Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures

Frank Nielsen [1]

Abstract

Data sets of multivariate normal distributions abound in many scientific areas like diffusion tensor medical imaging, structure tensor computer vision, radar signal processing, machine learning, etc. In order to process those data sets for downstream tasks like filtering, classification or clustering, one needs to define proper notions of dissimilarities and paths joining normal distributions. The Fisher-Rao distance defined as the Riemannian geodesic distance induced by the Fisher information is such a principled distance which however is not known in closed-form excepts on a few particular cases. We first report a fast and robust method to approximate arbitrarily finely the Fisher-Rao distance between normal distributions. Second, we introduce a distance based on a diffeomorphic embedding of the Gaussian manifold into a submanifold of the higher-dimensional symmetric positive-definite cone. We show that the projective Hilbert distance on the cone is a metric on the embedded Gaussian submanifold and pullback that distance with the straight line Hilbert cone geodesics to obtain a distance and paths between normal distributions. Compared to the Fisher-Rao distance approximation, the pullback Hilbert cone distance is computationally light since it requires to compute only extreme eigenvalues of matrices. Finally, we show how to use those distances in clustering tasks.

1. Introduction

Data sets of multivariate normal distributions (MVNs) are increasing frequent in many scientific areas like medical imaging (diffusion tensor imaging (Han & Park, 2014)), computer vision (image segmentation (Carson et al., 2002) or structure tensor imaging (Porikli et al., 2006)), signal processing (covariance matrices (Barbaresco, 2013) in radar or brain computer interfaces (Barachant et al., 2011)), and machine learning (Gaussian mixtures or kernel density estimators (Davis & Dhillon, 2006)). These data sets can be viewed as (weighted) point sets on a Gaussian manifold and Riemannian and information-geometric structures (Skovgaard, 1984; Yoshizawa & Tanabe, 1999) on that manifold allows one to define geodesics and distances or divergences which allows on to build algorithms like filtering, classification, clustering or optimization techniques (Tuzel et al., 2008; Absil et al., 2008; Hosseini & Sra, 2015). For example, we may simplify a Gaussian mixture model (Davis & Dhillon, 2006; Goldberger et al., 2008; Zhang & Kwok, 2010) (GMM) with n components by viewing the mixture as a weighted point set and simplify the mixture by clustering the point set into k clusters using k-means or k-medioids (Davis & Dhillon, 2006) (as known as discrete k-means). We may also consider n GMMs with potentially different components and build a codebook of all mixture components to quantize and compress the representation of these GMMs. In this work, we consider two kinds of metric distances and metric geodesics: The Fisher-Rao distance (Strapasson et al., 2016) and a new distance obtained by pulling back the Hilbert cone projective distance on an embedding of the Gaussian manifold into the higher-dimensional symmetric positive-definite matrix cone (Calvo & Oller, 1990).

The paper is organized as follows: In Section 2, we recall the Fisher-Rao geodesic distance and mention its lack of general closed-form formula on the Gaussian manifold. We then build on a recent breakthrough which studied the MVN Fisher-Rao geodesics (Kobayashi, 2023) (§2.2) to design an approximation method which upper bounds the true Fisher-Rao distance with arbitrary precision (§2.3). We present applications to simplification and quantization of GMMs in Section 3 using fast clustering methods relying on nearest neighbor query data structures (Yianilos, 1993) and smallest enclosing balls in metric spaces. In Section 4, we introduce the novel pullback Hilbert cone distance which is fast to compute and enjoys simple expression of geodesics.

[1] Sony Computer Science Laboratories Inc, Tokyo, Japan. Correspondence to: Frank Nielsen <Frank.Nielsen@acm.org>.

Proceedings of the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

2. Fisher-Rao geodesics and distances

A normal distribution $N(\mu,\Sigma)$ has probability density function (pdf) $p_{\mu,\Sigma}(x)$ defined on the full support $\mathbb{R}^d$ given by:

$$p_{\mu,\Sigma}(x)=\frac{(2\pi)^{-\frac{d}{2}}}{\sqrt{\det(\Sigma)}}\exp\left(-\frac{(x-\mu)^\top\Sigma^{-1}(x-\mu)}{2}\right).$$

Consider the statistical model consisting of all d-variate normal distributions:

$$\mathcal{N}(d)=\left\{N(\lambda)\ :\ \lambda=(\mu,\Sigma)\in\Lambda(d)=\mathbb{R}^d\times\mathrm{Sym}_+(d,\mathbb{R})\right\},$$

where $\mathrm{Sym}_+(d,\mathbb{R})$ denote the set of $d\times d$ positive-definite matrices. The dimension of $\mathcal{N}(d)$ is $m=\dim(\Lambda(d))=d+\frac{d(d+1)}{2}=\frac{d(d+3)}{2}$. When the dimension is clear from context, we omit to specify the dimension and write $\mathcal{N}$ for short.

Let $l_\lambda(x)=\log p_\lambda(x)$ denote the log-likelihood function. The MVN model is both identifiable (bijection between λ and p_λ) and regular (Calin & Udrişte, 2014; Amari, 2016), i.e., the Fisher information matrix $I(\theta)=-E_\theta[\nabla^2 l_\theta(x)]$ is positive-definite and defines a Riemannian metric g_{Fisher} called the Fisher information metric. The Riemannian manifold $(\mathcal{N},g_{\mathrm{Fisher}})$ is called the Fisher-Rao Gaussian manifold with squared infinitesimal length element (Skovgaard, 1984) $\mathrm{d}s^2_{\mathrm{Fisher}}$ at (μ,Σ) given by:

$$\mathrm{d}s^2_{\mathrm{Fisher}}=\mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu+\frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1}\mathrm{d}\Sigma\right)^2\right),$$

where $\mathrm{d}\mu\in\mathbb{R}^d$ and $\mathrm{d}\Sigma\in\mathrm{Sym}(d,\mathbb{R})$ (vector space of symmetric $d\times d$ matrices).

The Fisher-Rao length of a smooth curve $c(t)$ with $t\in[a,b]$ is defined by integrating the infinitesimal Fisher-Rao length element along the curve: $\mathrm{Len}(c)=\int_a^b \mathrm{d}s_{\mathrm{Fisher}}(c(t))\,\mathrm{d}t$, and the Fisher-Rao distance (Hotelling, 1930; Rao, 1945) between $N_0=N(\mu_0,\Sigma_0)$ and $N_1=N(\mu_1,\Sigma_1)$ is the geodesic distance on $(\mathcal{N},g_{\mathrm{Fisher}})$, i.e., the length of the Riemannian geodesic $\gamma^{\mathcal{N}}_{\mathrm{FR}}(N_0,N_1;t)$:

$$\rho_{\mathrm{FR}}(N_0,N_1)=\int_0^1 \mathrm{d}s_{\mathrm{Fisher}}(\gamma^{\mathcal{N}}_{\mathrm{FR}}(N_0,N_1;t))\,\mathrm{d}t. \quad (1)$$

In Riemannian geometry (Godinho & Natário, 2012), geodesics with boundary conditions $N_0=\gamma^{\mathcal{N}}_{\mathrm{FR}}(N_0,N_1;0)$ and $N_1=\gamma_{\mathrm{FR}}(N_0,N_1;1)$ are length minimizing curves among all curves $c(t)$ satisfying $c(0)=N(\mu_0,\Sigma_0)$ and $c(1)=N(\mu_1,\Sigma_1)$:

$$\rho_{\mathrm{FR}}(N_0,N_1)=\inf_{\substack{c(t)\\ c(0)=p_{\mu_0,\Sigma_0}\\ c(1)=p_{\mu_1,\Sigma_1}}}\left\{\mathrm{Len}(c)\right\}.$$

Riemannian geodesics are parameterized by (normalized) arc length t which ensures that

$$\rho(\gamma_\rho(P_0,P_1;s),\gamma_\rho(P_0,P_1;t))=|s-t|\,\rho(P_0,P_1). \quad (2)$$

More generally, geodesics in differential geometry (Calin & Udrişte, 2014) are auto-parallel curves with respect to an affine connection ∇: $\nabla_{\dot\gamma}\dot\gamma=0$, where $\dot{}=\frac{d}{\mathrm{d}t}$ and $\nabla_X Y$ is the covariant derivative induced by the connection. In Riemannian geometry, the default connection is the unique *Levi-Civita metric connection* (Godinho & Natário, 2012) ∇^g induced by the metric g.

