t)}$, and $\hat{\theta}_j^{(t)}$ denote the weight and the parameters of the j-th component at some iteration. Then, parameters of the two components are updated with MLE:

$$\hat{\theta}_c^{(t+1)} = \arg\max_{\theta_c \in \Theta_c} \sum_{x \in \mathcal{C}_c^{(t)} \setminus \{x_c\}} \log p_{F_c}(x; \theta_c) \tag{11.18}$$

$$\hat{\theta}_{j^*}^{(t+1)} = \arg\max_{\theta_{j^*} \in \Theta_{j^*}} \sum_{x \in \mathcal{C}_{j^*}^{(t)} \cup \{x_c\}} \log p_{F_{j^*}}(x; \theta_{j^*}). \tag{11.19}$$

The mixture weights $\hat{w}_c$ and $\hat{w}_{j^*}$ remain unchanged in this step (see line 9 of Algorithm 1). Consequently, $\tilde{\mathcal{L}}_{\mathbf{c}}$ increases by $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_{j^*})$ where $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j)$ is more generally defined as

$$\begin{aligned}\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j) = &\sum_{x \in \mathcal{C}_c^{(t)} \setminus \{x_c\}} \log p_{F_c}(x; \hat{\theta}_c^{(t+1)}) - \sum_{x \in \mathcal{C}_c^{(t)} \cup \{x_c\}} \log p_{F_c}(x; \hat{\theta}_c^{(t)}) - \log \frac{\hat{w}_c}{\hat{w}_j} \\ &+ \sum_{x \in \mathcal{C}_j^{(t)} \cup \{x_c\}} \log p_{F_j}(x; \hat{\theta}_j^{(t+1)}) - \sum_{x \in \mathcal{C}_j^{(t)} \setminus \{x_c\}} \log p_{F_j}(x; \hat{\theta}_j^{(t)}).\end{aligned} \tag{11.20}$$

This procedure is nothing more than a partial assignment (C-step) in the Lloyd method for k-MLE. This is an indirect way to reach our initial goal which is the maximization of $\tilde{\mathcal{L}}_{\mathbf{c}}$.

Following Telgarsky and Vattani [16], a better approach is to consider as "optimal" the assignment to cluster $\mathcal{C}_j$ which maximizes $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j)$

$$j^* = \arg\max_j \Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j). \tag{11.21}$$

Since $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_c) = 0$, such assignment satisfies $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_{j^*}) \geq 0$ and therefore the increase of $\tilde{\mathcal{L}}_{\mathbf{c}}$. As the optimization space is finite (partitions of $\{x_1, ..., x_N\}$), this procedure converges to a local maximum of $\tilde{\mathcal{L}}_{\mathbf{c}}$. There is no guarantee that $\mathcal{C}_{j^*}$ coincides with the MAP assignment for x_c.

For the k-means loss function, Hartigan method avoids empty clusters since any assignment to one of those empty clusters decreases it necessarily [16]. Analogous property will be now studied for k-MLE through the formulation of $\tilde{\mathcal{L}}_{\mathbf{c}}$ with η-coordinates:

$$\begin{aligned}&\tilde{\mathcal{L}}_{\mathbf{c}}(\{(w_j, \eta_j)\}_j; \{\mathcal{C}_j^{(t)}\}_j) \\ &= \sum_{j=1}^{K} \sum_{x \in \mathcal{C}_j^{(t)}} \left[F_j^*(\eta_j) + k_j(x) + \langle t_j(x) - \eta_j, \nabla F_j^*(\eta_j) \rangle + \log w_j \right].\end{aligned} \tag{11.22}$$

Recalling that the MLE satisfies $\hat{\eta}_j^{(t)} = |\mathcal{C}_j^{(t)}|^{-1} \sum_{x \in \mathcal{C}_j^{(t)}} t_j(x)$, dot product vanishes when $\eta_j = \hat{\eta}_j^{(t)}$ and it follows after small calculations

$$\begin{aligned}\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j) = &(|\mathcal{C}_c^{(t)}| - 1) F_c^*(\hat{\eta}_c^{(t+1)}) - |\mathcal{C}_c^{(t)}| F_c^*(\hat{\eta}_c^{(t)}) \\ &+ (|\mathcal{C}_j^{(t)}| + 1) F_j^*(\hat{\eta}_j^{(t+1)}) - |\mathcal{C}_j^{(t)}| F_j^*(\hat{\eta}_j^{(t)}) \\ &+ k_j(x_c) - k_c(x_c) - \log \frac{\hat{w}_c}{\hat{w}_j}.\end{aligned} \tag{11.23}$$

As far as F_j^* is known in closed-form, this criterion is faster to compute than Eq. (11.20) since updates of component parameters are immediate

$$\hat{\eta}_c^{(t+1)} = \frac{|\mathcal{C}_c^{(t)}| \hat{\eta}_c^{(t)} - t_c(x_c)}{|\mathcal{C}_c^{(t)}| - 1}, \quad \hat{\eta}_j^{(t+1)} = \frac{|\mathcal{C}_j^{(t)}| \hat{\eta}_j^{(t)} + t_j(x_c)}{|\mathcal{C}_j^{(t)}| + 1}. \tag{11.24}$$

When $\mathcal{C}_c^{(t)} = \{x_c\}$, there is no particular reason for $\Phi^{(t)}(x_c, \{x_c\}, \mathcal{C}_j)$ to be always negative. Simplications occurring for the k-means in euclidean case (e.g. $k_j(x_c) = 0$, clusters have equal weight $w_j = K^{-1}$, etc...) do not exist in this more general case.

Thus, in order to avoid to empty a cluster, it is mandatory to reject every outgoing transfer for a singleton cluster (cf. line 8).

Algorithm 2 details k-MLE algorithm with Hartigan method when F_j^* are available. When only F_j are known, $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j)$ can be computed with Eq. (11.20). In this case, the computation of MLE $\hat{\theta}_j$ is much slower and is an issue for a singleton cluster. Its existence and possible solutions will be discussed later for the Wishart distribution. Further remarks on Algorithm 2:

(line 1) When all $F_j^* = F^*$ are identical, this partitioning can be understood as geometric split in the expectation parameter space induced by divergence B_{F^*} and additive weight $-\log w_j$ (weighted Bregman Voronoi diagram [17]).

(line 4) This permutation avoids same ordering for each loop.

(line 6) A weaker choice may be done here: any cluster $\mathcal{C}_j$ (for instance the first) which satisfies $\Phi^{(t)}(x_i, \mathcal{C}_{z_i}, \mathcal{C}_j) > 0$ is a possible candidate still guaranteeing convergence of the algorithm. For such clusters, it may be also advisable to select $\mathcal{C}_j$ with maximum $\hat{z}_{ij}^{(t)}$.

(line 12) Obviously, this criterion is equivalent to local convergence of $\tilde{\mathcal{L}}_{\mathbf{c}}$.

As said before, this algorithm is faster when components parameters η_j can be updated in the expectation parameter space $\mathbf{H}_j$. But the price to pay is the memory needed to keep all sufficient statistics $t_j(x_i)$ for each observation x_i.

11.3.4 Initialization with DP-k-MLE++

To complete the description of k-MLE, it remains the problem of the initialization of the algorithm: choice of the exponential family for each component, initial values of $\{(\hat{w}_j^{(0)}, \eta_j^{(0)})\}_j$, number of components K. Ideally, a good initialization would be fast, select automatically the number of components (unknown for most applications) and provide initial mixture parameters not too far from a good local minimum of the clustering criterion. The choice of model complexity (i.e. the number K of groups) is a recurrent problem in clustering since a compromise has to be done between genericity and goodness of fit. Since the likelihood increases with K, many criteria such as BIC, NEC are based on the penalization of likelihood by a function of the degree of freedom of the model. Other approaches include MDL principle, Bayes factor or simply a visual inspection of some plottings (e.g. silhouette graph, dendrogram for hierarchical clustering, Gram matrix, etc...). The reader interested by this topic may refer to section M in the survey of Xu and Wunsch [18]. Proposed method, inspired by the algorithms k-MLE++ [3] and DP-means [19], will be described in the following.

At the beginning of a clustering, there is no particular reason for favoring one particular cluster among all others. Assuming uniform weighting for components, $\tilde{\mathcal{L}}_{\mathbf{c}}$ simplifies to

Algorithm 2: k-MLE (Hartigan method)

Input: Sample $\chi = \{x_1, .., x_N\}$, initial mixture parameters $\{(\hat{w}_j^{(0)}, \hat{\eta}_j^{(0)})\}_{j=1,..,K}$, $\{(t_j, F_j^*)\}_j$ sufficient statistics and dual log-normalizers of exponential families

Output: Ending values for $\{(\hat{w}_j^{(t)}, \hat{\eta}_j^{(t)})\}_j$ are estimates of mixture parameters, $\mathcal{C}_j^{(t)}$ a partition of χ

// `Partition` χ `in` K `disjoint subsets with MAP assignment`

foreach $x_i \in \chi$ **do** $\bar{z}_i^{(0)} = \arg\min_j (B_{F_j^*}(t_j(x_i) : \hat{\eta}_j^{(0)}) - \log \hat{w}_j^{(0)})$;

foreach $j \in 1, ..., K$ **do** $\mathcal{C}_j^{(0)} = \{x_i \in \chi | \bar{z}_i^{(0)} = j\}$;

repeat

  done_transfer = False;

  Random permute $(x_1, \bar{z}_1^{(t)}), ..., (x_N, \bar{z}_N^{(t)})$;

  foreach $x_i \in \chi$ *such that* $|\mathcal{C}_{\bar{z}_i^{(t)}}^{(t)}| > 1$ **do**

    // `Test optimal transfer for` x_i `(see Eqs. 11.23 or 11.20)`

    $j^* = \arg\min_j \Phi^{(t)}(x_i, \mathcal{C}_{\bar{z}_i^{(t)}}, \mathcal{C}_j)$;

    if $\Phi^{(t)}(x_i, \mathcal{C}_{\bar{z}_i^{(t)}}, \mathcal{C}_{j^*}) > 0$ **then**

      // `Update clusters and membership of` x_i

$$\mathcal{C}_{\bar{z}_i^{(t)}}^{(t+1)} = \mathcal{C}_{\bar{z}_i^{(t)}}^{(t)} \backslash \{x_i\}, \quad \mathcal{C}_{j^*}^{(t+1)} = \mathcal{C}_{j^*}^{(t)} \cup \{x_i\}, \quad \bar{z}_i^{(t+1)} = j^*$$

      // `Update only` $\eta_{\bar{z}_i}, \eta_{j^*}$ `with MLE` $(\{w_j\}_j$ `unchanged)`

$$\hat{\eta}_{\bar{z}_i^{(t)}}^{(t+1)} = \frac{|\mathcal{C}_{\bar{z}_i^{(t)}}^{(t)}| \hat{\eta}_{\bar{z}_i^{(t)}}^{(t)} - t_{\bar{z}_i^{(t)}}(x_i)}{|\mathcal{C}_{\bar{z}_i^{(t)}}^{(t+1)}|}, \quad \hat{\eta}_{z_i^{(t+1)}}^{(t+1)} = \frac{|\mathcal{C}_{\bar{z}_i^{(t+1)}}^{(t)}| \hat{\eta}_{\bar{z}_i^{(t+1)}}^{(t)} + t_{\bar{z}_i^{(t+1)}}(x_i)}{|\mathcal{C}_{\bar{z}_i^{(t+1)}}^{(t+1)}|}$$

      done_transfer = True; t = t +1;

  if *done_transfer is True* **then**

    // `Update mixture weights` $\{w_j\}_j$

    foreach $j \in 1, ..., K$ **do** $\hat{w}_j^{(t)} = N^{-1} |\mathcal{C}_j^{(t)}|$;

until *done_transfer is False*;

$$\mathring{\mathcal{L}}_{\mathbf{c}}(\{\theta_j\}_j; \{\mathcal{C}_j^{(t)}\}_j) = \sum_{j=1}^{K} \sum_{x \in \mathcal{C}_j^{(t)}} \log p_{F_j}(x; \theta_j) \qquad \textit{or equivalently to} \quad (11.25)$$

$$\mathring{\mathcal{L}}_{\mathbf{c}}(\{\eta_j\}_j; \{\mathcal{C}_j^{(t)}\}_j) = \sum_{j=1}^{K} \sum_{x \in \mathcal{C}_j^{(t)}} \left[F_j^*(\eta_j) + k_j(x) + \langle t_j(x) - \eta_j, \nabla F_j^*(\eta_j) \rangle \right]. \quad (11.26)$$

When all $F_j^* = F^*$ are identical and the partition $\{\mathcal{C}_j^{(t)}\}_j$ corresponds to MAP assignment, $\mathring{\mathcal{L}}$ is exactly the objective function $\breve{\mathcal{L}}$ for the Bregman k-means [2]. Rewriting $\breve{\mathcal{L}}$ as an equivalent criterion to be minimized, it follows

$$\breve{\mathcal{L}}(\{\eta_j\}_j) = \sum_{i=1}^{N} \min_{j=1}^{K} B_{F^*}(t(x_i) : \eta_j). \tag{11.27}$$

Bregman k-means++ [20, 21] provides initial centers $\{\eta_j^{(0)}\}_j$ which guarantee to find a clustering that is $\mathcal{O}(\log K)$-competitive to the optimal Bregman k-means clustering. The k-MLE++ algorithm amounts to use Bregman k-means++ on the dual log-normalizer F^* (see Algorithm 3).

Algorithm 3: k-MLE++

Input: A sample $\chi = \{x_1, ..., x_N\}$, t the sufficient statistics and F^* the dual log-normalizer of an exponential family, K the number of clusters

Output: Initial mixture parameters $\{(w_j^{(0)}, \eta_j^{(0)}\}_j)$

$w_1^{(0)} = 1/K$;

Choose first seed $\eta_1^{(0)} = t(x_i)$ for i uniformly random in $\{1, 2, \ldots, N\}$;

for $j = 2, ..., K$ **do**

  $w_j^{(0)} = 1/K$;

  // Compute relative contributions to $\breve{\mathcal{L}}(\{\eta_j\}_j)$

  foreach $x_i \in \chi$ **do** $p_i = \frac{\min_{j'=1}^{j} B_{F^*}(t(x_i):\eta_{j'})}{\sum_{i'=1}^{N} \min_{j'=1}^{j} B_{F^*}(t(x_{i'}):\eta_{j'})}$;

  Choose $\eta_j^{(0)} \in \{t(x_1), ..., t(x_N)\}$ with probability p_i;

When K is unknown, same strategy can still be applied but a stopping criterion has to be set. Probability p_i in Algorithm 3 is a relative contribution of observation x_i through $t(x_i)$ to $\breve{\mathcal{L}}(\{\eta_1, ..., \eta_K\})$ where K is the number of already selected centers. A high p_i indicates that x_i is relatively far from these centers, thus is atypic to the mixture $\{(w_1^{(0)}, \eta_1^{(0)}), ..., (w_K^{(0)}, \eta_K^{(0)}), \}$ for $w_j^{(0)} = w^{(0)}$ an arbitrary constant. When selecting a new center, p_i necessarily decreases in the next iteration. A good covering of χ is obtained when all p_i are lower than some threshold $\lambda \in [0, 1]$. Algorithm 4 describes the initialization named after DP-k-MLE++.

The higher the threshold λ, the lower the number of generated centers. In particular, the value $\frac{1}{N}$ should be considered as a reasonable minimum setting for λ. For $\lambda = 1$, the algorithm will simply return one center. Since $p_i = 0$ for already selected centers, this method guarantees all centers to be distinct.

11.3.5 Initialization with DP-comp-k-MLE

Although k-MLE can be used with component-wise exponential families, previous initialization methods yield components of same exponential family. Component distribution may be chosen simultaneously to a center selection when additional

Algorithm 4: DP-k-MLE++

Input: A sample $\chi = \{x_1, ..., x_N\}$, t the sufficient statistics and F^* the dual log-normalizer of an exponential family, $\lambda \in [0, 1]$

Output: Initial mixture parameters $\{w_j^{(0)}, \eta_j^{(0)}\}_j$, K the number of clusters

Choose first seed $\eta_1^{(0)} = t(x_i)$ for i uniformly random in $\{1, 2, \ldots, N\}$;
K=1;
repeat
  // Compute relative contributions to $\check{\mathcal{L}}(\{\eta_1, ..., \eta_K\})$
  foreach $x_i \in \chi$ **do** $p_i = \frac{\min_{j=1}^{K} B_{F^*}(t(x_i):\eta_j)}{\sum_{i'=1}^{N} \min_{j=1}^{K} B_{F^*}(t(x_{i'}):\eta_{j'})}$;
  if $\exists\, p_i > \lambda$ **then**
    K = K+1;
    // Select next seed
    Choose $\eta_K^{(0)} \in \{t(x_1), ..., t(x_N)\}$ with probability p_i;
until *all* $p_i \leq \lambda$;
for $j = 1, ..., K$ **do** $w_j^{(0)} = 1/K$;

knowledge ξ_i about x_i is available (see Sect. 11.5 for an example). Given such a choice function H, Algorithm 5 called "DP-comp-k-MLE" describes this new flexible initialization method. DP-comp-k-MLE is clearly a generalization of DP-k-MLE++ when H always returns the same exponential family. However, in the general case, it remains to be proved whether a DP-comp-k-MLE clustering is $\mathcal{O}(\log K)$-competitive to the optimal k-MLE clustering (with equal weight). Without this difficult theoretical study, suffix "++" is carefully omitted in the name DP-comp-k-MLE.

To end up with this section, let us recall that all we need to know for using proposed algorithms is the MLE for the considered exponential family, whether it is available in a closed-form or not. In many exponential families, all details (canonical decomposition, $F, \nabla F, F^*, \nabla F^* = (\nabla F)^{-1}$) are already known [10]. The next section focuses on the case of the Wishart distribution.

11.4 Learning Mixtures of Wishart with k-MLE

This section recalls the definition of Wishart distribution and proposes a maximum likelihood estimator for its parameters. Some known facts such as the Kullback-Leibler divergence between two Wishart densities are recalled. Its use with the above algorithms is also discussed.

11.4.1 Wishart Distribution

The Wishart distribution [5] is the multidimensional version of the chi-square distribution and it characterizes empirical scatter matrix estimator for the multivariate gaussian distribution. Let $\mathbb{X}$ be a n-sample consisting in independent realizations of

Algorithm 5: DP-comp-k-MLE

Input: A sample $\chi = \{x_1, ..., x_N\}$ with extra knowledge $\xi = \{\xi_1, ..., \xi_N\}$, H a choice function of an exponential family, $\lambda \in [0, 1]$

Output: Initial mixture parameters $\{(w_j^{(0)}, \eta_j^{(0)})\}_j$, $\{(t_j, F_j^*)\}_j$ sufficient statistics and dual log-normalizers of exponential families, K the number of clusters

// Select first seed and exponential family

for *i uniformly random in* $\{1, 2, \ldots, N\}$ **do**
  Obtain t_1, F_1^* from $H(x_i, \xi_i)$;
  Select first seed $\eta_1^{(0)} = t_1(x_i)$;
K=1;
repeat
  foreach $x_i \in \chi$ **do** $p_i = \frac{\min_{j=1}^K B_{F_j^*}(t_j(x_i):\eta_j)}{\sum_{i'=1}^N \min_{j=1}^K B_{F_j^*}(t_j(x_{i'}):\eta_{j'})}$;
  if $\exists\ p_i > \lambda$ **then**
    K = K+1;
    // Select next seed and exponential family
    for *i with probability* p_i *in* $\{1, 2, \ldots, N\}$ **do**
      Obtain t_K, F_K^* from $H(x_i, \xi_i)$;
      Select next seed $\eta_K^{(0)} = t_K(x_i)$;
until *all* $p_i \leq \lambda$;
for $j = 1, ..., K$ **do** $w_j^{(0)} = 1/K$;

a random gaussian vector with d dimensions, zero mean and covariance matrix S. Then scatter matrix $X = {}^t\mathbb{X}\mathbb{X}$ follows a central Wishart distribution with scale matrix S and degree of freedom n, denoted by $X \sim \mathcal{W}_d(n, S)$. Its density function is

$$\mathcal{W}_d(X; n, S) = \frac{|X|^{\frac{n-d-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(S^{-1}X)\right\}}{2^{\frac{nd}{2}}|S|^{\frac{n}{2}}\Gamma_d\left(\frac{n}{2}\right)},$$

where for $y > 0$, $\Gamma_d(y) = \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^d \Gamma\left(y - \frac{j-1}{2}\right)$ is the multivariate gamma function. Let us remark immediately that this definition implies that n is constrained to be strictly greater than $d - 1$.

Wishart distribution is an exponential family since

$$\mathcal{W}_d(X; \theta_n, \theta_S) = \exp\left\{< \theta_n, \log|X| >_{\mathbb{R}} + < \theta_S, -\frac{1}{2}X >_{HS} + k(X) - F(\theta_n, \theta_S)\right\},$$

where $(\theta_n, \theta_S) = (\frac{n-d-1}{2}, S^{-1}), t(X) = (\log|X|, -\frac{1}{2}X)$, $\langle , \rangle_{HS}$ denotes the Hilbert-Schmidt inner product and

$$F(\theta_n, \theta_S) = \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|) + \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right). \tag{11.28}$$

Note that this decomposition is not unique (see another one in [22]). Refer to Appendix A.1 for detailed calculations.

11.4.2 MLE for Wishart Distribution

Let us recall (see Sect. 11.2.1) that the MLE is obtained by mapping $(\nabla F)^{-1}$ on the average of sufficient statistics. Finding $(\nabla F)^{-1}$ amounts to solve here the following system (see Eqs. 11.5 and 11.28):

$$\begin{cases} d\log(2) - \log|\theta_S| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right) = \eta_n, \\ -\left(\theta_n + \frac{(d+1)}{2}\right)\theta_S^{-1} \qquad\qquad\qquad\quad = \eta_S. \end{cases} \tag{11.29}$$

with $\eta_n = \mathbb{E}[\log|X|]$ and $\eta_S = \mathbb{E}[-\frac{1}{2}X]$ the expectation parameters and Ψ_d the derivative of the $\log\Gamma_d$. Unfortunately, variables θ_n and θ_S are not separable so that no closed-form solution is known. Instead, as pointed out in [23], it is possible to adopt an iterative scheme that alternatively yields maximum likelihood estimate when the other parameter is fixed. This is equivalent to consider two sub-families $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$ of Wishart distribution $\mathcal{W}_d$ which are also exponential families. For the sake of simplicity, natural parameterizations and sufficient statistics of the decomposition in the general case are kept (see Appendices A.2 and A.3 for more details).

Distribution $\mathcal{W}_{d,\underline{n}}$ ($\underline{n} = 2\theta_{\underline{n}} + d + 1$): $k_{\underline{n}}(X) = \frac{\underline{n}-d-1}{2}\log|X|$ and

$$F_{\underline{n}}(\theta_S) = \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|\theta_S| + \log\Gamma_d\left(\frac{\underline{n}}{2}\right). \tag{11.30}$$

Using classical results for matrix derivatives, (Eq. 11.5) can be easily solved:

$$-\frac{\underline{n}}{2}\hat{\theta}_S^{-1} = \frac{1}{N}\sum_{i=1}^{N} -\frac{1}{2}X_i \implies \hat{\theta}_S = N\underline{n}\left(\sum_{i=1}^{N} X_i\right)^{-1}. \tag{11.31}$$

Distribution $\mathcal{W}_{d,\underline{S}}$ ($\underline{S} = \theta_{\underline{S}}^{-1}$): $k_{\underline{S}}(X) = -\frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X)$ and

$$F_{\underline{S}}(\theta_n) = \left(\theta_n + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| + \log\Gamma_d\left(\theta_n + \frac{d+1}{2}\right) \tag{11.32}$$

Again, Eq. (11.5) can be numerically solved:

$$\hat{\theta}_n = \Psi_d^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\log|X_i| - \log\left|2\underline{S}\right|\right) - \frac{d+1}{2}, \qquad \hat{\theta}_n > -1 \tag{11.33}$$

with Ψ_d^{-1} the functional reciprocal of Ψ_d. This latter can be computed with any optimization method on bounded domain (e.g. Brent method [24]). Let us mention that notation is simplified here since $\hat{\theta}_S$ and $\hat{\theta}_n$ should have been indexed by their corresponding family. Algorithm 6 summarizes the estimate $\hat{\theta}$ for parameters of the Wishart distribution. By precomputing $N\left(\sum_{i=1}^{N} X_i\right)^{-1}$ and $N^{-1}\sum_{i=1}^{N}\log|X_i|$, much computation time can be saved. The computation of the Ψ_d^{-1} remains an expensive part of the algorithm.

Let us now prove the convergence and the consistency of this method. Maximizing $\bar{l}$ amounts to minimize equivalently $E(\theta) = F(\theta) - \langle \frac{1}{N}\sum_{i=1}^{N} t(X_i), \theta\rangle$. The following properties are satisfied by E:

- The hessian $\nabla^2 E = \nabla^2 F$ of E is positive definite on Θ since F is convex.
- Its unique minimizer on Θ is the MLE $\hat{\theta} = \nabla F^*(\frac{1}{N}\sum_{i=1}^{N} t(X_i))$ whenever it exists (although F^* is not known for Wishart, and F is not separable).

Algorithm 6: MLE for the parameters of a Wishart distribution

Input: A sample $\chi = \{X_1, X_2, \ldots, X_N\}$ of $\mathcal{S}_{++}^d$ with $N > 1$

Output: Estimate $\hat{\theta}$ is the terminal values of MLE sequences $\{\hat{\theta}_n^{(t)}\}$ and $\{\hat{\theta}_S^{(t)}\}$

// Initialization of the $\{\hat{\theta}_n^{(t)}\}$ sequence

$\hat{\theta}_n^{(0)} = 1; t = 0;$

repeat

// Compute MLE in $\mathcal{W}_{d,\underline{n}}$ using Eq. 11.31

$$\hat{\theta}_S^{(t+1)} = N\underline{n}\left(\sum_{i=1}^{N} X_i\right)^{-1} \text{ with } \underline{n} = 2\hat{\theta}_n^{(t)} + d + 1$$

// Compute MLE in $\mathcal{W}_{d,\underline{S}}$ using Eq. 11.33

$$\hat{\theta}_n^{(t+1)} = \Psi_d^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\log|X_i| - \log\left|2\underline{S}\right|\right) - \frac{d+1}{2} \text{ with } \underline{S} = \left(\hat{\theta}_S^{(t+1)}\right)^{-1}$$

$t = t + 1;$

until *convergence of the likelihood*;

Therefore, Algorithm 6 is an instance of the group coordinate descent algorithm of Bezdek et al. (Theorem 2.2 in [25]) for $\theta = (\theta_n, \theta_S)$:

$$\hat{\theta}_S^{(t+1)} = \arg\max_{\theta_S} E(\hat{\theta}_n^{(t)}, \theta_S) \tag{11.34}$$

$$\hat{\theta}_n^{(t+1)} = \arg\max_{\theta_n} E(\theta_n, \hat{\theta}_S^{(t+1)}) \tag{11.35}$$

Resulting sequences $\{\hat{\theta}_n^{(t)}\}$ and $\{\hat{\theta}_S^{(t)}\}$ are shown to converge linearly to the coordinates of $\hat{\theta}$.

By looking carefully at the previous algorithms, let us remark that the initialization methods require to able to compute the divergence B_{F^*} between two elements η_1 and η_2 in the expectation space $\mathbf{H}$. Whereas F^* is known for $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$, Eq. (11.9) gives a potential solution for $\mathcal{W}_d$ by considering B_F on natural parameters θ_2 and θ_1 in Θ. Searching the correspondence $\mathbf{H} \mapsto \Theta$ is analogous to compute the MLE for a single observation...

The previous MLE procedure does not converge with a single observation X_1. Bogdan and Bogdan [26] proved that MLE exists and is unique in an exponential family off the affine envelope of the N points $t(X_1), ..., t(X_N)$ is of dimension D, the order of this exponential family. Since the affine envelope of $t(X_1)$ is of dimension $d \times d$ (instead of $\mathrm{D} = d \times d + 1$), the MLE does not exists and the likelihood function goes to infinity.[2] Unboundedness of likelihood function is well known problem that can be tackled by adding a penalty term to it [27]. A simpler solution is to take the MLE in family $\mathcal{W}_{d,\underline{n}}$ for some n (known or arbitrary fixed above $d-1$) instead of $\mathcal{W}_d$.

11.4.3 Divergences for Wishart Distributions

For two Wishart distributions $\mathcal{W}_d^1 = \mathcal{W}_d(X; n_1, S_1)$ and $\mathcal{W}_d^2 = \mathcal{W}_d(X; n_2, S_2)$, the KL divergence is known [22] (even if F^* is unknown):

$$\begin{aligned}\mathrm{KL}(\mathcal{W}_d^1||\mathcal{W}_d^2) = &-\log\left(\frac{\Gamma_d\left(\frac{n_1}{2}\right)}{\Gamma_d\left(\frac{n_2}{2}\right)}\right) + \left(\frac{n_1-n_2}{2}\right)\Psi_d\left(\frac{n_1}{2}\right)\\ &+\frac{n_1}{2}\left(-\log\frac{|S_1|}{|S_2|} + \mathrm{tr}(S_2^{-1}S_1) - d\right)\end{aligned} \tag{11.36}$$

Looking the KL divergences of the two Wishart sub-families $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$ gives an interesting perspective to this formula. Applying Eqs. 11.9 and 11.8, it follows

$$\mathrm{KL}(\mathcal{W}_{d,\underline{n}}^1||\mathcal{W}_{d,\underline{n}}^2) = \frac{n}{2}\left(-\log\frac{|S_1|}{|S_2|} + \mathrm{tr}(S_2^{-1}S_1) - d\right) \tag{11.37}$$

$$\mathrm{KL}(\mathcal{W}_{d,\underline{S}}^1||\mathcal{W}_{d,\underline{S}}^2) = -\log\left(\frac{\Gamma_d\left(\frac{n_1}{2}\right)}{\Gamma_d\left(\frac{n_2}{2}\right)}\right) + \left(\frac{n_1-n_2}{2}\right)\Psi_d\left(\frac{n_1}{2}\right) \tag{11.38}$$

Detailed calculations can be found in the Appendix. Notice that $\mathrm{KL}(\mathcal{W}_d^1||\mathcal{W}_d^2)$ is simply the sum of these two divergences

[2] Product $\hat{\theta}_n^{(t)}\hat{\theta}_S^{(t)}$ is constant through iterations.

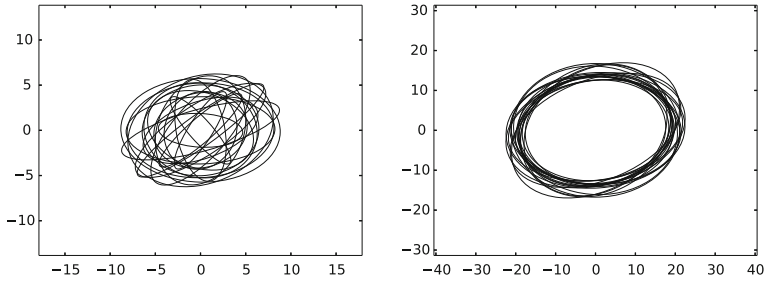


Fig. 11.1 20 random matrices from $\mathcal{W}_d(.; n, S)$ from $n = 5$ (*left*), $n = 50$ (*right*)

$$\mathrm{KL}(\mathcal{W}_d^1||\mathcal{W}_d^2) = \mathrm{KL}(\mathcal{W}_{d,\underline{S_1}}^1||\mathcal{W}_{d,\underline{S_1}}^2) + \mathrm{KL}(\mathcal{W}_{d,\underline{n_1}}^1||\mathcal{W}_{d,\underline{n_1}}^2) \tag{11.39}$$

and that $\mathrm{KL}(\mathcal{W}_{d,\underline{S}}^1||\mathcal{W}_{d,\underline{S}}^2)$ does not depend on $\underline{S}$.

