Cogn Comput (2009) 1:279–291
DOI 10.1007/s12559-009-9026-7

Learning the Fréchet Mean over the Manifold of Symmetric Positive-Definite Matrices

Simone Fiori

Published online: 15 October 2009

Abstract The present manuscript tackles the problem of learning the average of a set of symmetric positive-definite (SPD) matrices. Averages are computed via the notion of Fréchet mean, and the associated metric dispersion is interpreted as the variance of the patterns around the Fréchet mean. Also, the problem of continuous interpolation of two SPD patterns is tackled within the manuscript. The property of volume conservation of the Fréchet mean and of the considered interpolatory scheme for SPD matrices is discussed as well. The paper describes several applications where the technique could be readily exploited, including in machine learning, intelligent control, pattern classification, speech emotion classification and diffusion tensor data analysis in medicine.

Keywords Fréchet mean computation on curved spaces · Continuous interpolation · Learning by optimization on differential manifolds · Symmetric positive-definite matrices

Introduction

A branch of cognitive computation is about the computational foundations of intelligent behavior and about the development of theories and systems pertaining to intelligent behavior. Research on such branch of cognitive computation ranges from theoretical questions in machine learning to intelligent data processing and draws on methods from statistics, artificial intelligence and experimental computer science.

The ensemble statistical features of acquired data and the algorithms to estimate them are of prime importance in intelligent data processing. In particular, computing the mean value of a set of data, like a set of measures of a variable of a physical process, is a widely used technique to smooth out irregularities in the data and to filter out the noise and the measurement errors.

Let us consider a set of real scalar data. The average of a set of real numbers may be computed by the standard arithmetic average, which provides, as a mean value, a real number. If some little extra structure in the data is considered, we may observe a bigger richness of chances to define averages. Consider, for instance, a set of real scalar *positive* data: in this case, at least three kinds of averages are known and used in applications, namely, arithmetic, harmonic and geometric averages. No matter what kind of average is invoked for real positive numbers, they share a common feature: they return an average number which is real and positive. The definition of 'mean value' of a set of data becomes then richer and more complicated with the increasing amount of structure that the data to average possess.

From the above-mentioned examples, we may learn that the definition of mean value of a set of data is not unique nor straightforward and may depend of a series of factors. Every procedure for computing an average, however, should at least meet the following condition: *The returned mean value should be of the same nature of the data that it is computed from.*

As an instructive example, let us consider the problem of averaging over the general real linear group $\mathbb{G}(p)$ of invertible matrices, defined as $\mathbb{G}(p) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^{p\times p} | \det(y) \neq 0\}$. Let

S. Fiori (✉)
Dipartimento di Ingegneria Biomedica, Elettronica e Telecomunicazioni (DiBET), Facoltà di Ingegneria, Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy
e-mail: s.fiori@univpm.it

us suppose a data-set $\mathbb{S} = \{y_1, y_2, \ldots, y_N\}$ of $\mathbb{G}(p)$-matrices is available and an average matrix needs to be computed. It is instructive to note that average matrix may not be the sample arithmetic mean:

$$\frac{1}{N}(y_1 + y_2 + \cdots + y_N),$$

because, in general, a sum of invertible matrices is not necessarily invertible. Likewise, the geometric mean:

$$(y_1 \cdot y_2 \cdots y_N)^{\frac{1}{N}},$$

is not acceptable because it depends on the order of computation of matrix products (the group $\mathbb{G}(p)$ is not Abelian).

The question of defining a procedure to compute averages over curved spaces is intimately connected to a different question, namely, to the definition of a continuous interpolator between two objects on a manifold.

An example of interpolation readily arises in random number generation. Let us suppose that a random number generator is sought for that yields a scalar random variable $x \in \mathbb{R}$ which is neither Gaussian nor Laplacian, but whose distribution may vary with continuity between these limits. Let us imagine the Gaussian limit is specified by the distribution $\frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)$ and that the Laplacian limit is specified by the distribution $\frac{1}{6}\exp\left(-\frac{|x|}{3}\right)$. Then, a possible interpolation between the two distributions would be the mixture:

$$\frac{1-\theta}{2\pi}\exp\left(-\frac{x^2}{2}\right) + \frac{\theta}{6}\exp\left(-\frac{|x|}{3}\right),$$

with $\theta \in [0\,1]$. The above-mentioned function is indeed a probability distribution, because it appears non-negative and integrates to one, which varies with continuity between a Gaussian distribution ($\theta = 0$) and a Laplacian distribution ($\theta = 1$) as the parameter θ varies. Figure 1 illustrates an example of interpolation between the two probability distributions.

The present paper addresses the problem of learning the average of a set of symmetric positive-definite matrices (SPD) as well as their variance. The problem has wide ramifications and potential applications in several areas of artificial intelligence and cognition, some of which are identified below.

Symmetric positive-definite matrices find a wide range of applications in science. For instance:

- *Analysis of deformation* [15, 16]. Spatially organized data of the SPD-matrix type may arise from measurements of strain/stress and deformation in materials science and earth science.
- *Image analysis* [1]. Symmetric positive-definite matrices are now widely used in image analysis, in applications such as segmentation, grouping, motion analysis and texture segmentation.
- *Statistical analysis of diffusion tensor data in medicine* [7]. The tensors yield by diffusion tensor magnetic resonance imaging represent the covariance within a Brownian motion model of water diffusion. Current approaches to statistical analysis of diffusion tensor data take the SPD geometry into account, due to the fact that the space of diffusion tensors does not form a vector space.
- *Automatic and intelligent control* [3]. In robotics, the mass-inertia matrix of a robotic system is SPD and the accuracy of its estimate affects control performance. Likewise, controlling vibrations and precise positions of robotic structures often requires the estimation of the structure's mass-inertia and stiffness matrices. In intelligent control, control decisions are often made based on estimation of covariance matrices.

Also, symmetric positive-definite matrices with unitary trace are termed 'density matrices' and are used in quantum physics (see, e.g., [2]).

Symmetric positive-definite matrices plays an important role in machine intelligence, machine learning and cognition. For instance:

- *Human detection via classification on the space of SPD matrices* [20]. Human detection in still images is considered a challenging example of object detection problems, due to the structure and variable appearance of the human body as well as the variation of illumination. An approach in human detection is based on sequentially applying a classifier to all the possible sub-windows in a given image, where covariance features are used as human descriptors. As classical machine learning techniques to train the classifiers are not adequate, since the covariance matrices lie on a curved manifold, it is necessary to define an approach for clustering data points lying on the space of SPD matrices.
- *Pattern recognition* [13, 19]. Two important subclasses of pattern recognition problems can be reformulated as optimization on the manifold of SPD matrices: Data clustering and template matching. In data clustering, in order to solve the problem related to the use of the Euclidean metric in the data space that favors spherical clusters, a general (e.g., Mahalanobis) metric is used which bases on SPD kernel matrices to be optimized. In template matching, the matching of a pattern with a prototype should be discovered up to, e.g., rotations and dilations, which may be represented by SPD matrices to be estimated.
- *Speech emotion classification* [22]. Speech emotion classification [4] is a research domain in intelligent