In general, the Fisher-Rao distance between MVNs is not known in closed-form (Pinele et al., 2020; Nielsen, 2023). However, there are two main cases where closed-form formula are known:

- The case $d=1$: The Fisher-Rao distance between univariate normal distributions (Yoshizawa, 1972) $N_0=N(\mu_0,\sigma_0^2)$ and $N_1=N(\mu_1,\sigma_1^2)$ is

$$\rho_{\mathcal{N}}(N_0,N_1)=\sqrt{2}\log\left(\frac{1+\Delta(\mu_0,\sigma_0;\mu_1,\sigma_1)}{1-\Delta(\mu_0,\sigma_0;\mu_1,\sigma_1)}\right), \quad (3)$$

where for $(a,b,c,d)\in\mathbb{R}^4\backslash\{0\}$,

$$\Delta(a,b;c,d)=\sqrt{\frac{(c-a)^2+2(d-b)^2}{(c-a)^2+2(d+b)^2}} \quad (4)$$

is a Möbius distance (Burbea & Rao, 1982).

- The case where MVNs N_0 and N_1 share the same mean (James, 1973; Skovgaard, 1984), i.e., they belong to some submanifold $\mathcal{N}_\mu=\{N(\mu,\Sigma)\ :\ \Sigma\in\mathrm{Sym}_+(d,\mathbb{R})\}$. When $\mu=0$, we let $\mathcal{P}(d)=\mathcal{N}_0(d)$. We have:

$$\rho_{\mathcal{N}_\mu}(N_0,N_1)\ =\ \sqrt{\frac{1}{2}\sum_{i=1}^d \log^2\lambda_i(\Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}})},$$

where $\lambda_i(M)$ denotes the i-th largest eigenvalue of matrix M. Observe that matrix $\Sigma_0^{-1}\Sigma_1$ may not be symmetric but $\Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}}$ is always SPD and $\lambda_i(\Sigma_0^{-1}\Sigma_1)=\lambda_i(\Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}})$. The submanifolds $\mathcal{N}_\mu$ are totally geodesic in $\mathcal{N}$, and the Fisher-Rao geodesics are known in closed form: $\gamma^{\mathcal{N}}_{\mathrm{FR}}(N_0,N_1;t)=N(\mu,\Sigma_t)$ with $\Sigma_t=\Sigma_0^{\frac{1}{2}}\,(\Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}})^t\,\Sigma_0^{\frac{1}{2}}$. See also Appendix C.

Notice that all submanifolds $\mathcal{N}_\mu$ are non-positive curvature manifolds (NPC) (Bridson & Haefliger, 2013; Cheng et al., 2016), i.e. sectional curvatures are non-positive. However, $\mathcal{N}(d)$ is not a NPC manifold when $d>1$ since some sectional curvatures can be positive (Skovgaard, 1984). NPC property is important for designing optimization algorithms on manifolds with guaranteed convergence (Cheng et al., 2016). In a NPC manifold (M,g), we can write the Riemannian distance using the Riemannian logarithm map

$\mathrm{Log}_p : M \to T_pM$: $\rho_g(p_1, p_2) = \|\mathrm{Log}_{p_1}(p_2)\|_{p_1}$, where $\|v\|_p = \sqrt{g_p(v,v)}$. On the NPC SPD cone $\mathrm{Sym}_+(d,\mathbb{R})$ equipped with the trace metric (Moakher, 2005; Dolcetti & Pertici, 2021)

$$g_P^{\mathrm{trace}}(P_1, P_2) := \mathrm{tr}(P^{-1}P_1P^{-1}P_2),$$

the Riemannian logarithm map is expressed using the matrix logarithm Log, and we have

$$\rho_{g_{\mathrm{trace}}}(P_1, P_2) = \|\mathrm{Log}_{P_1}(P_2)\|_{P_1} = \|\mathrm{Log}(P_1^{-1}P_2)\|_F,$$

where $\|\cdot\|_F = \sqrt{\langle\cdot,\cdot\rangle_F}$ is the Fröbenius norm induced by the Fröbenius inner product: $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$ (Hilbert-Schmidt inner product). The SPD cone is also a Bruhat-Tits space (Lang & Lang, 1999).

Historically, the SPD Riemannian trace metric distance was studied by Siegel (Siegel, 1964) in the wider context of the complex manifold of symmetric complex square matrices with positive-definite imaginary part: The so-called Siegel upper half space (Friedland & Freitas, 2004) which generalizes the Poincaré upper plane. It was shown recently that the Siegel upper half space is NPC (Cabanes & Nielsen, 2021). Another popular distance in machine learning in the Wasserstein distance for which the underlying geometry on the Gaussian space was studied in (Takatsu, 2011).

2.1. Invariance under action of the positive affine group

The length element $\mathrm{d}s_{\mathrm{Fisher}}$ is invariant under the action of the positive affine group (Eriksen, 1986)

$$\mathrm{Aff}_+(d,\mathbb{R}) := \left\{(a, A) \ : \ a \in \mathbb{R}^d, A \in \mathrm{GL}_+(d,\mathbb{R})\right\},$$

where $\mathrm{GL}_+(d,\mathbb{R})$ denotes the group of $d \times d$ matrices with positive determinant. The group identity element of $\mathrm{Aff}_+(d,\mathbb{R})$ is $e = (0, I)$ and the group operation is $(a_1, A_1).(a_2, A_2) = (a_1 + A_1a_2, A_1A_2)$ with inverse operation $(a, A)^{-1} = (-A^{-1}a, A^{-1})$). The positive affine group may be handled as a matrix group by mapping elements (a, A) to $(d+1) \times (d+1)$ matrices $M_{(a,A)} := \begin{bmatrix} A & a \\ 0 & 1 \end{bmatrix}$. Then the matrix group operation is the matrix multiplication and inverse operation is given by the matrix inverse. Let us consider the following group action (denoted by the dot .) of the positive affine group on the Gaussian manifold $\mathcal{N}$:

$$(a, A).N(\mu, \Sigma) = N(a + A\mu, A\Sigma A^\top).$$

This action corresponds to the affine transformation of random variables: $Y = a + AX \sim N(a + A\mu, A\Sigma A^\top)$ where $X \sim N(\mu, \Sigma)$. The statistical model $\mathcal{N}$ can thus be interpreted as a group with identity element the standard MVN $N_{\mathrm{std}} = N(0, I)$: $\mathcal{N}(d) = \{(a, A).N_{\mathrm{std}} \ : \ (a, A) \in \mathrm{Aff}_+(d)\}$. We get a Lie group differential structure on $\mathcal{N}$ (Kwon et al., 2009) which moreover extends to a statistical Lie group structure in information geometry (Furuhata et al., 2021).

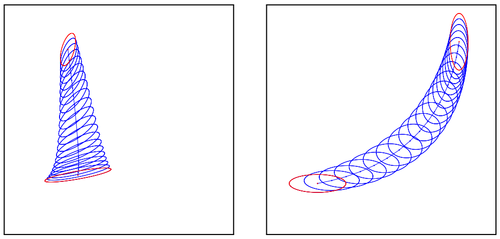

(a) (b)

Figure 1. Some Fisher-Rao geodesics with boundary conditions (displayed in red) on the bivariate Gaussian manifold. The sample space $\mathbb{R}^2$ is visualized for the range $[-0.3, 1.2] \times [-0.3, 1.2]$.