Divergence $\mathrm{KL}(\mathcal{W}_{d,\underline{n}}^1||\mathcal{W}_{d,\underline{n}}^2)$, commonly used as a dissimilarity measure between covariance matrices, is sometimes referred as the log-Det divergence due to the form of $\varphi(S) = F_{\underline{n}}(S) \propto \log|S|$ (see Eq. 11.30). However, the dependency on term $\underline{n}$ should be neglected only when the two empirical covariance matrices comes from samples of the same size. In this case, log-Det divergence between two covariance matrices is the KL divergence in the sub-family $\mathcal{W}_{d,\underline{n}}$.

11.4.4 Toy Examples

In this part, some simple simulations are given for $d = 2$. Since the observations are positive semi-definite matrices, it is possible to visualize them with ellipses parametrized by their eigen decompositions. For example, Fig. 11.1 shows 20 matrices generated from $\mathcal{W}_d(.; n, S)$ for $n = 5$ and for $n = 50$ with S having eigenvalues $\{2, 1\}$. This visualization highlights the difficulty for the estimation of the parameters (even for d small) when n is small.

Then, a dataset of 60 matrices is generated from a three components mixture with parameters $\mathcal{W}_d(.; 10, S_1)$, $\mathcal{W}_d(.; 20, S_2)$, $\mathcal{W}_d(.; 30, S_3)$ and equal weights $w_1 = w_2 = w_3 = 1/3$. The respective eigenvalues for S_1, S_2, S_3 are in turn $\{2, 1\}$, $\{2, 0.5\}$, $\{1, 1\}$. Figure 11.2 illustrates this dataset. To study the influence of a good initialization for k-MLE, the Normalized Mutual Information (NMI) [28] is computed between the final partition and the ground-truth partition for different initializations. This value between 0 and 1 is higher when the two partitions are more similar. Following table gives average and standard deviation of NMI over 30 runs:

	Rand. Init/Lloyd	Rand. Init/Hartigan	k-MLE++/Hartigan
NMI	0.229 ± 0.279	0.243 ± 0.276	0.67 ± 0.083

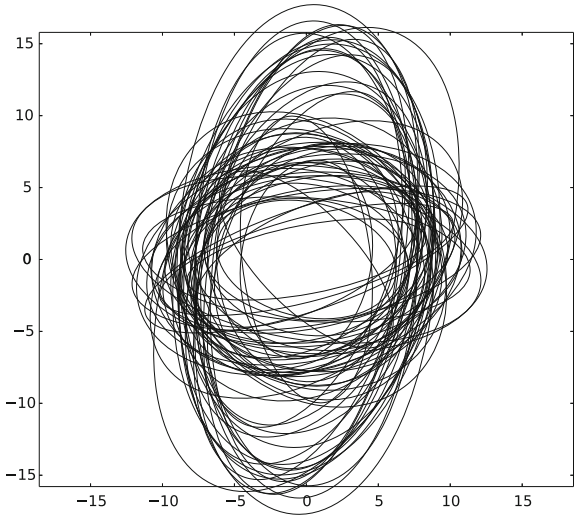


Fig. 11.2 Dataset: 60 matrices generated from a three components mixture of $\mathcal{W}_2$

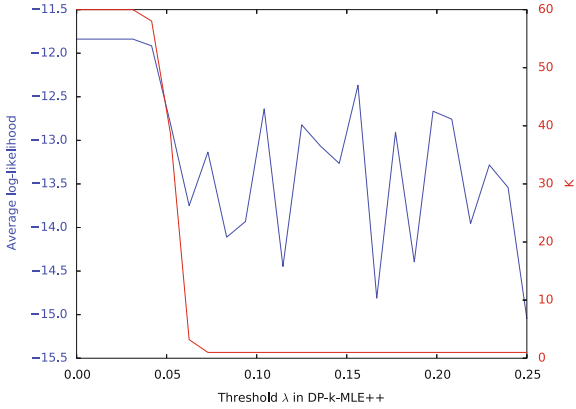


Fig. 11.3 DP-k-MLE++: Influence of λ on K and on the average log-likelihood

From this small experiment, we can easily verify the importance of a good initialization. Also, the partitions having the highest NMI are reported in Fig. 11.4 for each method. Let us mention that Hartigan method gives almost always a better partition than the Lloyd's one for the same initial mixture.

A last simulation indicates that the initialization with DP-k-MLE++ is very sensible to its parameter λ. Again with the same set of matrices, Fig. 11.3 shows how the number of generated clusters K and the average log-likelihood evolve with λ. Not surprisingly, both quantities decrease when λ increases.

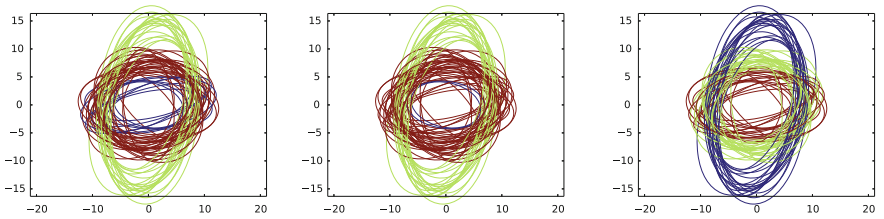


Fig. 11.4 Best partitions with Rand. Init/Lloyd (*left*), Rand. Init/Hartigan (*middle*), k-MLE++ Hartigan (*right*)

11.5 Application to Motion Retrieval

In this section, a potential application to motion retrieval is proposed following our previous work [23]. Raw motion-captured movement can be identifiable to a $n_i \times d$ matrix $\mathbb{X}_i$ where each row corresponds to captured locations of a set of sensors.

11.5.1 Movement Representation

When the aim is to provide a taxonomy of a set of movements, it is difficult to compare varying-size matrices. Cross-product matrices $X_i = {}^t\mathbb{X}_i\mathbb{X}_i$ is a possible descriptor[3] of $\mathbb{X}_i$. Denoting N the number of movements, set $\{X_1, ..., X_N\}$ of $d \times d$ matrix is exactly the input of k-MLE. Note that d can easily be of high dimension when the number of sensors is large.

The simplest way to initialize k-MLE in this setting is to apply DP-k-MLE++ for $\mathcal{W}_d$. But when n_i are known, it is better not to estimate them. In this case, DP-comp-k-MLE is appropriate for a function H selecting $\mathcal{W}_{d,\underline{n_i}}$ given $\xi_i = n_i$. When learning algorithm is fast enough, it is common practice to restart it for different initializations and to keep the best output (mixture parameters).

To enrich the description of a single movement, it is possible to define a mixture m_i per movement $\mathbb{X}_i$. For example, several subsets of successive observations with different sizes can be extracted and their cross-product matrices used as inputs for k-MLE (and DP-comp-k-MLE). Mixture m_i can be viewed as a sparse representation of local dynamics of $\mathbb{X}_i$ through their local second-order moments.

While these two representations are of different kind, it is possible to encompass both in a common framework for $\mathbb{X}_i$ described by a mixture of a single component $\{(w_{i,1} = 1, \eta_{i,1} = t(X_i))\}$. Algorithm k-MLE applied on such input for all movements (i.e. $\{t(X_i)\}_i$) provides then another set of mixture parameters $\{(\hat{w}_j, \hat{\eta}_j)\}_j$. Note that the general treatment of arbitrary mixtures of mixtures of Wishart is not claimed to be addressed here.

[3] For translation invariance, $\mathbb{X}_i$ are column centered before.

11.5.2 Querying with Cauchy-Schwartz Divergence

Let us consider a movement $\mathbb{X}$ (a $n \times d$ matrix) and its mixture representation m. Without loss of generality, let us denote $\{(w_j, \theta_j)\}_{j=1..K}$ the mixture parameters for m. The problem of comparing two movements amounts to compute a appropriate dissimilarity between m and another mixture m' of such a kind with parameters $\{(w'_j, \theta'_j)\}_{j=1..K'}$.

When both mixtures have a single component (K = K' = 1), an immediate solution is to consider the Kullback-Leibler divergence KL($m : m$) for two members of the same exponential family. Since it is the Bregman divergence on the swapped natural parameters $B_F(\theta' : \theta)$, a closed form is always available from Eq. (11.7). It is important to mention that this formula holds for θ and θ' viewed as parameters for $\mathcal{W}_d$ even if they are estimated in sub-families $W_{d,n}$ and $W_{d,n'}$.

For general mixtures of the same exponential family ($K > 1$ or $K' > 1$), KL divergence admits no more a closed form and has to be approximate with numerical methods. Recently, other divergences such as the Cauchy-Schwartz divergence (CS) [29] were shown to be available in a closed form:

$$\mathrm{CS}(m : m') = -\log \frac{\int m(x)m'(x)\mathrm{d}x}{\sqrt{\int m(x)^2\mathrm{d}x \int m'(x)^2\mathrm{d}x}}. \tag{11.40}$$

Within the same exponential family p_F, the integral of the product of mixtures is

$$\int m(x)m'(x)\mathrm{d}x = \sum_{j=1}^{K}\sum_{j'=1}^{K'} w_j w'_{j'} \int p_F(x;\theta_j)p_F(x;\theta'_{j'})\mathrm{d}x. \tag{11.41}$$

When carrier measure $k(X) = 0$, as it is for $\mathcal{W}_d$ but not for $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$, the integral can be further expanded as

$$\begin{aligned}
\int p_F(x;\theta_j)p_F(x;\theta'_{j'})\mathrm{d}X &= \int e^{\langle\theta_j, t(X)\rangle - F(\theta_j)} e^{\langle\theta'_j, t(X)\rangle - F(\theta'_{j'})}\mathrm{d}X \\
&= \int e^{\langle\theta_j+\theta'_{j'}, t(X)\rangle - F(\theta_j) - F(\theta'_{j'})}\mathrm{d}X \\
&= e^{F(\theta_j+\theta'_{j'}) - F(\theta_j) - F(\theta'_{j'})} \underbrace{\int e^{\langle\theta_j+\theta'_{j'}, t(X)\rangle - F(\theta_j+\theta'_{j'})}\mathrm{d}X}_{=1}.
\end{aligned}$$

Note that $\theta_j + \theta'_{j'}$ must be in the natural parameter space Θ to ensure that $F(\theta_j + \theta'_{j'})$ is finite. An equivalent condition is that Θ is a *convex cone*.

When $p_F = \mathcal{W}_d$, space $\Theta =]-1; +\infty[\times \mathcal{S}^p_{++}$ is not a convex cone since $\theta_{n_j} + \theta'_{n_{j'}} < -1$ for n_j and $n'_{j'}$ smaller than $d + 1$. Practically, this constraint is tested for each parameter pairs before going on with the computation the CS

divergence. A possible fix, not developed here, would be to constraint n to be greater than $d+1$ (or equivalently $\theta_n > 0$). Such a constraint amounts to take a convex subset $]0;+\infty[\times\mathcal{S}_{++}^{p}$ of Θ. Denoting $\Delta(\theta_j,\theta'_{j'}) = F(\theta_j+\theta'_{j'}) - F(\theta_j) - F(\theta'_{j'})$, the CS divergence is also

$$\begin{aligned}\mathrm{CS}(m:m') &= \frac{1}{2}\log\sum_{j=1}^{K}\sum_{j'=1}^{K}\left[w_j w_{j'}\exp^{\Delta(\theta_j,\theta_{j'})}\right] && (within\ m)\\ &+\frac{1}{2}\log\sum_{j=1}^{K'}\sum_{j'=1}^{K'}\left[w'_j w'_{j'}\exp^{\Delta(\theta'_j,\theta'_{j'})}\right] && (within\ m')\\ &-\log\sum_{j=1}^{K}\sum_{j'=1}^{K'}\left[w_j w'_{j'}\exp^{\Delta(\theta_j,\theta'_{j'})}\right] && (between\ m\ and\ m')\end{aligned} \tag{11.42}$$

Note that CS divergence is symmetric since $\Delta(\theta_j,\theta'_{j'})$ is. A numeric value of $\Delta(\theta_j,\theta'_{j'})$ can be computed for $\mathcal{W}_d$ from Eq. 11.28 (see Eq. 11.45 or 11.46 in the Appendix).

11.5.3 Summary of Proposed Motion Retrieval System

To conclude this section, let us recall the elements of our proposal for a motion retrieval system. Movement is represented by a Wishart mixture model learned by k-MLE initialized by DP-k-MLE++ or DP-comp-k-MLE. In the case of a mixture of a component, a simple application of the MLE for $\mathcal{W}_{d,\underline{n}}$ is sufficient. Although a Wishart distribution appears inadequate model for the scatter matrix X of a movement, it has been shown that this crude assumption provides a good classification rates on a real data set [23]. Learning representations of the movements may be performed offline since it is computational demanding. Using CS divergence as dissimilarity, we can then extract a taxonomy of movements with any spectral clustering algorithm. For a query movement, its representation by a mixture has to be computed first. Then it is possible to search the database for the most similar movements according to the CS divergence or to predict its type by a majority vote among them. More details of the implementation and results for the real dataset will be in a forthcoming technical report.

11.6 Conclusions and Perspectives

Hartigan's swap clustering method for k-MLE was studied for the general case of an exponential family. Unlike for k-means, this method does not guarantee to avoid empty clusters but achieves generally better performance than the Lloyd's heuristic.

Two methods DP-k-MLE and DP-comp-k-MLE are proposed to initialize k-MLE automatically by setting the number of clusters. While the former shares the good properties of k-MLE, the latter selects the component distributions given some extra knowledge. A small experiment indicates these methods appear to be quite sensible to their only parameter.

We recalled the definition and some properties of the Wishart distribution $\mathcal{W}_d$, especially its canonical decomposition as a member of an exponential family.By fixing either one of its two parameters n and S, two other (nested) exponential (sub-) families $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$ may be defined. From their respective MLEs, it is possible to define an iterative process which provably converges to the MLE for $\mathcal{W}_d$. For a single observation, the MLE does not exist.Then a crude solution is to replace the MLE in $\mathcal{W}_d$ by the MLE in one of the two sub-families.

The MLE is an example of a point estimator among many others (e.g. method of moments, minimax estimators, Bayesian point estimators). This suggests as future work many other learning algorithms such as k-MoM, k-Minimax [30], k-MAP following the same algorithmic scheme as k-MLE.

Finally, an application to the retrieval motion-captured motions is proposed.Each motion is described by a Wishart mixture model and the Cauchy-Schwarz divergence is used as a dissimilarity measure between two mixture models.As the CS divergence is always available in closed-form, such divergence is fast to compute compared to stochastic integration estimation schemes.This divergence can be used in spectral clustering methods and for visualization of a set of motions in an Euclidean embedding.

Another perspective is the connection between the closed-form divergences between mixtures and kernels based on divergences [31]: The CS divergence looks similar to the Normalized Correlation Kernel [32].This could lead to a broader class of methods (e.g., SVM) using these divergences.

Appendix A

This Appendix details some calculations for distributions $\mathcal{W}_d$, $\mathcal{W}_{d,\underline{n}}$, $\mathcal{W}_{d,\underline{S}}$.

A.1 Wishart Distribution $\mathcal{W}_d$

$$\begin{aligned}\mathcal{W}_d(X; n, S) &= \frac{|X|^{\frac{n-d-1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(S^{-1}X)\}}{2^{\frac{nd}{2}}|S|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})} \\ &= \exp\left\{\frac{n-d-1}{2}\log|X| - \frac{1}{2}\mathrm{tr}(S^{-1}X) - \frac{nd}{2}\log(2) - \frac{n}{2}\log|S| - \log\Gamma_d\left(\frac{n}{2}\right)\right\}\end{aligned}$$

Letting $(\theta_n, \theta_S) = (\frac{n-d-1}{2}, S^{-1}) \longleftrightarrow (n, S) = (2\theta_n + d + 1, \theta_S^{-1})$

$$\begin{aligned}
\mathcal{W}_d(X;\theta_n,\theta_S) &= \exp\left\{\frac{2\theta_n+d+1-d-1}{2}\log|X| - \frac{1}{2}\mathrm{tr}(\theta_S X) - \frac{(2\theta_n+d+1)d}{2}\log(2)\right.\\
&\quad\left. - \frac{(2\theta_n+d+1)}{2}\log|\theta_S^{-1}| - \log\Gamma_d\left(\frac{2\theta_n+d+1}{2}\right)\right\}\\
&= \exp\left\{\theta_n\log|X| - \frac{1}{2}\mathrm{tr}(\theta_S X) - \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|)\right.\\
&\quad\left. - \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right)\right\}\\
&= \exp\left\{< \theta_n, \log|X| >_{\mathbb{R}} + < \theta_S, -\frac{1}{2}X >_{HS} - F(\Theta)\right\}\\
&\quad \text{with } F(\Theta) = \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|) + \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right)\\
&= \exp\{< \Theta, t(X) > - F(\Theta) + k(X)\}\\
&\quad \text{with} t(X) = (\log|X|, -\frac{1}{2}X) \text{ and } k(X) = 0
\end{aligned}$$

$$F(\Theta) = \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|) + \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right)$$

$$\frac{\partial F}{\partial\theta_n}(\theta_n,\theta_S) = d\log(2) - \log|\theta_S| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right) \tag{11.43}$$

where Ψ_d is the multivariate Digamma function (or multivariate polygamma of order 0).

$$\frac{\partial F}{\partial\theta_S}(\theta_n,\theta_S) = -\left(\theta_n + \frac{(d+1)}{2}\right)\theta_S^{-1} \tag{11.44}$$

Dissimilarity $\Delta(\theta,\theta')$ between natural parameters $\theta = (\theta_n, \theta_S)$ and $\theta' = (\theta'_n, \theta'_S)$ is

$$\begin{aligned}
\Delta(\theta,\theta') &= F(\theta+\theta') - (F(\theta) + F(\theta')) = \left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S + \theta'_S|)\\
&\quad - \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|) - \left(\theta'_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta'_S|)\\
&\quad + \log\Gamma_d\left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right) - \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right) - \log\Gamma_d\left(\theta'_n + \frac{(d+1)}{2}\right)\\
&= -\frac{(d+1)}{2}d\log(2) + \left(\theta_n + \frac{(d+1)}{2}\right)\log|\theta_S| + \left(\theta'_n + \frac{(d+1)}{2}\right)\log|\theta'_S|\\
&\quad - \left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right)\log|\theta_S + \theta'_S| + \log\left(\frac{\Gamma_d\left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right)}{\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right)\Gamma_d\left(\theta'_n + \frac{(d+1)}{2}\right)}\right)
\end{aligned} \tag{11.45}$$

Remark $\Delta(\theta, \theta) \neq 0$. Same quantity with source parameters $\lambda = (n, S)$ and $\lambda' = (n', S')$ is

$$\Delta(\lambda, \lambda') = -\frac{(d+1)}{2} d \log(2) - \frac{n}{2} \log|S| - \frac{n'}{2} \log|S'| - \frac{n+n'-d-1}{2} \log|S^{-1} + S'^{-1}| + \log\left(\frac{\Gamma_d\left(\frac{n+n'-d-1}{2}\right)}{\Gamma_d\left(\frac{n}{2}\right)\Gamma_d\left(\frac{n'}{2}\right)}\right) \tag{11.46}$$

A.2 Distribution $\mathcal{W}_{d,\underline{n}}$

$$\mathcal{W}_d(X; \underline{n}, S) = \frac{|X|^{\frac{\underline{n}-d-1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(S^{-1}X)\}}{2^{\frac{\underline{n}d}{2}} |S|^{\frac{\underline{n}}{2}} \Gamma_d(\frac{\underline{n}}{2})}$$
$$= \exp\left\{\frac{\underline{n}-d-1}{2} \log|X| - \frac{1}{2}\mathrm{tr}(S^{-1}X) - \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|S| - \log\Gamma_d\left(\frac{\underline{n}}{2}\right)\right\}$$

Letting $\theta_S = S^{-1}$,

$$\mathcal{W}_d(X; \underline{n}, \theta_S) = \exp\left\{-\frac{1}{2}\mathrm{tr}(\theta_S X) + \frac{\underline{n}-d-1}{2}\log|X| - \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|\theta_S^{-1}| - \log\Gamma_d\left(\frac{\underline{n}}{2}\right)\right\}$$
$$= \exp\left\{< \theta_S, -\frac{1}{2}X >_{HS} + k(X) - F_{\underline{n}}(\theta_S)\right\}$$
$$\text{with } F_{\underline{n}}(\theta_S) = \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|\theta_S| + \log\Gamma_d\left(\frac{\underline{n}}{2}\right)$$
$$\text{with } k_{\underline{n}}(X) = \frac{\underline{n}-d-1}{2}\log|X|$$

Using the rule $\frac{\partial log|X|}{\partial X} =^t (X^{-1})$ [33] and the symmetry of θ_S, we get

$$\nabla_{\theta_S} F_{\underline{n}}(\theta_S) = -\frac{\underline{n}}{2}\theta_S^{-1}$$

The correspondence between natural parameter θ_S and expectation parameter η_S is

$$\eta_S = \nabla_{\theta_S} F_{\underline{n}}(\theta_S) = -\frac{\underline{n}}{2}\theta_S^{-1} \longleftrightarrow \theta_S = \nabla_{\eta_S} F^*_{\underline{n}}(\eta_S) = (\nabla_{\theta_S} F_{\underline{n}})^{-1}(\eta_S) = -\frac{\underline{n}}{2}\eta_S^{-1}$$

Finally, we obtain the MLE for θ_S in this sub family:

$$\hat{\theta}_S = -\frac{\underline{n}}{2}\left(\frac{1}{N}\sum_{i=1}^{N} -\frac{1}{2}X_i\right)^{-1} = \underline{n}N\left(\sum_{i=1}^{N} X_i\right)^{-1}$$

Same formulation with source parameter S:

$$\hat{S} = \hat{\theta}_S^{-1} = \left(\underline{n} N \left(\sum_{i=1}^{N} X_i \right)^{-1} \right)^{-1} = \frac{\sum_{i=1}^{N} X_i}{\underline{n} N}$$

Dual log-normalizer $F_{\underline{n}}^*$ for $\mathcal{W}_{d,\underline{n}}$ is

$$\begin{aligned}
F_{\underline{n}}^*(\eta_S) &= \langle (\nabla F_{\underline{n}})^{-1}(\eta_S), \eta_S \rangle - F_{\underline{n}}((\nabla F_{\underline{n}})^{-1}(\eta_S)) \\
&= \langle -\frac{\underline{n}}{2} \eta_S^{-1}, \eta_S \rangle - F_{\underline{n}}(-\frac{\underline{n}}{2} \eta_S^{-1}) \\
&= -\frac{\underline{n}}{2} \mathrm{tr}(\eta_S^{-1} \eta_S) - \frac{\underline{n} d}{2} \log(2) + \frac{\underline{n}}{2} \log \left[(\frac{\underline{n}}{2})^d |-\eta_S^{-1}| \right] - \log \Gamma_d \left(\frac{\underline{n}}{2} \right) \\
&= -\frac{\underline{n} d}{2} (1 + \log(2) - \log \underline{n} + \log 2) + \frac{\underline{n}}{2} \log |-\eta_S^{-1}| - \log \Gamma_d \left(\frac{\underline{n}}{2} \right) \\
&= \frac{\underline{n} d}{2} \log \left(\frac{\underline{n}}{4e} \right) + \frac{\underline{n}}{2} \log |-\eta_S^{-1}| - \log \Gamma_d \left(\frac{\underline{n}}{2} \right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}(\mathcal{W}_{d,\underline{n}}^1 || \mathcal{W}_{d,\underline{n}}^2) &= B_{F_{\underline{n}}}(\theta_{S_2} : \theta_{S_1}) \\
&= F_{\underline{n}}(\theta_{S_2}) - F_{\underline{n}}(\theta_{S_1}) - <\theta_{S_2} - \theta_{S_1}, \nabla_{\theta_S} F_{\underline{n}}(\theta_{S_1})> \\
&= \frac{\underline{n}}{2} \left(\log |\theta_{S_1}| - \log |\theta_{S_2}| \right) + \frac{\underline{n}}{2} \mathrm{tr}((\theta_{S_2} - \theta_{S_1}) \theta_{S_1}^{-1}) \\
&= \frac{\underline{n}}{2} \left(\log \frac{|\theta_{S_1}|}{|\theta_{S_2}|} + \mathrm{tr}(\theta_{S_2} \theta_{S_1}^{-1}) - d \right) \\
&= \frac{\underline{n}}{2} \left(-\log \frac{|\theta_{S_2}|}{|\theta_{S_1}|} + \mathrm{tr}(\theta_{S_2} \theta_{S_1}^{-1}) - d \right)
\end{aligned}$$

also with source parameter

$$\mathrm{KL}(\mathcal{W}_{d,\underline{n}}^1 || \mathcal{W}_{d,\underline{n}}^2) = \frac{\underline{n}}{2} \left(-\log \frac{|S_1|}{|S_2|} + \mathrm{tr}(S_2^{-1} S_1) - d \right)$$

Let's remark that KL divergence depends now on $\underline{n}$.

$$\begin{aligned}
B_{F_{\underline{n}}^*}(\eta_{S_1} : \eta_{S_2}) &= F_{\underline{n}}^*(\eta_{S_1}) - F_{\underline{n}}^*(\eta_{S_2}) - <\eta_{S_1} - \eta_{S_2}, \nabla F_{\underline{n}}^*(\eta_{S_2})>_{HS} \\
&= \frac{\underline{n}}{2} \left(\log |-\eta_{S_1}^{-1}| - \log |-\eta_{S_2}^{-1}| \right) - <\eta_{S_1} - \eta_{S_2}, -\frac{\underline{n}}{2} \eta_{S_2}^{-1}>_{HS} \\
&= \frac{\underline{n}}{2} \left(\log \frac{|-\eta_{S_1}^{-1}|}{|-\eta_{S_2}^{-1}|} + \mathrm{tr}(\eta_{S_1} \eta_{S_2}^{-1}) - d \right)
\end{aligned}$$

A.3 Distribution $\mathcal{W}_{d,\underline{S}}$

For fixed $\underline{S}$, the p.d.f of $\mathcal{W}_{d,\underline{S}}$ can be rewritten[4] as

$$\begin{aligned}\mathcal{W}_d(X;n,\underline{S}) &= \frac{|X|^{\frac{n-d-1}{2}}\exp\{-\frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X)\}}{|2\underline{S}|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})}\\ &= \exp\left\{\frac{n-d-1}{2}\log|X| - \frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X) - \frac{n}{2}\log|2\underline{S}| - \log\Gamma_d\left(\frac{n}{2}\right)\right\}\end{aligned}$$

Letting $\theta_n = \frac{n-d-1}{2}$ $(n = 2\theta_n + d + 1)$

$$\begin{aligned}\mathcal{W}_d(X;\theta_n,\underline{S}) &= \exp\left\{\theta_n\log|X| - \frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X) - \left(\theta_n + \frac{d+1}{2}\right)\log|2\underline{S}| - \log\Gamma_d\left(\theta_n + \frac{d+1}{2}\right)\right\}\\ &= \exp\left\{< \theta_n, \log|X| > + k_{\underline{S}}(X) - F_{\underline{S}}(\theta_n)\right\}\\ &\quad \text{with } F_{\underline{S}}(\theta_n) = \left(\theta_n + \frac{d+1}{2}\right)\log|2\underline{S}| + \log\Gamma_d\left(\theta_n + \frac{d+1}{2}\right)\\ &\quad \text{with } k_{\underline{S}}(X) = -\frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X)\end{aligned}$$

The correspondence between natural parameter θ_n and expectation parameter η_n is

$$\eta_n = \nabla_{\theta_n} F_{\underline{S}}(\theta_n) = \log|2\underline{S}| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right)$$

$$\begin{aligned}&\Leftrightarrow \quad \Psi_d\left(\theta_n + \tfrac{(d+1)}{2}\right) = \eta_n - \log|2\underline{S}|\\ &\Leftrightarrow \quad \theta_n + \tfrac{(d+1)}{2} = \Psi_d^{-1}\left(\eta_n - \log|2\underline{S}|\right)\\ &\Leftrightarrow \theta_n = \Psi_d^{-1}\left(\eta_n - \log|2\underline{S}|\right) - \tfrac{(d+1)}{2} = (\nabla F_{\underline{S}})^{-1}(\eta_n) = \nabla F^*_{\underline{S}}(\eta_n)\end{aligned}$$

Finally, we obtain the MLE for θ_n in this sub family:

$$\hat{\theta}_n = \Psi_d^{-1}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\log|X|\right] - \log|2\underline{S}|\right) - \frac{(d+1)}{2}$$

Same formulation with source parameter n:

[4] Since $|2S| = 2^d|S|$, we have $2^{\frac{nd}{2}}|S|^{\frac{n}{2}}$ that is equivalent to $|2S|^{\frac{n}{2}}$.

$$\frac{\hat{n}-d-1}{2} = \Psi_d^{-1}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\log|X|\right] - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}$$

$$\hat{n} = 2\Psi_d^{-1}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\log|X|\right] - \log\left|2\underline{S}\right|\right)$$

Dual log-normalizer $F^*_{\underline{S}}$ for $\mathcal{W}_{d,\underline{S}}$ is

$$\begin{aligned}
F^*_{\underline{S}}(\eta_n) &= \langle(\nabla F_{\underline{S}})^{-1}(\eta_n), \eta_n\rangle - F_{\underline{S}}((\nabla F_{\underline{S}})^{-1}(\eta_n)) \\
&= \langle\Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}, \eta_n\rangle \\
&\quad - \Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\log\left|2\underline{S}\right| - \log\Gamma_d\left(\Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\right) \\
&= \Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\left(\eta_n - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}\eta_n - \log\Gamma_d\left(\Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}(\mathcal{W}^1_{d,\underline{S}}||\mathcal{W}^2_{d,\underline{S}}) &= B_{F_{\underline{S}}}(\theta_{n_2} : \theta_{n_1}) = F_{\underline{S}}(\theta_{n_2}) - F_{\underline{S}}(\theta_{n_1}) - \langle\theta_{n_2} - \theta_{n_1}, \nabla F_{\underline{S}}(\theta_{n_1})\rangle \\
&= \left(\theta_{n_2} + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| + \log\Gamma_d\left(\theta_{n_2} + \frac{d+1}{2}\right) \\
&\quad - \left(\theta_{n_1} + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| - \log\Gamma_d\left(\theta_{n_1} + \frac{d+1}{2}\right) \\
&\quad - \langle\theta_{n_2} - \theta_{n_1}, \log\left|2\underline{S}\right| + \Psi_d\left(\theta_{n_1} + \frac{(d+1)}{2}\right)\rangle
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}(\mathcal{W}^1_{d,\underline{S}}||\mathcal{W}^2_{d,\underline{S}}) &= \log\frac{\Gamma_d\left(\theta_{n_2} + \frac{d+1}{2}\right)}{\Gamma_d\left(\theta_{n_1} + \frac{d+1}{2}\right)} - (\theta_{n_2} - \theta_{n_1})\Psi_d\left(\theta_{n_1} + \frac{(d+1)}{2}\right) \\
&= -\log\frac{\Gamma_d\left(\theta_{n_1} + \frac{d+1}{2}\right)}{\Gamma_d\left(\theta_{n_2} + \frac{d+1}{2}\right)} + (\theta_{n_1} - \theta_{n_2})\Psi_d\left(\theta_{n_1} + \frac{(d+1)}{2}\right)
\end{aligned}$$

also with source parameter

$$\mathrm{KL}(\mathcal{W}^1_{d,\underline{S}}||\mathcal{W}^2_{d,\underline{S}}) = -\log\left(\frac{\Gamma_d\left(\frac{n_1}{2}\right)}{\Gamma_d\left(\frac{n_2}{2}\right)}\right) + \left(\frac{n_1 - n_2}{2}\right)\Psi_d\left(\frac{n_1}{2}\right)$$