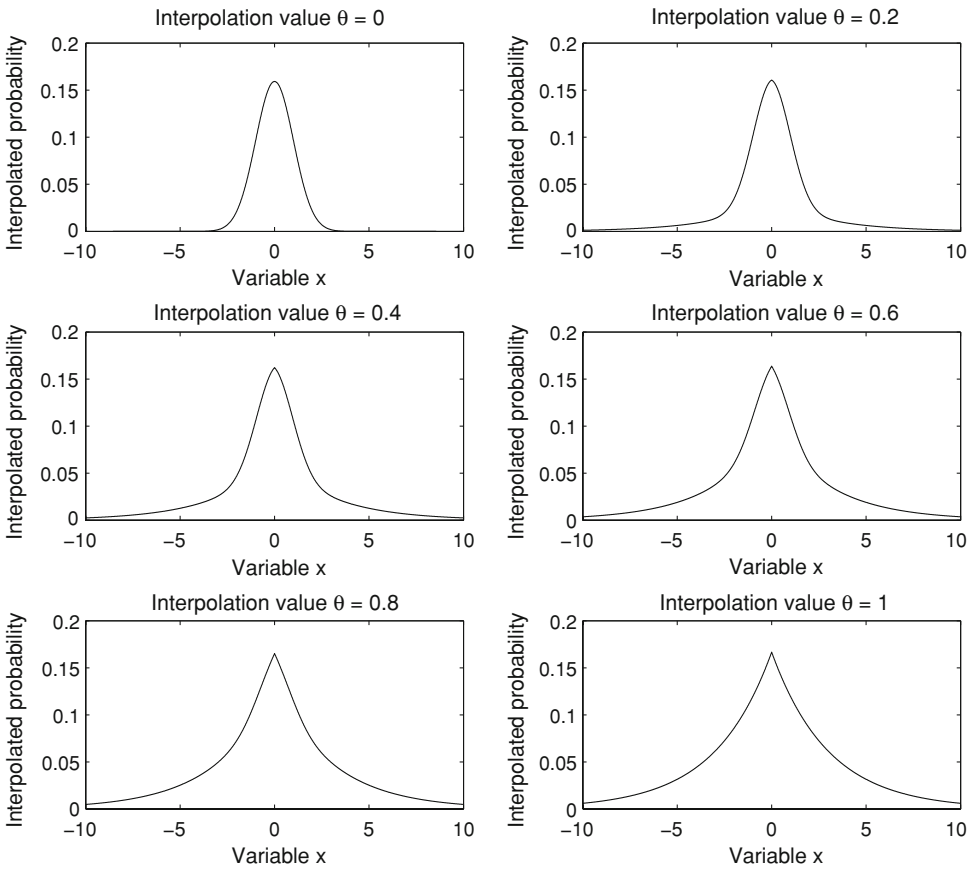


Fig. 1 Example of interpolation between a Gaussian and a Laplacian probability distributions

human–machine communication, with applications to tutoring, customer service, robotics and entertainment. Speech emotion classification may be performed by extracting features from speech signals and by detecting emotions with a classifier. Covariance matrices may be used as feature descriptors.

- *Modeling of cognitive evolution* [14]. Cognition is not directly measurable. It is assessed using psychometric tests which can be regarded as quantitative measures of cognition with error. Proust et al. [14] proposed a model to describe the evolution in continuous time of unobserved cognition in the elderly and assessed the impact of covariates directly on it. The latent cognitive process was defined using a linear mixed model including a Brownian motion and time-dependent covariates.
- *Analysis of the multifactorial nature of cognitive aging* [17]. Research has indicated that there may be age-related and Alzheimer's disease-related reductions in regional cerebral blood flow in the brain. Siedlecki et al. [17] explored differences in age-related and Alzheimer's disease-related reductions in regional cerebral blood flow patterns in the context of cognitive aging using a multivariate approach to the analysis of positron emission tomography data.
- *Modeling of functional brain imaging data* [9]. In neuroimaging studies of human cognitive abilities, brain activation patterns that include regions that are strongly interactive in response to experimental task demands are of particular interest. Existing analysis tools have been highly successful in identifying group differences in regional functional connectivity, including differences associated with states of awareness and normal aging. Habeck et al. [9] addresses the need for a within-group model that identifies patterns of regional functional connectivity that exhibit sustained activity across graduated changes in task parameters (e.g., task difficulty).
- *Design and analysis of wireless cognitive dynamic systems* [10]. A new discipline is emerging, called 'cognitive dynamic systems', which builds on ideas in statistical signal processing, stochastic control and information theory and weaves those well-developed ideas into new ones drawn from neuroscience, statistical learning theory and game theory. The discipline of cognitive dynamic systems will allegedly provide principled tools for the design and development of a new generation of wireless dynamic systems exemplified by cognitive radio and cognitive radar.

The space of symmetric positive-definite matrices may be regarded as a smooth manifold, that is, a curved space, and learning a mean symmetric positive-definite matrix falls in the field of computing mean values on curved spaces. Upon a proper metrization of the space, the concept of mean value may be defined as the Fréchet mean. In particular, we will deal with an intrinsic mean, namely, with an instance of Fréchet mean that does not depend on any embedding of the space of interest into an ambient Euclidean space. The formulation of learning in terms of Fréchet mean requires to solve an optimization problem over a manifold, which will be tackled by first computing the covariant derivative of the Fréchet criterion and then by setting up a numerical optimization algorithm over the space of symmetric positive-definite matrices. The volume-conservation feature of the solution will be studied as well.

Associated to the concept of Fréchet mean is the concept of metric dispersion or Fréchet mean variance, that, likewise covariance for random data that lie on a Euclidean space, measures the dispersion of the data around their mean value. Also, the notion of learning a Fréchet mean leads very naturally to the notion of learning a continuous interpolator between two SPD matrices on a manifold, which will be discussed within the present paper as well, along with its volume-conservation property.

Learning the Fréchet Mean over Riemannian Manifolds and Related Concepts

In the present section, we briefly survey the geometry of smooth manifolds by recalling concepts as tangent spaces, covariant derivatives, geodesic arcs and geodesic distance. For a reference on differential geometry, see e.g., Spivak [18]. We also recall the notion of Fréchet mean on a metrizable space and define the associated metric variance. The notion of interpolation over a smooth manifold will be treated as well.

A Survey of Some Geometrical Concepts

Let us denote the data space of interest as $\mathbb{Y}$, which is supposed to be a Riemannian manifold. Its tangent space at point $y \in \mathbb{Y}$ is denoted as $T_y\mathbb{Y}$.

Given any pair of points $v, w \in T_y\mathbb{Y}$, it is defined a inner product $\langle v, w\rangle_y \in \mathbb{R}$. The specification of a inner product for a Riemannian manifold turns it into a metric space. In fact, the length of a curve $c_{y,v} : [0\,1] \to \mathbb{Y}$, such that $c_{y,v}(0) = y \in \mathbb{Y}$ and $\dot{c}_{y,v}(0) = v \in T_y\mathbb{Y}$, is given by:

$$\ell(c_{y,v}) \stackrel{\text{def}}{=} \int_0^1 \langle \dot{c}_{y,v}(t), \dot{c}_{y,v}(t)\rangle^{\frac{1}{2}}_{c_{y,v}(t)} dt. \tag{1}$$

Given arbitrary $y \in \mathbb{Y}$ and $v \in T_y\mathbb{Y}$, the curve $g_{y,v} : [0\,1] \to \mathbb{Y}$ of shortest length is termed geodesic. Such minimal length, namely $\ell(g_{y,v})$, is termed geodesic distance between endpoints, namely between points $g_{y,v}(0) = y$ and $g_{y,v}(1)$. The Riemannian distance between endpoints is denoted by:

$$d(g_{y,v}(0), g_{y,v}(1)) \stackrel{\text{def}}{=} \ell(g_{y,v}). \tag{2}$$

It is clear that if a manifold $\mathbb{Y}$ is such that any pair of points may be connected by a geodesic arc, then it is possible to measure the distance between any given pair of points on it.

Given a regular function $f : \mathbb{Y} \to \mathbb{R}$, its covariant derivative $\nabla_y f$ in the direction of the vector $v \in T_y\mathbb{Y}$ measures the rate of change of the function f in the direction v. Namely, given any smooth curve $c_{y,v} : [0\,1] \to \mathbb{Y}$, such that $c_{y,v}(0) = y$ and $\dot{c}_{y,v}(0) = v$, the covariant derivative $\nabla_y f$ is the unique vector in $T_y\mathbb{Y}$ such that:

$$\langle v, \nabla_y f\rangle_y = \frac{d}{dt} f(c_{y,v}(t))\bigg|_{t=0}. \tag{3}$$

A common method to optimize on Euclidean spaces, namely gradient steepest descent, may be readily extended to smooth manifolds [21]. To this purpose, let us consider the differential equation on the manifold $\mathbb{Y}$:

$$\dot{y}(t) = -\nabla_{y(t)} f,\ y(0) = \bar{y} \in \mathbb{Y}. \tag{4}$$

In the function $f : \mathbb{Y} \to \mathbb{R}$ is bounded and the smooth manifold $\mathbb{Y}$ is compact, then the solution of such differential equation tends to a local minimum of the function f in $\mathbb{Y}$, depending on the boundary point $\bar{y}$. In fact, by definition of covariant derivative:

$$\frac{d}{dt} f(y(t)) = \langle \dot{y}(t), \nabla_{y(t)} f\rangle_{y(t)} = -\langle \nabla_{y(t)} f, \nabla_{y(t)} f\rangle_{y(t)} \leq 0. \tag{5}$$

The notion of pushforward map or tangent map is of primal importance in the following calculations. Let us consider two smooth manifolds $\mathbb{Y}$ and $\mathbb{U}$ and a smooth map $\varphi : \mathbb{Y} \to \mathbb{U}$. Let $y \in \mathbb{Y}$. A pushforward map:

$$\varphi'_y : T_y\mathbb{Y} \to T_{\varphi(y)}\mathbb{U} \tag{6}$$

is defined such that for every smooth curve $c_{y,v}(t) \in \mathbb{Y}$ with $t \in [-a\,a]$, $a > 0$, $c_{y,v}(0) = y \in \mathbb{Y}$ and $\dot{c}_{y,v}(0) = v \in T_y\mathbb{Y}$, it holds:

$$\varphi'_y(v) \stackrel{\text{def}}{=} \frac{d}{dt} \varphi(c_{y,v}(t))\bigg|_{t=0}. \tag{7}$$

The pushforward map is linear in the argument v.