It can be checked that the Fisher-Rao length element is invariant under the action of $\mathrm{Aff}_+(d,\mathbb{R})$ and therefore the Fisher-Rao distance is also invariant: $\rho_{\mathrm{FR}}((a, A).N_0 : (a, A).N_1) = \rho_{\mathrm{FR}}(N_0, N_1)$. It follows that the Fisher-Rao geodesics in $\mathcal{N}$ are equivariant (Eriksen, 1986; Kobayashi, 2023) $\gamma_{\mathrm{FR}}^{\mathcal{N}}(B.N_0, B.N_1; t) = B.\gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t)$ for any $B \in \mathrm{Aff}_+(d,\mathbb{R})$ and we can therefore consider without loss of generality that N_0 is the standard normal distribution and $N_1 \to N_1' = N\left(\Sigma_0^{-\frac{1}{2}}(\mu_1, -\mu_0), \Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}}\right)$.

2.2. Fisher-Rao geodesics with boundary conditions

The Fisher-Rao geodesic Ordinary Differential Equation (ODE) for MVNs was first studied by Skovgaard (Skovgaard, 1984):

$$\left\{\begin{array}{lcl} \ddot{\mu} - \dot{\Sigma}\Sigma^{-1}\dot{\mu} & = & 0, \\ \ddot{\Sigma} + \dot{\mu}\dot{\mu}^\top - \dot{\Sigma}\Sigma^{-1}\dot{\Sigma} & = & 0. \end{array}\right. \tag{5}$$

Eriksen (Eriksen, 1986) first reported a solution of the geodesic equation with initial conditions: That is Fisher-Rao geodesics emanating from source N_0 with initial prescribed tangent vector $v_0 = \dot{\gamma}(0)$ in the tangent plane T_{N_0}. Eriksen's solution required to compute a matrix exponential of a matrix of size $(2d+1) \times (2d+1)$ and the use of square matrices of dimension $2d+1$ was mysterious (Imai et al., 2011). Calvo and Oller (Calvo & Oller, 1991) later studied a more general differential equation system than in Eq. 5 and reported a closed-form solution without using extra dimensions (see Appendix A). For many years, the Fisher-Rao geodesics with boundary conditions N_0 and N_1 were not known in closed-form and had to be approximated using geodesic shooting methods (Han & Park, 2014; Pilté & Barbaresco, 2016): Those geodesic shooting methods were time consuming and numerically unstable, thus limiting their use in applications (Han & Park, 2014). A recent breakthrough by Kobayashi (Kobayashi, 2023) full explains and extends geometrically the rationale of Eriksen and obtains a method to compute in closed-form the

Fisher-Rao geodesic with boundary conditions. Namely, Kobayashi (Kobayashi, 2023) proved that the Fisher-Rao geodesics can be obtained by a Riemannian submersion of horizontal geodesics of the non-compact Riemannian symmetric space of dimension $2d+1$. We report concisely below the recipe which we extracted from Kobayashi's principled geometric method to derive $N_t = \gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t)$ as follows:

Fisher-Rao geodesic $N_t = N(\mu(t), \Sigma(t)) = \gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t)$:

- For $i \in \{0, 1\}$, let $G_i = M_i\, D_i\, M_i^\top$, where

$$M_i = \begin{bmatrix} \Sigma_i^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \Sigma_i \end{bmatrix}, \tag{6}$$

$$D_i = \begin{bmatrix} I_d & 0 & 0 \\ \mu_i^\top & 1 & 0 \\ 0 & -\mu_i & I_d \end{bmatrix}, \tag{7}$$

  where I_d denotes the identity matrix of shape $d \times d$. That is, matrices G_0 and $G_1 \in \mathrm{Sym}_+(2d+1, \mathbb{R})$ can be expressed by *block Cholesky factorizations*.

- Consider the Riemannian geodesic in $\mathrm{Sym}_+(2d+1, \mathbb{R})$ with respect to the trace metric:

$$G(t) = G_0^{\frac{1}{2}} \left(G_0^{-\frac{1}{2}} G_1 G_0^{-\frac{1}{2}} \right)^t G_0^{\frac{1}{2}}.$$

  In order to compute the matrix power G^p for $p \in \mathbb{R}$, we first calculate the Singular Value Decomposition (SVD) of G: $G = O\, L\, O^\top$ (where O is an orthogonal matrix and $L = \mathrm{diag}(\lambda_1, \ldots, \lambda_{2d+1})$ a diagonal matrix) and then get the matrix power as $G^p = O\, L^p\, O^\top$ with $L^p = \mathrm{diag}(\lambda_1^p, \ldots, \lambda_{2d+1}^p)$.

- Retrieve $N(t) = \gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t) = N(\mu(t), \Sigma(t))$ from $G(t)$:

$$\Sigma(t) = [G(t)]_{1:d,1:d}^{-1}, \tag{8}$$

$$\mu(t) = \Sigma(t)\, [G(t)]_{1:d,d+1}, \tag{9}$$

  where $[G]_{1:d,1:d}$ denotes the block matrix with rows and columns ranging from 1 to d extracted from $(2d+1) \times (2d+1)$ matrix G, and $[G]_{1:d,d+1}$ is similarly the column vector of $\mathbb{R}^d$ extracted from G.

Note that this technique also proves that the MVN geodesics are unique although $\mathcal{N}$ is not NPC. It is proven in (Furuhata et al., 2021) that the Gaussian manifold admits a solvable Lie group and hence is diffeomorphic to some Euclidean space. Figure 1 displays several bivariate normal Fisher-Rao geodesics with boundary conditions obtained by implementing this method (Kobayashi, 2023). We display $N(\mu, \Sigma)$ by an ellipse $\mathcal{E} = \{\mu + Lx \ :\ \|x\|_2 = 1\}$ where $\Sigma = LL^\top$ (Cholesky decomposition).

2.3. Fisher-Rao distances

The previous section reported the closed-form solutions for the Fisher-Rao geodesics $\gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t)$. We shall now explain a method to approximate their lengths and hence the Fisher-Rao distances:

$$\rho_{\mathrm{FR}}(N_0, N_1) = \mathrm{Len}(\gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t)).$$

Consider discretizating regularly $t \in [0,1]$ using $T+1$ steps: $\frac{0}{T} = 0, \frac{1}{T}, \ldots, \frac{T-1}{T}, \frac{T}{T} = 1$. Since geodesics are totally 1D submanifolds, we have

$$\rho_{\mathrm{FR}}(N_0, N_1) = \sum_{i=0}^{T-1} \rho_{\mathrm{FR}}(\gamma_{\mathrm{FR}} \left(N_{\frac{i}{T}}, N_{\frac{i+1}{T}} \right)).$$

By choosing T large enough, we have $N = N_{\frac{i}{T}}$ close to $N' = N_{\frac{i+1}{T}}$, and we can approximate the geodesic distance as follows:

$$\rho_{\mathrm{FR}}(N, N') \approx \mathrm{d}s_{\mathrm{Fisher}}(N) \approx \sqrt{\frac{2}{f''(1)} I_f(N, N')},$$

where $I_f(p,q)$ is *any* f-divergence (Ali & Silvey, 1966; Csiszár, 1967) between pdfs $p(x)$ and $q(x)$ induced by a strictly convex generator $f(u)$ satisfying $f(1) = 0$:

$$I_f(p,q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x.$$

Indeed, we have for two close distributions p_θ and $p_{\theta+\mathrm{d}\theta}$ (Amari, 2016): $I_f(p_\theta, p_{\theta+\mathrm{d}\theta}) \approx \frac{f''(1)}{2} \mathrm{d}s_{\mathrm{Fisher}}^2$.