Let us remark that this quantity does not depend on $\underline{S}$.

$$\begin{aligned}
B_{F^*_{\underline{S}}}(\eta_{n_1} : \eta_{n_2}) &= F^*_{\underline{S}}(\eta_{n_1}) - F^*_{\underline{S}}(\eta_{n_2}) - <\eta_{n_1} - \eta_{n_2}, \nabla F^*_{\underline{S}}(\eta_{n_2})>_{HS} \\
&= \Psi_d^{-1}\left(\eta_{n_1} - \log\left|2\underline{S}\right|\right)\left(\eta_{n_1} - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}\eta_{n_1} \\
&\quad - \log\Gamma_d\left(\Psi_d^{-1}\left(\eta_{n_1} - \log\left|2\underline{S}\right|\right)\right) \\
&\quad - \Psi_d^{-1}\left(\eta_{n_2} - \log\left|2\underline{S}\right|\right)\left(\eta_{n_2} - \log\left|2\underline{S}\right|\right) + \frac{(d+1)}{2}\eta_{n_2}
\end{aligned}$$

$$
\begin{aligned}
&+ \log \Gamma_d \left(\Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) \right) \\
&- \langle \eta_{n_1} - \eta_{n_2}, \Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) - \frac{(d+1)}{2} \rangle_{HS} \\
B_{F_{\underline{S}}^*}(\eta_{n_1} : \eta_{n_2}) &= \log \frac{\Gamma_d \left(\Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) \right)}{\Gamma_d \left(\Psi_d^{-1} \left(\eta_{n_1} - \log \left| 2\underline{S} \right| \right) \right)} \\
&- \left[\Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) - \Psi_d^{-1} \left(\eta_{n_1} - \log \left| 2\underline{S} \right| \right) \right] \left(\eta_{n_1} - \log \left| 2\underline{S} \right| \right)
\end{aligned}
$$

References

1. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley Series in Probability and Statistics. Wiley-Interscience, New York (2008)
2. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**, 1705–1749 (2005)
3. Nielsen, F.: k-MLE: a fast algorithm for learning statistical mixture models. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 869–872 (2012). Long version as arXiv:1203.5181
4. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. **31**, 651–666 (2010)
5. Wishart, J.: The generalised product moment distribution in samples from a Normal multivariate population. Biometrika **20**(1/2), 32–52 (1928)
6. Tsai, M.-T.: Maximum likelihood estimation of Wishart mean matrices under Lwner order restrictions. J. Multivar. Anal. **98**(5), 932–944 (2007)
7. Formont, P., Pascal, T., Vasile, G., Ovarlez, J.-P., Ferro-Famil, L.: Statistical classification for heterogeneous polarimetric SAR images. IEEE J. Sel. Top. Sign. Proces. **5**(3), 567–576 (2011)
8. Jian, B., Vemuri, B.: Multi-fiber reconstruction from diffusion MRI using mixture of wisharts and sparse deconvolution. In: Information Processing in Medical Imaging, pp. 384–395, Springer, Berlin (2007)
9. Cherian, A., Morellas, V., Papanikolopoulos, N., Bedros, S.: Dirichlet process mixture models on symmetric positive definite matrices for appearance clustering in video surveillance applications. In: Computer Vision and Pattern Recognition (CVPR), pp. 3417–3424 (2011)
10. Nielsen, F., Garcia, V.: Statistical exponential families: a digest with flash cards. http://arxiv.org/abs/0911.4863.. Accessed Nov 2009
11. Rockafellar, R.T.: Convex Analysis, vol. 28. Princeton University Press, Princeton (1997)
12. Wainwright, M.J., Jordan, M.J.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. **1**(1–2), 1–305 (2008)
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. (Methodological). **39** 1–38 (1977)
14. Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. **14**(3), 315–332 (1992)
15. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. J. Roy. Stat. Soc. C (Applied Statistics). **28**(1), 100–108 (1979)
16. Telgarsky, M., Vattani, A.: Hartigan's method: k-means clustering without Voronoi. In: Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 820–827 (2010)

17. Nielsen, F., Boissonnat, J.D., Nock, R.: On Bregman Voronoi diagrams. In: ACM-SIAM Symposium on Discrete Algorithms, pp. 746–755 (2007)
18. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Networks **16**(3), 645–678 (2005)
19. Kulis, B., Jordan, M.I.: Revisiting k-means: new algorithms via Bayesian nonparametrics. In: International Conference on Machine Learning (ICML) (2012)
20. Ackermann, M.R.: Algorithms for the Bregman K-median problem. PhD thesis. Paderborn University (2009)
21. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
22. Ji, S., Krishnapuram, B., Carin, L.: Variational Bayes for continuous hidden Markov models and its application to active learning. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 522–532 (2006)
23. Hidot, S., Saint-Jean, C.: An Expectation-Maximization algorithm for the Wishart mixture model: application to movement clustering. Pattern Recogn. Lett. **31**(14), 2318–2324 (2010)
24. Brent. R.P.: Algorithms for Minimization Without Derivatives. Courier Dover Publications, Mineola (1973)
25. Bezdek, J.C., Hathaway, R.J., Howard, R.E., Wilson, C.A., Windham, M.P.: Local convergence analysis of a grouped variable version of coordinate descent. J. Optim. Theory Appl. **54**(3), 471–477 (1987)
26. Bogdan, K., Bogdan, M.: On existence of maximum likelihood estimators in exponential families. Statistics **34**(2), 137–149 (2000)
27. Ciuperca, G., Ridolfi, A., Idier, J.: Penalized maximum likelihood estimator for normal mixtures. Scand. J. Stat. **30**(1), 45–59 (2003)
28. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
29. Nielsen, F.: Closed-form information-theoretic divergences for statistical mixtures. In: International Conference on Pattern Recognition (ICPR), pp. 1723–1726 (2012)
30. Haff, L.R., Kim, P.T., Koo, J.-Y., Richards, D.: Minimax estimation for mixtures of Wishart distributions. Ann. Stat. **39**(6), 3417–3440 (2011)
31. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. J. Mach. Learn. Res. **5**, 819–844 (2004)
32. Moreno, P.J., Ho, P., Vasconcelos, N.: A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In: Advances in Neural Information Processing Systems (2003)
33. Petersen, K.B., Pedersen, M.S.: The matrix cookbook. http://www2.imm.dtu.dk/pubdb/p.php?3274. Accessed Nov 2012

Chapter 12
Morphological Processing of Univariate Gaussian Distribution-Valued Images Based on Poincaré Upper-Half Plane Representation

Jesús Angulo and Santiago Velasco-Forero

Abstract Mathematical morphology is a nonlinear image processing methodology based on the application of complete lattice theory to spatial structures. Let us consider an image model where at each pixel is given a univariate Gaussian distribution. This model is interesting to represent for each pixel the measured mean intensity as well as the variance (or uncertainty) for such measurement. The aim of this work is to formulate morphological operators for these images by embedding Gaussian distribution pixel values on the Poincaré upper-half plane. More precisely, it is explored how to endow this classical hyperbolic space with various families of partial orderings which lead to a complete lattice structure. Properties of order invariance are explored and application to morphological processing of univariate Gaussian distribution-valued images is illustrated.

Keywords Ordered Poincaré half-plane · Hyperbolic partial ordering · Hyperbolic complete lattice · Mathematical morphology · Gaussian-distribution valued image · Information geometry image filtering

12.1 Introduction

This work is motivated by the exploration of a mathematical image model f where instead of having a scalar intensity $t \in \mathbb{R}$ at each pixel p, i.e., $f(p) = t$, we have, a univariate Gaussian probability distribution of intensities $N(\mu, \sigma^2) \in \mathcal{N}$, i.e., image f is defined as the function

J. Angulo (✉)
CMM-Centre de Morphologie Mathématique, Mathématiques et Systèmes,
MINES ParisTech, Paris, France
e-mail: jesus.angulo@mines-paristech.fr

S. Velasco-Forero
Department of Mathematics, National University of Singapore, Buona Vista, Singapore
e-mail: matsavf@nus.edu.sg

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_12,

$$f: \begin{cases} \Omega \to \mathcal{N} \\ p \mapsto N(\mu, \sigma^2) \end{cases}$$

where Ω is the support space of pixels p (e.g., for 2D images $\Omega \subset \mathbb{Z}^2$) and $\mathcal{N}$ denotes the family of univariate Gaussian probability distribution functions (pdf). Nowadays most of imaging sensors only produce single scalar values since the CCD (charge-coupled device) cameras typically integrates the light (arriving photons) during a given exposure time τ. To increase the signal-to-noise ratio (SNR), exposure time is increased to $\tau' = \alpha\tau$, $\alpha > 1$. Let suppose that α is a positive integer number, this is equivalent to a multiple acquisition of α frames during τ for each frame (i.e., a kind of temporal oversampling). The standard approach only considers the sum (or average) of the multiple intensities [28], without taking into account the variance which is a basic estimator of the noise useful for probabilistic image processing. Another example of such a representation from a gray scale image consists in considering that each pixel is described by the mean and the variance of the intensity distribution from its centered neighboring patch. This model has been for instance recently used in [10] for computing local estimators which can be interpreted as pseudo-morphological operators.

Let us consider the example of gray scale image parameterized by the mean and the standard deviation of patches given in Fig. 12.1. We observe that the underlying geometry of this space of patches is not Euclidean, e.g., the geodesics are clearly curves. In fact, as we discuss in the paper, this parametrization corresponds to one of the models of hyperbolic geometry.

Henceforth, the corresponding image processing operators should be able to deal with Gaussian distributions-valued pixels. In particular, morphological operators for images $f \in \mathcal{F}(\Omega, \mathcal{N})$ involve that the space of Gaussian distributions $\mathcal{N}$ must be endowed of a partial ordering leading to a complete lattice structure. In practice, it means that given a set of Gaussian pdfs, as the example given in Fig. 12.2, we need to be able to define a Gaussian pdf which corresponds to the infimum (inf) of the set and another one to the supremum (sup). Mathematical morphology is a nonlinear image processing methodology based on the computation of sup/inf-convolution filters (i.e., dilation/erosion operators) in local neighborhoods [31]. Mathematical morphology is theoretically formulated in the framework of complete lattices and operators defined on them [21, 29]. When only the supremum or the infimum are well defined, other morphological operators can be formulated in the framework of complete semilattices [22, 23]. Both cases are considered here for images $f \in \mathcal{F}(\Omega, \mathcal{N})$.

A possible way to deal with the partial ordering problem of $\mathcal{N}$ can be founded on stochastic ordering (or stochastic dominance) [30] which is basically defined in terms of majorization of cumulative distribution functions.

However, we prefer to adopt here an information geometry approach [4], which is based on considering that the univariate Gaussian pdfs are points in a hyperbolic space [3, 9]. More generally, Fisher geometry amounts to hyperbolic geometry of constant curvature for other location-scale families of probability distributions (Cauchy, Laplace, elliptical) $p(x; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where curvature depends on

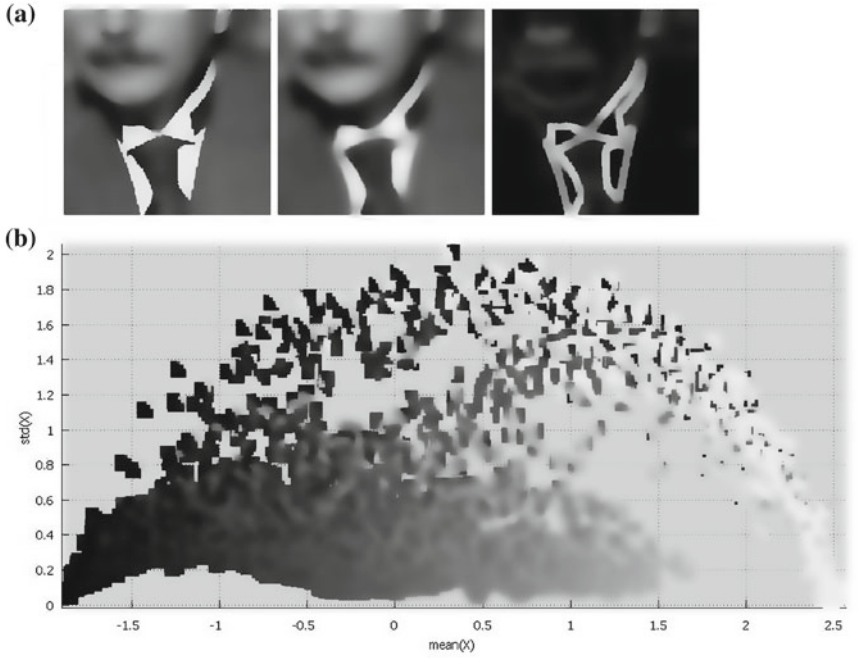


Fig. 12.1 Parametrization of a *gray scale* image where each pixel is described by the mean and the variance of the intensity distribution from its centered neighboring patch of 5×5 pixels: **a** *Left*, original *gray scale* image; *center*, image of mean of each patch; *right*, image of standard deviation of each patch. **b** Visualization of the patch according to their coordinates in the space mean/std.dev

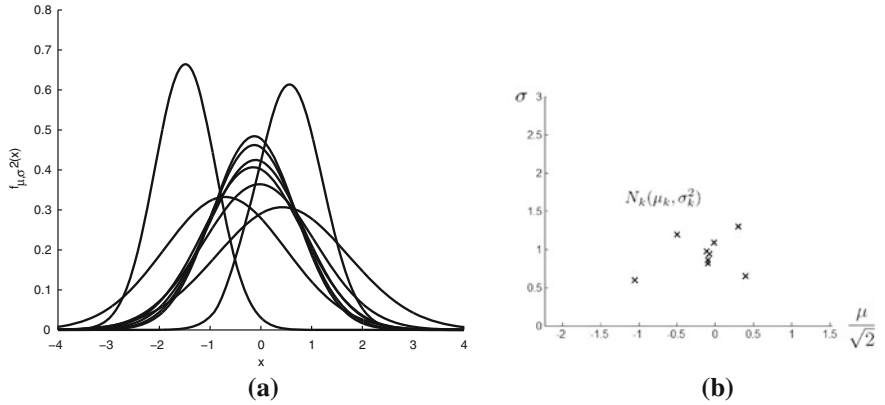


Fig. 12.2 **a** Example of a set of nine univariate Gaussian pdfs, $N_k(\mu_k, \sigma_k^2)$, $1 \leq k \leq 9$. **b** Same set of Gaussian pdfs represented as points of coordinates $(x_k = \mu_k/\sqrt{2}, y_k = \sigma_k)$ in the upper-half plane

the dimension and the density profile [3, 14, 15]. For a deep flavor on hyperbolic geometry see [12]. There are several models representing the hyperbolic space in $\mathbb{R}^d$, $d > 1$, such as the three following ones: the (Poincaré) upper half-space model $\mathcal{H}^d$, the Poincaré disk model $\mathcal{P}^d$ and the Klein disk model $\mathcal{K}^d$.

1. The (Poincaré) upper half-space model is the domain $\mathcal{H}^d = \{(x_1, \cdots, x_d) \in \mathbb{R}^d \mid x_d > 0\}$ with the Riemannian metric $ds^2 = \frac{dx_1^2+\cdots+dx_d^2}{x_d^2}$;
2. The Poincaré disk model is the domain $\mathcal{P}^d = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid x_1^2 + \cdots + x_d^2 < 1\}$ with the Riemannian metric $ds^2 = 4\frac{dx_1^2+\cdots+dx_d^2}{(1-x_1^2-\cdots-x_d^2)^2}$;
3. The Klein disk model is the space $\mathcal{K}^d = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid x_1^2+\cdots+x_d^2 < 1\}$ with the Riemannian metric $ds^2 = \frac{dx_1^2+\cdots+dx_d^2}{1-x_1^2-\cdots-x_d^2} + \frac{(x_1dx_1+\cdots+x_ddx_d)^2}{(1-x_1^2-\cdots-x_d^2)^2}$.

These models are isomorphic between them in the sense that one-to-one correspondences can be set up between the points and lines in one model to the points and lines in the other so as to preserve the relations of incidence, betweenness and congruence. In particular, there exists an isometric mapping between any pair among these models and analytical transformations to convert from one to another are well known [12].

Klein disk model has been considered for instance in computational information geometry (Voronoï diagrams, clustering, etc.) [26] and Poincaré disk model in information geometric radar processing [5–7]. In this paper, we focus on the Poincaré half-plane model, $\mathcal{H}^2$, which is sufficient for our practical purposes of manipulating Gaussian pdfs. Figure 12.2b illustrates the example of a set of nine Gaussian pdfs $N_k(\mu_k, \sigma_k^2)$ represented as points of coordinates $(\mu_k/\sqrt{2}, \sigma_k)$ in the upper-half plane as follows:

$$(\mu, \sigma) \mapsto z = \frac{\mu}{\sqrt{2}} + i\sigma.$$

The rationale behind the scaling factor $\sqrt{2}$ is given in Sect. 12.3.

In summary, from a theoretical viewpoint, the aim of this paper is to endow $\mathcal{H}^2$ with partial orderings which lead to useful invariance properties in order to formulate appropriate morphological operators for images $f\colon \Omega \to \mathcal{H}^2$. This work is an extension to the conference paper [1]. The rest of the paper is organized as follows. Section 12.2 reminds the basics on the geometry of Poincaré half-plane model. The connection between Poincaré half-plane model of hyperbolic geometry and Fisher Information geometry of Gaussian distributions is briefly recalled in Sect. 12.3. Then, various partial orderings on $\mathcal{H}^2$ are studied in Sect. 12.4. Based on the corresponding complete lattice structure of $\mathcal{H}^2$, Sect. 12.5 presents definition of morphological operators for images on $\mathcal{F}(\Omega, \mathcal{H}^2)$ and its application to morphological processing univariate Gaussian distribution-valued images. Section 12.6 concludes the paper with the perspectives of the present work.

12.2 Geometry of Poincaré Upper-Half Plane $\mathcal{H}^2$

In complex analysis, the upper-half plane is the set of complex numbers with positive imaginary part:

$$\mathcal{H}^2 = \{z = x + iy \in \mathbb{C} \mid \ y > 0\}. \tag{12.1}$$

We also use the notation $x = \Re(z)$ and $y = \Im(z)$. The boundary of upper-half plane (called sometimes circle at infinity) is the real axis together with the infinity, i.e., $\partial\mathcal{H}^2 = \mathbb{R} \cup \infty = \{z = x + iy \mid \ y = 0, x = \pm\infty, y = \infty\}$.

12.2.1 Riemannian Metric, Angle and Distance

In hyperbolic geometry, the Poincaré upper-half plane model (originated with Beltrami and also known as Lobachevskii space in Soviet scientific literature) is the space $\mathcal{H}^2$ together with the Poincaré metric

$$(g_{kl})_{k,l=1,2} = \begin{pmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix} \tag{12.2}$$

such that the hyperbolic arc length is given by

$$ds^2 = \sum_{k,l=1,2} g_{kl}\, dx\, dy = \frac{dx^2 + dy^2}{y^2} = \frac{|dz|^2}{y^2} = y^{-1} dz\, y^{-1} dz^*. \tag{12.3}$$

With this metric, the Poincaré upper-half plane is a complete Riemannian manifold of constant sectional curvature K equal to -1. We can consider a continuum of other hyperbolic spaces by multiplying the hyperbolic arc length (12.3) by a positive constant k which leads to a metric of constant Gaussian curvature $K = -1/k^2$. The tangent space to $\mathcal{H}^2$ at a point z is defined as the space of tangent vectors at z. It has the structure of a 2-dimensional real vector space, $T_z\mathcal{H}^2 \simeq \mathbb{R}^2$. The Riemannian metric (12.3) is induced by the following inner product on $T_z\mathcal{H}^2$: for $\zeta_1, \zeta_2 \in T_z\mathcal{H}^2$, with $\zeta_k = (\xi_k, \eta_k)$, we put

$$\langle \zeta_1, \zeta_2 \rangle_z = \frac{(\zeta_1, \zeta_2)}{\Im(z)^2} \tag{12.4}$$

which is a scalar multiple of the Euclidean inner product $(\zeta_1, \zeta_2) = \xi_1\xi_2 + \eta_1\eta_2$.

The angle θ between two geodesics in $\mathcal{H}^2$ at their intersection point z is defined as the angle between their tangent vectors in $T_z\mathcal{H}^2$, i.e.,

$$\cos\theta = \frac{\langle \zeta_1, \zeta_2 \rangle_z}{\|\zeta_1\|_z \|\zeta_2\|_z} = \frac{(\zeta_1, \zeta_2)}{\sqrt{(\zeta_1, \zeta_1)}\sqrt{(\zeta_2, \zeta_2)}}. \tag{12.5}$$

We see that this notion of angle measure coincides with the Euclidean angle measure. Consequently, the Poincaré upper-half plane is a conformal model.

The distance between two points $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ in $\left(\mathcal{H}^2, ds^2\right)$ is the function

$$\text{dist}_{\mathcal{H}^2}(z_1, z_2) = \cosh^{-1}\left(1 + \frac{(x_1 - x_2)^2 + (y_1 - y_2)^2}{2y_1 y_2}\right) \tag{12.6}$$

Distance (12.6) is derived from the logarithm of the cross-ratio between these two points and the points at the infinity, i.e., $\text{dist}_{\mathcal{H}^2}(z_1, z_2) = \log D(z_1^\infty, z_1, z_2, z_2^\infty)$ where $D(z_1^\infty, z_1, z_2, z_2^\infty) = \frac{z_1 - z_2^\infty}{z_1 - z_1^\infty} \frac{z_2 - z_1^\infty}{z_2 - z_2^\infty}$. To obtain their equivalence, we remind that $\cosh^{-1}(x) = \log(x + \sqrt{x^2 - 1})$. From this formulation it is easy to check that for two points with $x_1 = x_2$ the distance is $\text{dist}_{\mathcal{H}^2}(z_1, z_2) = \left|\log\left(\frac{y_1}{y_2}\right)\right|$.

To see that $\text{dist}_{\mathcal{H}^2}(z_1, z_2)$ is a metric distance in $\mathcal{H}^2$, we first notice the argument of $\cosh^{-1}$ always lies in $[1, \infty)$ and $\cosh(x) = \frac{e^x + e^{-x}}{2}$, so cosh is increasing and concave on $[0, \infty)$. Thus $\cosh^{-1}(1) = 0$ and $\cosh^{-1}$ is increasing and concave down on $[1, \infty)$, growing logarithmically. The properties required to be a metric (non-negativity, symmetry and triangle inequality) are proven using the cross-ratio formulation of the distance.

We note that the distance from any point $z \in \mathcal{H}^2$ to $\partial\mathcal{H}^2$ is infinity.

12.2.2 Geodesics

The geodesics of $\mathcal{H}^2$ are the vertical lines, $VL(a) = \{z \in \mathcal{H}^2 \mid \Re(z) = a\}$, and the semi-circles in $\mathcal{H}^2$ which meet the horizontal axis $\Re(z) = 0$ orthogonally, $SC_r(a) = \{z \in \mathcal{H}^2 \mid |z - z'| = r;\ \Re(z') = a \text{ and } \Im(z') = 0\}$; see Fig. 12.3a. Thus given any pair $z_1, z_2 \in \mathcal{H}^2$, there is a unique geodesic connecting them, parameterized for instance in polar coordinates by the angle, i.e.,

$$\gamma(t) = a + re^{it} = (a + r\cos t) + i(r\sin t), \quad \theta_1 \leq t \leq \theta_2 \tag{12.7}$$

where $z_1 = a + re^{i\theta_1}$ and $z_2 = a + re^{i\theta_2}$.

A more useful expression of the geodesics involves explicitly the cartesian coordinates of the pair of points. First, given the pair $z_1, z_2 \in \mathcal{H}^2, x_1 \neq x_2$, the semi-circle orthogonal to x-axis connecting them has a center $c = (a_{1\frown 2}, 0)$ and a radius $r_{1\frown 2}$, $(z_1, z_2) \mapsto SC_{r_{1\frown 2}}(a_{1\frown 2})$, where

$$a_{1\frown 2} = \frac{x_2^2 - x_1^2 + y_2^2 - y_1^2}{2(x_2 - x_1)};\quad r_{1\frown 2} = \sqrt{(x_1 - a_{1\frown 2})^2 + y_1^2} = \sqrt{(x_2 - a_{1\frown 2})^2 + y_2^2}. \tag{12.8}$$

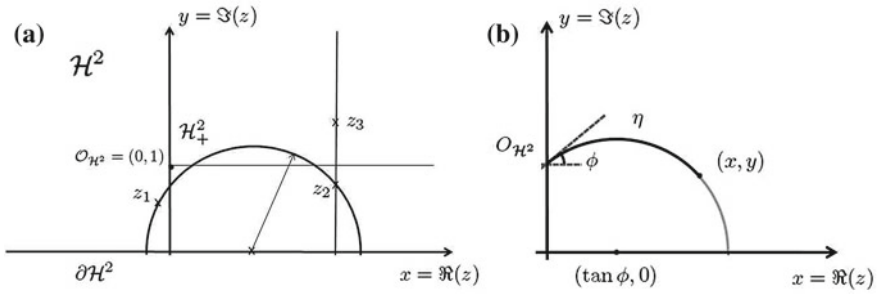


Fig. 12.3 **a** Geodesics of $\mathcal{H}^2$: z_1 and z_2 are connected by a unique *semi-circle*; the geodesic between z_2 and z_3 is a segment of *vertical line*. **b** Hyperbolic polar coordinates

Then, the unique geodesic parameterized by the length, $t \mapsto \gamma(z_1, z_2; t)$, $\gamma : [0, 1] \to \mathcal{H}^2$ joining two points $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ such as $\gamma(z_1, z_2; 0) = z_1$ and $\gamma(z_1, z_2; 1) = z_2$ is given by

$$\gamma(z_1, z_2; t) = \begin{cases} x_1 + ie^{\xi t + t_0} & \text{if } x_1 = x_2 \\ [r \tanh(\xi t + t_0) + a] + i\left[\frac{r}{\cosh(\xi t + t_0)}\right] & \text{if } x_1 \neq x_2 \end{cases} \tag{12.9}$$

with a and r given in (12.8) and where for $x_1 = x_2$, $t_0 = \log(y_1)$, $\xi = \log \frac{y_2}{y_1}$ and for $x_1 \neq x_2$

$$t_0 = \cosh^{-1}\left(\frac{r}{y_1}\right) = \sinh^{-1}\left(\frac{x_1 - a}{y_1}\right), \quad \xi = \log\left(\frac{y_1}{y_2}\frac{r + \sqrt{r^2 - y_2^2}}{r + \sqrt{r^2 - y_1^2}}\right).$$

If we take the parameterized smooth curve $\gamma(t) = x(t) + iy(t)$, where $x(t)$ and $y(t)$ are continuously differentiable for $b \leq t \leq c$, then the hyperbolic length along the curve is determined by integrating the metric (12.3) as:

$$L(\gamma) = \int_\gamma ds = \int_a^b |\dot{\gamma}(t)|_\gamma dt = \int_a^b \frac{\sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}}{y(t)} dt = \int_a^b \frac{|\dot{z}(t)|}{y(t)} dt.$$

Note that this expression is independent from the parameter choice. Hence, using the polar angle parametrization (12.7), we obtain an alternative expression of the geodesic distance given

$$\text{dist}_{\mathcal{H}^2}(z_1, z_2) = \inf_\gamma L(\gamma) = \int_{\theta_1}^{\theta_2} \frac{r}{r \sin t} dt = \left|\log\left(\cot\frac{\theta_2}{2}\right) - \log\left(\cot\frac{\theta_1}{2}\right)\right|$$

which is independent of r and consequently, as described below, dilation is an isometry.

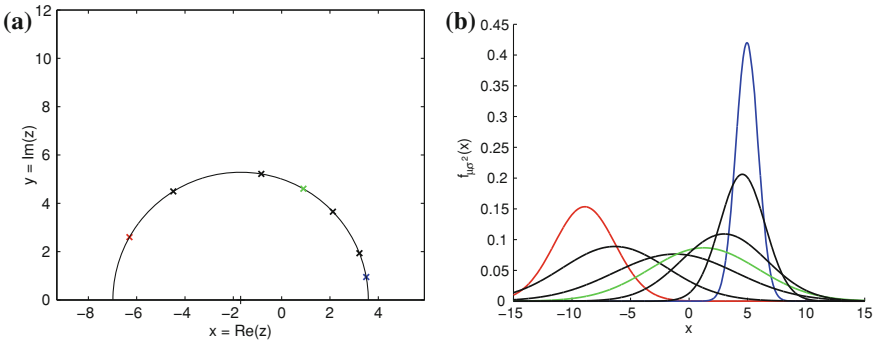


Fig. 12.4 **a** Example of interpolation of 5 points in $\mathcal{H}^2$ between points $z_1 = -6.3 + i2.6$ (in *red*) and $z_2 = 3.5 + i0.95$ (in *blue*) using their geodesic $t \mapsto \gamma(z_1, z_2; t)$, with $t = 0.2, 0.4, 0.5, 0.6, 0.8$. The average point (in *green*) corresponds just to $\gamma(z_1, z_2; 0.5) = 0.89 + i4.6$. **b** Original (in *red* and *blue*) univariate Gaussian pdfs and corresponding interpolated ones

Remark Interpolation between two univariate normal distributions. Using the closed-form expression of geodesics $t \mapsto \gamma(z_1, z_2; t)$, given in (12.9), it is possible to compute the average univariate Gaussian pdf between $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, with $(\mu_k = \sqrt{2}x_k, \sigma_k = y_k)$, by taking $t = 0.5$. More generally, we can interpolate a series of distributions between them by discretizing t between 0 and 1. An example of such a method is given in Fig. 12.4. We note in particular that the average Gaussian pdf can have a variance bigger than σ_1^2 and σ_2^2. We note also that, due to the "logarithmic scale" of imaginary axis, equally spaces points in t do not have equal Euclidean arc-length in the semi-circle.

12.2.3 Hyperbolic Polar Coordinates

The position of point $z = x + iy$ in $\mathcal{H}^2$ can be given either in terms of Cartesian coordinates (x, y) or by means of polar hyperbolic coordinates (η, ϕ), where η represents the distance of the point from the origin $\mathcal{O}_{\mathcal{H}^2} = (0, 1)$ and ϕ represents the slope of the tangent in $\mathcal{O}_{\mathcal{H}^2}$ to the geodesic (i.e., semi-circle) joining the point (x, y) with the origin. The formulas which relate the hyperbolic coordinates (η, ϕ) to the Cartesian ones (x, y) are [11]

$$\begin{cases} x = \frac{\sinh\eta\cos\phi}{\cosh\eta - \sinh\eta\sin\phi}, \\ y = \frac{1}{\cosh\eta - \sinh\eta\sin\phi}, \end{cases} \quad \begin{array}{l} \eta > 0 \\ -\frac{\pi}{2} < \phi < \frac{\pi}{2} \end{array} \quad \begin{cases} \eta = \mathrm{dist}_{\mathcal{H}^2}(\mathcal{O}_{\mathcal{H}^2}, z) \\ \phi = \arctan\frac{x^2+y^2-1}{2x} \end{cases} \tag{12.10}$$

We notice that the center of the geodesic passing through (x, y) from $\mathcal{O}_{\mathcal{H}^2}$ has Cartesian coordinates given by $(\tan\phi, 0)$; see Fig. 12.3b.