In the present paper, the smooth manifolds of interest are of matrix type, therefore maps between manifolds are matrix-to-matrix functions. In this case, if the function φ

above is analytic about a matrix-point $y_0 \in \mathbb{Y}$, namely, if it may be expressed as:

$$\varphi(y) = \sum_{k=0}^{\infty} a_k (y - y_0)^k, \tag{8}$$

then the tangent map $\varphi'_y(v)$ in a point $y \in \mathbb{Y}$ applied to the tangent direction $v \in T_y\mathbb{Y}$ may be expressed as:

$$\varphi'_y(v) = \sum_{k=1}^{\infty} a_k \sum_{r=1}^{k} (y - y_0)^{r-1} v (y - y_0)^{k-r}. \tag{9}$$

Below it is recalled the analytic expansion of three matrix-to-matrix maps of particular interest in the context of the present paper:

- Matrix inverse function, $\varphi : \mathbb{G}(p) \to \mathbb{G}(p)$, $\varphi(y) = y^{-1}$. The expansion has center $y_0 = e$ and coefficients $a_k = (-1)^k$.
- Matrix exponential, $\varphi : \mathfrak{gl}(p) \to \mathbb{G}(p)$, $\varphi(y) = \exp(y)$. The expansion has center $y_0 = 0$ and coefficients $a_k = (k!)^{-1}$.
- Matrix principal logarithm $\varphi : \mathbb{B}(p) \to \mathfrak{gl}(p)$, $\varphi(y) = \log(y)$. The expansion has center $y_0 = e$ and coefficients $a_0 = 0$, $a_k = -(-1)^k k^{-1}$ for $k \geq 1$. The subset $\mathbb{B}$ is defined by $\mathbb{B} \stackrel{\text{def}}{=} \{y \in \mathbb{G}(p) | \|y - e\| < 1\}$, for any matrix norm $\|\cdot\|$.

In the expressions mentioned above, symbol $\mathfrak{gl}(p)$ denotes the Lie algebra associated to the Lie group $\mathbb{G}(p)$, and symbol e denotes the identity element of the Lie group $\mathbb{G}(p)$.

Learning a Sample Fréchet Mean and its Associated Variance

In order to define the notion of 'mean value' in a metrizable matrix set $\mathbb{Y}$, we should consider the following requirements to be fundamental:

- The mean value of a set of objects in a space $\mathbb{Y}$ must be of the same nature of the objects that it is compute from, namely, it must belong to the same space $\mathbb{Y}$;
- The notion of mean value of a set of objects in a metrizable space should embody the intuitive understanding that it must locate as close as possible to all the objects. Therefore, a fundamental notion in the definition of mean value is a measure of how far two elements in the space $\mathbb{Y}$ fall apart.

Accordingly, the notion of variance of objects in a metrizable space will be defined in a way that accounts for the amount of dispersion of the objects about the mean value and also depends on how the dissimilarity of such objects is measured.

A way of defining the mean value of a set of objects $y_1, \ldots, y_N \in \mathbb{Y}$ is provided by the notion of Fréchet mean. The Fréchet mean and associated variance [8] may be defined as:

$$\mu \stackrel{\text{def}}{=} \arg\min_{y \in \mathbb{Y}} \frac{1}{N} \sum_{n=1}^{N} d^2(y, y_n), \tag{10}$$

$$\sigma^2 \stackrel{\text{def}}{=} \min_{y \in \mathbb{Y}} \frac{1}{N} \sum_{n=1}^{N} d^2(y, y_n), \tag{11}$$

where operator $d : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}_0^+$ denotes a distance function in the metrizable space $\mathbb{Y}$. There is no warranty, in general, that the optimization problem (10) admits a unique solution. In the case of interest in the present manuscript, however, uniqueness of the solution is ensured.

In the present manuscript, we assume that the space of interest $\mathbb{Y}$ is a Riemannian manifold, therefore, the tools introduced in "A survey of some geometrical concepts" may be taken advantage of in the definition and in the learning of a Fréchet mean and associated variance.

In order to find a minimizer for the criterion $\frac{1}{N}\sum_{n=1}^{N} d^2(y, y_n)$, we may make use of the differential equation (4). The (unique, by hypothesis) minimum is apparently achieved for a value μ such that:

$$\nabla_y \sum_{n=1}^{N} d^2(y, y_n)\Big|_{y=\mu} = 0 \in T_\mu \mathbb{Y}. \tag{12}$$

Once the space of interest $\mathbb{Y}$ and its geometrical features are set, it is necessary to solve the optimization problem (10) by a numerical optimization algorithm that takes the geometry of the space $\mathbb{Y}$ into account. In particular, starting from an initial guess $\mu_0 \in \mathbb{Y}$, it is possible to envisage an iterative learning algorithm that generates a pattern $\mu_s \in \mathbb{Y}$ at any learning step $s \in \mathbb{N}$. Such sequence may be generated by moving from each point $\mu_s \in \mathbb{Y}$ to the next point μ_{s+1} along a short geodesic arc [11] in the opposite direction of the covariant derivative of the criterion function evaluated at point μ_s. Namely, if we set:

$$v_s \stackrel{\text{def}}{=} \frac{1}{2} \nabla_{\mu_s} \sum_{n=1}^{N} d^2(\mu_s, y_n), \tag{13}$$

then the learning algorithm may be expressed as:

$$\mu_{s+1} = g_{\mu_s, -v_s}(t_s), \tag{14}$$

where $t_s \in [0\ 1]$ denotes any suitable learning stepsize schedule that drives the iterative learning algorithm (14) to convergence and $s = 0,\ldots, S$, with the number of iterations S being sufficiently large.

A learning stepsize schedule t_s may be defined as follows. Let us denote by σ_s^2 the variance of the samples around the mean value μ_s at iteration s, namely:

$$\sigma_s^2 = \frac{1}{N}\sum_{n=1}^{N} d^2(\mu_s, y_n). \tag{15}$$

The stepsize t_s should be evaluated in such a way that at step $s+1$ the variance σ_{s+1}^2 is reduced as much as possible with respect to the variance σ_s^2. We may regard the variance σ_{s+1}^2 as a function of the stepsize t_s, in fact:

$$\sigma_{s+1}^2 = \frac{1}{N}\sum_{n=1}^{N} d^2(g_{\mu_s,-v_s}(t_s), y_n). \tag{16}$$

We might thus optimize the value of t_s so that the difference $|\sigma_{s+1}^2 - \sigma_s^2|$ be as large as possible. Clearly, this is a difficult non-linear problem to solve and in practice, it is advisable to choose some sub-optimal approximation. Under the hypothesis that the stepsize value t_s is small enough, we may invoke the expansion of the function σ_{s+1}^2 around the point $t_s = 0$:

$$\sigma_{s+1}^2 \approx \sigma_s^2 + C_{1,s}t_s + \frac{1}{2}C_{2,s}t_s^2, \tag{17}$$

so that the optimal stepsize $t_s^{\star}$ that maximizes the difference $|\sigma_{s+1}^2 - \sigma_s^2|$ is readily found to be:

$$t_s^{\star} \stackrel{\text{def}}{=} -\frac{C_{1,s}}{C_{2,s}}. \tag{18}$$

The coefficients $C_{1,s}, C_{2,s} \in \mathbb{R}$ may be calculated as the first-order and second-order derivatives of the function $\sigma_{s+1}^2(t_s)$ with respect to the parameter t_s in the point $t_s = 0$.

After the mean value $\mu \in \mathbb{Y}$ has been learnt by the algorithm (14), the associated Fréchet variance may be computed as:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} d^2(\mu, y_n). \tag{19}$$

As expected, the variance σ^2 measures the dispersion of the points $y_n \in \mathbb{Y}$ around the mean value $\mu \in \mathbb{Y}$ according to the predefined distance function $d(\cdot, \cdot)$.

Continuous Binary Interpolation

The same considerations made about the definition of a mean value hold true for the definition of the notion of continuous interpolation of two objects in a matrix manifold.