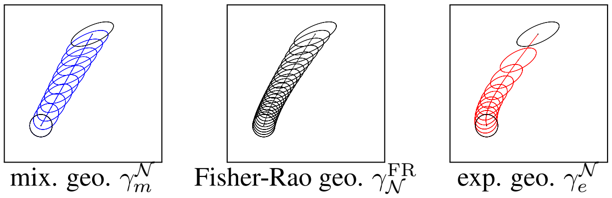
We choose the Jeffreys f-divergence which is the arithmetic symmetrization of the Kullback-Leibler divergence obtained for the generator $f_J(u) = (u-1)\log u$ with $f_J''(1) = 2$. It follows that $D_J(N_1, N_2) = I_{f_J}(N_1, N_2) = \mathrm{tr}\left(\frac{\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2}{2} - I\right) + (\mu_2 - \mu_1)^\top \frac{\Sigma_1^{-1} + \Sigma_2^{-1}}{2} (\mu_2 - \mu_1)$.

Thus we get the following overall approximation of the Fisher-Rao distance:

$$\tilde{\rho}_T(N_0, N_1) = \sum_{i=0}^{T-1} \sqrt{D_J\left(N_{\frac{i}{T}}, N_{\frac{i+1}{T}}\right)} \approx \rho_{\mathrm{FR}}(N_0, N_1). \tag{10}$$

In (Gao & Chaudhari, 2021), the authors choose $\mathrm{d}s_{\mathrm{Fisher}}(p) = \sqrt{2 D_{\mathrm{KL}}(p_\theta, p_{\theta+\mathrm{d}\theta})}$ where $D_{\mathrm{KL}} = I_{f_{\mathrm{KL}}}$ is the Kullback-Leibler divergence, a f-divergence obtained for $f_{\mathrm{KL}}(u) = -\log u$.

Property 1 (Fisher-Rao upper bound). The Fisher-Rao distance between normal distributions is upper bounded by the square root of the Jeffreys divergence: $\rho_{\mathrm{FR}}(N_0, N_1) \leq \sqrt{D_J(N_0, N_1)}$.



mix. geo. $\gamma_m^{\mathcal{N}}$ Fisher-Rao geo. $\gamma_{\mathcal{N}}^{\mathrm{FR}}$ exp. geo. $\gamma_e^{\mathcal{N}}$

Figure 2. Visualizing geodesics with respect to the mixture, Levi-Civita (Fisher-Rao), and exponential connections.

The proof can be found in many places, e.g. (Grosse et al., 2013; Amari, 2016; Rong et al., 2017). Yet we report another proof in Appendix B.

Notice that we have $\rho_{\mathrm{FR}}(N_0, N_1) \leq \tilde{\rho}_T(N_0, N_1)$ for all $T > 1$. Define the energy of a curve $c(t)$ with $t \in [a, b]$ by $E(c) = \int_a^b \mathrm{d}s^2_{\mathrm{Fisher}}(t)\mathrm{d}t$. We have

$$E(\gamma_e^{\mathcal{N}}(N_0, N_1; t)) = E(\gamma_m^{\mathcal{N}}(N_0, N_1; t)) = D_J(N_0, N_1),$$

where $\gamma_e^{\mathcal{N}}(N_0, N_1; t) = N(\mu_t^e, \Sigma_t^e)$ and $\gamma_m^{\mathcal{N}}(N_0, N_1; t) = N(\mu_t^m, \Sigma_t^m)$ are the exponential and mixture geodesics in information geometry (Yoshizawa & Tanabe, 1999) given by $\mu_t^m = \bar{\mu}_t$ and $\Sigma_t^m = \bar{\Sigma}_t + t\mu_1\mu_1^\top + (1-t)\mu_2\mu_2^\top - \bar{\mu}_t\bar{\mu}_t^\top$ where $\bar{\mu}_t = t\mu_1 + (1-t)\mu_2$ and $\bar{\Sigma}_t = t\Sigma_1 + (1-t)\Sigma_2$, and $\mu_t^e = \bar{\Sigma}_t^H(t\Sigma_1^{-1}\mu_1 + (1-t)\Sigma_2^{-1}\mu_2)$ and $\Sigma_t^e = \bar{\Sigma}_t^H$ where $\bar{\Sigma}_t^H = (t\Sigma_1^{-1} + (1-t)\Sigma_2^{-1})^{-1}$ is the matrix harmonic mean. See Figure 2. The mixture, Fisher-Rao, and exponential geodesics are α-connection geodesics (Furuhata et al., 2021) for $\alpha = -1$, $\alpha = 0$ and $\alpha = 1$, respectively. Notice that these e/m geodesics are computationally less intensive to evaluate than the Fisher-Rao geodesics.

Since the upper bound of Property 1 is tight infinitesimally, we get in the limit convergence to the Fisher-Rao distance:

$$\lim_{T\to\infty} \tilde{\rho}_T(N_0, N_1) = \rho_{\mathrm{FR}}(N_0, N_1).$$

Example 1. Let us consider the example of Han and Park (Han & Park, 2014) (displayed in Figure 1(b)): $N_0 = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}\right)$ and $N_1 = N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}\right)$. The time consuming geodesic shooting algorithm of (Han & Park, 2014) evaluates the Fisher-Rao distance to $\rho_{\mathcal{N}}(N_0, N_1) \approx 3.1329$. We get the following approximations: $\tilde{\rho}_T(N_0, N_1) = 3.1996$ for $T = 100$. See Figure 3 for the convergence curve of $\tilde{\rho}_T(N_0, N_1)$ as a function of T.

3. Fisher-Rao clustering

We shall consider two applications of the Fisher-Rao distance between MVNs using clustering:

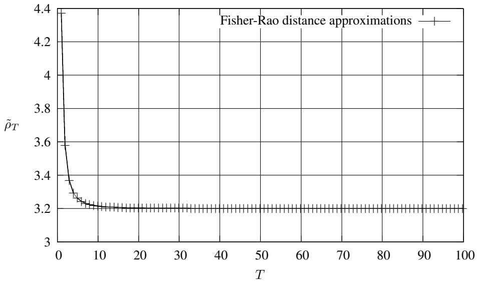


Figure 3. Convergence of $\tilde{\rho}_T(N_0, N_1)$ to $\rho_{\mathrm{FR}}(N_0, N_1)$.

The first application considers clustering weighted MVNs which is useful to simplify Gaussian Mixture Models (Davis & Dhillon, 2006; Strapasson et al., 2016) (GMMs): A GMM $m(x) = \sum_{i=1}^n w_i p_{\mu_i,\Sigma_i}(x)$ with n components is a weighted set of n MVNs $N(\mu_i, \Sigma_i)$ and clustering this set into k-clusters allows one to simplify the GMM $m(x)$. For this task, we may use the k-means clustering (Lloyd, 1982) when centroids are available in closed-form (Davis & Dhillon, 2006) (using the Kullback-Leibler divergence) or the k-medioid clustering (Kaufman, 1990) when we choose the representative of clusters from the input otherwise (using the Fisher-Rao distance).