12.2.4 Invariance and Isometric Symmetry

Let the projective special linear group defined by $\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm \mathrm{I}\}$ where the special linear group $\mathrm{SL}(2,\mathbb{R})$ consists of 2×2 matrices with real entries which determinant equals $+1$, i.e.,

$$g \in \mathrm{SL}(2,\mathbb{R}): \quad g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \; ad - bc = 1;$$

and I denotes the identity matrix. This defines the group of Möbius transformations $M_g : \mathcal{H}^2 \to \mathcal{H}^2$ by setting for each $g \in \mathrm{SL}(2,\mathbb{R})$,

$$z \mapsto M_g(z) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az+b}{cz+d} = \frac{ac|z|^2 + bd + (ad+bc)\Re(z) + i\Im(z)}{|cz+d|^2},$$

such that $\Im\left(M_g(z)\right) = (y(ad-bc)) \,/\, \left((cx+d)^2 + (cy)^2\right) > 0$. The inverse map is easily computed, i.e., $z \mapsto M_g^{-1}(z) = (dz-b)/(-cz+a)$. Since Möbius transformations are well defined in $\mathcal{H}^2$ and map $\mathcal{H}^2$ to $\mathcal{H}^2$ homeomorphically.

The Lie group $\mathrm{PSL}(2,\mathbb{R})$ acts on the upper half-plane by preserving the hyperbolic distance, i.e.,

$$\mathrm{dist}_{\mathcal{H}^2}(M_g(z_1), M_g(z_2)) = \mathrm{dist}_{\mathcal{H}^2}(z_1, z_2), \; \forall g \in \mathrm{SL}(2,\mathbb{R}), \; \forall z_1, z_2 \in \mathcal{H}^2.$$

This includes three basic types of transformations: (1) translations $z \mapsto z + \alpha$, $\alpha \in \mathbb{R}$; (2) scaling $z \mapsto \beta z$, $\beta \in \mathbb{R} \setminus 0$; (3) inversion $z \mapsto z^{-1}$. More precisely, any generic Möbius transformation $M_g(z)$, $c \neq 0$, can be decomposed into the following four maps: $f_1 = z + d/c$, $f_2 = -1/z$, $f_3 = z(ad-bc)/c^2$, $f_4 = z + a/c$ such that $M_g(z) = (f_1 \circ f_2 \circ f_3 \circ f_4)(z)$. If $c = 0$, we have $M_g(z) = (h_1 \circ h_2 \circ h_3)(z)$ where $h_1(z) = az$, $h_2(z) = z + b$, $h_3(z) = z/d$. It can be proved that Möbius transformations take circles to circles. Hence, given a circle in the complex plane $\mathbb{C}$ of radius r and center c, denoted by $C_r(c)$, we have its following mappings [27]: a translation $z \mapsto z + \alpha$, such as the functions f_1, f_4 and h_2 maps $C_r(c)$ to $C_r(c+\alpha)$; a scaling $z \mapsto \beta z$, such as the functions f_3, h_1 and h_3, maps $C_r(c)$ to $C_{\beta r}(\beta c)$; for inversion $z \mapsto z^{-1}$, $C_r(c)$ maps to $C_{r/|cz|}(-1/c)$.

Let $H \in \mathcal{H}^2$ be a geodesic of the upper half-plane, which is described uniquely by its endpoints in $\partial\mathcal{H}^2$, there exists a Möbius transformation M_g such that M_g maps H bijectively to the imaginary axis, i.e., $VL(0)$. If H is the vertical line $VL(a)$, the transformation is the translation $z \mapsto M_g(z) = z - a$. If H is the semi-circle $SC_r(a)$ with endpoints in real axis being $\zeta_-, \zeta_+ \in \mathbb{R}$, where $\zeta_- = a - r$ and $\zeta_+ = a + r$, the map is given by $M_g(z) = \frac{z-\zeta_-}{z-\zeta_+}$, such that $M_g(\zeta_-) = 0$, $M_g(\zeta_+) = \infty$ and $M_g(a+ir) = i$.

The unit-speed geodesic going up vertically, through the point $z = i$ is given by

$$\gamma(t) = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} \cdot i = ie^t.$$

Because PSL(2, $\mathbb{R}$) acts transitively by isometries of the upper half-plane, this geodesic is mapped into other geodesics through the action of PSL(2, $\mathbb{R}$). Thus, the general unit-speed geodesic is given by

$$\gamma(t) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} \cdot i = \frac{aie^t + b}{cie^t + d}. \tag{12.11}$$

12.2.5 Hyperbolic Circles and Balls

Let consider an Euclidean circle of center $c = (x_c, y_c) \in \mathcal{H}^2$ and radius r in the upper-half plane, defined as $C_r(c) = \{z \in \mathcal{H}^2 | \sqrt{(x_c - a)^2 + (y_c - b)^2} = r\}$, such that it is contained in the upper-half plane, i.e., $C_r(c) \subset \mathcal{H}^2$. The corresponding hyperbolic circle $C_{\mathcal{H}^2, r_h}(c_h)$ $\{z \in \mathcal{H}^2 | \text{dist}_{\mathcal{H}^2}(c_h, z) = r_h\}$ is geometrically equal to $C_r(c)$ but its hyperbolic center and radius are given by

$$c_h = (x_c, \sqrt{y_c^2 - r^2}); \quad r_h = \tanh^{-1}\left(\frac{r}{y_c}\right).$$

We note that the hyperbolic center is always below the Euclidean center. The inverse equations are

$$c = (x_c = x_h, \; y_c = y_h \cosh r_h); \quad r = y_h \sinh r_h. \tag{12.12}$$

Naturally, the hyperbolic ball of center c_h and radius r_h is defined by $B_{\mathcal{H}^2, r_h}(c_h)$ $\{z \in \mathcal{H}^2 | \text{dist}_{\mathcal{H}^2}(c_h, z) \leq r_h\}$. Let us consider a hyperbolic ball centered at the origin $B_{\mathcal{H}^2, r_h}(0, 1)$, parameterized by its boundary curve ∂B in Euclidean coordinates:

$$x = r \cos\theta; \quad y = b + r \sin\theta$$

where using (12.12), we have $b = \cosh r_h$ and $r = \sinh r_h$. The length of the boundary and area of this ball are respectively given by [32]:

$$L(\partial B) = \int_0^{2\pi} \frac{r}{b + r\sin\theta} d\theta = 2\pi \sinh r_h, \tag{12.13}$$

$$\text{Area}(B) = \int\int_B \frac{dxdy}{y^2} = \oint_\gamma \frac{dx}{y} = 2\pi(\cosh r_h - 1). \tag{12.14}$$

Comparing the values of an Euclidean ball which has area πr_h^2 and length of its boundary circle $2\pi r_h$, and considering the Taylor series $\sinh r_h = r_h + \frac{r_h^3}{3!} + \frac{r_h^5}{5!} + \cdots$ and $\cosh r_h = 1 + \frac{r_h^2}{2!} + \frac{r_h^4}{4!} + \cdots$, one can note that the hyperbolic space is much larger than the Euclidean one. Curvature is defined through derivatives of the metric, but the fact that infinitesimally the hyperbolic ball grows faster than the Euclidean balls can be used a measure of the curvature of the space at the origin $(0, 1)$ [32]: $K = \lim_{r_h \to 0} \frac{3[2\pi r_h - L(\partial B)]}{\pi r_h^3} = -1$. Since there is an isometry that maps the neighborhood of any point to the neighborhood of the origin, the curvature of hyperbolic space is identically constant to -1.

Remark Minimax center in $\mathcal{H}^2$. Finding the smallest circle that contains the whole set of points $x_1, x_2, \ldots, x_N$ in the Euclidean plane is a classical problem in computational geometry, called the minimum enclosing circle MEC. It is also relevant its statistical estimation since the unique center of the circle c^∞ (called 1-center or minimax center) is defined as the L^∞ center of mass, i.e., for $\mathbb{R}^2$, $c^\infty = \arg\min_{x \in \mathbb{R}^2} \max_{1 \leq i \leq N} \|x_i - x\|_2$. Computing the smallest enclosing sphere in Euclidean spaces is intractable in high dimensions, but efficient approximation algorithms have been proposed. The Bădoiu and Clarkson algorithm [8] leads to a fast and simple approximation (of known precision ϵ after a given number of iterations $\left\lceil \frac{1}{\epsilon^2} \right\rceil$ using the notion of core-set, but independent from dimensionality n). The computation of the minimax center is particularly relevant in information geometry (smallest enclosing information disk [25]) and has been considered for hyperbolic models such as the Klein disk, using a Riemannian extension of Bădoiu and Clarkson algorithm [2], which only requires a closed-form of the geodesics. Figure 12.5 depicts an example of minimax center computation using Bădoiu and Clarkson algorithm for a set of univariate Gaussian pdfs represented in $\mathcal{H}^2$. We note that, using this property of circle preservation, the computation the minimal enclosing hyperbolic circle of a given set of points $Z = \{z_k\}_{1 \leq k \leq K}$, $z_k \in \mathcal{H}^2$, denoted $MEC_{\mathcal{H}^2}(Z)$ is equivalent to computing the corresponding minimal enclosing circle $MEC(Z)$ if and only if we have $MEC(Z) \subset \mathcal{H}^2$. This is the case for the example given in Fig. 12.5.

12.3 Fisher Information Metric and α-Order Entropy Metric of Univariate Normal Distributions

In information geometry, the Fisher information metric is a particular Riemannian metric which can be associated to a smooth manifold whose points are probability measures defined on a common probability space [3, 4]. It can be obtained as the infinitesimal form of the Kullback–Leibler divergence (relative entropy). An alternative formulation is obtained by computing the negative of the Hessian of the Shannon entropy.

Given an univariate probability distribution $p(x|\theta)$, $x \in X$, it can be viewed as a point on a statistical manifold with coordinates given by $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$. The

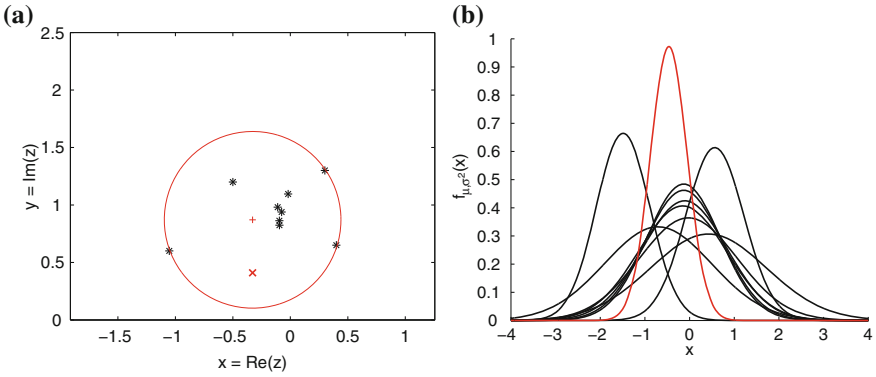


Fig. 12.5 **a** Example of minimax center (x_h, y_h) (*red* $\times$) of a set of nine points $Z = \{z_k\}_{1 \leq k \leq 9}$ in $\mathcal{H}^2$ (original points $*$ in *black*), the minimal enclosing *circle* $MEC_{\mathcal{H}^2}(Z)$ is also depicted (in *red*). **b** Corresponding minimax center Gaussian set $N(\mu = \sqrt{2}x_h, \sigma^2 = y_h^2)$ of nine univariate Gaussian pdfs, $N_k(\mu_k, \sigma_k^2)$, $1 \leq k \leq 9$

Fisher information matrix then takes the form:

$$g_{kl}(\theta)_{k,l=1,2} = \int_X \frac{\partial \log p(x,\theta)}{\partial \theta_k} \frac{\partial \log p(x,\theta)}{\partial \theta_l} p(x,\theta)\, dx.$$

The corresponding positive definite form

$$ds^2(\theta) = \sum_{k,l=1}^{n} g_{kl}(\theta) d\theta_k d\theta_l$$

is defined as the Fisher information metric. In the univariate Gaussian distributed case $p(x|\theta) \equiv N(\mu, \sigma^2)$, we have in particular $\theta = (\mu, \sigma)$ and it can be easily deduced that the Fisher information matrix is

$$(g_{kl}(\mu, \sigma)) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix} \tag{12.15}$$

and the corresponding metric is

$$ds^2((\mu, \sigma)) = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2} = 2\sigma^{-2}\left(\frac{d\mu^2}{\sqrt{2}} + d\sigma^2\right). \tag{12.16}$$

Therefore, the Fisher information geometry of univariate normal distribution is essentially the geometry of the Poincaré upper-half plane with the following change of variables:

$$x = \mu/\sqrt{2}, \quad y = \sigma$$

Hence, given two univariate Gaussian pdfs $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, the Fisher distance between them, $\text{dist}_{Fisher} : \mathcal{N} \times \mathcal{N} \to \mathbb{R}_+$, defined from the Fisher information metric is given by [9, 14]:

$$\text{dist}_{Fisher}\left((\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)\right) = \sqrt{2}\text{dist}_{\mathcal{H}^2}\left(\frac{\mu_1}{\sqrt{2}} + i\sigma_1, \frac{\mu_2}{\sqrt{2}} + i\sigma_2\right). \quad (12.17)$$

The change of variable also involves that the geodesics in the hyperbolic Fisher space of normal distributions are half-lines and half-ellipses orthogonal at $\sigma = 0$, with eccentricity $1/\sqrt{2}$.

The canonic approach can be generalized according to Burbea and Rao geometric framework [9], which is based on replacing the Shannon entropy by the notion of α-order entropy, which associated Hessian metric leads to an extended large class of information metric geometries. Focussing on the particular case of univariate normal distributions, $p(x|\theta) \equiv N(\mu, \sigma^2)$, we consider again points in the upper half-plane, $z = x + iy \in \mathcal{H}^2$ and for a given $\alpha > 0$ the $\alpha-$order entropy metric is given by [9]:

$$\begin{cases} x &= [A(\alpha)]^{-1/2}\mu, \quad y = \sigma; \\ ds_\alpha &= B(\alpha)y^{-(\alpha+1)}(dx^2 + dy^2); \end{cases} \quad (12.18)$$

where

$$A(\alpha) = (\alpha^{1/2} - \alpha^{-1/2})^2 + 2\alpha^{-1};\ B(\alpha) = \alpha^{-3/2}(2\pi)^{(1-\alpha)/2}A(\alpha);\ \alpha > 0. \quad (12.19)$$

The metric in (12.18) constitutes a Kähler metric on $\mathcal{H}^2$ and when $\alpha = 1$ reduces to the Poincaré metric (12.3). Its Gaussian curvature is $K_\alpha(z) = -(\alpha + 1)[B(\alpha)]^{-1} y^{\alpha-1}$; being always negative (hyperbolic geometry). In particular, for $\alpha = 1$ we recover the particular constant case $K_1(z) = -1$.

The geodesics of the Burbea-Rao α-order entropy metric can be written in its parametric polar form as [9]:

$$\gamma(\theta) = x(\theta) + iy(\theta), \quad 0 < \theta < \pi, \quad \text{with} \quad (12.20)$$

$$\begin{cases} x(\theta) = a + r^{1/\beta}F_{1/\beta}(\theta), \\ y(\theta) = r^{1/\beta}\sin^{1/\beta}\theta, \end{cases} \quad (12.21)$$

where

$$\begin{aligned} \beta &= (\alpha + 1)/2, \quad r > 0, \quad a \in \mathbb{R}, \\ F_\gamma(\theta) &= -\gamma \int_{\pi/2}^{\theta} \sin^\gamma t dt. \end{aligned}$$

Figure 12.6 shows examples of the geodesics from the Burbea-Rao α-order entropy metric for $a = 0$, $r = 1$ and $\alpha = 0.01$, 0.5, 1, 5 and 20.

By integration of the metric, it is obtained the Burbea-Rao α-order entropy geodesic distance for $z_1, z_2 \in \mathcal{H}^2$ [9]:

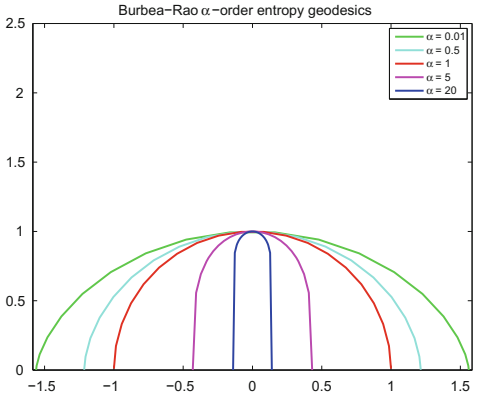


Fig. 12.6 Examples of the geodesics from the Burbea-Rao $\alpha-$order entropy metric obtained using (12.20) and (12.21) for $a = 0$, $r = 1$ and $\alpha = 0.01, 0.5, 1, 5$ and 20

$$\mathrm{dist}_{\mathcal{H}^2}(z_1, z_2;\ \alpha) = \frac{2\sqrt{B(\alpha)}}{|1-\alpha|}\left|\frac{x_1 - x_2}{r} + y_1^{(1-\alpha)/2}\sqrt{1 - r^{-2}y_1^{\alpha+1}} - y_2^{(1-\alpha)/2}\sqrt{1 - r^{-2}y_2^{\alpha+1}}\right|, \tag{12.22}$$

which unfortunately depends on the value of r. This quantity should be determined by solving a system of three nonlinear equations for the unknown variables θ_1, θ_2 and r:

$$\begin{cases} x_1 - x_2 = r^{1/\beta}\left(F_{1/\beta}(\theta_1) - F_{1/\beta}(\theta_2)\right), \\ y_1 = r^{1/\beta}\sin^{1/\beta}\theta_1, \\ y_2 = r^{1/\beta}\sin^{1/\beta}\theta_2. \end{cases}$$

An alternative solution to compute a closed form distance between two univariate normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ according to the Burbea-Rao α-deformed geometry is based on the α-order Hellinger distance [9]:

$$\mathrm{dist}_{\mathrm{Hellinger}}\left((\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2);\ \alpha\right) = \frac{2(2\pi)^{(1-\alpha)/4}}{\alpha^{5/4}}\left[\left(\sigma_1^{(1-\alpha)/2} - \sigma_2^{(1-\alpha)/2}\right)^2 + 2(\sigma_1\sigma_2)^{(1-\alpha)/2}\left(1 - \left(\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\right)^{1/2}\exp\left(\frac{-\alpha(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right)\right)\right]^{1/2}. \tag{12.23}$$

In particular, when $\alpha = 1$ this formula reduces to

$$\mathrm{dist}_{\mathrm{Hellinger}}\left((\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)\right) = 2^{3/2}\left(1 - \left(\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\right)^{1/2}\exp\left(\frac{-(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right)\right)^{1/2}. \tag{12.24}$$

12.4 Endowing $\mathcal{H}^2$ with Partial Ordering and Its Complete (inf-semi) Lattice Structures

The notion of ordering invariance in the Poincaré upper-half plane was considered in the Soviet literature [19, 20]. Ordering invariance with respect to simple transitive subgroup T of the group of motions was studied, i.e., group T consists of transformations t of the form:

$$z = x + iy \mapsto z' = (\lambda x + \alpha) + i\lambda y,$$

where $\lambda > 0$ and α are real numbers. We named T the Guts group. We note that T is just the composition of a translation and a scaling in $\mathcal{H}^2$, and consequently, T is an isometric group (see Sect. 2.4).

Nevertheless, up to the best of our knowledge, the formulation of partial orders on Poincaré upper-half plane has not been widely studied. We introduce here partial orders in $\mathcal{H}^2$ and study invariance properties to transformations of Guts group or to other subgroups of $\mathrm{SL}(2, \mathbb{R})$ (Möbius transformations).

12.4.1 Upper Half-Plane Product Ordering

A real vector space E on which a partial order $\leq$ is given (reflexive, transitive, antisymmetric) is called an ordered vector space if (i) $x, y, z \in E$ and $x \leq y$ implies $x + z \leq y + z$; (ii) $x, y \in E$, $0 \leq \lambda \in \mathbb{R}$, and $x \leq y$ implies $\lambda x \leq \lambda y$. Let us consider that the partial order $\leq$ is the product order. An element $x \in E$ with $x \geq 0$ (it means that all the vector components are positive) is said to be positive. The set $E_+ = \{x \in E \mid x \geq 0\}$ for all positive elements is called the cone of positive elements. It turns out that the order of an ordered vector space is determined by the set of positive elements. Let E be a vector space and $C \subset E$ a cone. Then, $x \leq y$ if $x - y \in C$ defines an order on E such that E is an ordered vector space with $E_+ = C$. The notion of partially ordered vector space is naturally extended to partially ordered groups [18]. An ordered vector space E is called a vector lattice $(E, \leq)$ if $\forall x, y \in E$ there exists the joint (supremum or least upper bound) $x \vee y = \sup(x, y) \in E$ and the meet (infimum or greatest lower bound) $x \wedge y = \inf(x, y) \in E$. A vector lattice is also called a Riesz space.

Thus, we can introduce a similar order structure in $\mathcal{H}^2$ as a product order of $\mathbb{R} \times \mathbb{R}_+$. To achieve this goal, we need to define, on the one hand, the equivalent of ordering preserving linear combination. More precisely, given three points $z_1, z_2, z_3 \in \mathcal{H}^2$ and a scalar positive number $0 \leq \lambda \in \mathbb{R}$ we say that

$$z_1 \leq_{\mathcal{H}^2} z_2 \text{ implies } \lambda \boxdot z_1 \boxplus z_3 \leq_{\mathcal{H}^2} \lambda \boxdot z_2 \boxplus z_3,$$

where we have introduced the following pair of operations in $\mathcal{H}^2$:

$$\lambda \boxdot z = \lambda x + iy^{\lambda} \text{ and } z_1 \boxplus z_2 = (x_1 + x_2) + i(y_1 y_2).$$

On the other hand, the corresponding partial ordering $\leq_{\mathcal{H}^2}$ will be determined by the positive cone in $\mathcal{H}^2$ defined by $\mathcal{H}^2_+ = \{z \in \mathcal{H}^2 \mid x \geq 0 \text{ and } y \geq 1\}$, i.e.,

$$z_1 \leq_{\mathcal{H}^2} z_2 \Leftrightarrow z_2 \boxminus z_1 \in \mathcal{H}^2_+, \tag{12.25}$$

with $z_2 \boxminus z_1 = (x_2 - x_1) + i(y_2^{-1} y_1)$. According to this partial ordering the corresponding supremum and infimum for any pair of points z_1 and z_2 in $\mathcal{H}^2$ are formulated as follows

$$z_1 \vee_{\mathcal{H}^2} z_2 = (x_1 \vee x_2) + i \exp\left(\log(y_1) \vee \log(y_2)\right), \tag{12.26}$$

$$z_1 \wedge_{\mathcal{H}^2} z_2 = (x_1 \wedge x_2) + i \exp\left(\log(y_1) \wedge \log(y_2)\right). \tag{12.27}$$

Therefore $\mathcal{H}^2$ endowed with partial ordering (12.25) is a complete lattice, but it is not bounded since the greatest (or top) and least (or bottom) elements are in the boundary $\partial\mathcal{H}^2$. We also have a duality between supremum and infimum, i.e.,

$$z_1 \vee_{\mathcal{H}^2} z_2 = \complement\left(\complement z_1 \wedge_{\mathcal{H}^2} \complement z_2\right); \quad z_1 \wedge_{\mathcal{H}^2} z_2 = \complement\left(\complement z_1 \vee_{\mathcal{H}^2} \complement z_2\right),$$

with respect to the following involution

$$z \mapsto \complement z = (-1) \boxdot z = -x + iy^{-1}. \tag{12.28}$$

We easily note that, in fact, $\exp\left(\log(y_1) \vee \log(y_2)\right) = y_1 \vee y_2$ and similarly for the infimum, since the logarithm is an isotone mapping (i.e., monotone increasing) and therefore order-preserving. Therefore, the partial ordering $\leq_{\mathcal{H}^2}$ does not involve any particular structure for $\mathcal{H}^2$ and does not take into account the Riemannian nature of the upper half plane. According to that, we note also that the partial ordering $\leq_{\mathcal{H}^2}$ is invariant to the Guts group of transforms, i.e.,

$$z_1 \leq_{\mathcal{H}^2} z_2 \Leftrightarrow T(z_1) \leq_{\mathcal{H}^2} T(z_2).$$

12.4.2 Upper Half-Plane Symmetric Ordering

Let us consider a symmetrization of the product ordering with respect to the origin in the upper half-plane. Given any pair of points $z_1, z_2 \in \mathcal{H}^2$, we define the upper half-plane symmetric ordering as

$$z_1 \preceq_{\mathcal{H}^2} z_2 \Leftrightarrow \begin{cases} 0 \leq x_1 \leq x_2 \text{ and } 0 \leq \log(y_1) \leq \log(y_2) \text{ or} \\ x_2 \leq x_1 \leq 0 \text{ and } 0 \leq \log(y_1) \leq \log(y_2) \text{ or} \\ x_2 \leq x_1 \leq 0 \text{ and } \log(y_2) \leq \log(y_1) \leq 0 \text{ or} \\ 0 \leq x_1 \leq x_2 \text{ and } \log(y_2) \leq \log(y_1) \leq 0 \end{cases} \tag{12.29}$$

The four conditions of this partial ordering entails that only points belonging to the same quadrant of $\mathcal{H}^2$ can be ordered, where the four quadrants $\{\mathcal{H}^2_{++}, \mathcal{H}^2_{-+}, \mathcal{H}^2_{--}, \mathcal{H}^2_{+-}\}$ are defined with respect to the origin $\mathcal{O}_{\mathcal{H}^2} = (0, 1)$ which corresponds to the pure imaginary complex $z_0 = i$. In other words, we can summarize the partial ordering (12.29) by saying that if z_1 and z_2 belong to the same $\mathcal{O}$-quadrant of $\mathcal{H}^2$ we have $z_1 \preceq_{\mathcal{H}^2} z_2 \Leftrightarrow |x_1| \leq |x_2|$ and $|\log(x_1)| \leq |\log(x_2)|$. Endowed with the partial ordering (12.29), $\mathcal{H}^2$ becomes a partially ordered set (poset) where the bottom element is z_0, but we notice that there is no top element. In addition, for any pair of point z_1 and z_2, the infimum $\curlywedge_{\mathcal{H}^2}$ is given by

$$z_1 \curlywedge_{\mathcal{H}^2} z_2 \Leftrightarrow \begin{cases} (x_1 \wedge x_2) + i(y_1 \wedge y_2) & \text{if } z_1, z_2 \in \mathcal{H}^2_{++} \\ (x_1 \vee x_2) + i(y_1 \wedge y_2) & \text{if } z_1, z_2 \in \mathcal{H}^2_{-+} \\ (x_1 \vee x_2) + i(y_1 \vee y_2) & \text{if } z_1, z_2 \in \mathcal{H}^2_{--} \\ (x_1 \wedge x_2) + i(y_1 \vee y_2) & \text{if } z_1, z_2 \in \mathcal{H}^2_{+-} \\ z_0 & \text{otherwise} \end{cases} \tag{12.30}$$

The infimum (12.30) extends naturally to any finite set of points in $\mathcal{H}^2$, $Z = \{z_k\}_{1 \leq k \leq K}$, and will be denoted by $\bigwedge_{\mathcal{H}^2} Z$. However, the supremum $z_1 \curlyvee_{\mathcal{H}^2} z_2$ is not defined; or more precisely, it is defined if and only if z_1 and z_2 belong to the same quadrant, i.e., similarly to (12.30) *mutatis mutandis* $\wedge$ by $\vee$ with the "otherwise" case as "non existent". Consequently, the poset $(\mathcal{H}^2, \preceq_{\mathcal{H}^2})$ is only a complete inf-semilattice. The fundamental property of such infimum (12.30) is its self-duality with respect to involution (12.28), i.e.,

$$z_1 \curlywedge_{\mathcal{H}^2} z_2 = \complement\left(\complement z_1 \curlywedge_{\mathcal{H}^2} \complement z_2\right). \tag{12.31}$$

Due to the strong dependency of partial ordering $\preceq_{\mathcal{H}^2}$ with respect to $\mathcal{O}_{\mathcal{H}^2}$, it is easy to see that such ordering is only invariant to transformations that does not move points from one quadrant to another one. This is the case typically for mappings as $z \mapsto \lambda \boxdot z$, $\lambda > 0$.

12.4.3 Upper Half-Plane Polar Ordering

Previous order $\preceq_{\mathcal{H}^2}$ is only a partial ordering, and consequently given any pair of points z_1 and z_2, the infimum $z_1 \curlywedge_{\mathcal{H}^2} z_2$ can be different from z_1 and z_2. In addition, the supremum is not always defined. Let us introduce a total ordering in $\mathcal{H}$ based on hyperbolic polar coordinates, which also takes into account an ordering relationship with respect to $\mathcal{O}_{\mathcal{H}^2}$. Thus, given two points $\forall z_1, z_2 \in \mathcal{H}$ the upper half-plane polar ordering states

$$z_1 \preceq^{pol}_{\mathcal{H}^2} z_2 \Leftrightarrow \begin{cases} \eta_1 < \eta_2 & \text{or} \\ \eta_1 = \eta_2 & \text{and} \ \ \tan\phi_1 \leq \tan\phi_2 \end{cases} \tag{12.32}$$

where (η, ϕ) are defined in Eq. (12.10). The polar supremum $z_1 \vee^{pol}_{\mathcal{H}^2} z_2$ and infimum $z_1 \wedge^{pol}_{\mathcal{H}^2} z_2$ are naturally obtained from the order (12.32) for any subset of points Z, denoted by $\bigvee^{pol}_{\mathcal{H}^2} Z$ and $\bigwedge^{pol}_{\mathcal{H}^2} Z$. Total order $\preceq^{pol}_{\mathcal{H}^2}$ leads to a complete lattice, bounded from the bottom (i.e., the origin $\mathcal{O}_{\mathcal{H}^2}$) but not from the top. Furthermore, as $\preceq^{pol}_{\mathcal{H}^2}$ is a total ordering, the supremum and the infimum will be either z_1 or z_2.

Polar total order is invariant to any Möbius transformation M_g which preserves the distance to the origin (isometry group) and more generally to isotone maps in distance, i.e., $\eta(z_1) \leq \eta(z_2) \Leftrightarrow \eta(M_g(z_1)) \leq \eta(M_g(z_2))$ but which also preserves the orientation order, i.e., order on the polar angle. This is for instance the case of orientation group SO(2) and the scaling maps $z \mapsto M_g(z) = \lambda z$, $0 < \lambda \in \mathbb{R}$.

We note also that instead of considering as the origin $\mathcal{O}_{\mathcal{H}^2}$, the polar hyperbolic coordinates can be defined with respect to a different origin $z_0^{'}$ and consequently, the total order is adapted to the new origin (i.e., bottom element is just $z_0^{'}$).