Continuous geometric interpolation of two (or more) objects has the noticeable purpose of filling-in missing data in curved spaces. A continuous interpolation $\mu_\theta \in \mathbb{Y}$ between two points $y_1, y_2 \in \mathbb{Y}$ may be defined and learnt through the optimization problem [12]:

$$\mu_\theta \stackrel{\text{def}}{=} \arg\min_{y\in\mathbb{Y}}[(1-\theta)d^2(y, y_1) + \theta d^2(y, y_2)], \tag{20}$$

with the variable $\theta \in [0\,1]$ providing a parametrization for the interpolation between the two given points.

A Learning Algorithm for Averaging SPD Matrices

We may now proceed to develop an algorithm to learn an average SPD matrix, its associate variance and an interpolation matrix for SPD matrices.

Design of an Averaging Algorithm over the Space $\mathbb{S}^+(p)$

Let us consider the manifold of symmetric positive-definite matrices, defined implicitly by $\mathbb{S}^+(p) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^{p\times p} | y^T = y, y > 0\}$. We recall that a matrix $y \in \mathbb{R}^{p\times p}$ is termed positive definite if for every non-zero vector $\xi \in \mathbb{R}^p$ it holds $\xi^T y \xi > 0$.

The tangent space at a point $y \in \mathbb{S}^+(p)$ is given by $T_y\mathbb{S}^+(p) = \{v \in \mathbb{R}^{p\times p} | v^T = v\}$. The canonical inner product in the space $\mathbb{S}^+(p)$ is defined as:

$$\langle w, v\rangle_y \stackrel{\text{def}}{=} \text{tr}[wy^{-1}vy^{-1}], \tag{21}$$

for every $y \in \mathbb{S}^+(p)$ and $w, v \in T_y\mathbb{S}^+(p)$, where symbol tr[·] denotes matrix-trace. With the above setting, the geodesic $g_{y,v}(t)$ and the associated (squared) Riemannian distance between two points $x, y \in \mathbb{S}^+(p)$ write:

$$g_{y,v}(t) = y^{\frac{1}{2}}\exp(ty^{-\frac{1}{2}}vy^{-\frac{1}{2}})y^{\frac{1}{2}} \tag{22}$$

$$d^2(x, y) = \text{tr}[\log^2(yx^{-1})]. \tag{23}$$

The function to optimize in order to learn the Fréchet mean of a set of SPD matrices y_n is, therefore, $\frac{1}{2}\sum_{n=1}^{N}\text{tr}[\log^2(yy_n^{-1})]$.

A preliminary observation about the matrix-variable at hand is in order. In the following equations, the product $x^{-1}y$ appears often, where $x, y \in \mathbb{S}^+(p)$: It should be noted that, in general, the quantity $x^{-1}y \neq \mathbb{S}^+(p)$, however, it is true that $x^{-1}y \in \mathbb{G}(p)$; also, note that $(x^{-1}y)^T = yx^{-1}$.

With the aim to employ the learning algorithm (13)- (14), we first need to compute the covariant derivative of the criterion to optimize. In the definition of covariant derivative (3), the generic smooth curve $c_{y,v}$ may be replaced with a geodesic. Let us, therefore, consider the following problem: Solve for the covariant derivative $\nabla_y f$ within:

$$\text{tr}[vy^{-1}(\nabla_y f)y^{-1}] = \frac{d}{dt}f(y^{\frac{1}{2}}\exp(ty^{-\frac{1}{2}}vy^{-\frac{1}{2}})y^{\frac{1}{2}})\Big|_{t=0}, \tag{24}$$

$$f(z) \stackrel{\text{def}}{=} \frac{1}{2}\text{tr}\left[\log^2(zx^{-1})\right], z \in \mathbb{S}^+(p), \tag{25}$$

for arbitrary $x, y \in \mathbb{S}^+(p)$ and $v \in T_y\mathbb{S}^+(p)$. Let us set $w \stackrel{\text{def}}{=} y^{-\frac{1}{2}}vy^{-\frac{1}{2}}$ and note, first, that:

$$\frac{d}{dt}f(y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}) = \frac{1}{2}\frac{d}{dt}\mathrm{tr}\left[\log^2(y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}x^{-1})\right]$$
$$= \mathrm{tr}\left[\left(\frac{d}{dt}\log(y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}x^{-1})\right) \times \log(y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}x^{-1})\right].$$

$$\log'_{yx^{-1}}(vx^{-1}) = \sum_{k=1}^{\infty} a_k \sum_{r=1}^{k}(yx^{-1}-e)^{r-1}(vx^{-1})(yx^{-1}-e)^{k-r},$$

with $a_k = -(-1)^k/k$, consequently, by further making use of the analytic expansion of the matrix logarithm and of the properties of the trace operator, we obtain:

$$\frac{d}{dt}f(y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}})\Big|_{t=0} = \mathrm{tr}\left[\sum_{k=1}^{\infty} a_k \sum_{r=1}^{k}(yx^{-1}-e)^{r-1}(vx^{-1})(yx^{-1}-e)^{k-r}\log(yx^{-1})\right]$$
$$= \mathrm{tr}\left[(vx^{-1})\sum_{k=1}^{\infty} a_k \sum_{r=1}^{k}(yx^{-1}-e)^{k-r}\log(yx^{-1})(yx^{-1}-e)^{r-1}\right]$$
$$= \mathrm{tr}\left[(vx^{-1})\sum_{k=1}^{\infty} a_k \sum_{r=1}^{k}\sum_{h=1}^{\infty} a_h(yx^{-1}-e)^{k-r}(yx^{-1}-e)^{h}(yx^{-1}-e)^{r-1}\right]$$
$$= \mathrm{tr}\left[(vx^{-1})\sum_{k=1}^{\infty}(ka_k)\sum_{h=1}^{\infty} a_h(yx^{-1}-e)^{k+h-1}\right]$$
$$= \mathrm{tr}\left[(vx^{-1})\left(\sum_{k=1}^{\infty}(-1)^{k+1}(yx^{-1}-e)^{k-1}\right)\left(\sum_{h=1}^{\infty} a_h(yx^{-1}-e)^{h}\right)\right].$$

Thanks to the concept of pushforward map, we may compute the inner derivative in the last line as follows:

$$\frac{d}{dt}\log(y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}x^{-1}) = \log'_{y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}x^{-1}} \times \left(y^{\frac{1}{2}}\left(\frac{d}{dt}\exp(tw)\right)y^{\frac{1}{2}}x^{-1}\right)$$
$$= \log'_{y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}x^{-1}}(y^{\frac{1}{2}}\exp'_{tw}(w)y^{\frac{1}{2}}x^{-1}).$$

Now, by setting $t = 0$, we obtain:

$$\frac{d}{dt}f\left(y^{\frac{1}{2}}\exp(tw)y^{\frac{1}{2}}\right)\Big|_{t=0} = \mathrm{tr}\left[\left(\log'_{y^{\frac{1}{2}}y^{\frac{1}{2}}x^{-1}}\left(y^{\frac{1}{2}}\exp'_0(w)y^{\frac{1}{2}}x^{-1}\right)\right) \times \log\left(y^{\frac{1}{2}}y^{\frac{1}{2}}x^{-1}\right)\right]$$
$$= \mathrm{tr}\left[\left(\log'_{yx^{-1}}\left(y^{\frac{1}{2}}wy^{\frac{1}{2}}x^{-1}\right)\right)\log\left(yx^{-1}\right)\right]$$
$$= \mathrm{tr}\left[\left(\log'_{yx^{-1}}\left(y^{\frac{1}{2}}y^{-\frac{1}{2}}vy^{-\frac{1}{2}}y^{\frac{1}{2}}x^{-1}\right)\right) \times \log\left(yx^{-1}\right)\right]$$
$$= \mathrm{tr}\left[\log'_{yx^{-1}}(vx^{-1})\log(yx^{-1})\right].$$

From the expansion formula (9), it is immediate to verify that:

According to the expansions recalled in "A survey of some geometrical concepts", the first infinite sum is equal to $(yx^{-1})^{-1}$, while the second sum is equal to $\log(yx^{-1})$. The covariant derivative $\nabla_y f$, as a solution of the problem (24–25), should thus satisfy equation:

$$\mathrm{tr}\left[vy^{-1}(\nabla_y f)y^{-1}\right] = \mathrm{tr}\left[(vx^{-1})(yx^{-1})^{-1}\log(yx^{-1})\right], \quad (26)$$

which readily leads to the result:

$$\nabla_y\left(\frac{1}{2}\mathrm{tr}\left[\log^2(yx^{-1})\right]\right) = y\log(x^{-1}y). \quad (27)$$

The above-mentioned result may be easily made use of in order to compute the covariant derivative of the total variance in the expression of the Fréchet mean in (10), namely:

$$\nabla_y\left(\frac{1}{2}\sum_{n=1}^{N}\mathrm{tr}\left[\log^2(yy_n^{-1})\right]\right) = y\sum_{n=1}^{N}\log(y_n^{-1}y). \quad (28)$$

As the total variance is a sum of convex functions, its minimum is achieved for a value of $y \in \mathbb{S}^+(p)$ that make its covariant derivative vanish to zero.