The second application considers the quantization of sets of MVNs which is useful to further compress a set $\{m_1, \ldots, m_n\}$ of n GMMs $m_i(x) = \sum_{j=1}^{n_i} w_{i,j} p_{\mu_{i,j},\Sigma_{i,j}}(x)$ with overall $N = \sum_{i=1}^n n_i$ MVNs $N(\mu_{i,j}, \Sigma_{i,j})$. We build a codebook of k MVNs $N(m_i, S_i)$ by quantizing the N non-weighted MVNs using the guaranteed k-center clustering of (Gonzalez, 1985) (also called k-centers clustering (Dueck & Frey, 2007)). Then each mixture $m_i(x)$ is quantized into a mixture $\tilde{m}_{w_i}(x) = \sum_{i=1}^k w_{i,j} p_{m_i,S_i}(x)$. The advantage of quantization is that the original set $\{m_1, \ldots, m_n\}$ of GMMs is compactly represented by n points in the $(k-1)$-dimensional standard simplex Δ_{k-1} encoding $\{\tilde{m}_1, \ldots, \tilde{m}_n\}$ since they share the same components. The set $\{\tilde{m}_w : w \in \Delta_{k-1}\}$ form a mixture family in information geometry (Amari, 2016; Nielsen & Hadjeres, 2019) with a dually flat space which can be exploited algorithmically.

Notice that minimizing the objective functions of these k-means, k-medioid and k-center clustering objective are NP-hard when dealing with MVNs.

3.1. Nearest neighbor queries

In order to speed up these center-based clustering, we shall find for a given MVN $N(\mu, \Sigma)$ (a query) its closest cluster center among k MVNs $\{N(m_i, S_i)\}$ using Nearest Neighbor (NN) query search (Andoni, 2009; Bhatia et al., 2010). There exist many data-structures for exact and approximate

NN queries. For example, the vantage point (VP) tree structure is well-suited in metric spaces (Yianilos, 1993) and has also been considered for NN queries with respect to the Kullback-Leibler divergence between MVNs (Nielsen et al., 2009). Although NN queries based on VP-trees still require linear time in the worst-case, they can also achieve logarithmic time in best cases. At the heart of NN search using VP-trees, we are given a query ball $\mathrm{Ball}(p,r)$ with center p and radius r, and we need to find potential intersections with balls $\mathrm{Ball}(v,r_v)$ stored at nodes v of the VP tree. Thus when using the (fine approximation $\tilde{\rho}_T$) Fisher-Rao metric distance, we need to answer predicates of whether two Fisher-Rao balls $\mathrm{Ball}_{\mathrm{FR}}(N,r)$ and $\mathrm{Ball}_{\mathrm{FR}}(N',r')$ intersect or not: This can be done by determining the sign of $\rho_{\mathrm{FR}}(N,N')-(r+r')$. When positive the balls do not intersect and when negative the balls intersect. Since we handle some approximation errors by using $\tilde{\rho}_T$ instead of ρ_{FR}, but since $\tilde{\rho}_T \geq \rho_{\mathrm{FR}}$ we need to explore both branches of a VP-tree if the balls $\mathrm{Ball}_{\mathrm{FR}}(N,r)$ and $\mathrm{Ball}_{\mathrm{FR}}(N',r')$ stored at the two siblings of a node v are such that $\tilde{\rho}_T(N,N') \leq r+r'$.

3.2. k-center and miniball

For the quantization tasks, the k-center clustering heuristic of Gonzalez (Gonzalez, 1985) guarantees to find a good k-center clustering in metric spaces with an approximation factor upper bounded by 2. We can further refine the cluster representative of each cluster by computing approximations of the *smallest enclosing Fisher-Rao balls* (miniballs) of clusters.

A simple Riemannian approximation technique has been reported for approximating the smallest enclosing ball of n points $\{p_1,\ldots,p_n\}$ on a Riemannian manifold (M,g) with geodesic distance $\rho_g(p,p')$ and geodesics $\gamma_g(p,p';t)$ in (Arnaudon & Nielsen, 2013):

```
Miniball($\{p_1,\ldots,p_n\},\rho_g,T$):

• Let $c_1 \leftarrow p_1$
• For $t=1$ to $T$
    – Compute the index of the point which is farthest to current circumcenter $c_t$:
          $f_t = \arg\max_{i\in\{1,\ldots,n\}} \rho_g(c_t,p_i)$
    – Update the circumcenter by walking along the geodesic linking $c_t$ to $p_{f_t}$:
          $c_{t+1} = \gamma_g\left(c_t,p_{f_t};\frac{1}{t+1}\right)$
      Recall that geodesics are parameterized by normalized arc length so that $\rho_g(c_t,c_{t+1}) = \frac{1}{t+1}\rho_g(c_t,p_{f_t})$.
• Return $c_T$
```

Conditions of convergence are analyzed in (Arnaudon & Nielsen, 2013): For example, it always converge for Cartan-Hadamard manifolds (complete simply connected NPC manifolds like the SPD cone).

The Fisher-Rao distance $\rho_{\mathrm{FR}}(N_\Sigma(\mu_0),N_\Sigma(\mu_1))$ between two MVNs with same covariance matrix Σ is

$$\rho_{\mathrm{FR}}(N_\Sigma(\mu_0),N_\Sigma(\mu_1)) = \sqrt{2}\,\mathrm{arccosh}\left(1+\frac{1}{4}\Delta_\Sigma^2(\mu_0,\mu_1)\right),$$

where $\Delta_\Sigma(\mu_0,\mu_1)=\sqrt{(\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_20-\mu_1)}$ is the Mahalanobis distance. Therefore when all MVNs belong to the non-totally flat submanifold $\mathcal{N}_\Sigma=\{N(\mu,\Sigma)\ :\ \mu\in\mathbb{R}^d\}$, the smallest enclosing ball amounts to an Euclidean smallest enclosing ball (Welzl, 2005) since in that case ρ_{FR} is an increasing function of the Mahalanobis distance.

Since the computations of $\tilde{\rho}_T$ approximating ρ_{FR} is costly, the following section shall consider a new fast metric distance on $\mathcal{N}$ which further relates to the Fisher-Rao distance.

4. Pullback Hilbert cone distance

Let us define dissimilarities and paths on $\mathcal{N}(d)$ from dissimilarities and geodesics on $\mathcal{P}(d+1)=\mathcal{N}_0(d)$ by considering the following family of *diffeomorphic embeddings* $f_a:\mathcal{N}(d)\to\mathcal{P}(d+1)$ for $a\in\mathbb{R}_{>0}$ proposed in (Calvo & Oller, 1990):

$$f_a(N(\mu,\Sigma)):=\begin{bmatrix}\Sigma+a\mu\mu^\top & a\mu\\ a\mu^\top & a\end{bmatrix}\in\mathcal{P}(d+1). \quad (11)$$

Let $\overline{\mathcal{N}}_a(d)=\{f_a(N)\ :\ N\in\mathcal{N}(d)\}\subset\mathcal{P}(d+1)$ denote the embedded Gaussian submanifold in $\mathcal{P}(d+1)$ of codimension 1. We let $f_a^{\mathrm{inv}}:\overline{\mathcal{N}}_a(d)\to\mathcal{N}(d)$ denote the functional inverse so that $f_a\circ f_a^{\mathrm{inv}}=id_{\mathcal{N}}$ is the identity function $\mathrm{id}_{\mathcal{N}}:\mathcal{N}\to\mathcal{N}$. The notation inv in f_a^{inv} is chosen to avoid confusion with the matrix inverse $f_a(N(\mu,\Sigma))^{-1}$:

$$f_a(N(\mu,\Sigma))^{-1}=\begin{bmatrix}\Sigma^{-1} & -\Sigma^{-1}\mu\\ -\mu^\top\Sigma^{-1} & \mu^\top\Sigma^{-1}\mu+\frac{1}{a}\end{bmatrix}.$$

The open SPD cone $\mathcal{P}(d+1)$ can thus be foliated by the family of submanifolds $\overline{\mathcal{N}}_a$ (Calvo & Oller, 1990): $\mathcal{P}(d+1)=\left\{a\times\overline{\mathcal{N}}_a\ :\ a\in\mathbb{R}_{>0}\right\}$. We let $f=f_1$ and $f^{\mathrm{inv}}=f_1^{\mathrm{inv}}$, and $\overline{\mathcal{N}}=\overline{\mathcal{N}}_1$.