One can replace in the polar ordering the distance $\text{dist}_{\mathcal{H}^2}(\mathcal{O}_{\mathcal{H}^2}, z)$ by the α-order Hellinger distance to obtain now the total ordering $\preceq^{\alpha-pol}_{\mathcal{H}^2}$ parametrized by α:

$$z_1 \preceq^{\alpha-pol}_{\mathcal{H}^2} z_2 \Leftrightarrow \begin{cases} \text{dist}_{\text{Hellinger}}\left(\mathcal{O}_{\mathcal{H}^2}, z_1;\ \alpha\right) < \text{dist}_{\text{Hellinger}}\left(\mathcal{O}_{\mathcal{H}^2}, z_2;\ \alpha\right) & \text{or} \\ \text{dist}_{\text{Hellinger}}\left(\mathcal{O}_{\mathcal{H}^2}, z_1;\ \alpha\right) = \text{dist}_{\text{Hellinger}}\left(\mathcal{O}_{\mathcal{H}^2}, z_2;\ \alpha\right) & \text{and} \ \ \tan\phi_1 \leq \tan\phi_2 \end{cases} \tag{12.33}$$

As we illustrate in Sect. 12.5, the "deformation" of the distance driven by α can significantly change the supremum and infimum from a set of points Z. Obviously, the properties of invariance of $\preceq^{\alpha-pol}_{\mathcal{H}^2}$ are related to the isometries of the α-order Hellinger distance.

12.4.4 Upper Half-Plane Geodesic Ordering

As discussed above, there is a unique hyperbolic geodesic joining any pair of points. Given two points $z_1, z_2 \in \mathcal{H}^2$ such that $x_1 \neq x_2$, let $SC_{r_{1\frown 2}}(a_{1\frown 2})$ be the semi-circle defining their geodesic, where the center $a_{1\frown 2}$ and the radius $r_{1\frown 2}$ are given by Eq. (12.8). Let us denote by $z_{1\frown 2}$ the point of $SC_{r_{1\frown 2}}(a_{1\frown 2})$ having the maximal imaginary part, i.e., its imaginary part is equal to the radius: $z_{1\frown 2} = a_{1\frown 2} + ir_{1\frown 2}$.

The upper half-plane geodesic ordering $\preceq^{geo}_{\mathcal{H}^2}$ defines an order for points being in the same half of their geodesic semi-circle as follows,

$$z_1 \preceq^{geo}_{\mathcal{H}^2} z_2 \Leftrightarrow \begin{cases} a_{1\frown 2} \leq x_1 < x_2 & \text{or} \\ x_2 < x_1 \leq a_{1\frown 2} \end{cases} \tag{12.34}$$

The property of transitivity of this partial ordering, i.e., $z_1 \preceq^{geo}_{\mathcal{H}^2} z_2, z_2 \preceq^{geo}_{\mathcal{H}^2} z_3 \Rightarrow z_1 \preceq^{geo}_{\mathcal{H}^2} z_3$, holds for points belonging to the same geodesic. For two points in a geodesic vertical line, $x_1 = x_2$, we have $z_1 \preceq^{geo}_{\mathcal{H}^2} z_2 \Leftrightarrow y_2 \leq y_1$. We note that considering the duality with respect to the involution (12.28), one has

$$z_1 \preceq^{geo}_{\mathcal{H}^2} z_2 \Leftrightarrow \complement z_1 \succeq^{geo}_{\mathcal{H}^2} \complement z_2.$$

According to this partial ordering, we define the geodesic infimum, denoted by $\curlywedge^{geo}_{\mathcal{H}^2}$, as the point on the geodesic joining z_1 and z_2 with maximal imaginary part, i.e., for any $z_1, z_2 \in \mathcal{H}^2$, with $x_1 \neq x_2$, we have

$$z_1 \curlywedge^{geo}_{\mathcal{H}^2} z_2 \Leftrightarrow \begin{cases} (x_1 \vee x_2) + i(y_1 \vee y_2) & \text{if } \ x_1, x_2 \leq a_{1\frown 2} \\ (x_1 \wedge x_2) + i(y_1 \vee y_2) & \text{if } \ x_1, x_2 \geq a_{1\frown 2} \\ z_{1\frown 2} & \text{otherwise} \end{cases} \tag{12.35}$$

If $x_1 = x_2$, we have that $z_1 \curlywedge^{geo}_{\mathcal{H}^2} z_2 = x_1 + i(y_1 \vee y_2)$. In any case, we have that $\text{dist}_{\mathcal{H}^2}(z_1, z_2) = \text{dist}_{\mathcal{H}^2}(z_1, z_1 \curlywedge^{geo}_{\mathcal{H}^2} z_2) + \text{dist}_{\mathcal{H}^2}(z_1 \curlywedge^{geo}_{\mathcal{H}^2} z_2, z_2)$. Intuitively, we notice that the geodesic infimum is the point of the geodesic farthest from the real line.

We observe that if one attempts to define the geodesic supremum from the partial ordering $\preceq^{geo}_{\mathcal{H}^2}$, it results that the supremum is not defined for any pair of points, i.e., supremum between z_1 and z_2 is defined if and only if both points are in the same half of its semi-circle. To tackle this limitation, we propose to define the geodesic supremum $z_1 \curlyvee^{geo}_{\mathcal{H}^2} z_2$ by duality with respect to the involution $\complement z$, i.e.,

$$z_1 \curlyvee^{geo}_{\mathcal{H}^2} z_2 = \complement\left(\complement z_1 \curlywedge^{geo}_{\mathcal{H}^2} \complement z_2\right) \Leftrightarrow \begin{cases} (x_1 \wedge x_2) + i(y_1 \wedge y_2) & \text{if } \ x_1, x_2 \leq a_{1\frown 2} \\ (x_1 \vee x_2) + i(y_1 \wedge y_2) & \text{if } \ x_1, x_2 \geq a_{1\frown 2} \\ \complement z_{\complement 1\frown \complement 2} & \text{otherwise} \end{cases} \tag{12.36}$$

where $\complement z_{\complement 1\frown \complement 2}$ is the dual point associated to the semi-circle defined by dual points $\complement z_1$ and $\complement z_2$.

Nevertheless, in order to have a structure of complete lattice for $(\mathcal{H}^2, \preceq^{geo}_{\mathcal{H}^2})$, it is required that the infimum and the supremum of any set of points $Z = \{z_k\}_{1 \leq k \leq K}$ with $K \geq 2$, are well defined. Namely, according to (12.35), the geodesic infimum of Z, denoted $\bigwedge^{geo}_{\mathcal{H}^2} Z$, corresponds to the point z_{inf} with the maximal imaginary part on all possible geodesics joining any pair of points $z_n, z_m \in Z$. In geometric terms, it means that between all these geodesics, there exists one which gives z_{inf}. Instead of computing all the geodesics, we propose to define the infimum $\bigwedge^{geo}_{\mathcal{H}^2} Z$ as the point $z_{\text{inf}} = a_{\text{inf}} + ir_{\text{inf}}$, where a_{inf} is the center of the smallest semi-circle in $\mathcal{H}^2$ of radius r_{inf} which encloses all the points in the set Z. We have the following property

$$\bigwedge^{geo}_{\mathcal{H}^2} Z = z_{\text{inf}} \preceq^{geo}_{\mathcal{H}^2} z_k, \ \ 1 \leq k \leq K,$$

which geometrically means that the geodesic connecting $z_{\inf}$ to any point z_k of Z lies always in one of the half part of the semi-circle defined by $z_{\inf}$ and z_k.

In practice, the minimal enclosing semi-circle defining $z_{\inf}$ can be easily computed by means of the following algorithm based on the minimum enclosing Euclidean circle *MEC* of a set of points: (1) Working on $\mathbb{R}^2$, define a set of points given, on the one hand, by Z and, on the other hand, by Z^* which corresponds to the reflected points with respect to x-axis (complex conjugate), i.e., points $Z = \{(x_k, y_k)\}$ and points $Z^* = \{(x_k, -y_k)\}$, $1 \leq k \leq K$; (2) Compute the $MEC(Z \cup Z^*) \mapsto C_r(c)$, in such a way that, by a symmetric point configuration, we necessarily have the center on x-axis, i.e., $c = (x_c, 0)$; (3) The infimum $\bigwedge_{\mathcal{H}^2}^{geo} Z = z_{\inf}$ is given by $z_{\inf} = x_c + ir$. Figure 12.7a–b gives an example of computation of the geodesic infimum from a set of points in $\mathcal{H}^2$.

As for the case of two points, the geodesic supremum of Z is defined by duality with respect to involution (12.28), i.e.,

$$z_{\sup} = \bigvee_{\mathcal{H}^2}^{geo} Z = \complement \left(\bigwedge_{\mathcal{H}^2}^{geo} \complement Z \right) = a_{\sup} + i r_{\sup}, \tag{12.37}$$

with $a_{\sup} = -x_c^{dual}$ and $r_{\sup} = 1/r^{dual}$, where $SC_{r^{dual}}(x_c^{dual})$ is the minimal enclosing semi-circle from dual set of points $\complement Z$. An example of computing the geodesic supremum $z_{\sup}$ is also given in Fig. 12.7a–b. It is easy to see that geodesic infimum and supremum have the following properties for any $Z \subset \mathcal{H}^2$:

1. $z_{\inf} \preceq_{\mathcal{H}^2}^{geo} z_{\sup}$;
2. $\Im(z_{\inf}) \geq \Im(z_k)$ and $\Im(z_{\sup}) \leq \Im(z_k)$, $\forall z_k \in Z$;
3. $\bigvee_{1 \leq k \leq K} \Re(z_k) < \{\Re(z_{\inf}), \Re(z_{\sup})\} < \bigwedge_{1 \leq k \leq K} \Re(z_k)$.

The proofs are straightforward from the notion of minimal enclosing semi-circle and the fact that $z_{\sup}$ lies inside the semi-circle defined by $z_{\inf}$.

Geodesic infimum and supremum being defined by minimal enclosing semi-circles, their invariance properties are related to translation and scaling of points in set Z as defined in Sect. 2.4, but not to inversion. This invariance domain just corresponds to the Guts group of transformations, i.e.,

$$\bigwedge_{\mathcal{H}^2}^{geo} \{T(z_k)\}_{1 \leq k \leq K} = T \left(\bigwedge_{\mathcal{H}^2}^{geo} \{z_k\}_{1 \leq k \leq K} \right).$$

As we discussed in Sect. 12.3, we do not have an explicit algorithm to compute the Burbea-Rao α-order entropy geodesic and consequently, our framework based on computing the minimum enclosing geodesic to define the infimum cannot be extended to this general case. We can nevertheless consider the example depicted in Fig. 12.8, where we have computed such smallest Burbea-Rao α-order geodesic enclosing the set of points Z. Indeed, the example is useful to identify the limit cases with respect to α. In fact, we note that if $\alpha \to 0$, the corresponding α-geodesic

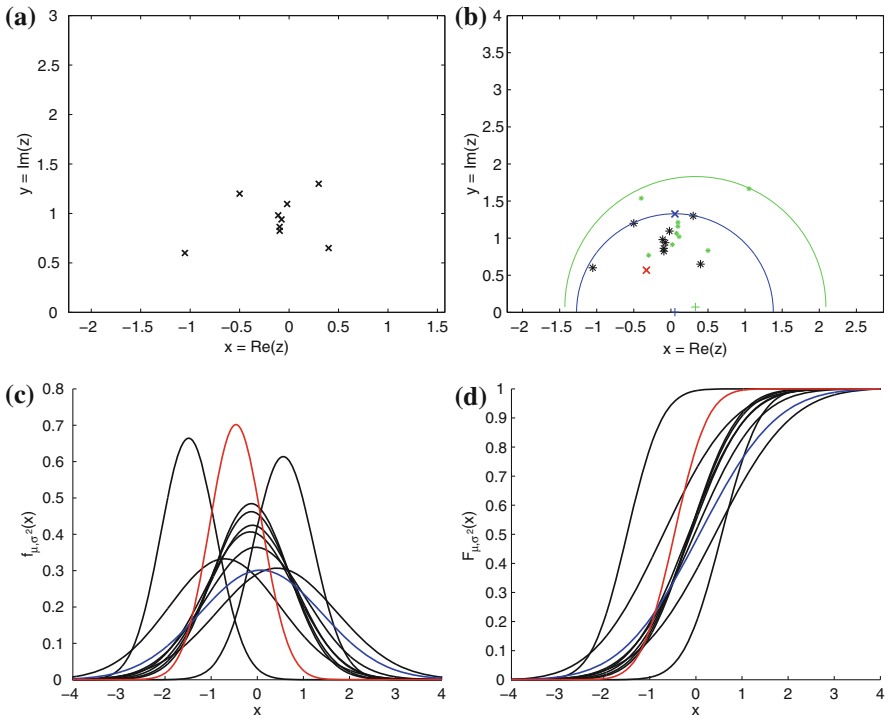


Fig. 12.7 **a** Set of nine points in $\mathcal{H}^2$, $Z = \{z_k\}_{1 \leq k \leq 9}$. **b** Computation of infimum $\bigwedge_{\mathcal{H}^2}^{geo} Z = z_{\inf}$ (*blue* "×") and supremum $\bigvee_{\mathcal{H}^2}^{geo} Z = z_{\sup}$ (*red* "×"). *Black*"∗" are the original points and *green* "∗" the corresponding dual ones. **c** In *black*, set of Gaussian pdfs associated to Z, i.e., $N_k(\mu = \sqrt{2}x_k, \sigma^2 = y_k^2)$; in *blue*, infimum Gaussian pdf $N_{\inf}(\mu = \sqrt{2}x_{\inf}, \sigma^2 = y_{\inf}^2)$; in *red*, supremum Gaussian pdf $N_{\sup}(\mu = \sqrt{2}x_{\sup}, \sigma^2 = y_{\sup}^2)$. **d** Cumulative distribution functions of Gaussian pdfs from **c**

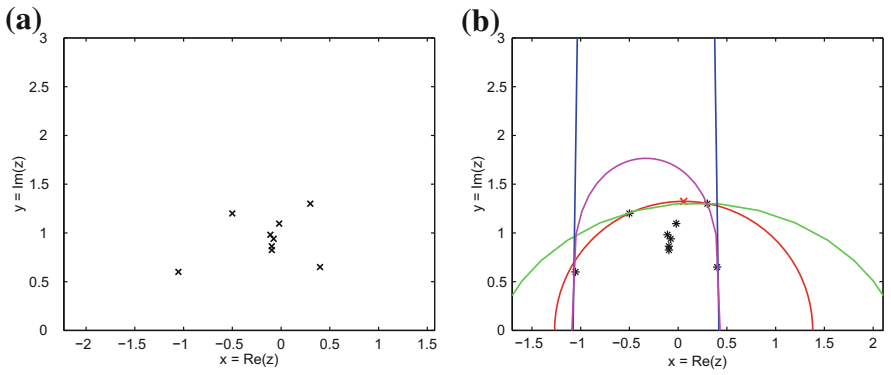


Fig. 12.8 **a** Set of nine points in $\mathcal{H}^2$, $Z = \{z_k\}_{1 \leq k \leq 9}$. **b** Computation of the smallest Burbea-Rao α-order geodesic enclosing the set Z, for $\alpha = 0.01$ (in *green*), $\alpha = 1$ (in *red*), $\alpha = 5$ (in *magenta*), $\alpha = 20$ (in *blue*)

infimum will correspond to the z_k having the largest imaginary part, and dually for the supremum, i.e., z_k having the smallest imaginary part. In the case of large α, we note that the real part of both, the α-geodesic infimum and supremum equals $\left(\vee_{1\leq k\leq K}\Re(z_k) - \wedge_{1\leq k\leq K}\Re(z_k)\right)/2$, and the imaginary part of the infimum goes to $+\infty$ and of the supremum to 0 when $\alpha \to +\infty$.

12.4.5 Upper Half-Pane Aymmetric Godesic Infimum/Supremum

According to the properties of geodesic infimum z_{inf} and supremum z_{sup} discussed above, we note that their real parts $\Re(z_{\text{inf}})$ and $\Re(z_{\text{sup}})$ belong to the interval bounded by the real parts of points of set Z. Moreover, $\Re(z_{\text{inf}})$ and $\Re(z_{\text{sup}})$ are not ordered between them. Therefore, the real part of supremum one can be smaller than that of the infimum one. For instance, in the extreme case of a set Z where all the imaginary parts are equal, the real part of its geodesic infimum and supremum are both equal to the average of the real parts of points, i.e., given $Z = \{z_k\}_{1\leq k\leq K}$, if $y_k = y, 1 \leq k \leq K$, then $\Re(z_{\text{inf}}) = \Re(z_{\text{sup}}) = 1/K \sum_{k=1}^{K} x_k$. From the viewpoint of morphological image filtering, it can be potentially interesting to impose an asymmetric behavior for the infimum and supremum such that $\Re(z_{\text{inf}}^{-\rightarrow+}) \leq z_k \leq \Re(z_{\text{sup}}^{-\rightarrow+}), 1 \leq k \leq K$. Note that the proposed notation $-\rightarrow+$ indicates a partially ordered set on x-axis. In order to fulfil these requirements, we can geometrically consider the rectangle bounding the minimal enclosing semi-circle, which is just of dimensions $2r_{\text{inf}} \times r_{\text{inf}}$, and use it to define the asymmetric infimum $z_{\text{inf}}^{-\rightarrow+}$ as the upper-left corner of the rectangle. The asymmetric supremum $z_{\text{sup}}^{-\rightarrow+}$ is similarly defined from the bounding rectangle of the dual minimal enclosing semi-circle. Mathematically, given the geodesic infimum z_{inf} and supremum z_{sup}, we have the following definitions for the asymmetric geodesic infimum and supremum (Fig. 12.9):

$$\begin{cases} z_{\text{inf}}^{-\rightarrow+} = \bigvee_{\mathcal{H}^2}^{-\rightarrow+} Z = (a_{\text{inf}} - r_{\text{inf}}) + ir_{\text{inf}}; \\ z_{\text{sup}}^{-\rightarrow+} = \bigwedge_{\mathcal{H}^2}^{-\rightarrow+} Z = -(x_c^{dual} - r^{dual}) + i\frac{1}{r^{dual}}. \end{cases} \tag{12.38}$$

Remark Geodesic infimum and supremum of Gaussian distributions. Let us consider their interpretation as infimum and supremum of a set of univariate Gaussian pdfs, see example depicted in Fig. 12.7. Given a set of K Gaussian pdfs $N_k(\mu = \sqrt{2}x_k, \sigma^2 = y_k^2), 1 \leq k \leq K$, we observe that the Gaussian pdf associated to the geodesic infimum $N_{\text{inf}}(\mu = \sqrt{2}x_{\text{inf}}, \sigma^2 = y_{\text{inf}}^2)$ has a variance larger than any Gaussian of the set and its mean is a kind of barycenter between the Gaussian pdfs having a larger variance. The supremum Gaussian pdf $N_{\text{sup}}(\mu = \sqrt{2}x_{\text{sup}}, \sigma^2 = y_{\text{sup}}^2)$ has a smaller variance than the K Gaussian pdfs and its mean is between the ones of smaller variance. In terms of the corresponding cumulative distribution functions, we observe that geodesic supremum/infimum do not have a natural interpretation. In the case of the asymmetric Gaussian geodesic infimum $N_{\text{inf}}^{-\rightarrow+}(\mu = \sqrt{2}x_{\text{inf}}^{-\rightarrow+}, \sigma^2 = (y_{\text{inf}}^{-\rightarrow+})^2)$ and Gaussian supremum $N_{\text{sup}}^{-\rightarrow+}(\mu = \sqrt{2}x_{\text{sup}}^{-\rightarrow+}, \sigma^2 = (y_{\text{sup}}^{-\rightarrow+})^2)$, we observe how the means are

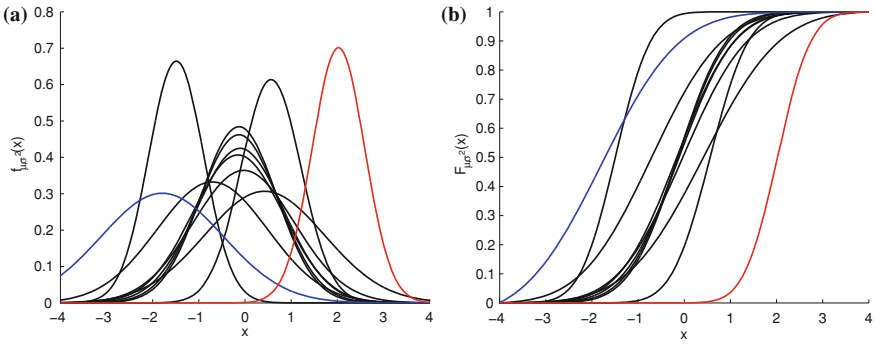


Fig. 12.9 **a** Infimum and supremum Gaussian pdfs (in *green* and *red* respectively) from asymmetric geodesic infimum $z_{\text{inf}}^{-\rightarrow+}$ and $z_{\text{sup}}^{-\rightarrow+}$ from set of Fig. 12.7. **b** Cumulative distribution functions of Gaussian pdfs from (**a**)

ordered with respect to the K others, which also involves that the corresponding cdfs are ordered. The latter is related to the notion of stochastic dominance [30] and will be explored in detail in ongoing research.

12.5 Morphological Operators on $\mathcal{F}(\Omega, \mathcal{H}^2)$ for Processing Univariate Gaussian Distribution-Valued Images

Let consider that $\mathcal{H}^2$ has been endowed with one of the partial orderings discussed above, generally denoted by $\leq$. Hence $(\mathcal{H}^2, \leq)$ is a poset, which has also a structure of complete lattice since we consider that infimum $\bigwedge$ and supremum $\bigvee$ are defined for any set of points in $\mathcal{H}^2$.

12.5.1 Adjunction on Complete Lattice $(\mathcal{H}^2, \leq)$

The operators $\varepsilon : \mathcal{H}^2 \rightarrow \mathcal{H}^2$ and $\delta : \mathcal{H}^2 \rightarrow \mathcal{H}^2$ are an erosion and a dilation if they commute respectively with the infimum and the supremum: $\varepsilon\left(\bigwedge_k z_k\right) = \bigwedge_k \varepsilon(z_k)$ and $\delta\left(\bigvee_k z_k\right) = \bigvee_k \delta(z_k)$, for every set $\{z_k\}_{1 \leq k \leq K}$. Erosion and dilation are increasing operators, i.e., $\forall z, z' \in \mathcal{H}^2$, if $z \leq z'$ then $\varepsilon(z) \leq \varepsilon(z')$ and $\delta(z) \leq \delta(z')$. Erosion and dilation are related by the notion of adjunction [21, 29], i.e.,

$$\delta(z) \leq z' \Leftrightarrow z \leq \varepsilon(z'); \quad \forall z, z' \in \mathcal{H}^2. \tag{12.39}$$

Adjunction law (12.39) is of fundamental importance in mathematical morphology since it allows to define a unique dilation δ associated to a given erosion ε,

i.e., $\delta(z') = \bigwedge\{z \in \mathcal{H}^2 : \ z' \leq \varepsilon(z)\}$, $z' \in \mathcal{H}^2$. Similarly one can define a unique erosion from a given dilation: $\varepsilon(z) = \bigvee\{z' \in \mathcal{H}^2 : \ \delta(z') \leq z\}$, $z \in \mathcal{H}^2$. Given an adjunction (ε, δ), their composition product operators, $\gamma(z) = \delta(\varepsilon(z))$ and $\varphi(z) = \varepsilon(\delta(z))$ are respectively an opening and a closing, which are the basic morphological filters having very useful properties [21, 29]: idempotency $\gamma\gamma(z) = \gamma(z)$, anti-extensivity $\gamma(z) \leq z$ and extensivity $z \leq \varphi(z)$, and increaseness. Another relevant result is the fact, given an erosion ε, the opening and closing by adjunction are exclusively defined in terms of erosion [21] as $\gamma(z) = \bigwedge\left\{z' \in \mathcal{H}^2 : \ \varepsilon(z) \leq \varepsilon(z')\right\}$, $\varphi(z) = \bigwedge\left\{\varepsilon(z') : \ z' \in \mathcal{H}^2, \ z \leq \varepsilon(z')\right\}$, $\forall z \in \mathcal{H}^2$.

In the case of complete inf-semilattice $(\mathcal{H}^2, \leq)$, where infimum $\bigwedge$ is defined but supremum $\bigvee$ is not necessarily, we have the following particular results [22, 23]: (a) it is always possible to associate an opening γ to a given erosion ε by means of $\gamma(z) = \bigwedge\left\{z' \in \mathcal{H}^2 : \ \varepsilon(z) \leq \varepsilon(z')\right\}$, (b) even though the adjoint dilation δ is not well-defined in $\mathcal{H}^2$, it is always well-defined for elements on the image of $\mathcal{H}^2$ by ε, and (c) $\gamma = \delta\varepsilon$. The closing defined by $\varphi = \varepsilon\delta$ is only partially defined. Obviously, in the case of inf-semilattice, it is still possible to define δ such that $\bigvee \delta(z_k) = \delta \bigvee (z_k)$ for families for which supremum $\bigvee$ exist.

12.5.2 Erosion and Dilation in $\mathcal{F}(\Omega, \mathcal{H}^2)$

If $(\mathcal{H}^2, \leq)$ is a complete lattice, the set of images $\mathcal{F}(\Omega, \mathcal{H}^2)$ is also a complete lattice defined as follows: for all $f, g \in \mathcal{F}(\Omega, \mathcal{H}^2)$, (i) $f \leq g \Leftrightarrow f(p) \leq g(p)$, $\forall p \in \Omega$; (ii) $(f \wedge g)(p) = f(p) \wedge g(p)$, $\forall p \in \Omega$; (iii) $(f \vee g)(p) = f(p) \vee g(p)$, $\forall p \in \Omega$, where $\wedge$ and $\vee$ are the infimum and supremum in $\mathcal{H}^2$. One can now define the following adjoint pair of flat erosion $\varepsilon_B(f)$ and flat dilation $\delta_B(f)$ of each pixel p of image f [21, 29]:

$$\varepsilon_B(f)(p) = \bigwedge_{q \in B(p)} f(p+q), \tag{12.40}$$

$$\delta_B(f)(p) = \bigvee_{q \in B(p)} f(p-q), \tag{12.41}$$

such that

$$\delta_B(f)(p) \leq g(p) \Leftrightarrow f(p) \leq \varepsilon_B(g)(p); \ \ \forall f, g \in \mathcal{F}(\Omega, \mathcal{H}^2). \tag{12.42}$$

where set B is called the structuring element, which defines the set of points in Ω when it is centered at point p, denoted $B(p)$ [31]. These operators, which are translation invariant, can be seen as constant-weight (this is the reason why they are called flat) inf/sup-convolutions, where the structuring element B works as a moving window.

The above erosion (resp. dilation) moves object edges within the image in such a way that it expands image structures with values in $\mathcal{H}^2$ close to the bottom element

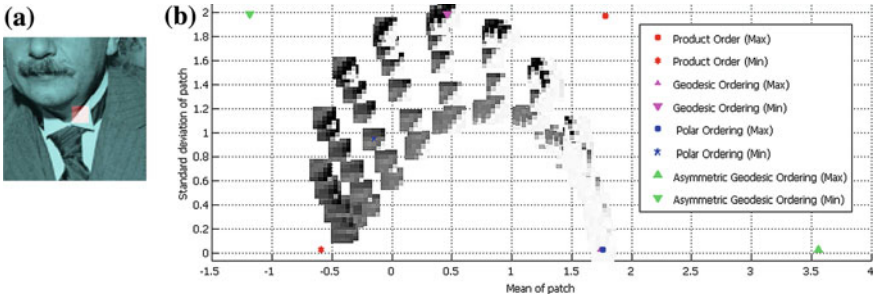


Fig. 12.10 Supremum and infimum of a set of 25 patches parameterized by their mean and standard deviation: **a** in *red* the region where the overlapped patches are taken; **b** embedding into the space $\mathcal{H}^2$ according to the coordinates $\mu/\sqrt{2}$ and σ and corresponding sup and inf for the different ordering strategies

(resp. close to the top) of the lattice $\mathcal{F}(\Omega, \mathcal{H}^2)$ and shrinks the objects with values close to the top element (resp. close to the bottom).

Let us consider now the various cases of supremum and infimum that we have introduced above. In order to support the discussion, we have included an example in Fig. 12.10. In fact, we have taken all the patches of size 5×5 pixels surrounding one of the pixels from image of Fig. 12.1. The 25 patches are then embedded into the space $\mathcal{H}^2$ according to the coordinates $\mu/\sqrt{2}$ and σ. Finally, the supremum and infimum of this set of points are computed for the different cases. It just corresponds to the way to obtain respectively the dilation and erosion for the current pixel center of the red region in image Fig. 12.10a.

Everything works perfectly for the supremum and infimum in the upper half-plane product ordering $\bigvee_{\mathcal{H}^2}$ and $\bigwedge_{\mathcal{H}^2}$, which consequently can be used to construct dilation and erosion operators in $\mathcal{F}(\Omega, \mathcal{H}^2)$. In fact, this is exactly equivalent to the classical operators applied on the real and imaginary parts separately.

Similarly, the ones for the upper half-plane polar ordering $\bigvee^{pol}_{\mathcal{H}^2}$ and $\bigwedge^{pol}_{\mathcal{H}^2}$, based on a total partial ordering, also lead respectively to dilation and erosion operators. The erosion produces a point which corresponds here to the patch closer to the origin. That means a patch of intermediate mean and standard deviation intensity since the image intensity is normalized, see Sect. 5.4. On the contrary, the dilation gives a point associated to the farthest patch from the origin. In this example, an homogenous bright patch. Note that patches of great distance correspond to the most "contrasted ones" on the image: either homogeneous patches of dark or bright intensity or patches with a strong variation in intensity (edge patches).

We note that for the symmetric ordering $\preceq_{\mathcal{H}^2}$ one only has an inf-semilattice structure associated to $\curlywedge_{\mathcal{H}^2}$. However, in the case of the upper half-plane geodesic ordering, the pair of operators (12.40) and (12.41) associated to our supremum $\bigvee^{geo}_{\mathcal{H}^2}$ and infimum $\bigwedge^{geo}_{\mathcal{H}^2}$ will not verify the adjunction (12.42). Same limitation also holds for the upper half-plane asymmetric geodesic supremum and infimum. Hence, the geodesic supremum and infimum do not strictly involve a pair of dilation and erosion

in the mathematical morphology sense. Nevertheless, we can compute both operators and use them to filter out images in $\mathcal{F}(\Omega, \mathcal{H}^2)$ without problem. From the example of Fig. 12.10 we observe that the geodesic infimum gives a point with a standard deviation equal to or larger than any of the patches and a mean intermediate between the patches of high standard deviation. The supremum involves a point of standard deviation smaller than (or equal to) than the others, and the mean is obtained by averaging around the mean of the ones with a small standard deviation. Consequently the erosion involves a nonlinear filtering which enhances the image zones of high standard deviation, typically the contours. The dilation enhances the homogenous zones. The asymmetrization produces operators where the dilation and erosion have the same interpretation for the mean as the classical ones but the filtering effects are driven by the zones of low or high standard deviation.

12.5.3 Opening and Closing in $\mathcal{F}(\Omega, \mathcal{H}^2)$

Given the adjoint image operators $(\varepsilon_B, \delta_B)$, the opening and closing by adjunction of image f, according to structuring element B, are defined as the composition operators [21, 29]:

$$\gamma_B(f) = \delta_B\left(\varepsilon_B(f)\right), \tag{12.43}$$

$$\varphi_B(f) = \varepsilon_B\left(\delta_B(f)\right). \tag{12.44}$$

Openings and closings are referred to as morphological filters, which remove objects of image f that do not comply with a criterion related, on the one hand, to the invariance of the object support to the structuring element B and, on the other hand, to the values of the object on $\mathcal{H}^2$ which are far from (in the case of the opening) or near to (in the case of the closing) the bottom element of $\mathcal{H}^2$ according to the given partial ordering $\leq$.

Once the pairs of dual operators $(\varepsilon_B, \delta_B)$ and (γ_B, φ_B) are defined, the other morphological filters and transformation can be naturally defined [31] for images in $\mathcal{F}(\Omega, \mathcal{H}^2)$. We limit here the illustrative examples to the basic ones.