In the present case, the optimization algorithm (14) becomes:

$$\begin{aligned}\mu_{s+1} &= \mu_s^{\frac{1}{2}} \exp\left(-t_s \mu_s^{-\frac{1}{2}} \left(\mu_s \sum_{n=1}^{N} \log\left(y_n^{-1}\mu_s\right)\right)\mu_s^{-\frac{1}{2}}\right)\mu_s^{\frac{1}{2}} \\ &= \mu_s^{\frac{1}{2}} \exp\left(-t_s \mu_s^{\frac{1}{2}} \left(\sum_{n=1}^{N} \log\left(y_n^{-1}\mu_s\right)\right)\mu_s^{-\frac{1}{2}}\right)\mu_s^{\frac{1}{2}} \\ &= \mu_s \exp\left(-t_s \sum_{n=1}^{N} \log\left(y_n^{-1}\mu_s\right)\right), \\ &= \mu_s \exp\left(t_s \sum_{n=1}^{N} \log\left(\mu_s^{-1} y_n\right)\right), \end{aligned} \tag{29}$$

because it holds $\exp(a^{-1}ba) = a^{-1}\exp(b)a$, whenever matrices a and b are such that the above expressions make sense.

Once a suitably accurate mean special-orthogonal-matrix-type connection-pattern $\mu \in \mathbb{S}^+(p)$ has been computed, its associated variance may be calculated by:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} \mathrm{tr}\left[\log^2(\mu^{-1} y_n)\right]. \tag{30}$$

Optimal Stepsize Schedule for the Learning Algorithm on $\mathbb{S}^+(p)$

With the aim to compute the optimal learning stepsize schedule for the averaging algorithm (29), it is necessary to compute the first-order and second-order derivative of the variance (16) with respect to the parameter t_s in the point $t_s = 0$.

In order to accomplish such calculation, let us rewrite the iterative learning algorithm (29) as follows:

$$\mu_{s+1} = \mu_s \exp(t_s w_s), \tag{31}$$

$$w_s \stackrel{\mathrm{def}}{=} -\sum_{n=1}^{N} \log\left(y_n^{-1}\mu_s\right). \tag{32}$$

According to the expression (16), the variance $\sigma_{s+1}^2(t_s)$ may thus be customized as:

$$\sigma_{s+1}^2(t_s) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{tr}\left[\log^2\left(y_n^{-1}\mu_s \exp(t_s w_s)\right)\right]. \tag{33}$$

The calculation of the derivatives of the function $\sigma_{s+1}^2(t_s)$ may thus be performed by studying aside the function:

$$F(t) \stackrel{\mathrm{def}}{=} \frac{1}{2}\mathrm{tr}\left[\log^2(m\exp(tw))\right], \tag{34}$$

with $m \in \mathbb{G}(p)$, $t \in [0\ 1]$ and $\exp(tw) \in \mathbb{S}^+(p)$. For the first-order derivative, it holds:

$$\begin{aligned}\frac{dF}{dt} &= \mathrm{tr}\left[\left(\frac{d}{dt}\log(m\exp(tw))\right)\log(m\exp(tw))\right] \\ &= \mathrm{tr}\left[\log'_{m\exp(tw)}\left(m\frac{d}{dt}\exp(tw)\right)\log(m\exp(tw))\right] \\ &= \mathrm{tr}\left[\log'_{m\exp(tw)}\left(m\exp'_{tw}(w)\right)\log(m\exp(tw))\right].\end{aligned}$$

Now, by plugging the series expansion of the pushforward map $\log'$ into the last line above, we obtain:

$$\begin{aligned}&\mathrm{tr}\left[\log'_{m\exp(tw)}\left(m\exp'_{tw}(w)\right)\log(m\exp(tw))\right] \\ &\quad = \mathrm{tr}\left[\log(m\exp(tw))(m\exp(tw))^{-1}(m\exp'_{tw}(w))\right],\end{aligned}$$

where, by the series expansion of the pushforward map $\exp'$:

$$\exp'_{tw}(w) = \exp(tw)w,$$

therefore, the first-order derivative of function (34) assumes the expression:

$$\dot{F}(t) = \mathrm{tr}[w\log(m\exp(tw))]. \tag{35}$$

The first-order derivative $\dot{F}(t)$ in the point $t = 0$ has thus the value:

$$\dot{F}(0) = \mathrm{tr}[w\log m]. \tag{36}$$

On the basis of the first-order derivative (35), the second-order derivative $\ddot{F}$ is readily found, in fact:

$$\frac{d^2F}{dt^2} = \mathrm{tr}\left[w\log'_{m\exp(tw)}(m\exp'_{tw}(w))\right],$$

which, in the point $t = 0$, assumes the value:

$$\ddot{F}(0) = \mathrm{tr}[w\log'_m(mw)]. \tag{37}$$

Under the hypothesis that the mean value μ_s is close enough to all available SPD matrices y_n, we may consider $m \approx e$ in the above formulas, which allows to come to an approximation of the optimal learning stepsize as a constant value. First, note that:

$$\log'_m(mw) = (m-e)\log'_m(w) + \log'_m(w),$$

and that a first-order approximation of $\log'_m(w)$ is w. As a consequence, we may assume that the following approximation is acceptable:

$$\mathrm{tr}[w\log'_m(mw)] \approx \mathrm{tr}[w^2]. \tag{38}$$

On the basis of the above findings, we may compute the following coefficients for the expansion of the variance $\sigma_{s+1}^2(t_s)$ (16):

$$\begin{aligned}C_{1,s} &= \frac{d}{dt_s}\frac{1}{N}\sum_{n=1}^{N} d^2(g_{\mu_s,-v_s}(t_s), y_n)\Big|_{t_s=0} \\ &= -\frac{2}{N}\mathrm{tr}\left[\left(\sum_{n=1}^{N}\log\left(y_n^{-1}\mu_s\right)\right)^2\right],\end{aligned} \tag{39}$$

$$\begin{aligned}C_{2,s} &= \frac{d^2}{dt_s^2}\frac{1}{N}\sum_{n=1}^{N} d^2(g_{\mu_s,-v_s}(t_s), y_n)\Big|_{t_s=0} \\ &\approx 2\mathrm{tr}\left[\left(\sum_{n=1}^{N}\log(y_n^{-1}\mu_s)\right)^2\right],\end{aligned} \tag{40}$$

therefore, it holds $C_{1,s} \approx -C_{2,s}/N$. The optimal stepsize schedule has then the approximate constant value:

$$t_s^{\star} \approx \frac{1}{N}. \tag{41}$$

The iteration algorithm (29) with stepsize (41) may be rewritten explicitly as:

$$\mu_{s+1} = \mu_s \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log\left(\mu_s^{-1} y_n\right)\right). \tag{42}$$

When computing the average of real-world data, it is customary to normalize the data in a way that brings all data points in the vicinity of some particular point of the space. If we make the hypothesis that all data points $y_1, \ldots, y_N \in \mathbb{S}^+(p)$ belong to a neighborhood of the identity $e \in \mathbb{S}^+(p)$, we may start the iteration of the algorithm (29) with $\mu_0 = e$. In this case, it is possible to calculate explicitly the first approximate solution μ_1, along with the associate variance σ_1^2, that write, respectively:

$$\mu_1 = \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log y_n\right), \tag{43}$$

$$\sigma_1^2 = \frac{1}{N}\sum_{n=1}^{N} \mathrm{tr}\left[(\log y_n)^2\right]. \tag{44}$$

The above-mentioned quantities play an important role in the theory of averaging of SPD matrices. As a matter of fact, in some papers they are used as mean value and associated variance, respectively, of a set of SPD matrices [1].

On the Feature of Volume Conservation

The iteration algorithm (29) endowed with the stepsize (41) enjoys an interesting property that we refer to as 'volume conservation'. Namely, at each iteration the determinant of the mean value μ_s equals the geometric average of the determinants of data points y_n. Volume conservation is particularly important in applications such as diffusion magnetic resonance imaging [1, 7].