Calvo and Oller (Calvo & Oller, 1990) proved that $(\mathcal{N}(d),g_{\mathrm{Fisher}})$ is isometrically embedded into $(\mathcal{P}(d+1),\frac{1}{2}g_{\mathrm{trace}})$ but that $\overline{\mathcal{N}}$ is not totally geodesic. Thus we have

$$\begin{aligned}\rho_{\mathrm{CO}}(N_0,N_1) &= \rho_{\mathrm{FR}}(N(0,f(N_0)),N(0,f(N_1))),\\ &= \rho_{\mathcal{P}}(\overline{\mathcal{N}}_0,\overline{\mathcal{N}}_1)\geq\rho_{\mathrm{FR}}(N_0,N_1),\end{aligned}$$

where $\overline{\mathcal{N}}_i = f(N_i)$. See Eq. 5. It follows that we get a series of lower bound for $\rho_{\mathrm{FR}}(N_0, N_1)$:

$$\rho_{\mathrm{CO},T}(N_0, N_1) = \sum_{i=0}^{T-1} \rho_{\mathcal{P}}\left(N_{\frac{i}{T}}, N_{\frac{i+1}{T}}\right),$$

such that for all T, $\tilde{\rho}_T \geq \rho_{\mathrm{FR}} \geq \rho_{\mathrm{CO},T}$.

We can also approximate the smallest enclosing Fisher-Rao ball of $\{N(\mu_i, \Sigma_i)\}$ on $\mathcal{N}(d)$ by embedding the normals into $\overline{\mathcal{N}}$ as $\{\bar{P}_i = f(N(\mu_i, \Sigma_i))\}$. We then apply the above iterative smallest enclosing ball approximation Miniball (Arnaudon & Nielsen, 2013) to get $\tilde{C}_T \in \mathcal{P}(d+1)$ after T iterations. Then we project orthogonally with respect to the trace metric $\tilde{C}_T$ onto $\overline{\mathcal{N}}$ as $\bar{C}_T = \mathrm{proj}_{\overline{\mathcal{N}}}(\tilde{C}_T)$ and maps back to the Gaussian manifold using f^{inv} to get the approximate normal circumcenter.

The following proposition describes the orthogonal projection operation $\bar{P}_\perp = \mathrm{proj}_{\overline{\mathcal{N}}}(P)$ of $P = [P_{i,j}] \in \mathcal{P}(d+1)$ onto $\overline{\mathcal{N}}$ based on the analysis reported in the Appendix of (Calvo & Oller, 1990) (page 239):

Proposition 4.1. *Let $\beta = P_{d+1,d+1}$ and write $P = \begin{bmatrix} \Sigma + \beta\mu\mu^\top & \beta\mu \\ \beta\mu^\top & \beta \end{bmatrix}$. Then the orthogonal projection at $P \in \mathcal{P}$ onto $\overline{\mathcal{N}}$ is:*

$$\bar{P}_\perp := \mathrm{proj}_{\overline{\mathcal{N}}}(P) = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu^\top \\ \mu & 1 \end{bmatrix}, \tag{12}$$

and the SPD trace distance between P and $\bar{P}_\perp$ is

$$\rho_{\mathcal{P}}(P, \bar{P}_\perp) = |\log\beta|. \tag{13}$$

Consider pulling back SPD cone dissimilarities and geodesics of $\mathcal{P}(d+1)$ onto $\mathcal{N}(d)$ as follows:

Definition 4.2 (Pullback dissimilarities)**.** A dissimilarity $D(N_0, N_1)$ (not necessarily be a metric distance nor a smooth divergence) on $\mathcal{N}(d)$ (with $N_0{:=}N(\mu_0, \Sigma_0)$ and $N_1{:=}N(\mu_1, \Sigma_1)$) can be obtained from any dissimilarity $D(\cdot,\cdot)$ on the SPD cone by pulling back the SPD matrix cone dissimilarity using f:

$$D(N_0, N_1){:=}D(f(N_0), f(N_1)). \tag{14}$$

Similarly, we pullback cone geodesics onto $\mathcal{N}$:

Definition 4.3 (Pullback curves)**.** A path $c_\gamma(N_0, N_1; t)$ joining $N_0 = c_\gamma(N_0, N_1; 0)$ and $N_1 = c_\gamma(N_0, N_1; 1)$ can be defined by the pullback of any geodesic $\gamma(f(N_0), f(N_1); t)$ on the SPD cone:

$$c_\gamma(N_0, N_1; t){:=}f^{\mathrm{inv}}(\gamma(f(N_0), f(N_1); t)). \tag{15}$$

Hence, we can leverage the rich literature on dissimilarities and geodesics on the SPD cone (e.g., (Hero et al., 2001; Chebbi & Moakher, 2012; Sra, 2016; Baggio et al., 2018; Chen et al., 2021)). Note that the Riemannian SPD trace metric geodesic is also the geodesic for Finslerian distances $\rho_h(P_0, P_1){:=}\left\|\mathrm{Log}\left(P_0^{-\frac{1}{2}}\, P_1\, P_0^{-\frac{1}{2}}\right)\right\|_h$ where h is a totally symmetric gauge function (i.e., $h(x_1, \ldots, x_n) = h(\sigma(x_1, \ldots, x_n))$ for any permutation σ)) and $\|P\|_h{:=}h(\lambda_1(P), \ldots, \lambda_d(P))$. When $h(x) = h_p(x) = \|x\|_p = \left(\sum_{i=1}^d x_i^p\right)^{\frac{1}{p}}$ is the p-norm for $1 \leq p < \infty$, we get the Schatten matrix p-norms (Bhatia, 2009).

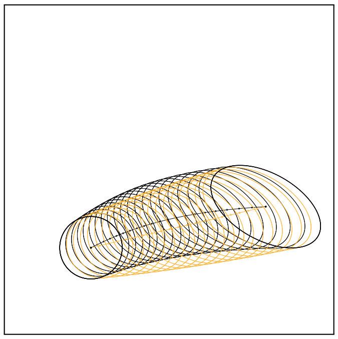


Figure 4. Comparing the pullback Hilbert geodesic (orange, coinciding with the mixture geodesic) with the exact Fisher-Rao geodesic displayed in black.

The *Hilbert projective cone distance* (Hilbert, 1895; Birkhoff, 1957; Chen et al., 2021) on the SPD cone $\mathrm{Sym}_+(d, \mathbb{R})$ is defined by

$$\begin{aligned}\rho_{\mathrm{Hilbert}}(P_0, P_1) &= \log\left(\frac{\lambda_{\max}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}{\lambda_{\min}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}\right), \\ &= \log\left(\frac{\lambda_{\max}(P_0^{-1} P_1)}{\lambda_{\min}(P_0^{-1} P_1)}\right).\end{aligned}$$

It is a *projective distance* (or quasi-metric distance) because it is symmetric and satisfies the triangular inequality but we have $\rho_{\mathrm{Hilbert}}(P_0, P_1) = 0$ if and only if $P_0 = \lambda P_1$ for some $\lambda > 0$. However, the pullback Hilbert distance on $\mathcal{N}$, $\rho_{\mathrm{Hilbert}}(N_0, N_1) := \rho_{\mathrm{Hilbert}}(f(N_0), f(N_1))$, is a proper metric distance on $\overline{\mathcal{N}}$ since $f(N_0) = f(N_1)$ if and only if $\lambda = 1$ because the array element at last row and last column $[f(N_0)]_{d+1,d+1} = [f(N_1)]_{d+1,d+1} = 1$ is identical. Thus $f(N_0) = \lambda f(N_1)$ for $\lambda = 1$. The pullback Hilbert cone distance only requires to calculate the *extreme eigenvalues* of the matrix product $f(N_0)^{-1} f(N_1)$. Thus we can bypass a costly SVD and compute approximately these extreme eigenvalues using the power method (Trevisan, 2017) (Appendix D).