Following our analysis on the particular cases of ordering and supremum/infimum in $\mathcal{H}^2$, we can conclude that opening and closing in $\mathcal{F}(\Omega, \mathcal{H}^2)$ are well formulated for the upper half-plane product ordering and the upper half-plane polar ordering. In the case of the upper half-plane symmetric ordering, the opening is always defined and the closing cannot be computed. Again, we should insist on the fact that for the upper half-plane geodesic ordering, the composition operators obtained by supremum $\bigvee_{\mathcal{H}^2}^{geo}$ and infimum $\bigwedge_{\mathcal{H}^2}^{geo}$ will not produce opening and closing *stricto sensu*. Notwithstanding, the corresponding composition operators yield a regularization effect of $\mathcal{F}(\Omega, \mathcal{H}^2)$-images which can be of interest for practical applications.

12.5.4 Application to Morphological Processing Univariate Gaussian Distribution Valued Images

Example 1 A first example of morphological processing for images in $\mathcal{F}(\Omega, \mathcal{H}^2)$ is given in Fig. 12.11. The starting point is a standard gray-level image $g \in \mathcal{F}(\Omega, \mathbb{R})$, which is mapped to the image $f(p) = f_x(p) + if_y(p)$ by the following transformations: (1) the image is normalized to have zero mean and a unit variance; (2) the real and imaginary components of $f(p)$ are obtained by computing respectively the mean and standard deviation over a patch centered at p of radius W pixels (in the example $W = 4$); i.e.,

$$g(p) \mapsto \hat{g}(p) = \frac{g(p) - \text{Mean}(g)}{\sqrt{\text{Var}(g)}} \mapsto f(p) = \text{Mean}_W(\hat{g})(p) + i\sqrt{\text{Var}_W(\hat{g})(p)}.$$

We note that definition of our representation space $f_y(p) > 0$. It means that the variance of each patch should always be bigger than zero and obviously this is not the case in constant patches. In order to cope with this problem, we propose to add a ϵ to the value of the standard deviation.

Figure 12.11 gives a comparison of morphological erosions $\varepsilon_B(f)(p)$ and openings $\gamma_B(f)(p)$ on this image f using the five complete (inf-semi)lattice of $\mathcal{H}^2$ considered in the paper. We have included also the pseudo-erosions and pseudo-openings associated to the geodesic supremum and infimum and the asymmetric geodesic ones. The same structuring element B, a square of 5×5 pixels, has been used for all the examples. First of all, we remind again that working on the product complete lattice $(\mathcal{H}^2, \leq_{\mathcal{H}^2})$ is equivalent to a marginal processing of real and imaginary components. As expected, the symmetric ordering-based inf-semilattice $(\mathcal{H}^2, \preceq_{\mathcal{H}^2})$ and polar ordering-based lattice $(\mathcal{H}^2, \leq_{\mathcal{H}^2}^{pol})$ produce rather similar results for openings. We observe that in both cases the opening produces a symmetric filtering effect between bright/dark intensity in the mean and standard deviation component. But it is important to remark that the processing effects depend on how image components are valued with respect to the origin $z_0 = (0, 1)$. This is the reason why it is proposed to always normalize by mean/variance the image.

The results of the pseudo-openings produced by working on geodesic lattice $(\mathcal{H}^2, \bigvee_{\mathcal{H}^2}^{geo}, \bigwedge_{\mathcal{H}^2}^{geo})$ and asymmetric geodesic lattice $(\mathcal{H}^2, \bigwedge_{\mathcal{H}^2}^{\overrightarrow{\ }+}, \bigvee_{\mathcal{H}^2}^{\overrightarrow{\ }+})$ involves a processing which is mainly driven by the values of the standard deviation. Hence, the filtering effects are potentially more interesting for applications requiring to deal with pixel uncertainty, either in a symmetric processing of both bright/dark mean values with $(\mathcal{H}^2, \bigvee_{\mathcal{H}^2}^{geo}, \bigwedge_{\mathcal{H}^2}^{geo})$ or in a more classical morphological asymmetrization with $(\mathcal{H}^2, \bigwedge_{\mathcal{H}^2}^{\overrightarrow{\ }+}, \bigvee_{\mathcal{H}^2}^{\overrightarrow{\ }+})$.

Example 2 Figure 12.12 illustrates a comparative example of erosions $\varepsilon_B(f)(p)$ on a very noisy image $g(p)$. We note that $g(p)$ is mean centered. The "noise" is related to an acquisition at the limit of exposure time/spatial resolution. We consider an image model $f(p) = f_x(p) + if_y(p)$, where $f_x(p) = g(x)$ and $f_y(p)$ is the standard deviation of intensities in a patch of radius equal to 4 pixels. In fact, the results of erosion obtained

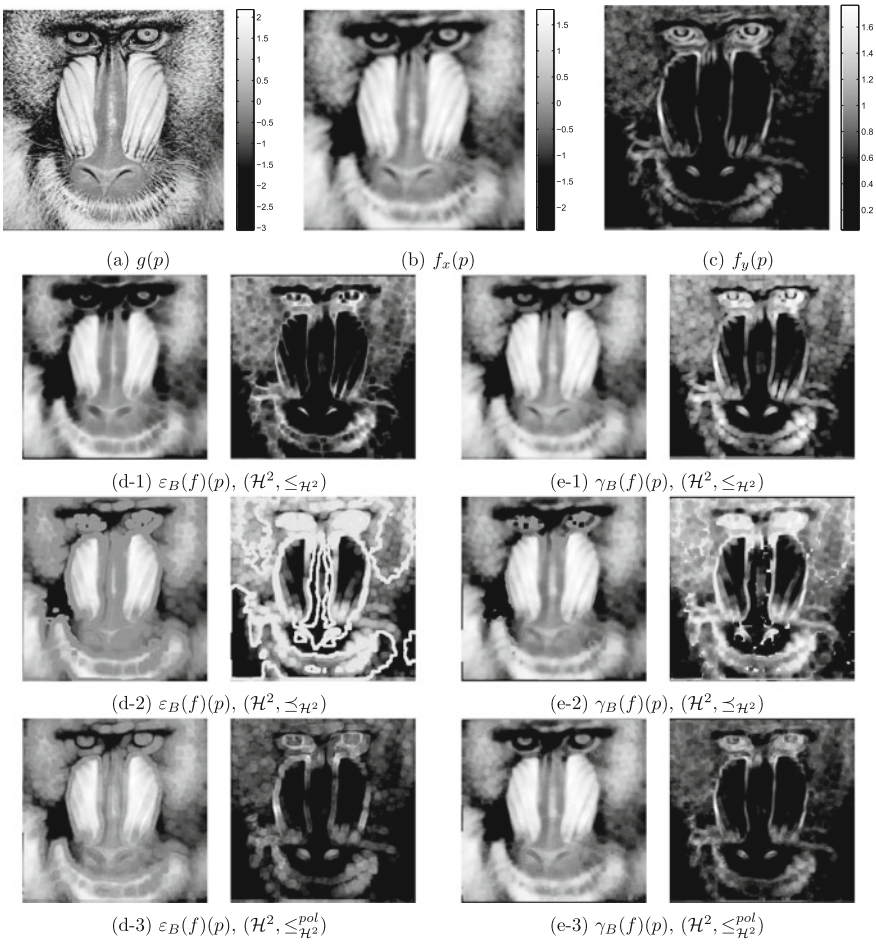
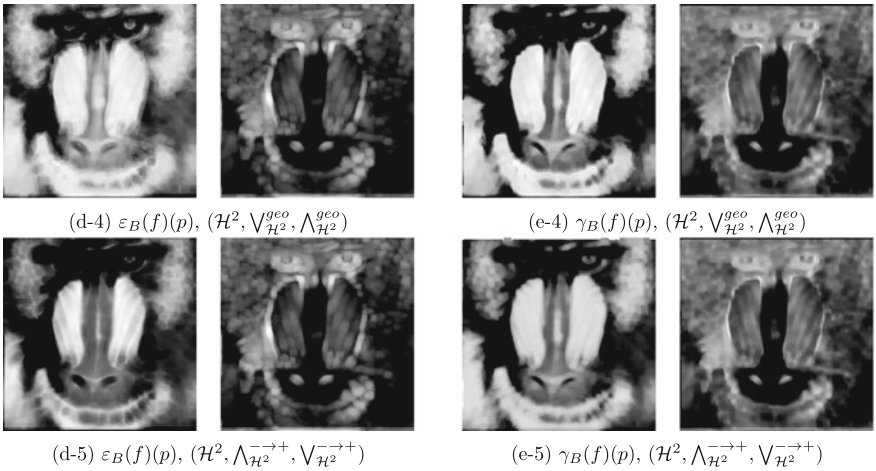


Fig. 12.11 Comparison of morphological erosions and openings of an image $f \in \mathcal{F}(\Omega, \mathcal{H}^2)$: **a** Original real-valued image $g(p) \in \mathcal{F}(\Omega, \mathbb{R})$ used to simulate (see the text) the image $f(p) = f_x(p) + if_y(p)$, where **b** and **c** gives respectively the real and imaginary components. **d**- and **e**- depict respectively the erosion $\varepsilon_B(f)(p)$ and opening $\gamma_B(f)(p)$ of image $f(p)$ for five orderings on the upper half-plane. The structuring element B is a window of 5×5 pixels. (Continued) in next figure

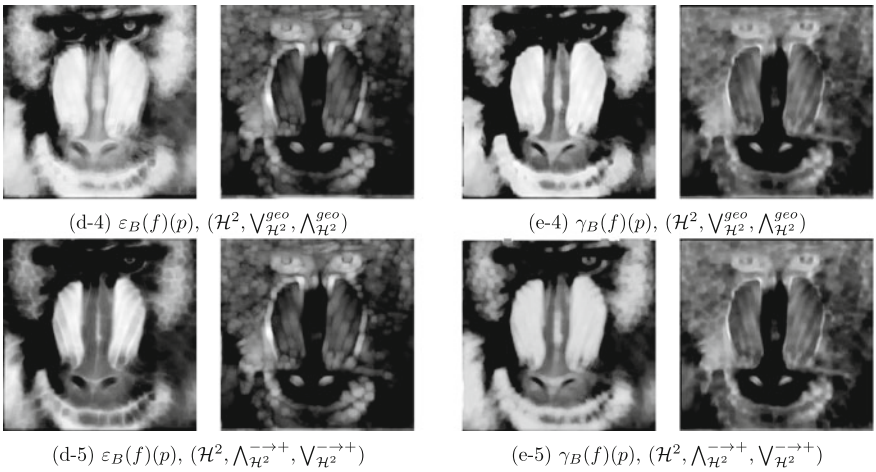
by the product and symmetric partial orderings, are compared to ones obtained by polar ordering and more generally by the α-polar ordering with four values of α. We observe, on the one hand, polar orderings are more relevant than the product or symmetric ones. As expected, the α-polar erosion with $\alpha = 1$ is almost equivalent to the hyperbolic polar ordering. We note, on the other hand, the interest of the limit cases of α-polar erosion. The erosion for small α produces a strongly regularized image where the bright/dark objects with respect to the background has been nicely



(d-4) $\varepsilon_B(f)(p)$, $(\mathcal{H}^2, \bigvee_{\mathcal{H}^2}^{geo}, \bigwedge_{\mathcal{H}^2}^{geo})$ (e-4) $\gamma_B(f)(p)$, $(\mathcal{H}^2, \bigvee_{\mathcal{H}^2}^{geo}, \bigwedge_{\mathcal{H}^2}^{geo})$

(d-5) $\varepsilon_B(f)(p)$, $(\mathcal{H}^2, \bigwedge_{\mathcal{H}^2}^{\overrightarrow{-}+}, \bigvee_{\mathcal{H}^2}^{\overrightarrow{-}+})$ (e-5) $\gamma_B(f)(p)$, $(\mathcal{H}^2, \bigwedge_{\mathcal{H}^2}^{\overrightarrow{-}+}, \bigvee_{\mathcal{H}^2}^{\overrightarrow{-}+})$

Fig. 12.11 continued



(d-4) $\varepsilon_B(f)(p)$, $(\mathcal{H}^2, \bigvee_{\mathcal{H}^2}^{geo}, \bigwedge_{\mathcal{H}^2}^{geo})$ (e-4) $\gamma_B(f)(p)$, $(\mathcal{H}^2, \bigvee_{\mathcal{H}^2}^{geo}, \bigwedge_{\mathcal{H}^2}^{geo})$

(d-5) $\varepsilon_B(f)(p)$, $(\mathcal{H}^2, \bigwedge_{\mathcal{H}^2}^{\overrightarrow{-}+}, \bigvee_{\mathcal{H}^2}^{\overrightarrow{-}+})$ (e-5) $\gamma_B(f)(p)$, $(\mathcal{H}^2, \bigwedge_{\mathcal{H}^2}^{\overrightarrow{-}+}, \bigvee_{\mathcal{H}^2}^{\overrightarrow{-}+})$

Fig. 12.11 (continued)

enhanced. In the case of large α, the background (i.e., pixels values close to the origin in $\mathcal{H}^2$) is enhanced, which involves removing all the image structures smaller than the structuring element B.

Example 3 In Fig. 12.13 a limited comparison for the case of dilation $\delta_B(f)(p)$ is depicted. The image $f(p) = f_x(p) + if_y(p)$ is obtained similarly to the case of Example 1. We can compare the supremum by product ordering with those obtained by the polar supremum and the α-polar supremum, with $\alpha = 0.01$. The analysis is similar to the previous case.

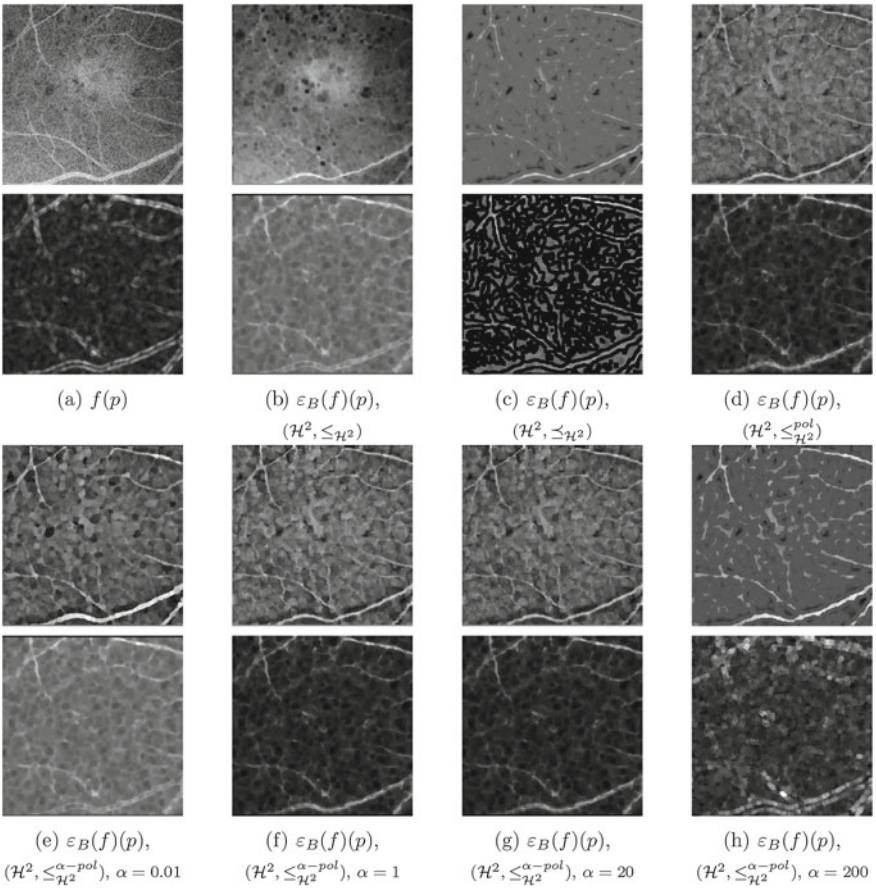


Fig. 12.12 Comparison of erosion of Gaussian distribution-valued noisy image $\varepsilon_B(f)(p)$: **a** Original image $f \in \mathcal{F}(\Omega, \mathcal{H}^2)$, showing both the real and the imaginary components; **b** *upper* half-plane product ordering (equivalent to standard processing); **c** *upper* half-space symmetric ordering; **d** *upper* half-plane polar ordering; **e**–**h** *upper* half-plane α-polar ordering, with four values of α. In all the cases the structuring element B is also a square of 5×5 pixels

Example 4 Figure 12.14 involves again the noisy retinal image, and it shows a comparison of results from (pseudo-)opening $\gamma_B(f)(p)$ and (pseudo-)closing $\varphi_B(f)(p)$ obtained for the product ordering, the geodesic lattice $(\mathcal{H}^2, \bigvee^{geo}_{\mathcal{H}^2}, \bigwedge^{geo}_{\mathcal{H}^2})$ and the asymmetric geodesic lattice $(\mathcal{H}^2, \bigwedge^{\rightarrow +}_{\mathcal{H}^2}, \bigvee^{\rightarrow +}_{\mathcal{H}^2})$. The structuring element B is a square of 5×5 pixels. In order to be able to compare their enhancement effects with an averaging operator, it is also given the result of filtering by computing the minimax center in a square of 5×5 pixels [2, 8], see Remark in Sect. 2.5. We note that operators associated to the asymmetric geodesic supremum and infimum yield mean images relatively similar to the standard ones underlaying the supremum and infimum in the product lattice. However, including the information given by the local standard deviation, the contrast of the structures is better in the asymmetric geodesic

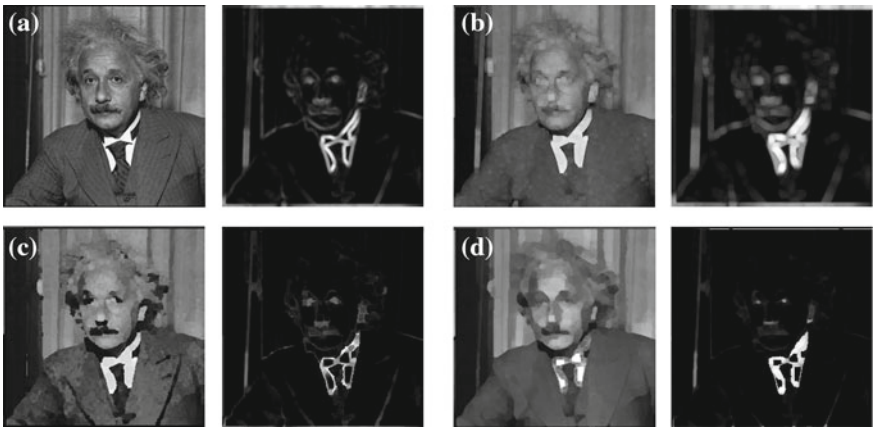


Fig. 12.13 Comparison of dilation of Gaussian distribution-valued image $\delta_B(f)(p)$: **a** Original image $f \in \mathcal{F}(\Omega, \mathcal{H}^2)$, showing both the real and the imaginary components; **b** *upper* half-plane product ordering (equivalent to standard processing); **c** *upper* half-plane polar ordering; **e** half-plane α-polar ordering, with $\alpha = 0.01$. In all the cases the structuring element B is also a square of 5×5 pixels

supremum and infimum. Nevertheless, we observe that the operators by geodesic supremum and infimum also produce in this example a significant regularization of the image. By the way, we note that the corresponding geodesic pseudo-opening and pseudo-closing give rather similar mean images but different standard deviation images, as expected by the formulation of the geodesic supremum and infimum.

Example 5 The example given in Fig. 12.15 corresponds to an image $f(p) = f_x(p) + if_y(p)$ obtained by multiple acquisition of a sequence of 100 frames, where $f_x(p)$ represents the mean intensity at each pixel and $f_y(p)$ the standard deviation of intensity along the sequence. The 100 frames have been taken from a stationary camera.

The goal of the example is to show how to extract image objects of large intensity and support size smaller than the structuring element (here a square of 7×7 pixels) using the residue between the original image $f(p)$ and its filtered image by opening $\gamma_B(f)$. In the case of images on $\mathcal{F}(\Omega, \mathcal{H}^2)$, the residue is defined as the pixelwise hyperbolic distance between them. In this case study, results on processing on polar ordering-based lattice versus asymmetric geodesic lattice are compared.

12.5.5 Conclusions on Morphological Operators for $\mathcal{F}(\Omega, \mathcal{H}^2)$ images

Based on the discussion given in Sect. 5.2 as well as on the examples from Sect. 5.4, we can draw some conclusions on the experimental part of this chapter.

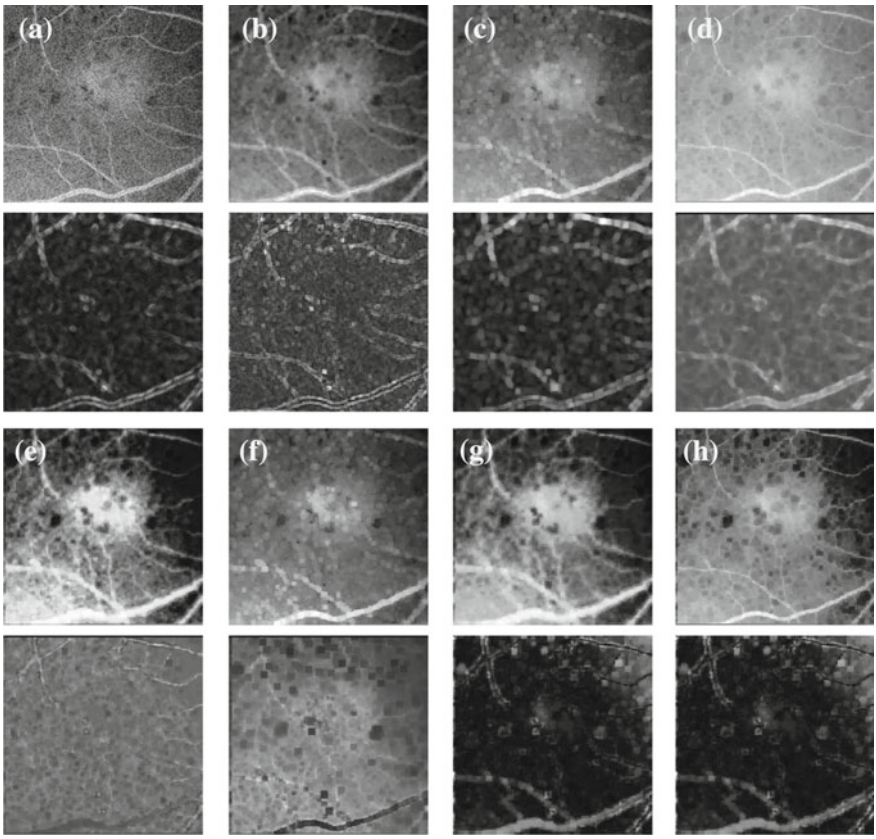


Fig. 12.14 Morphological processing of Gaussian distribution-valued noisy image: **a** Original image $f \in \mathcal{F}(\Omega, \mathcal{H}^2)$, showing both the real and the imaginary components; **b** filtered image by computing the minimax center in a square of 5×5 pixels; **c** morphological opening working on the product lattice; **d** morphological closing working on the product lattice; **e** morphological pseudo-opening working on the geodesic lattice; **f** morphological pseudo-opening on the asymmetric geodesic lattice; **g** morphological pseudo-closing working on the geodesic lattice; **h** morphological pseudo-closing on the asymmetric geodesic lattice. In all the cases the structuring element B is also a square of 5×5 pixels

- First of all, we note that the examples considered here are only a preliminary exploration on the potential applications of morphological processing univariate Gaussian distribution-valued images.
- We have two main case studies. First, standard images which are embedded into the Poincaré upper-half plane representation by parameterization of each local patch by its mean and standard deviation. Second, images which naturally involves a distribution of values at each pixel. Note that in the first case, the information of standard deviation is mainly associated to discriminate between homogenous zones and inhomogeneous ones (textures or contours). In the second case, the

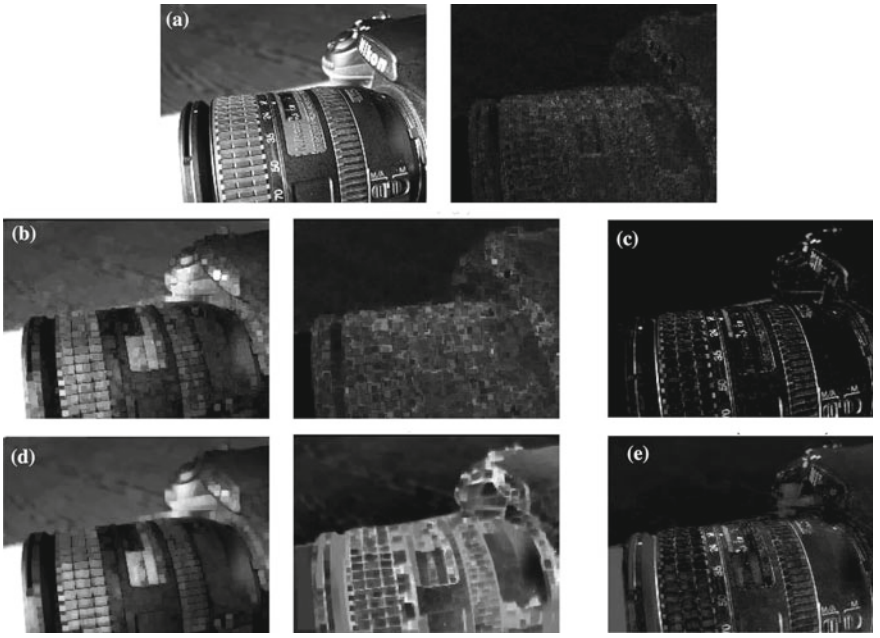


Fig. 12.15 Morphological detail extraction of multiple acquisition image modeled as a Gaussian distribution-valued: **a** Original image $f \in \mathcal{F}(\Omega, \mathcal{H}^2)$, showing both the real and the imaginary components; **b** morphological opening $\gamma_B(f)$ working on polar ordering-based lattice; **c** corresponding residue (pixelwise hyperbolic difference) between the original and the opened image; **d** morphological pseudo-opening $\gamma_B(f)$ working on the asymmetric geodesic lattice; **e** corresponding residue. In both cases the structuring element B is also a square of 7×7 pixels

standard deviation involves relevant information on the nature of the noise during the acquisition.

- For any of these two cases, we should remark that different alternatives of ordering and derived operators considered in the paper will produce nonlinear processing its main property is that filtering effects are strongly driven by the standard deviation.
- Upper half-plane product ordering is nothing more than standard processing of mean and standard deviation separately. The symmetric ordering leading to an inf-semilattice has a limited interest since similar effects are obtained by the polar ordering.
- Upper half-plane polar ordering using standard hyperbolic polar coordinates or the α-order Hellinger distance produces morphological operators appropriate for image regularization and enhancement. We remind that points close to the origin (selected by the erosion) correspond in the case of the patches to those of intermediate mean and standard deviation intensity after normalization. On the contrary, patches far from the origin correspond to the "contrasted" ones: either homogeneous patches of dark or bright intensity or patches with a strong variation in intensity (edge patches).

We note that with respect to filters based on averaging, the half-plane polar dilation/erosion as well as their product operators, produces strong simplified images where the edges and the main objects are enhanced without any blurring effect.
From our viewpoint this is useful for both cases of images. Then, the choice of a high or low value for α will depend on the particular nature of the features to be enhanced. In any case, this parameter can be optimized.

- Upper half-plane geodesic ordering involves a nonlinear filtering framework which takes into account the intrinsic geometry of $\mathcal{H}^2$. It is mainly based on the notion of minimal enclosing geodesic which covers the set of points.

In practice, the geodesic infimum gives a point with a standard deviation equal to or larger than any of the point and a mean which can be seen as intermediate between the mean values of high standard deviation. The supremum produces a point of standard deviation equal to or smaller than the others, and the mean is obtained by averaging around the mean of the ones having a small standard deviation.

Consequently the erosion involves a nonlinear filtering which enhances the image zones of high standard deviation, typically the contours. The dilation enhances the homogenous zones. We should note that the processed mean images by the composition of these two operators (i.e., openings and closings) are strongly enhanced by increasing their bright/dark contrast. Therefore, it should be considered as an appropriate tool for contrast structure enhancement on irregular backgrounds.

The asymmetric version of the geodesic ordering involves that dilation and erosion have the same interpretation for the mean as the classical ones but the filtering effects are driven by the zones of low or high standard deviation. These operators are potentially useful for object extraction by residue between the original image and the opening/closing. In comparison with classical residues, the new ones produce sharper extracted objects.

12.6 Perspectives

Levelings are a powerful family self-dual morphological operators which have been also formulated in vector spaces [24], using geometric notions as minimum enclosing balls and half-planes intersection. We intend to explore the formulation of levelings in the upper half-plane in a future work.

The complete lattice structures for the Poincaré upper-half plane introduced in this work, and corresponding morphological operators, can be applied to process other hyperbolic-valued images. For instance, on the one hand, it was proven in [13] that the structure tensor for 2D images, i.e., at each pixel is given a 2×2 symmetric positive definite matrix which determinant is equal to 1, are isomorphic to the Poincaré unit disk model. On the other hand, polarimetric images [17] where at each pixel is given a partially polarized state can be embedded in the Poincaré unit disk model. In both cases, we only need the mapping from the Poincaré disk model to the Poincaré

half-plane, i.e.,

$$z \mapsto -i\frac{z+1}{z-1}.$$

We have considered here the case of Gaussian distribution-valued images. It should be potentially interesting for practical applications to consider that the distribution of intensity at a given pixel belongs to a more general distributions compared to the Gaussian one. In particular, the case of the Gamma distribution seems an appropriate framework. The information geometry of the gamma manifold has been studied in the past [16] and some of the ideas developed in this work can be revisited for the case of Gamma-distribution valued images by endowing the gamma manifold of complete lattice structure.

Previous extension only concerns the generalization of ordering structure for univariate distributions. In the case of multivariate Gaussian distributions, we can consider to replace the Poincaré upper-half plane by the Siegel upper-half space [7].