In order to prove volume conservation, let us recall two properties of determinant operator det, namely, for each $a, b \in \mathbb{G}(p)$, it holds:

$$\det(ab) = \det(a)\det(b), \tag{45}$$

$$\det(\exp(a)) = \exp(\mathrm{tr}(a)), \tag{46}$$

$$\det(a^{-1}) = (\det(a))^{-1}. \tag{47}$$

By applying the determinant operator to both sides of equation (42), we obtain:

$$\begin{aligned}
\det(\mu_{s+1}) &= \det(\mu_s)\exp\left(\frac{1}{N}\sum_{n=1}^{N}\mathrm{tr}[\log(\mu_s^{-1}y_n)]\right)\\
&= \det(\mu_s)\prod_{n=1}^{N}\exp\left(\frac{1}{N}\mathrm{tr}[\log(\mu_s^{-1}y_n)]\right)\\
&= \det(\mu_s)\prod_{n=1}^{N}\exp\left(\mathrm{tr}[\log(\mu_s^{-1}y_n)^{\frac{1}{N}}]\right)\\
&= \det(\mu_s)\prod_{n=1}^{N}\det\exp\left(\log(\mu_s^{-1}y_n)^{\frac{1}{N}}\right)\\
&= \det(\mu_s)\prod_{n=1}^{N}\left(\frac{\det y_n}{\det \mu_s}\right)^{\frac{1}{N}}.
\end{aligned}$$

Therefore, the following property holds true:

$$\det(\mu_s) = \left(\prod_{n=1}^{N}\det y_n\right)^{\frac{1}{N}},$$

for every $s = 1,\ldots, S$. It obviously holds true for the final average matrix $\mu \in \mathbb{S}^+(p)$. An important consequence of this property is that whenever the SPD matrices to average come with approximately the same determinant values, then it holds:

$$\det(\mu) \approx \det(y_1) \approx \cdots \approx \det(y_N), \tag{48}$$

namely, the averaging algorithm does not cause any inflation or deflation of volumes and truly obeys to the basic principle that the computed mean value is of the same nature of the data that it is computed from.

Interpolation over the Space of SPD Matrices

In the present section, we deal with the problem of learning a matrix that interpolates two SPD matrices and study its volume-conservation feature.

Let us assume two points $y_1, y_2 \in \mathbb{S}^+(p)$ are given and that we are looking for a geometrically sound interpolation of such two points in $\mathbb{S}^+(p)$. The solution to such problem may be looked for in the sense of the optimization problem (20). Let us define the criterion $C : \mathbb{S}^+(p) \to \mathbb{R}_0^+$ as:

$$C(y) = (1-\theta)\mathrm{tr}[\log^2(yy_1^{-1})] + \theta\mathrm{tr}[\log^2(yy_2^{-1})], \tag{49}$$

with $\theta \in [0\,1]$ assigned.

As the criterion $C(\cdot)$ is a convex combination of two convex function, it is convex as well. Its minimum may be thus determined by setting to zero its covariant derivative, which has the expression:

$$\nabla_y C = (1-\theta)y\log(y_1^{-1}y) + \theta y\log(y_2^{-1}y). \tag{50}$$

Setting the above-mentioned covariant derivative to zero leads to the expression for the interpolation matrix:

$$\mu_\theta = y_1 \exp(\theta \log(y_1^{-1} y_2)). \tag{51}$$

Note that $\mu_0 = y_1$ and $\mu_1 = y_2$, coherently with the optimization principle (20).

Let us now investigate the volume-conservation property of the interpolating scheme (51). By taking the determinant of both sides of equation (51) and by using the already-mentioned properties of the determinant operator, it is straightforward to show that:

$$\det(\mu_\theta) = (\det(y_1))^{1-\theta} (\det(y_2))^{\theta}. \tag{52}$$

Whenever the two SPD matrices to interpolate come with approximately the same determinant values, it holds:

$$\det(\mu_\theta) \approx \det(y_1) \approx \det(y_2), \tag{53}$$

for every $\theta \in [0\,1]$. We should, therefore, conclude that the interpolatory scheme (51) does not cause any inflation or deflation of data volumes for equal-volume data.

Examples and Numerical Experiments

In the present section, the behavior of the proposed algorithms is illustrated via examples and numerical experiments.

Numerical Tests on Learning Averages over the Space of SPD Matrices

In order to test numerically the learning algorithm (42), we may generate N random matrices $y_n \in \mathbb{S}^+(p)$ by exploiting geodesic arcs departing from the identity $e \in \mathbb{S}^+(p)$:

$$y_n = g_{e,v_n}(\zeta_n), \tag{54}$$

where v_n are random symmetric matrices, $g_{\cdot,\cdot}(\cdot)$ denotes the geodesic form (22) and $\zeta_n \in [0\,1]$ are randomly generated with uniform distribution. Symmetric matrices v_n may be generated by the rule $v_n \stackrel{\text{def}}{=} \frac{1}{2}(u_n + u_n^T)$, with matrices $u_n \in \mathbb{R}^{p\times p}$ being random with each entry being a normal random variable. Note that geodesic arcs departing from the identity have the simple expression $g_{e,v}(t) = \exp(tv)$ for $v \in T_e\mathbb{S}^+(p)$ and $t \in [0\ 1]$.

Figure 2 displays the result of a single run obtained with $N = 99$ patterns and $p = 10$. The number of iterations of the algorithm was set to $S = 10$.

The initial guess μ_0 was chosen randomly in $\mathbb{S}^+(10)$.

As it is readily seen from the top panel of Fig. 2, the learning algorithm converges steadily. Also, convergence was achieved in a few iterations. The middle panel of Fig. 2 shows that the distance $d(e, \mu)$ is almost zero. The bottom panel of Fig. 2 also confirms that the determinant

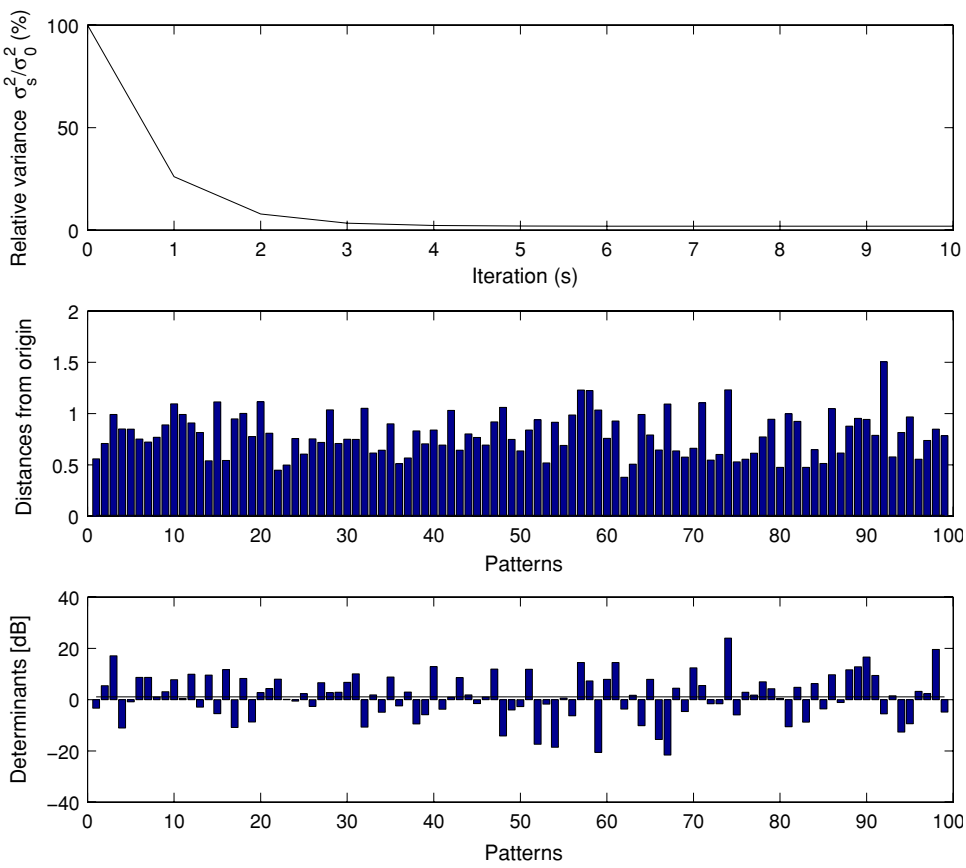


Fig. 2 Experiment on averaging over the space $\mathbb{S}^+(5)$. *Top panel* relative variance σ_s^2/σ_0^2 versus iteration index s. *Middle panel* distances of the patterns y_n to the identity e (*bars*) compared to the distance of the computed mean matrix μ to the identity (*horizontal solid line*). *Bottom panel* determinants of the patterns y_n (*bars*) compared with the determinant of the computed mean value μ (*horizontal solid line*)

$\det(\mu)$ tends to align to most of the similar determinants $\det(y_n)$.