The geodesic in the Hilbert SPD cone are straight lines (Nussbaum, 1994) parameterized as follows:

$$\gamma_{\mathrm{Hilbert}}(P_0, P_1; t) := \left(\frac{\beta\alpha^t - \alpha\beta^t}{\beta - \alpha}\right) P_0 + \left(\frac{\beta^t - \alpha^t}{\beta - \alpha}\right) P_1,$$

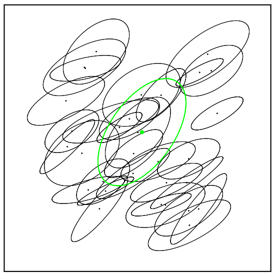


Figure 5. Approximating the Hilbert smallest enclosing ball of a set of bivariate normal distributions. The approximated minimax center is shown in green.

where $\alpha = \lambda_{\min}(P_1^{-1}P_0)$ and $\beta = \lambda_{\max}(P_1^{-1}P_0)$. Figure 4 compares the pullback Hilbert geodesic curve with the Fisher-Rao geodesic.

A pregeodesic is a geodesic which may be arbitrarily reparameterized by another parameter $u = r(t)$ for some smooth function r. That is, a pregeodesic is not necessarily parameterized by arc length. Let us notice that the weighted arithmetic mean $\mathrm{LERP}(P_0, P_1; u) = (1-u)P_0 + uP_1$ is a pregeodesic of $\gamma_{\mathrm{Hilbert}}(P_0, P_1; t)$. Although Hilbert SPD space is not a Riemannian space, it enjoys non-positive curvature properties according to various definitions of curvatures (Alabdulsada & Kozma, 2019; Karlsson & Noskov, 2000).

We can adapt the approximation of the minimum enclosing Hilbert ball by replacing ρ_{FR} by ρ_{Hilbert} and cutting metric geodesic $\gamma_{\mathrm{Hilbert}}$ instead of geodesics $\gamma_{\mathrm{FR}}^{\mathcal{N}}$ (see Figure 5).

First, the diffeomorphic embedding f exhibits several interesting features:

Proposition 4.4. *The Jeffreys divergence between p_{μ_1,Σ_1} and p_{μ_2,Σ_2} amounts to the Jeffreys divergence between $q_{\bar{P}_1} = p_{0,f(\mu_1,\Sigma_1)}$ and $q_{\bar{P}_2} = p_{0,f(\mu_2,\Sigma_2)}$ where $\bar{P}_i = f(\mu_i, \Sigma_i)$: $D_J(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) = D_J(q_{\bar{P}_1}, q_{\bar{P}_2})$.*

Proof. Since $D_J(p, q) = D_{\mathrm{KL}}(p, q) + D_{\mathrm{KL}}(q, p)$, we shall prove that $D_{\mathrm{KL}}(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) = D_{\mathrm{KL}}(q_{\bar{P}_1}, q_{\bar{P}_2})$. The KLD between two centered $(d+1)$-variate normals $q_{P_1} = p_{0,P_1}$ and $q_{P_2} = p_{0,P_2}$ is

$$D_{\mathrm{KL}}(q_{P_1}, q_{P_2}) = \frac{1}{2}\left(\mathrm{tr}(P_2^{-1}P_1) - d - 1 + \log\frac{|P_2|}{|P_1|}\right).$$

This divergence can be interpreted as the matrix version of the Itakura-Saito divergence (Davis & Dhillon, 2006). It is a matrix spectral distance since we can write $D_{\mathrm{KL}}(q_{P_1}, q_{P_2}) = (h_{\mathrm{KL}} \circ \lambda^{\mathrm{sp}})(\Sigma_2^{-1}\Sigma_1)$, where $\lambda^{\mathrm{sp}}(S) = (\lambda_1(S), \ldots, \lambda_d(S))$ and $h_{\mathrm{KL}}(u_1, \ldots, u_d) = \frac{1}{2}(u_i - 1 - \log u_i)$ (a gauge function). Similarly, the Jeffreys divergence between two centered MVNs is a matrix spectral distance with gauge function $h_J(u) = \sum_{i=1}^d \left(\sqrt{u_i} - \frac{1}{\sqrt{u_i}}\right)^2$.

The SPD cone equipped with $\frac{1}{2}$ of the trace metric can be interpreted as Fisher-Rao centered normal manifolds (isometry): $\forall\mu, (\mathcal{N}_\mu, g_{\mathcal{N}_\mu}^{\mathrm{Fisher}}) \cong (\mathcal{P}, \frac{1}{2}g^{\mathrm{trace}})$.

Since the determinant of a block matrix is $\det\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \det\left(A - BD^{-1}C\right)$, we get with $D = 1$: $\det(f(\mu, \Sigma)) = \det(\Sigma + \mu\mu^\top - \mu\mu^\top) = \det(\Sigma)$.

Let $\bar{P}_1 = f(\mu_1, \Sigma_1)$ and $\bar{P}_2 = f(\mu_2, \Sigma_2)$. Checking $D_{\mathrm{KL}}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] = D_{\mathrm{KL}}[q_{\bar{P}_1} : q_{\bar{P}_2}]$ where $q_{\bar{P}} = p_{0,\bar{P}}$ amounts to verify that $\mathrm{tr}(\bar{P}_2^{-1}\bar{P}_1) = 1 + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1 + \Delta_\mu^\top\Sigma_2^{-1}\Delta_\mu)$. Indeed, using the inverse matrix

$$f(\mu, \Sigma)^{-1} = \begin{bmatrix} \Sigma^{-1} & -\Sigma^{-1}\mu \\ -\mu^\top\Sigma^{-1} & 1 + \mu^\top\Sigma^{-1}\mu \end{bmatrix},$$

we have $\mathrm{tr}(\bar{P}_2^{-1}\bar{P}_1) = \mathrm{tr}\left(\begin{bmatrix} \Sigma_2^{-1} & -\Sigma_2^{-1}\mu_2 \\ -\mu_2^\top\Sigma_2^{-1} & 1 + \mu_2^\top\Sigma_2^{-1}\mu_2 \end{bmatrix}\begin{bmatrix} \Sigma_1 + \mu_1\mu_1^\top & \mu_1 \\ \mu_1^\top & 1 \end{bmatrix}\right) = 1 + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1 + \Delta_\mu^\top\Sigma_2^{-1}\Delta_\mu)$. Thus even if the dimension of the sample spaces of $p_{\mu,\Sigma}$ and $q_{\bar{P}=f(\mu,\Sigma)}$ differs by one, we get the same KLD and Jeffreys divergence by Calvo and Oller's isometric mapping f. □

Second, the mixture geodesics are preserved by the embedding f:

Proposition 4.5. *The mixture geodesics are preserved by the embedding f: $f(\gamma_m^{\mathcal{N}}(N_0, N_1; t)) = \gamma_m^{\mathcal{P}}(f(N_0), f(N_1); t)$.*

We check that $f(\mathrm{LERP}(N_0, N_1; t)) = \mathrm{LERP}(\bar{P}_0, \bar{P}_1; t)$. Thus the pullback of the Hilbert cone geodesics are thus coinciding with the mixture geodesics on $\mathcal{N}$.

Therefore all algorithms on $\mathcal{N}$ which only require m-geodesics or m-projections (Amari, 2016) by minimizing the right-hand side of the KLD can be implemented by algorithms on $\mathcal{P}$ by using the f-embedding. On $\mathcal{P}$, the minimizing problems amounts to a logdet minimization problem well-studied in the both optimization community and information geometry community information projections (Tsuda et al., 2003).