References

1. Angulo, J., Velasco-Forero, S.: Complete lattice structure of Poincaré upper-half plane and mathematical morphology for hyperbolic-valued images. In: Nielsen, F., Barbaresco, F. (eds.) Proceedings of First International Conference Geometric Science of Information (GSI'2013), vol. 8085, pp. 535–542. Springer LNCS (2013)
2. Arnaudon, M., Nielsen, F.: On approximating the riemannian 1-center. Comput. Geom. **46**(1), 93–104 (2013)
3. Amari, S.-I., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L., Rao, C.R.: Differential geometry in statistical inference. Lecture Notes-Monograph Series, vol. 10, pp. 19–94, Institute of Mathematical Statistics, Hayward (1987)
4. Amari, S.-I., Nagaoka, H.: Methods of information geometry, translations of mathematical monographs. Am. Math. Soc. **191**, (2000)
5. Barbaresco, F.: Interactions between symmetric cone and information geometries: Bruhat-Tits and siegel spaces models for high resolution autoregressive doppler imagery. In: Nielsen, F. (eds.) Emerging Trends in Visual Computing (ETCV'08), Springer LNCS, Heidelberg vol. 5416, pp. 124–163, (2009)
6. Barbaresco, F.: Geometric radar processing based on Fréchet distance: information geometry versus optimal transport theory. In: Proceedings of IEEE International Radar Symposium (IRS'2011), pp. 663–668 (2011)
7. Barbaresco, F.: Information geometry of covariance matrix: cartan-siegel homogeneous bounded domains, Mostow/Berger fibration and fréchet median. In: Nielsen, F., Bhatia, R. (eds.) Matrix Information Geometry, pp. 199–255, Springer, Heidelberg (2013)
8. Bădoiu, M., Clarkson, K.L.: Smaller core-sets for balls. In: Proceedings of the Fourteenth annual ACM-SIAM Symposium on Discrete Algorithms (SIAM), pp. 801–802, ACM, New York(2003)
9. Burbea, J., Rao, C.R.: Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. J. Multivar. Anal. **12**(4), 575–96 (1982)
10. Căliman, A., Ivanovici, M., Richard, N.: Probabilistic pseudo-morphology for grayscale and color images. Pattern Recogn. **47**, 721–35 (2004)
11. Cammarota, V., Orsingher, E.: Travelling randomly on the poincaré half-plane with a pythagorean compass. J. Stat. Phys. **130**(3), 455–82 (2008)

12. Cannon, J.W., Floyd, W.J., Kenyon, R., Parry, W.R.: Hyperbolic geometry. Flavors of Geometry, vol. 31, MSRI Publications, Cambridge (1997)
13. Chossat, P., Faugeras, O.: Hyperbolic planforms in relation to visual edges and textures perception. PLoS Comput. Biol. **5**(12), p1 (2009)
14. Costa, S.I.R., Santos, S.A., Strapasson, J.E.: Fisher information matrix and hyperbolic geometry. In: Proc. of IEEE ISOC ITW2005 on Coding and Complexity, pp. 34–36, (2005)
15. Costa, S.I.R., Santos, S.A., Strapasson, J.E.: Fisher information distance: a geometrical reading, arXiv:1210:2354v1, p. 15 (2012)
16. Dodson, C.T.J., Matsuzoe, H.: An affine embedding of the gamma manifold. Appl. Sci. **5**(1), 7–12 (2003)
17. Frontera-Pons, J., Angulo, J.: Morphological operators for images valued on the sphere. In: Proceedings of IEEE ICIP'12 (IEEE International Conference on Image Processing), pp. 113–116, Orlando (Florida), USA, October (2012)
18. Fuchs, L.: Partially Ordered Algebraic Systems. Pergamon, Oxford (1963)
19. Guts, A.K.: Mappings of families of oricycles in lobachevsky space. Math. USSR-Sb. **19**, 131–8 (1973)
20. Guts, A.K.: Mappings of an ordered lobachevsky space. Siberian Math. J. **27**(3), 347–61 (1986)
21. Heijmans, H.J.A.M.: Morphological Image Operators. Academic Press, Boston (1994)
22. Heijmans, H.J.A.M., Keshet, R.: Inf-semilattice approach to self-dual morphology. J. Math. Imaging Vis. **17**(1), 55–80 (2002)
23. Keshet, R.: Mathematical morphology on complete semilattices and its applications to image processing. Fundamenta Informaticæ **41**, 33–56 (2000)
24. Meyer, F.: Vectorial Levelings and Flattenings. In: Mathematical Morphology and its Applications to Image and Signal Processing (Proc. of ISMM'02), pp. 51–60, Kluwer Academic Publishers, Dordrecht (2000)
25. Nielsen, F., Nock, R.: On the smallest enclosing information disk. Inform. Process. Lett. **105**, 93–7 (2008)
26. Nielsen, F., Nock. R.: Hyperbolic voronoi diagrams made easy. In: Proceedings of the 2010 IEEE International Conference on Computational Science and Its Applications, pp. 74–80, IEEE Computer Society, Washington (2010)
27. Sachs, Z.: Classification of the isometries of the upper half-plane, p. 14. University of Chicago, VIGRE REU (2011)
28. Sbaiz, L., Yang, F., Charbon, E., Süsstrunk, S., Vetterli, M.: The gigavision camera. In: Proceedings of IEEE ICASSP'09, pp. 1093–1096 (2009)
29. Serra, J.: Image Analysis and Mathematical Morphology. Vol II: theoretical advances, Academic Press, London (1988)
30. Shaked, M., Shanthikumar, G.: Stochastic Orders and Their Applications. Associated Press, New York (1994)
31. Soille, P.: Morphological Image Analysis. Springer-Verlag, Berlin (1999)
32. Treibergs, A.: The hyperbolic plane and its immersions into $\mathbb{R}^3$, Lecture Notes in Department of Mathematics, p. 13. University of Utah (2003)

Chapter 13
Dimensionality Reduction for Classification of Stochastic Texture Images

C. T. J. Dodson and W. W. Sampson

Abstract Stochastic textures yield images representing density variations of differing degrees of spatial disorder, ranging from mixtures of Poisson point processes to macrostructures of distributed finite objects. They arise in areas such as signal processing, molecular biology, cosmology, agricultural spatial distributions, oceanography, meteorology, tomography, radiography and medicine. The new contribution here is to couple information geometry with multidimensional scaling, also called dimensionality reduction, to identify small numbers of prominent features concerning density fluctuation and clustering in stochastic texture images, for classification of groupings in large datasets. Familiar examples of materials with such textures in one dimension are cotton yarns, audio noise and genomes, and in two dimensions paper and nonwoven fibre networks for which radiographic images are used to assess local variability and intensity of fibre clustering. Information geometry of trivariate Gaussian spatial distributions of mean pixel density with the mean densities of its first and second neighbours illustrate features related to sizes and density of clusters in stochastic texture images. We derive also analytic results for the case of stochastic textures arising from Poisson processes of line segments on a line and rectangles in a plane. Comparing human and yeast genomes, we use 12-variate spatial covariances to capture possible differences relating to secondary structure. For each of our types of stochastic textures: analytic, simulated, and experimental, we obtain dimensionality reduction and hence 3D embeddings of sets of samples to illustrate the various features that are revealed, such as mean density, size and shape of distributed objects, and clustering effects.

C. T. J. Dodson (✉)
School of Mathematics, University of Manchester,
Manchester, M13 9PL, UK
e-mail: ctdodson@manchester.ac.uk

W. W. Sampson
School of Materials, University of Manchester, Manchester, M13 9PL, UK
e-mail: william.sampson@manchester.ac.uk

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_13,

Keywords Dimensionality reduction · Stochastic texture · Density array · Clustering · Spatial covariance · Trivariate Gaussian · Radiographic images · Genome · Simulations · Poisson process

13.1 Introduction

The new contribution in this paper is to couple information geometry with dimensionality reduction, to identify small numbers of prominent features concerning density fluctuation and clustering in stochastic texture images, for classification of groupings in large datasets. Our methodology applies to any stochastic texture images, in one, two or three dimensions, but to gain an impression of the nature of examples we analyse some familiar materials for which we have areal density arrays, and derive analytic expressions of spatial covariance matrices for Poisson processes of finite objects in one and two dimensions. Information geometry provides a natural distance structure on the textures via their spatial covariances, which allows us to obtain multidimensional scaling or dimensionality reduction and hence 3D embeddings of sets of samples. See Mardia et al. [14] for an account of the original work on multidimensional scaling.

The simplest one-dimensional stochastic texture arises as the density variation along a cotton yarn, consisting of a near-Poisson process of finite length cotton fibres on a line, another is an audio noise drone consisting of a Poisson process of superposed finite length notes or chords. A fundamental microscopic 1-dimensional stochastic process is the distribution of the 20 amino acids along protein chains in a genome [1, 3]. Figure 13.1 shows a sample of such a sequence of the 20 amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y mapped onto the 20 grey level values 0.025, 0.075, . . . , 0.975 from the database [19], so yielding a grey-level barcode as a 1-dimensional texture. We analyse such textures in Sect. 13.6.5.

The largest 3-dimensional stochastic structure is the cosmological void distribution, which is observable via radio astronomy [1]. More familiar three-dimensional stochastic porous materials include metallic (Fig. 13.2) and plastic solid foams, geological strata and dispersions in gels, observable via computer tomography [1]. Near-planar, non-woven stochastic fibre networks are manufactured for a variety of applications such as, at the macroscale for printing, textiles, reinforcing, and filtration and at the nanoscale in medicine. Figure 13.3 shows a selection of electron micrographs for networks at different scales. Radiography or optical densitometry yield areal density images of the kinds shown in Fig. 13.4.

Much analytic work has been done on modelling of the statistical geometry of stochastic fibrous networks [1, 6, 7, 17]. Using complete sampling by square cells, their areal density distribution is typically well represented by a log-gamma or a (truncated) Gaussian distribution of variance that decreases monotonically with increasing cell size; the rate of decay is dependent on fibre and fibre cluster dimensions. They have gamma void size distributions with a long tail. Clustering of fibres is well-approximated by Poisson processes of Poisson clusters of differing density and size.

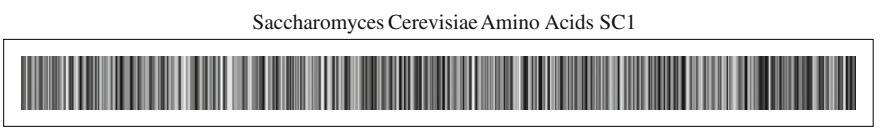


Fig. 13.1 Example of a 1-dimensional stochastic texture, a *grey* level barcode for the amino acid sequence in a sample of the *Saccharomyces cerevisiae* yeast genome from the database [19]

Fig. 13.2 Aluminium foam with a narrow Gaussian-like distribution of void sizes of around 1 cm diameter partially wrapped in fragmented metallic shells, used as crushable buffers inside vehicle bodies. The cosmological void distribution is by contrast gamma-like with a long tail [8], interspersed with 60 % of galaxies in large-scale sheets, 20 % in rich filaments and 20 % in sparse filaments [12]. Such 3D stochastic porous materials can both be studied by tomographic methods, albeit at different scales by different technologies, yielding sequences of 2D stochastic texture images

An unclustered Poisson process of single fibres is the standard reference structure for any given size distribution of fibres; its statistical geometry is well-understood for finite and infinite fibres. Note that any skewness associated with the underlying point process of fibre centres becomes negligible through the process of sampling by square cells [18].

Many stochastic textures arise from spatial processes that may be approximated by mixtures of Poisson or other distributions of finite objects or clusters of objects, in an analogous way to that which has been used for the past 50 years for the study of fibre networks. The Central Limit Theorem suggests that often such spatial processes may be represented by Gaussian pixel density distributions, with variance decreasing as pixel size increases, the gradient of this decrease reflecting the size distributions and abundances of the distributed objects and clusters, hence indicating the appropriate pixel size to choose for feature extraction. Once a pixel size has been chosen then we are interested in the statistics of three random variables: the mean density in such pixels, and the mean densities of its first and second neighbouring pixels. The correlations among these three random variables reflect the size and distribution of

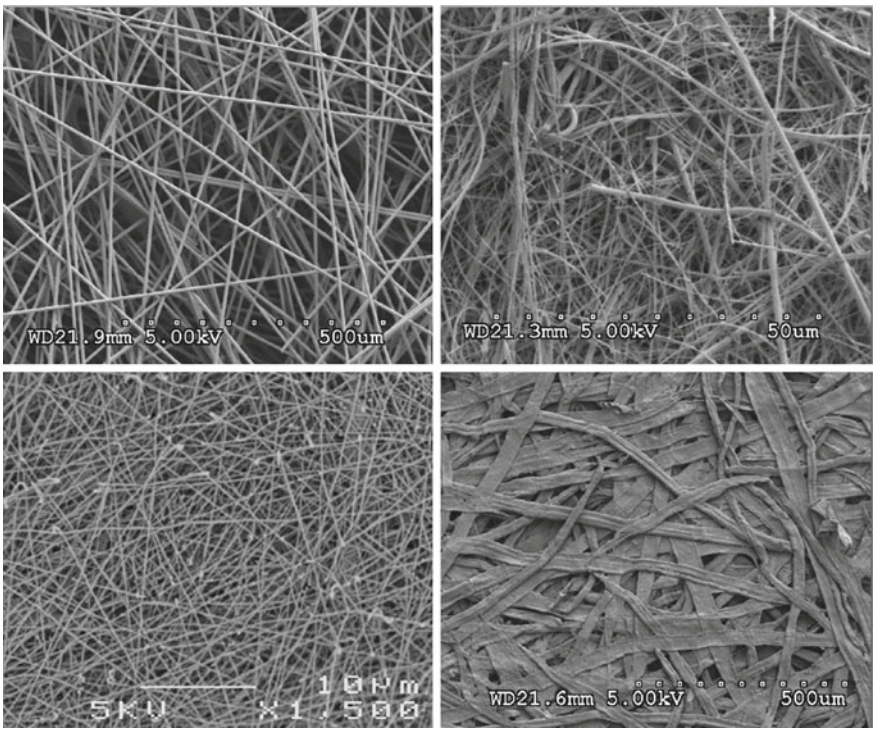


Fig. 13.3 Electron micrographs of four stochastic fibrous materials. *Top left* Nonwoven carbon fibre mat; *top right* glass fibre filter; *bottom left* electrospun nylon nanofibrous network (Courtesy S. J. Eichhorn and D. J. Scurr); *bottom right* paper using wood cellulose fibres—typically flat ribbonlike, of length 1–2 mm and width 0.02–0.03 mm

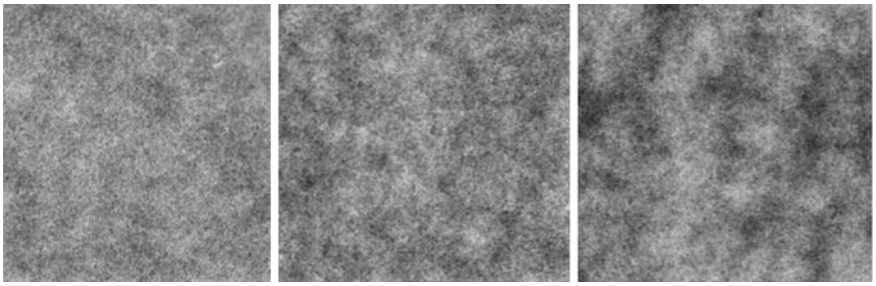


Fig. 13.4 Areal density radiographs of three paper networks made from natural wood cellulose fibres, with constant mean coverage, $\bar{c} \approx 20$ fibres, but different distributions of fibres. Each *image* represents a *square region* of side length 5 cm; *darker regions* correspond to higher coverage. The *left image* is similar to that expected for a Poisson process of the same fibres, so typical real samples exhibit clustering of fibres

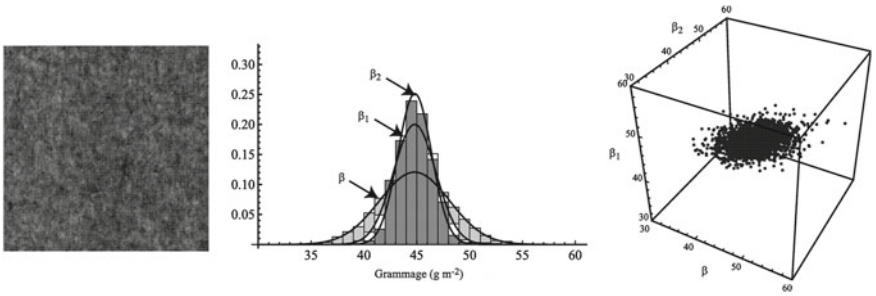


Fig. 13.5 Trivariate distribution of pixel density values for radiograph of a 5 cm square newsprint sample. *Left* source density map; *centre* histogram of $\tilde{\beta}_i$, $\tilde{\beta}_{1,i}$ and $\tilde{\beta}_{2,i}$; *right* 3D scatter plot of $\tilde{\beta}_i$, $\tilde{\beta}_{1,i}$ and $\tilde{\beta}_{2,i}$

density clusters; this may be extended to more random variables by using also third, fourth, etc., neighbours. In some cases, of course, other pixel density distributions may be more appropriate, such as mixtures of Gaussians.

13.2 Spatial Covariance

The mean of a random value p is its average value, $\bar{p}$, over the population. The covariance $\mathrm{Cov}(p, q)$ of a pair of random variables, p and q is a measure of the degree of association between them, the difference between their mean product and the product of their means:

$$\mathrm{Cov}(p, q) = \overline{p\,q} - \bar{p}\,\bar{q}\ . \tag{13.1}$$

In particular, the covariance of a variable with itself is its variance. From the array of local average pixel density values $\tilde{\beta}_i$, we generate two numbers associated with each: the average density of the six first-neighbour pixels, $\tilde{\beta}_{1,i}$ and the average density of the 16 second-neighbour pixels, $\tilde{\beta}_{2,i}$. Thus, we have a trivariate distribution of the random variables $(\tilde{\beta}_i, \tilde{\beta}_{1,i}, \tilde{\beta}_{2,i})$ with $\bar{\beta}_2 = \bar{\beta}_1 = \bar{\beta}$.

Figure 13.5 provides an example of a typical data set obtained from a radiograph of a 5 cm square commercial newsprint sample; the histogram and three-dimensional scatter plot show data obtained for pixels of side 1 mm.

From the Central Limit Theorem, we expect the marginal distributions of $\tilde{\beta}_i$, $\tilde{\beta}_{1,i}$ and $\tilde{\beta}_{2,i}$ to be well approximated by Gaussian distributions. For the example in Fig. 13.5, these Gaussians are represented by the solid lines on the histogram; this Gaussian approximation holds for all samples investigated in this study.

We have a simulator for creating stochastic fibre networks [10]. The code works by dropping clusters of fibres within a circular region where the centre of each cluster is distributed as a point Poisson process in the plane and the number of fibres per cluster,

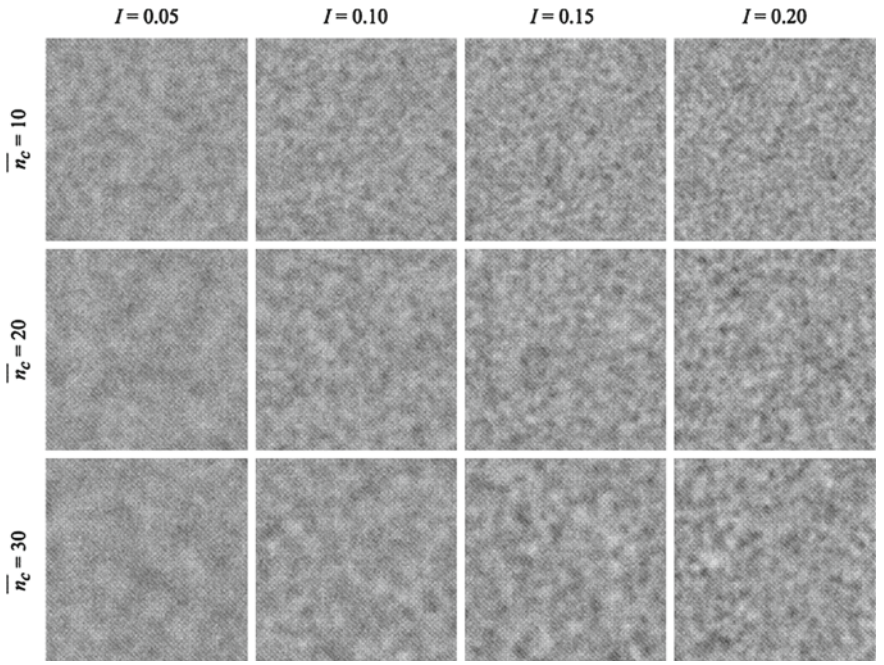


Fig. 13.6 Simulated areal density maps each representing a 4 cm × 4 cm region formed from fibres with length $\lambda = 1$ mm, to a mean coverage of 6 fibres

n_c, is a Poisson distributed random variable. The size of each cluster is determined by an intensity parameter, $0 < I \leq 1$ such that the mean mass per unit area of the cluster is constant and less than the areal density of a fibre. Denoting the length and width of a fibre by λ and ω respectively, the radius of a cluster containing n_c fibre centres is

$$r = \sqrt{\frac{n_c\,\lambda\,\omega}{\pi\,I}}\,. \tag{13.2}$$

Figure 13.6 shows examples of density maps generated by the simulator. We observe textures that increase in 'cloudyness' with n_c and increase in 'graininess' with I.

13.3 Analytic Covariance for Spatial Poisson Processes of Finite Objects

Consider a Poisson process in the plane for finite rectangles of length λ and width $\omega \leq \lambda$, with uniform orientation of rectangle axes to a fixed direction. The covariance or autocorrelation function for such objects is known and given by [7]:

For $0 < r \leq \omega$

$$\alpha_1(r) = 1 - \frac{2}{\pi}\left(\frac{r}{\lambda} + \frac{r}{\omega} - \frac{r^2}{2\omega\lambda}\right). \tag{13.3}$$

For $\omega < r \leq \lambda$

$$\alpha_2(r) = \frac{2}{\pi}\left(\arcsin\left(\frac{\omega}{r}\right) - \frac{\omega}{2\lambda} - \frac{r}{\omega} + \sqrt{\frac{r^2}{\omega^2} - 1}\right). \tag{13.4}$$

For $\lambda < r \leq \sqrt{(\lambda^2 + \omega^2)}$

$$\alpha_3(r) = \frac{2}{\pi}\left(\arcsin\left(\frac{\omega}{r}\right) - \arccos\left(\frac{\lambda}{r}\right) - \frac{\omega}{2\lambda} - \frac{\lambda}{2\omega} - \frac{r^2}{2\lambda\omega} + \sqrt{\frac{r^2}{\lambda^2} - 1} + \sqrt{\frac{r^2}{\omega^2} - 1}\right). \tag{13.5}$$

Then, the coverage c at a point is the number of rectangles overlapping that point, a Poisson variable with grand mean value $\bar{c}$, and the average coverage or density in finite pixels $\tilde{c}$ tends to a Gaussian random variable. For sampling of the process using, say square inspection pixels of side length x, the variance of their density $\tilde{c}(x)$ is

$$Var(\tilde{c}(x)) = Var(c(0)) \int_0^{\sqrt{2}x} \alpha(r, \omega, \lambda)\, b(r)\, dr \tag{13.6}$$

where b is the probability density function for the distance r between two points chosen independently and at random in the given type of pixel; it was derived by Ghosh [13].

Using square pixels of side length x, for $0 \leq r \leq x$

$$b(r, x) = \frac{4r}{x^4}\left(\frac{\pi x^2}{2} - 2rx + \frac{r^2}{2}\right). \tag{13.7}$$

For $x \leq r \leq \sqrt{2}x$

$$b(r, x) = \frac{4r}{x^4}\left(x^2\left(\arcsin\left(\frac{x}{r}\right) - \arccos\left(\frac{x}{r}\right)\right)\right) + \frac{4r}{x^4}\left(2x\sqrt{(r^2 - x^2)} - \frac{1}{2}(r^2 + 2x^2)\right). \tag{13.8}$$

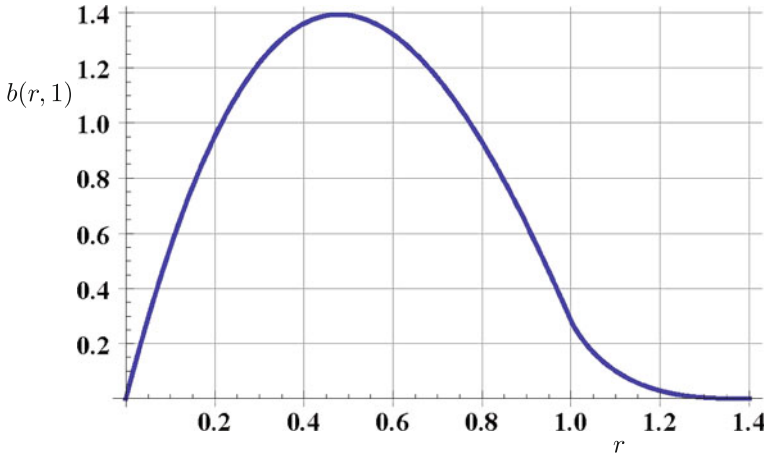


Fig. 13.7 Probability density function $b(r, 1)$ from Eqs. (13.7), (13.8) for the distance r between two points chosen independently and at random in a unit square

A plot of this function is given in Fig. 13.7. Observe that, for vanishingly small pixels, that is points, b degenerates into a delta function on $r = 0$. Ghosh [13] gave also the form of b for other types of pixels; for arbitrary rectangular pixels those expressions can be found in [7]. For small values of r, so $r \ll D$, the formulae for convex pixels of area A and perimeter P all reduce to

$$b(r, A, P) = \frac{2\pi r}{A} - \frac{2Pr^2}{A^2}$$

which would be appropriate to use when the rectangle dimensions ω, λ are small compared with the dimensions of the pixel.

It helps to visualize practical variance computations by considering the case of sampling using large square pixels of side mx say, which themselves consist of exactly m^2 small square pixels of side x. The variance $Var(\tilde{c}(mx))$ is related to $Var(\tilde{c}(x))$ through the covariance $Cov(x, mx)$ of x-pixels in mx-pixels [7]:

$$Var(\tilde{c}(mx)) = \frac{1}{m^2} Var(\tilde{c}(x)) + \frac{m^2 - 1}{m^2} Cov(x, mx).$$

As $m \to \infty$, the small pixels tend towards points, $\frac{1}{m^2} Var(\tilde{c}(x)) \to 0$ so $Var(\tilde{c}(mx))$ admits interpretation as $Cov(0, mx)$, the covariance among points inside mx-pixels, the intra-pixel covariance, precisely $Var(\tilde{c}(mx))$ from Eq. (13.6).

The fractional between pixel variance for x-pixels is

$$\tilde{\rho}(x) = \frac{Cov(0, x)}{Var(c(0))} = \frac{Var(\tilde{c}(x))}{Var(c(0))}$$

which increases monotonically with λ and with ω but decreases monotonically with mx, see Deng and Dodson [6] for more details. In fact, for a Poisson process of rectangles the variance of coverage at points is precisely the mean coverage, $Var(c(0)) = \bar{c}$, so if we agree to measure coverage as a fraction of the mean coverage then Eq. (13.6) reduces to the integral

$$\frac{Var(\tilde{c}(x))}{\bar{c}} = \int_0^{\sqrt{2}x} \alpha(r, \omega, \lambda)\, b(r)\, dr = \tilde{\rho}(x). \tag{13.9}$$

Now, the covariance among points inside mx-pixels, $Cov(0, mx)$, is the expectation of the covariance between pairs of points separated by distance r, taken over the possible values for r in an mx-pixel; that amounts to the integral in Eq. (13.6). By this means we have continuous families of 2×2 covariance matrices for $x \in \mathbb{R}^+$ and $2 < m \in \mathbb{Z}^+$ given by

$$\begin{aligned} \Sigma^{x,m} &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} Var(\tilde{c}(x)) & Cov(x, mx) \\ Cov(x, mx) & Var(\tilde{c}(x)) \end{pmatrix} \\ &= \begin{pmatrix} \tilde{\rho}(x) & \tilde{\rho}(mx) \\ \tilde{\rho}(mx) & \tilde{\rho}(x) \end{pmatrix}. \end{aligned} \tag{13.10}$$

which encodes information about the spatial structure formed from the Poisson process of rectangles, for each choice of rectangle dimensions $\omega \leq \lambda \in \mathbb{R}^+$. This can be extended to include mixtures of different rectangles with given relative abundances and processes of more complex objects such as Poisson clusters of rectangles.

There is a one dimensional version of the above that is discussed in [6, 7], with point autocorrelation calculated easily as

$$\alpha(r) = \begin{cases} 1 - \frac{r}{\lambda} & 0 \leq r \leq \lambda \\ 0 & \lambda < r. \end{cases} \tag{13.11}$$

Also, the probability density function for points chosen independently and at random with separation r in a pixel, which is here an interval of length x, is

$$b(r) = \frac{2}{x}\left(1 - \frac{r}{x}\right) \quad (0 \leq r \leq x). \tag{13.12}$$

Then the integral (13.6) gives the fractional between pixel variance as

$$\tilde{\rho}(x, \lambda) = \begin{cases} 1 - \frac{x}{3\lambda} & 0 \leq x \leq \lambda \\ \frac{\lambda}{x}\left(1 - \frac{\lambda}{3x}\right) & \lambda < x. \end{cases} \tag{13.13}$$

So in the case of a one dimensional stochastic texture from a Poisson process of segments of length λ we have the explicit expression for the covariance matrices in

Eq. (13.10):

$$\Sigma^{x,m}(\lambda) = \begin{pmatrix} \tilde{\rho}(x,\lambda) & \tilde{\rho}(mx,\lambda) \\ \tilde{\rho}(mx,\lambda) & \tilde{\rho}(x,\lambda) \end{pmatrix}. \tag{13.14}$$

In particular, if we take unit length intervals as the base pixels, for the Poisson process of unit length line segments, $x = \lambda = 1$ we obtain

$$\Sigma^{1,m}(1) = \begin{pmatrix} (1-\frac{1}{3}) & \frac{1}{m}\left(1-\frac{1}{3m}\right) \\ \frac{1}{m}\left(1-\frac{1}{3m}\right) & \left(1-\frac{1}{3}\right) \end{pmatrix} \quad \text{for } m = 2, 3, \ldots. \tag{13.15}$$

13.4 Information Distance

Given the family of pixel density distributions, with associated spatial covariance structure among neighbours, we can use the Fisher metric [1] to yield an arc length function on the curved space of parameters which represent mean and covariance matrices. Then the information distance between any two such distributions is given by the length of the shortest curve between them, a geodesic, in this space. The computational difficulty is in finding the length of this shortest curve since it is the infimum over all curves between the given two points. Fortunately, in the cases we need, multivariate Gaussians, this problem has been largely solved analytically by Atkinson and Mitchell [2].

Accordingly, some of our illustrative examples use information geometry of trivariate Gaussian spatial distributions of pixel density with covariances among first and second neighbours to reveal features related to sizes and density of clusters, which could arise in one, two or three dimensions. For isotropic spatial processes, which we consider here, the variables are means over shells of first and second neighbours, respectively. For anisotropic networks the neighbour sets would be split into more new variables to pick up the spatial anisotropy in the available spatial directions.

Other illustrations will use the analytic bivariate covariances given in Sect. 13.3 by Eq. (13.10).

What we know analytically is the geodesic distance between two multivariate Gaussians, A, B, of the same number n of variables in two particular cases [2]:

1. $\mu^{\mathbf{A}} \neq \mu^{\mathbf{B}}$, $\mathbf{\Sigma^A} = \mathbf{\Sigma^B} = \mathbf{\Sigma} : f^A = (n, \mu^A, \Sigma)$, $f^B = (n, \mu^B, \Sigma)$

$$D_\mu(f^A, f^B) = \sqrt{\left(\mu^A - \mu^B\right)^T \cdot \Sigma^{-1} \cdot \left(\mu^A - \mu^B\right)}. \tag{13.16}$$

2. $\mu^{\mathbf{A}} = \mu^{\mathbf{B}} = \mu$, $\mathbf{\Sigma^A} \neq \mathbf{\Sigma^B} : f^A = (n, \mu, \Sigma^A)$, $f^B = (n, \mu, \Sigma^B)$

$$D_\Sigma(f^A, f^B) = \sqrt{\frac{1}{2}\sum_{j=1}^{n} \log^2(\lambda_j)}, \tag{13.17}$$

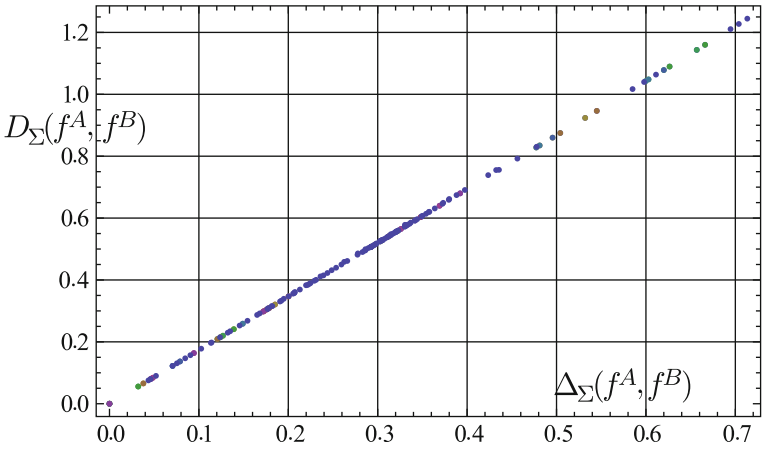


Fig. 13.8 Plot of $D_\Sigma(f^A, f^B)$ from (13.17) against $\Delta_\Sigma(f^A, f^B)$ from (13.18) for 185 different trivariate Gaussian covariance matrices

$$\text{with } \{\lambda_j\} = \text{Eig}(\Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}}).$$

In the present paper we use Eqs. (13.16) and (13.17) and take the simplest choice of a linear combination of both when both mean and covariance are different.