In order to get a better insight of the numerical behavior of the averaging method (42), we may consider the case $p = 3$. In this case, the matrices y_n and μ possess three positive eigenvalues that can be represented by points of the space $\mathbb{R}^3$. The algorithm was tested with $N = 49$ points, $S = 10$ iterations. The initial guess μ_0 was chosen randomly in $\mathbb{S}^+(3)$. First, it is necessary to make sure the learning algorithm achieved convergence, as can be seen from the Fig. 3.

Now, we may consider this representation in terms of eigenvalues, which is used here as a graphical representation of the behavior of the averaging algorithm only. The closer the eigenvalues of the mean matrix μ look to the eigenvalues of the identity, the better the averaging is. Figure 4 illustrates the obtained result.

The numerical test confirms that the eigenvalue-coordinates of the point μ is closer to those of the identity than to those of most of the points y_n.

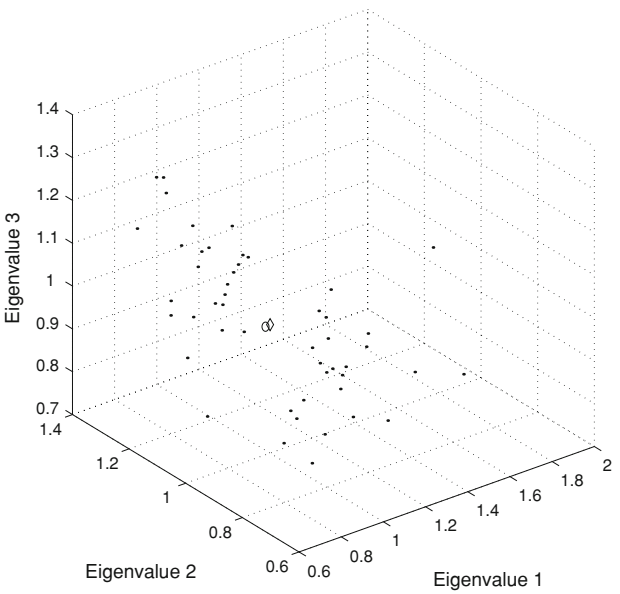


Fig. 4 Experiment on averaging over the space $\mathbb{S}^+(3)$. The *open circle* represents the computed average matrix μ, the *square* denotes the identity matrix, while *dots* represent the matrices y_n, all in terms of matrices' eigenvalues

Numerical Tests on Interpolating Two SPD Matrices

A test on the interpolation of two SPD matrices was conducted as well in order to gain insights into the behavior of the rule (51). The two matrices $y_1, y_2 \in \mathbb{S}^+(p)$ to average may again be generated via the rule (54).

Again, if we choose $p = 3$, it is possible to give a graphical representation of the matrices at hand. For example, each column of a $\mathbb{S}^+(3)$ matrix may be

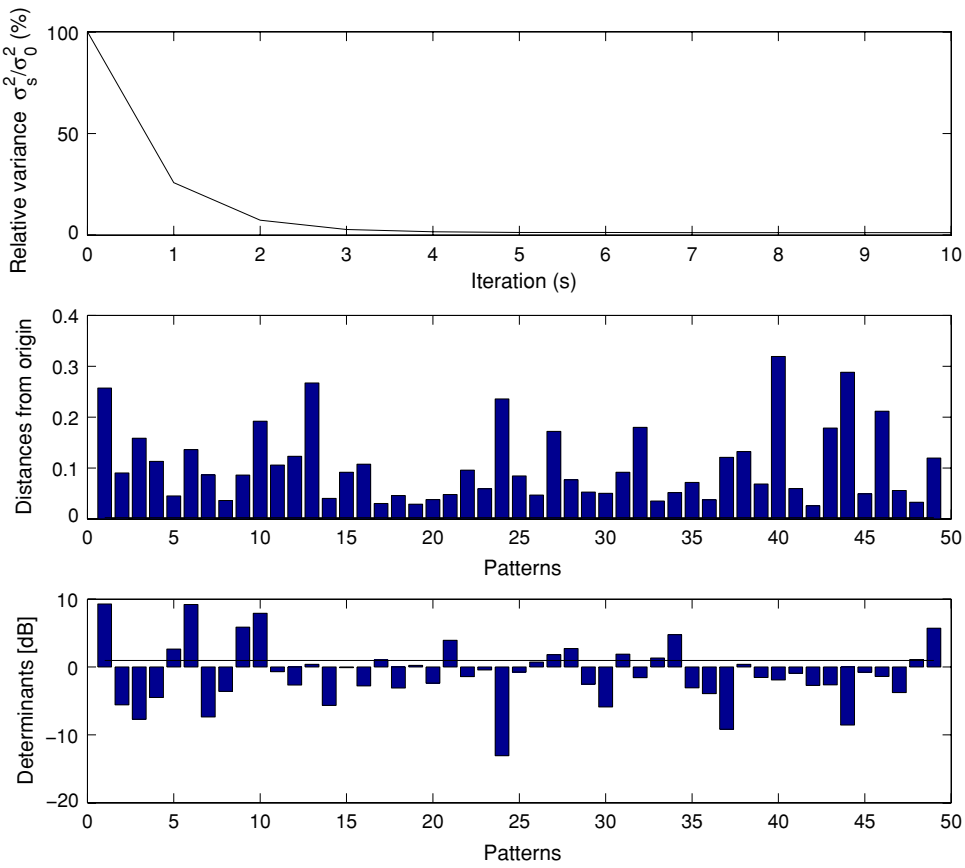


Fig. 3 Test on averaging over the space $\mathbb{S}^+(3)$. *Top panel* relative variance σ_s^2/σ_0^2 versus iteration index s. *Middle panel* distances of the patterns y_n to the identity e (*bars*) compared to the distance of the computed mean matrix μ to the identity (*horizontal solid line*). *Bottom panel* determinants of the patterns y_n (*bars*) compared with the determinant of the computed mean value μ (*horizontal solid line*)

represented as a vector of proportional length, so that a $\mathbb{S}^+(3)$ matrix may be represented by a frame of three (non-orthogonal, non-unitary) three-dimensional vectors. Figure 5 illustrates the result of interpolation of two matrices. As it is easily observed, the interpolation rule (51) provides a set of learnt matrices that vary with continuity between the templates y_1 and y_2.

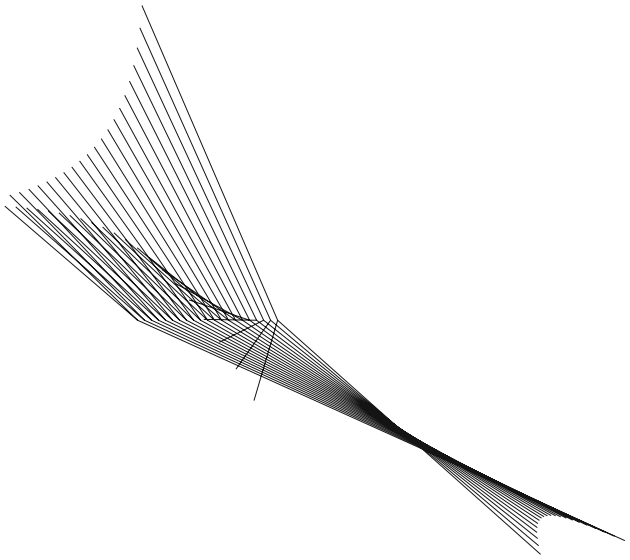


Fig. 5 Experiment on learning interpolates over the space $\mathbb{S}^+(3)$

Also, Fig. 6 illustrates each of the nine entries of the matrix μ_θ, denoted as $\mu_\theta^{(r,c)}$, versus the variable θ. Of course, Fig. 6 is redundant because the nine panel are symmetric.

To conclude, Fig. 7 illustrates the function $\det(\mu_\theta)$ versus the variable θ. As it is readily seen, the function $\theta \mapsto \mu_\theta$ provides a matrix that varies with continuity in the interval $[\min\{\det(y_1), \det(y_2)\} \quad \max\{\det(y_1), \det(y_2)\}]$.