However, from the form of $D_\Sigma(f^A, f^B)$ in (13.17) we deduce that an approximate monotonic relationship arises with a more easily computed symmetrized log-trace function given by

$$\Delta_\Sigma(f^A, f^B) = \sqrt{\log\left(\frac{1}{2n}\left(Tr(\Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}}) + Tr(\Sigma^{B^{-1/2}} \cdot \Sigma^A \cdot \Sigma^{B^{-1/2}})\right)\right)}. \tag{13.18}$$

This is illustrated by the plot in Fig. 13.8 of $D_\Sigma(f^A, f^B)$ from Eq. (13.17) on $\Delta_\Sigma(f^A, f^B)$ from Eq. (13.18) for 185 trivariate Gaussian covariance matrices, where we see that

$$D_\Sigma(f^A, f^B) \approx 1.7\Delta_\Sigma(f^A, f^B).$$

A commonly used approximation for information distance is obtained from the Kullback–Leibler divergence, or relative entropy. Between two multivariate Gaussians $f^A = (n, \mu^A, \Sigma^A)$, $f^B = (n, \mu^B, \Sigma^B)$ with the same number n of variables, its square root gives a separation measurement [16]:

$$KL(f^A, f^B) = \frac{1}{2}\log\left(\frac{\det \Sigma^B}{\det \Sigma^A}\right) + t\frac{1}{2}\mathrm{Tr}[\Sigma^{B^{-1}} \cdot \Sigma^A] + \frac{1}{2}\left(\mu^A - \mu^B\right)^T \cdot \Sigma^{B^{-1}} \cdot \left(\mu^A - \mu^B\right) - \frac{n}{2}. \tag{13.19}$$

This is not symmetric, so to obtain a distance we take the average KL-distance in both directions:

$$D_{KL}(f^A, f^B) = \sqrt{\frac{|KL(f^A, f^B)| + |KL(f^B, f^A)|}{2}} \tag{13.20}$$

The Kullback–Leibler distance tends to the information distance as two distributions become closer together; conversely it becomes less accurate as they move apart.

For comparing relative proximity, $\Delta_\Sigma(f^A, f^B)$ is a better measure near zero than the symmetrized Kullback–Leibler $D_{KL}(f^A, f^B)$ distance in those multivariate Gaussian cases so far tested and may be computationally quicker for handling large batch processes.

13.5 Dimensionality Reduction of Spatial Density Arrays

We shall illustrate the differences of spatial features in given data sets obtained from the distribution of local density for real and simulated planar stochastic fibre networks. In such cases there is benefit in mutual information difference comparisons of samples in the set but the difficulty is often the large number of samples in a set of interest—perhaps a hundred or more. Human brains can do this very well; the enormous numbers of optical sensors that stream information from the eyes into the brain with the result that we have a 3-dimensional reduction which serves to help us 'see' the external environment. We want to see a large data set organised in such a way that natural groupings are revealed and quantitative dispositions among groups are preserved. The problem is how to present the information contained in the whole data set, each sample yielding a 3×3 covariance matrix Σ and mean μ. The optimum presentation is to use a 3-dimensional plot, but the question is what to put on the axes.

To solve this problem we use multi-dimensional scaling, or dimensionality reduction, to extract the three most significant features from the set of samples so that all samples can be displayed graphically in a 3-dimensional plot. The aim is to reveal groupings of data points that correspond to the prominent characteristics; in our context we have different former types, grades and differing scales and intensities of fibre clustering. Such a methodology has particular value in the quality control for processes with applications that frequently have to study large data sets of samples from a trial or through a change in conditions of manufacture or constituents. Moreover, it can reveal anomalous behaviour of a process or unusual deviation in a

product. The raw data of one sample from a study of spatial variability might typically consist of a spatial array of 250×250 pixel density values, so what we solve is a problem in classification for stochastic image textures.

The method, which we introduced in a preliminary report [11], depends on extracting the three largest eigenvalues and their eigenvectors from a matrix of mutual information distances among distributions representing the samples in the data set. The number in the data set is unimportant, except for the computation time in finding eigenvalues. This follows the methods described by Carter et al. [4, 5]. Our study is for datasets of pixel density arrays from complete sampling of density maps of stochastic textures which incorporate spatial covariances. We report the results of such work on a large collection of radiographs from commercial papers made from continuous filtration of cellulose and other fibres, [9].

The series of computational stages is as follows:

1. Obtain mutual 'information distances' $D(i, j)$ among the members of the data set of N textures $X_1, X_2, \ldots, X_N$ using the fitted trivariate Gaussian pixel density distributions.
2. The array of $N \times N$ differences $D(i, j)$ is a real symmetric matrix with zero diagonal. This is centralized by subtracting row and column means and then adding back the grand mean to give $CD(i, j)$.
3. The centralized matrix $CD(i, j)$ is again a real symmetric matrix with zero diagonal. We compute its N eigenvalues $ECD(i)$, which are necessarily real, and the N corresponding N-dimensional eigenvectors $VCD(i)$.
4. Make a 3×3 diagonal matrix A of the first three eigenvalues of largest absolute magnitude and a $3 \times N$ matrix B of the corresponding eigenvectors. The matrix product $A \cdot B$ yields a $3 \times N$ matrix and its transpose is an $N \times 3$ matrix T, which gives us N coordinate values (x_i, y_i, z_i) to embed the N samples in 3-space.

Example: Bivariate Gaussians

$$
\begin{aligned}
f(x, y) &= \frac{1}{2\pi\sqrt{\Delta}} exp \frac{-1}{\Delta^2}(y-\mu_2)^2\sigma_{11} + (x-\mu_1)[(x-\mu_1)\sigma_{22} + 2(-y+\mu_2))\sigma_{12}], \\
\mu &= (\mu_1, \mu_2), \\
\Delta &= Det[\Sigma] = \sigma_{11}\sigma_{22} - \sigma_{12}^2, \\
\Sigma &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \sigma_{11}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \sigma_{12}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \sigma_{22}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \\
\Sigma^{-1} &= \begin{pmatrix} \frac{\sigma_{22}}{\Delta} & -\frac{\sigma_{12}}{\Delta} \\ -\frac{\sigma_{12}}{\Delta} & \frac{\sigma_{11}}{\Delta} \end{pmatrix}.
\end{aligned}
$$

Put $\delta\mu_i = (\mu_i^A - \mu_i^B)$.

Then we have

$$D_\mu(f^A, f^B) = \sqrt{\delta\mu^T \cdot \Sigma^{-1} \cdot \delta\mu}$$
$$= \sqrt{\frac{\delta\mu_2\,(\sigma_{11}\delta\mu_2 - \sigma_{12}\delta\mu_1)}{\Delta} + \frac{\delta\mu_1\,(\sigma_{22}\delta\mu_1 - \sigma_{12}\delta\mu_2)}{\Delta}}.$$

Numerical example:

$$\Sigma^A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma^B = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \quad \Sigma^{B^{-1}} = \begin{pmatrix} 3/7 & -1/7 \\ -1/7 & 3/14 \end{pmatrix}$$

$$\Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix},$$

with eigenvalues: $\lambda_1 = 7$, $\lambda_2 = 2$.

$$D_\Sigma(\Sigma^A, \Sigma^B) = \sqrt{\frac{1}{2}\sum_{j=1}^{n} \log^2(\lambda_j)} \approx 1.46065$$

$$\Delta_\Sigma(\Sigma^A, \Sigma^B) = \sqrt{\log \frac{7+2}{4}} \approx 0.9005.$$

For comparison, the symmetrized Kullback–Leibler distance [16] is given by

$$D_{KL}(\Sigma^A, \Sigma^B) = \frac{1}{2}\left(\sqrt{\frac{1}{2}\log 14 - \frac{19}{28}} + \sqrt{\frac{1}{2}\log\frac{1}{14} + \frac{7}{2}}\right) \approx 1.1386.$$

13.6 Analysis of Samples

13.6.1 Analytic Results for Poisson Processes of Line Segments and Rectangles

We provide here some graphics showing three dimensional embeddings of Poisson processes that yield stochastic textures of pixel density, using the analysis in Sect. 13.3.

Figure 13.9 shows an embedding of 20 samples calculated for a Poisson line process of line segments, (13.15), with $x = \lambda = 1$ and $m = 2, 3, \ldots 21$. The starting green point in the lower right is for $m = 2$ and the red end point is for $m = 21$. Figure 13.10 shows an embedding of 18 samples calculated for a planar Poisson process of unit squares, from (13.10), with $\omega = \lambda = 1$. It shows the separation into two groups of samples: analysed with small base pixels, $x = 0.1$ right, and with large base pixels, $x = 1$ left. Figure 13.11 shows an embedding of 18 samples calculated

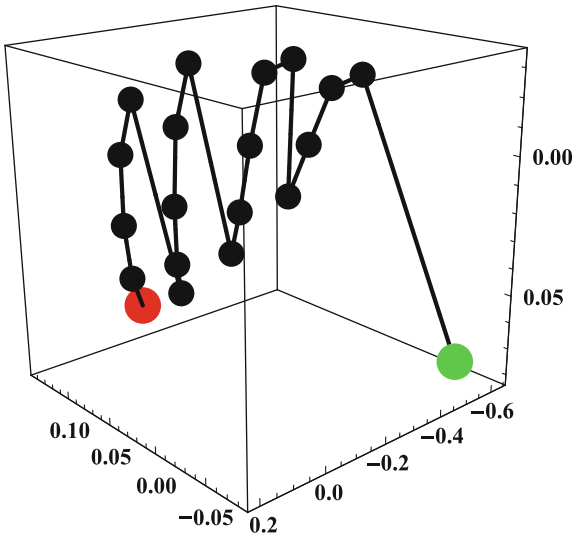


Fig. 13.9 Embedding of 20 evaluations of information distance for the bivariate covariances arising from a Poisson line process of line segments, (13.15), with $x = \lambda = 1$ and $m = 2, 3, \ldots 21$. The starting *green point* in the *lower right* is for $m = 2$ and the *red end point* is for $m = 21$

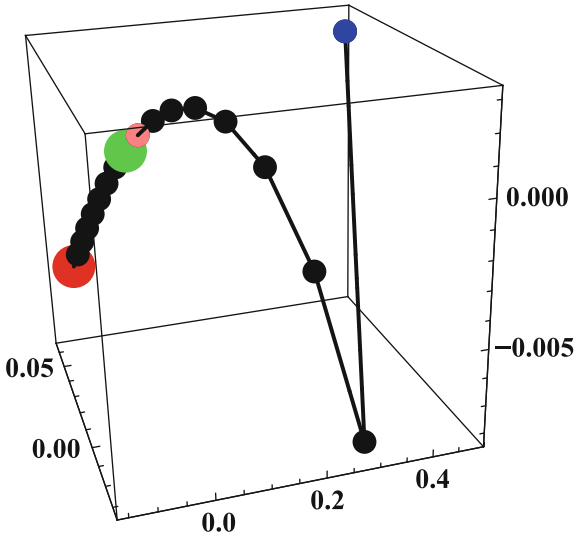


Fig. 13.10 Embedding of 18 evaluations of information distance for the bivariate covariances arising from a planar Poisson process of squares, (13.10), with $\omega = \lambda = 1$. The two groups arise from different schemes of inspection pixels. *Right group* used small base pixels with $x = 0.1$, from *blue* to *pink* $m = 2, 3, \ldots, 10$; *left group* used large base pixels with $x = 1$, from *green* to *red* $m = 2, 3, \ldots, 10$

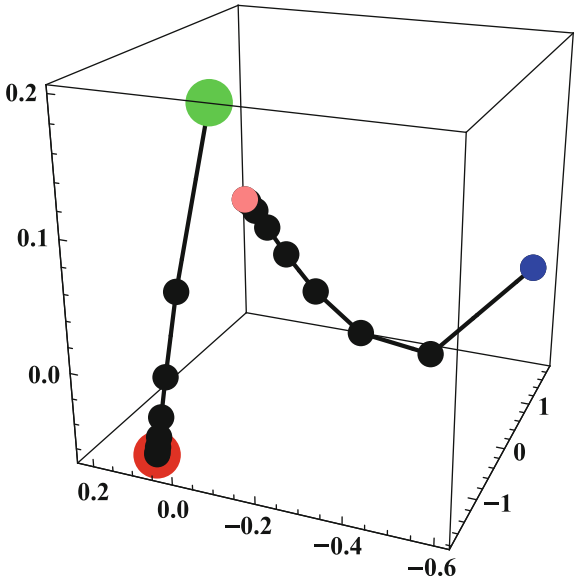


Fig. 13.11 Embedding of 22 evaluations of information distance for the bivariate covariances arising from a planar Poisson process of rectangles, (13.10), with $\omega = 0.2$, $\lambda = 1$. The two groups arise from different schemes of inspection pixels. *Left group* used large base pixels $x = 1$, from *green* to *red* $m = 2, 3, \ldots, 10$; *right group* used small base pixels $x = 0.1$, from *blue* to *pink* $m = 2, 3, \ldots, 10$

for a planar Poisson process of rectangles with aspect ratio 5:1, from (13.10), with $\omega = 0.2$, $\lambda = 1$. Again it shows the separation into two groups of samples analysed with small pixels, right, and with large pixels, left.

13.6.2 Deviations from Poisson Arising from Clustering

Our three spatial variables for each spatial array of data are the mean density in a central pixel, mean of its first neighbours, and mean of its second neighbours. We begin with analysis of a set of 16 samples of areal density maps for simulated stochastic fibre networks made from the same number of 1 mm fibres but with differing scales (clump sizes) and intensities (clump densities) of fibre clustering. Among these is the standard unclustered Poisson fibre network; all samples have the same mean density.

Figure 13.12 gives analyses for spatial arrays of pixel density differences from Poisson networks. It shows a plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface (left), and the same data grouped by numbers of fibres in clusters and cluster densities (right), for geodesic information distances among 16 datasets of 1 mm pixel density differences between a Poisson network and simulated networks made from 1 mm fibres. Each network has the same mean density but with different scales and densities

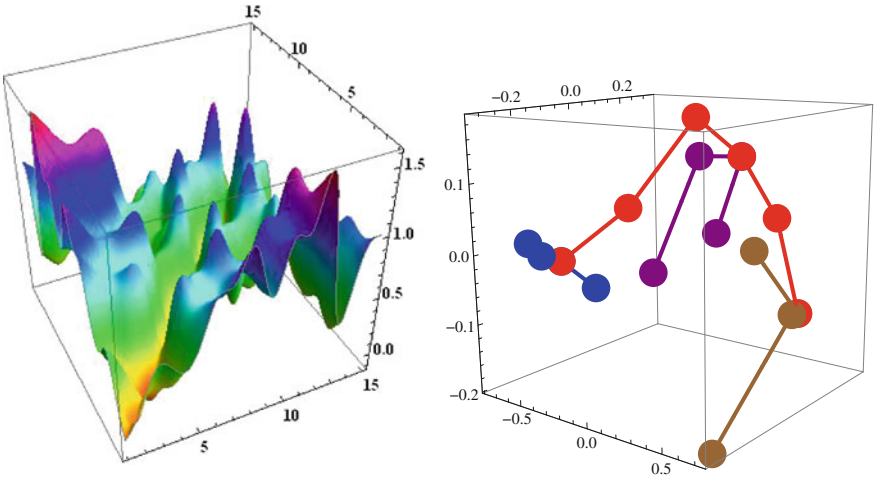


Fig. 13.12 Pixel density differences from Poisson networks. *Left* plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface, for trivariate Gaussian information distances among 16 datasets of 1 mm pixel density differences between a Poisson network and simulated networks made from 1 mm fibres, each network has the same mean density but with different clustering. *Right* embedding of the same data grouped by numbers of fibres in clusters and cluster densities

of clustering; thus the mean difference is zero in this case. Using pixels of the order of fibre length is appropriate for extracting information on the sizes of typical clusters. The embedding reveals the clustering features as orthogonal subgroups.

Next, Fig. 13.13 gives analyses for pixel density arrays of the clustered networks. It shows on the left the plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface (left) for trivariate Gaussian information distances among the 16 datasets of 1 mm pixel densities for simulated networks made from 1 mm fibres, each network with the same mean density but with different clustering. In this case the trivariate Gaussians all have the same mean vectors. Shown on the right is the dimensionality reduction embedding of the same data grouped by numbers of fibres in clusters and cluster densities; the solitary point is a Poisson network of the same fibres.

13.6.3 Effect of Mean Density in Poisson Structures

Figure 13.14 gives analyses for pixel density arrays for Poisson networks of different mean density. It shows the plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface (left), for trivariate Gaussian information distances among 16 simulated Poisson networks made from 1 mm fibres, with different mean density, using pixels at 1 mm scale. Also shown is, (right) dimensionality reduction embedding of the same Poisson network data, showing the effect of mean network density.

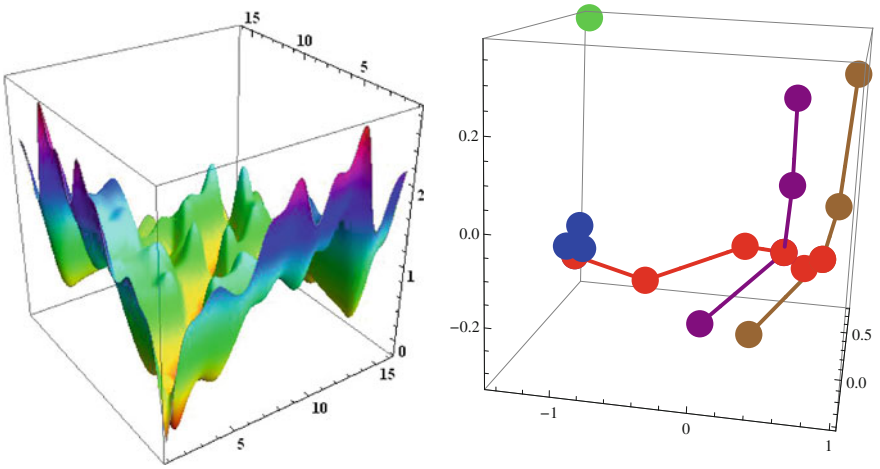


Fig. 13.13 Pixel density arrays for clustered networks: *Left* plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface, for trivariate Gaussian information distances among 16 datasets of 1 mm pixel density arrays for simulated networks made from 1 mm fibres, each network with the same mean density but with different clustering. *Right* embedding of the same data grouped by numbers of fibres in clusters and cluster densities; the solitary point is an unclustered Poisson network

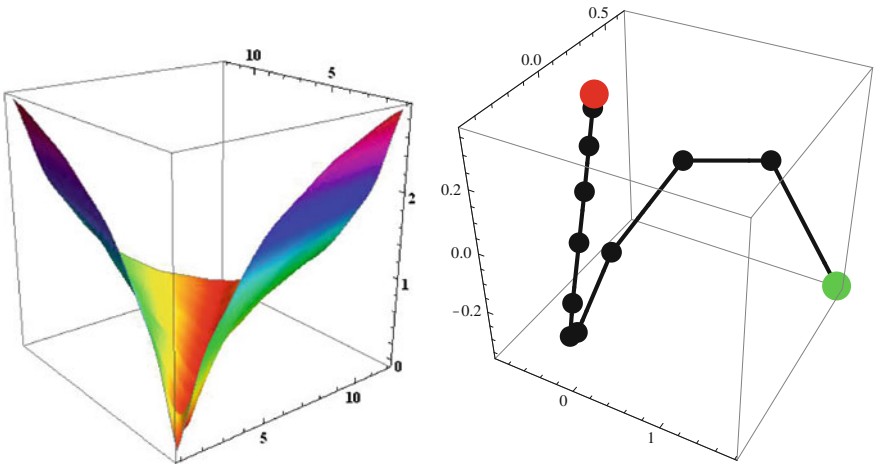


Fig. 13.14 Pixel density arrays for Poisson networks of different mean density. *Left* plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface (*left*), for trivariate Gaussian information distances among 16 simulated Poisson networks made from 1 mm fibres, with different mean density, using pixels at 1 mm scale. *Right* embedding of the same Poisson network data, showing the effect of mean network density

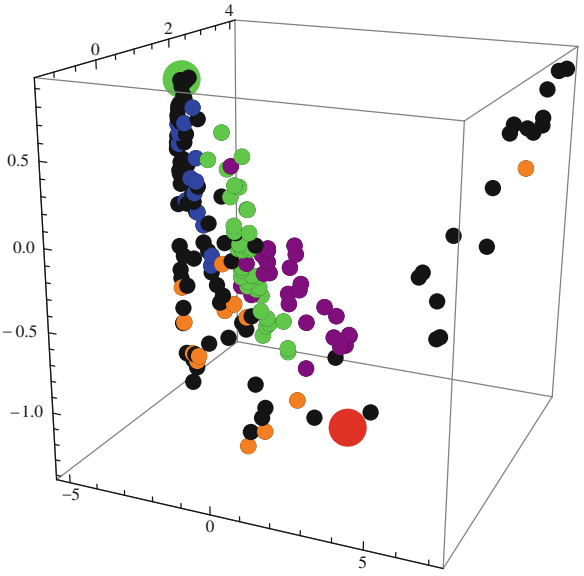


Fig. 13.15 Embedding using 182 trivariate Gaussian distributions for samples from the data set [9]. *Blue points* are from gap formers; *orange* are various handsheets, *purple* are from pilot paper machines and *green* are from hybrid formers. The embedding separates these different forming methods into subgroups

13.6.4 Analysis of Commercial Samples

Figure 13.15 shows a 3-dimensional embedding for a data set from [9] including 182 paper samples from gap formers, handsheets, pilot machine samples and hybrid formers. We see that to differing degrees the embedding separates these different and very disparate forming methods by assembling them into subgroups. This kind of discrimination could be valuable in evaluating trials, comparing different installations of similar formers and for identifying anomalous behaviour.

The benefit from these analyses is the representation of the important structural features of number of fibres per cluster and cluster density, by almost orthogonal subgroups in the embedding.

13.6.5 Analysis of Saccharomyces Cerevisiae Yeast and Human Genomes

This yeast is the genome studied in [3] for which we showed that all 20 amino acids along the protein chains exhibited mutual clustering, and separations of 3–12 are generally favoured between repeated amino acids, perhaps because this is the

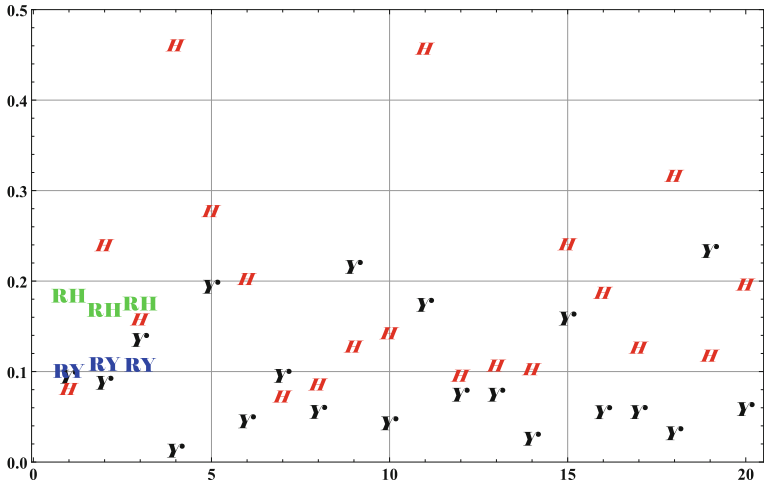


Fig. 13.16 Determinants of 12-variate spatial covariances for 20 samples of yeast amino acid sequences, *black* **Y**, together with three Poisson sequences of 100,000 amino acids with the yeast relative abundances, *blue* **RY**. Also shown are 20 samples of human sequences, *red* **H**, and three Poisson sequences of 100,000 amino acids with the human relative abundances, *green* **RH**.

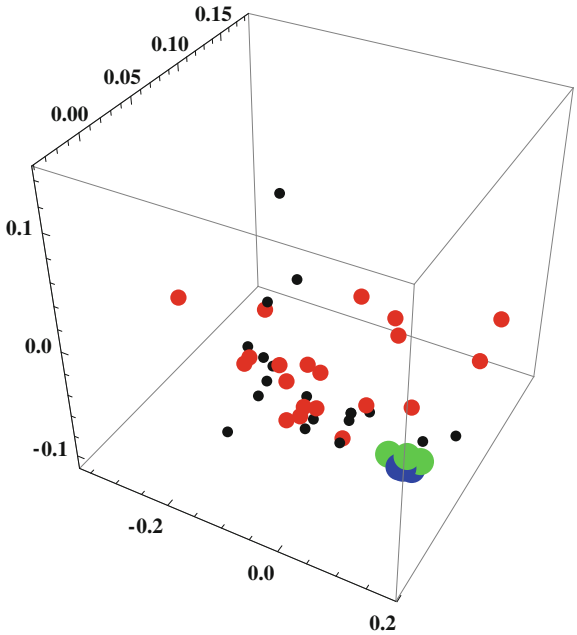


Fig. 13.17 Twelve-variate spatial covariance embeddings for 20 samples of yeast amino acid sequences, *small black points*, together with three Poisson sequences of 100,000 amino acids with the yeast relative abundances, *large blue points*. Also shown are 20 human DNA sequences, *medium red points*, and three Poisson sequences of 100,000 amino acids with the human relative abundances, *large green points*

usual length of secondary structure, cf. also [1]. The database of sample sequences is available on the Saccharomyces Genome Database [19]. Here we mapped the sequences of the 20 amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y onto the 20 grey-level values 0.025, 0.075, . . . , 0.975 so yielding a grey-level barcode for each sequence, Fig. 13.1. Given the usual length of secondary structure to range from 3 to 12 places along a sequence, we used spatial covariances between each pixel and its successive 12 neighbours. Figure 13.16 plots the determinants of the 12-variate spatial covariances of 20 for yeast, black **Y**, together with three Poisson random sequences of 100,000 amino acids with the yeast relative abundances, blue **RY**. Also shown are 20 samples of human sequences, red **H**, and three Poisson sequences of 100,000 amino acids with the human relative abundances, green **RH**. Figure 13.17 shows an embedding of these 20 12-variate spatial covariances for yeast, small black points, together with three Poisson sequences of 100,000 amino acids with the yeast relative abundances, large blue points, and 20 human DNA sequences, medium red points using data from the NCBI Genbank Release 197.0 [15], and three Poisson sequences of 100,000 amino acids with the human relative abundances, large green points. The sequences ranged in length from 340 to 1,900 amino acids. As with the original analysis of recurrence spacings [3] which revealed clustering, the difference of the yeast and human sequence structures from Poisson is evident. However, it is not particularly easy to distinguish yeast from human sequences by this technique, both lie in a convex region with the Poisson sequences just outside, but there is much scatter. Further analyses of genome structures will be reported elsewhere.

References

1. Arwini, K., Dodson, C.T.J.: Information geometry near randomness and near independence. In: Sampson, W.W. (eds.) Stochasic Fibre Networks (Chapter 9), pp. 161–194. Lecture Notes in Mathematics, Springer-Verlag, New York, Berlin (2008)
2. Atkinson, C., Mitchell, A.F.S.: Rao's distance measure. Sankhya: Indian J. Stat. Ser. A **48**(3), 345–365 (1981)
3. Cai, Y., Dodson, C.T.J .Wolkenhauer, O. Doig, A.J.: Gamma distribution analysis of protein sequences shows that amino acids self cluster. J. Theor. Biol. **218**(4), 409–418 (2002)
4. Carter, K.M., Raich, R., Hero, A.O.:. Learning on statistical manifolds for clustering and visualization. In 45th Allerton Conference on Communication, Control, and Computing, Monticello, Illinois. (2007). https://wiki.eecs.umich.edu/global/data/hero/images/c/c6/Kmcarter-learnstatman.pdf
5. Carter, K.M.: Dimensionality reduction on statistical manifolds. Ph.D. thesis, University of Michigan (2009). http://tbayes.eecs.umich.edu/kmcarter/thesis
6. Deng, M., Dodson, C.T.J.: Paper: An Engineered Stochastic Structure. Tappi Press, Atlanta (1994)
7. Dodson, C.T.J.: Spatial variability and the theory of sampling in random fibrous networks. J. Roy. Statist. Soc. B **33**(1), 88–94 (1971)
8. Dodson, C.T.J.: A geometrical representation for departures from randomness of the intergalactic void probablity function. In: Workshop on Statistics of Cosmological Data Sets NATO-ASI Isaac Newton Institute, 8–13 August 1999. http://arxiv.org/abs/0811.4390

9. Dodson, C.T.J., Ng, W.K., Singh, R.R.: Paper: stochastic structure analysis archive. Pulp and Paper Centre, University of Toronto (1995) (3 CDs)
10. Dodson, C.T.J., Sampson, W.W.: In Advances in Pulp and Paper Research, Oxford 2009. In: I'Anson, S.J., (ed.) Transactions of the XIVth Fundamental Research Symposium, pp. 665–691. FRC, Manchester (2009)
11. Dodson, C.T.J., Sampson, W.W.: Dimensionality reduction for classification of stochastic fibre radiographs. In Proceedings of GSI2013—Geometric Science of Information, Paris, 28–30: Lecture Notes in Computer Science 8085. Springer-Verlag, Berlin (August 2013)
12. Doroshkevich, A.G., Tucker, D.L., Oemler, A., Kirshner, R.P., Lin, H., Shectman, S.A., Landy, S.D., Fong, R.: Large- and superlarge-scale structure in the las campanas redshift survey. Mon. Not. R. Astr. Soc. **283**(4), 1281–1310 (1996)
13. Ghosh, B.: Random distances within a rectangle and between two rectangles. Calcutta Math. Soc. **43**(1), 17–24 (1951)
14. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press, London (1980)
15. NCBI Genbank of The National Center for Biotechnology Information. Samples from CCDS_protein. 20130430.faa.gz. ftp://ftp.ncbi.nlm.nih.gov/genbank/README.genbank
16. Nielsen, F., Garcia, V., Nock, R.: Simplifying Gaussian mixture models via entropic quantization. In: Proceedings of 17th European Signal Processing Conference, Glasgow, Scotland 24–28 August 2009, pp. 2012–2016
17. Sampson, W.W.: Modelling Stochastic Fibre Materials with Mathematica. Springer-Verlag, Berlin, New York (2009)
18. Sampson, W.W.: Spatial variability of void structure in thin stochastic fibrous materials. Mod. Sim. Mater. Sci. Eng. 20:015008 pp13 (2012). doi:10.1088/0965-0393/20/1/015008
19. Saccharomyces Cerevisiae Yeast Genome Database. http://downloads.yeastgenome.org/sequence/S288C_reference/orf_protein/

Index

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2,