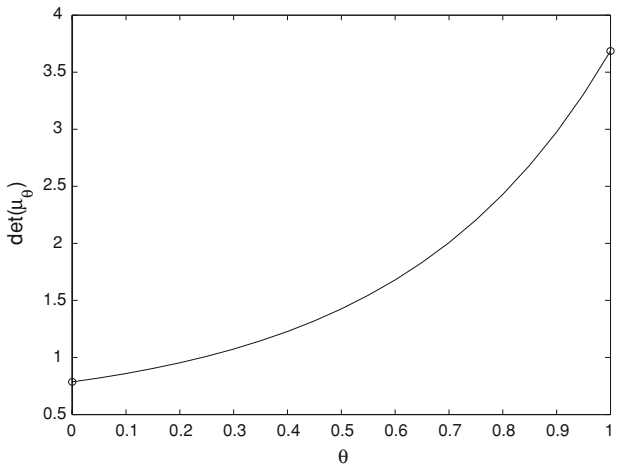
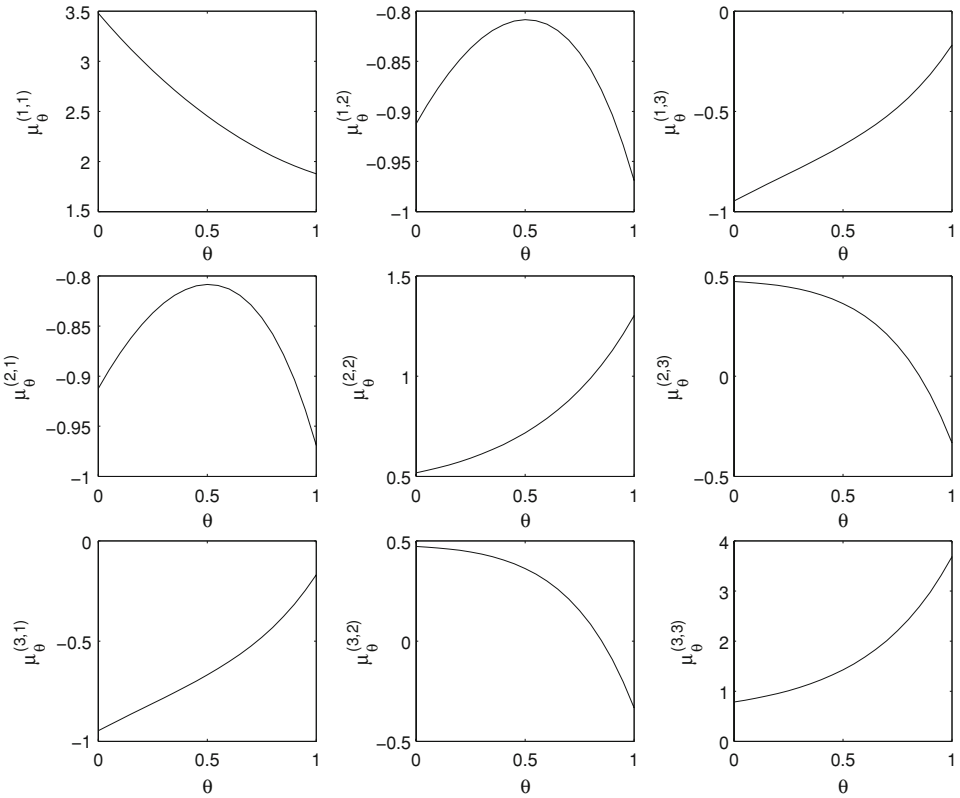


Fig. 7 Experiment on interpolating over $\mathbb{S}^+(3)$. Function $\det(\mu_\theta)$ versus the variable θ

Fig. 6 Experiment on learning interpolates over the space $\mathbb{S}^+(3)$. The *nine* entries of the interpolation matrix μ_θ versus the variable θ



Conclusions

The present manuscript is devoted to the problem of learning the average of a set of symmetric positive-definite matrices as well as their variance. Averaging over the curved manifold of symmetric positive-definite matrices was defined via the notion of Fréchet mean and the associated metric dispersion was interpreted as the variance of a set of symmetric positive-definite matrices around their Fréchet mean. The problem of continuous interpolation of two symmetric positive-definite matrices was considered as well. The property of volume conservation of the Fréchet mean and of the considered interpolatory scheme for symmetric positive-definite matrices was also discussed. The manuscript described several applications where the considered learning algorithm could be readily exploited, including in machine learning, pattern classification, speech emotion classification, diffusion tensor data analysis in medicine and intelligent control.

The behavior of the learning algorithm was tested numerically. Numerical experiments show that the averaging algorithm converges steadily and in a few iterations.

The present research study may be framed into a more general investigation about the learning of the statistical features of patterns belonging to curved manifolds [5, 6]. Such broader investigation field may include the learning of probability density functions and related statistical descriptors on manifold, the generation of (pseudo) random events on manifolds as well as dimensionality reduction of high-dimensional data on manifolds.

The above-mentioned statistical techniques might be applied in the future to artificial intelligence and cognition, as for example in pattern recognition/detection from camera imagery and automatic emotion classification. Some related statistical techniques on manifolds are under investigation, as for instance a manifold-valued-data dimensionality reduction technique based on multi-dimensional scaling (known in the scientific literature with the acronym MDS) adapted to metrizable data manifolds.

Acknowledgments The preparation of the present paper was consistently advanced during my Summer 2008 stay at the Image and Signal Processing laboratory of the Department of Electrical and Electronic Engineering of the Tokyo University of Agriculture and Technology (TUAT) for research purposes, thanks to a grant from the International Information Science Foundation (IISF). I wish to gratefully thank Prof. Toshihisa Tanaka, who made this fruitful visit possible, as well as all the laboratory members for their warm hospitality.

References

1. Arsigny V, Fillard P, Pennec X, Ayache N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J Matrix Anal Appl (submitted).
2. Cerf NJ, Adam C. Quantum extension of conditional probability. Phys Rev A 1999;60(2):893–7.
3. Chen Y, McInroy JE. Estimation of symmetric positive-definite matrices from imperfect measurements. IEEE Trans Automat Contr. 2002;47(10):1721–5.
4. Chetouani M, Mahdhaoui A, Ringeval F. Time-scale feature extractions for emotional speech characterization. Cogn Comput. 2009;1:194–201.
5. Fiori S. On vector averaging over the unit hyphersphere. Digit Signal Process. 2009;9(4):715–25.
6. Fiori S, Tanaka T. An algorithm to compute averages on matrix Lie groups. IEEE Trans Signal Process. Accepted for publication.
7. Fletcher PT, Joshi S. Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Process. 2007; 87(2): 250–62.
8. Fréchet M. Les élements aléatoires de nature quelconque dans un espace distancié. Annales de l'Institut Henri Poincaré 1948;10: 215–310.
9. Habeck C, Krakauer JW, Ghez C, Sackeim HA, Eidelberg D, Stern Y, Moeller JR. A new approach to spatial covariance modeling of functional brain imaging data: ordinal trend analysis. Neural Comput. 2005;17(7):1602–45.
10. Haykin S. Foundations of cognitive dynamic systems. Cambridge: Cambridge University Press.
11. Luenberger DG. The gradient projection methods along geodesics. Manage Sci. 1972;18:620–31.
12. McGraw T, Vemuri BC, Yezierski B, Mareci T. Von Mises–Fisher mixture model of the diffusion ODF. In Proceedings of the 3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano (ISBI 2006). 2006. pp. 65–8.
13. Prabhu N, Chang H-C, Deguzman M. Optimization on Lie manifolds and pattern recognition. Pattern Recognit. 2005; 38(12):2286–300.
14. Proust C, Jacqmin-Gadda H, Taylor JMG, Ganiayre J, Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. Biometrics 2006; 62(4):1014–24.
15. Rahman IU, Drori I, Stodden VC, Donoho DL, Schröder P. Multiscale representations for manifold-valued data. Multiscale Model Simul. 2005;4(4):1201–32.
16. Salencon J. Handbook of continuum mechanics. Berlin: Springer; 2001.
17. Siedlecki KL, Habeck CG, Brickman AM, Gazes Y, Stern Y. Examining the multifactorial nature of cognitive aging with covariance analysis of positron emission tomography data. J Int Neuropsychol Soc (in press).
18. Spivak M. A comprehensive introduction to differential geometry, vol 1. 2nd edn. Berkeley, CA: Publish or Perish Press; 1979.
19. Tsuda K, Rätsch G, Warmuth MK. Matrix exponentiated gradient updates for on-line learning and Bregman projection. J Mach Learn Res. 2005;6:995–1018.
20. Tuzel O, Porikli F, Meer P. Region covariance: a fast descriptor for detection and classification. In: Proceedings of European Conference on Computer Vision, vol. 2. Graz, Austria; 2006. pp. 589–600.
21. Udrişte C. Convex functions and optimization methods on riemannian manifolds. Dordrecht: Kluwer; 1994.
22. Ye C, Liu J, Chen C, Song M, Bu J. Speech emotion classification on a Riemannian manifold. In: Proceedings of Advances in Multimedia Information Processing (PCM 2008), Lecture Notes in Computer Science, vol 5353/2008. Berlin/Heidelberg: Springer. pp. 61–9.