However, the exponential geodesics are preserved only for submanifolds $\mathcal{N}_\mu$ of $\mathcal{N}$ with fixed mean μ. Thus $\overline{\mathcal{N}}_\mu$ preserve both mixture and exponential geodesics: The submanifolds $\overline{\mathcal{N}}_\mu$ are said to be *doubly auto-parallel* (Ohara, 2019).

Online materials are available at `https://franknielsen.github.io/FisherRaoMVN/`

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Alabdulsada, L. M. and Kozma, L. On non-positive curvature properties of the hilbert metric. *The Journal of Geometric Analysis*, 29:569–576, 2019.

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

Alman, J. and Williams, V. V. A refined laser method and faster matrix multiplication. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 522–539. SIAM, 2021.

Amari, S.-i. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016. ISBN 9784431559771.

Andoni, A. *Nearest neighbor search: The old, the new, and the impossible*. PhD thesis, Massachusetts Institute of Technology, 2009.

Arnaudon, M. and Nielsen, F. On approximating the Riemannian 1-center. *Computational Geometry*, 46(1):93–104, 2013.

Baggio, G., Ferrante, A., and Sepulchre, R. Conal distances between rational spectral densities. *IEEE Transactions on Automatic Control*, 64(5):1848–1857, 2018.

Ballard, G., Kolda, T., and Plantenga, T. Efficiently computing tensor eigenvalues on a GPU. In *2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum*, pp. 1340–1348. IEEE, 2011.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2011.

Barbaresco, F. Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Fréchet median. In *Matrix information geometry*, pp. 199–255. Springer, 2013.

Bhatia, N. et al. Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085*, 2010.

Bhatia, R. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

Birkhoff, G. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

Bridson, M. R. and Haefliger, A. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Burbea, J. and Rao, C. R. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12 (4):575–596, 1982.

Cabanes, Y. and Nielsen, F. Classification in the Siegel Space for Vectorial Autoregressive Data. In *5th International Conference on Geometric Science of Information (GSI)*, pp. 693–700. Springer, 2021.

Calin, O. and Udrişte, C. *Geometric modeling in probability and statistics*, volume 121. Springer, 2014.

Calvo, M. and Oller, J. M. A distance between multivariate normal distributions based in an embedding into the Siegel group. *Journal of multivariate analysis*, 35(2): 223–242, 1990.

Calvo, M. and Oller, J. M. An explicit solution of information geodesic equations for the multivariate normal model. *Statistics & Risk Modeling*, 9(1-2):119–138, 1991.

Carson, C., Belongie, S., Greenspan, H., and Malik, J. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on pattern analysis and machine intelligence*, 24(8):1026–1038, 2002.

Chebbi, Z. and Moakher, M. Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function. *Linear Algebra and its Applications*, 436(7):1872–1889, 2012.

Chen, Y., Georgiou, T. T., and Pavon, M. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrodinger bridge. *Siam Review*, 63(2):249–313, 2021.

Cheng, G., Ho, J., Salehian, H., and Vemuri, B. C. Recursive computation of the Fréchet mean on non-positively curved Riemannian manifolds with applications. *Riemannian Computing in Computer Vision*, pp. 21–43, 2016.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to algorithms*. MIT press, 2022.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Davis, J. and Dhillon, I. Differential entropic clustering of multivariate Gaussians. *Advances in Neural Information Processing Systems*, 19, 2006.

Dolcetti, A. and Pertici, D. Elliptic isometries of the manifold of positive definite real matrices with the trace metric. *Rendiconti del Circolo Matematico di Palermo Series 2*, 70(1):575–592, 2021.

Dueck, D. and Frey, B. J. Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Eriksen, P. S. Geodesics connected with the Fisher metric on the multivariate normal manifold. Technical Report 86-13, Institute of Electronic Systems, Aalborg University Centre, Denmark, 1986.

Friedland, S. and Freitas, P. J. Revisiting the Siegel upper half plane I. *Linear algebra and its applications*, 376: 19–44, 2004.

Furuhata, H., Inoguchi, J.-i., and Kobayashi, S. A characterization of the alpha-connections on the statistical manifold of normal distributions. *Information Geometry*, 4(1):177–188, 2021.

Gao, Y. and Chaudhari, P. An information-geometric distance on the space of tasks. In *International Conference on Machine Learning*, pp. 3553–3563. PMLR, 2021.

Godinho, L. and Natário, J. An introduction to Riemannian geometry. *With Applications*, 2012.

Goldberger, J., Greenspan, H. K., and Dreyfuss, J. Simplifying mixture models using the unscented transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1496–1502, 2008.

Gonzalez, T. F. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38: 293–306, 1985.

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. R. Annealing between distributions by averaging moments. *Advances in Neural Information Processing Systems*, 26, 2013.

Han, M. and Park, F. C. DTI segmentation and fiber tracking using metrics on multivariate normal distributions. *Journal of mathematical imaging and vision*, 49:317–334, 2014.

Hero, A. O., Ma, B., Michel, O., and Gorman, J. Alpha-divergence for classification, indexing and retrieval. *Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich*, 2001.

Hilbert, D. Über die Gerade Linie als Kurzeste Verbindung Zweier Punkte. *Mathematische Annalen*, 46(1):91–96, 1895.

Hosseini, R. and Sra, S. Matrix manifold optimization for Gaussian mixtures. *Advances in Neural Information Processing Systems*, 28, 2015.

Hotelling, H. Spaces of statistical parameters. *Bull. Amer. Math. Soc*, 36:191, 1930.

Imai, T., Takaesu, A., and Wakayama, M. Remarks on geodesics for multivariate normal models. Technical report, Faculty of Mathematics, Kyushu University, 2011.

James, A. T. The variance information manifold and the functions on it. In *Multivariate Analysis–III*, pp. 157–169. Elsevier, 1973.

Karlsson, A. and Noskov, G. A. *The Hilbert metric and Gromov hyperbolicity*. Sonderforschungsbereich 343, 2000.

Kaufman, L. Partitioning around medoids (program PAM). *Finding groups in data*, 344:68–125, 1990.

Kobayashi, S. Geodesics of multivariate normal distributions and a Toda lattice type Lax pair. *arXiv preprint arXiv:2304.12575*, 2023.

Kwon, J., Lee, K. M., and Park, F. C. Visual tracking via geometric particle filtering on the affine group with optimal importance functions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 991–998. IEEE, 2009.

Lang, S. and Lang, S. Bruhat-Tits spaces. *Math Talks for Undergraduates*, pp. 79–100, 1999.

Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Moakher, M. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 26(3): 735–747, 2005.

Nakamura, Y. Algorithms associated with arithmetic, geometric and harmonic means and integrable systems. *Journal of computational and applied mathematics*, 131(1-2): 161–174, 2001.

Nielsen, F. A Simple Approximation Method for the Fisher–Rao Distance between Multivariate Normal Distributions. *Entropy*, 25(4):654, 2023.

Nielsen, F. and Hadjeres, G. Monte Carlo information-geometric structures. *Geometric Structures of Information*, pp. 69–103, 2019.

Nielsen, F., Piro, P., and Barlaud, M. Bregman vantage point trees for efficient nearest neighbor queries. In *2009 IEEE International Conference on Multimedia and Expo*, pp. 878–881. IEEE, 2009.

Nussbaum, R. D. Finsler structures for the part metric and Hilbert's projective metric and applications to ordinary differential equations. *Differential and Integral Equations*, 7(6):1649–1707, 1994.

Ohara, A. Doubly autoparallel structure on positive definite matrices and its applications. In *International Conference on Geometric Science of Information*, pp. 251–260. Springer, 2019.

Pilté, M. and Barbaresco, F. Tracking quality monitoring based on information geometry and geodesic shooting. In *2016 17th International Radar Symposium (IRS)*, pp. 1–6. IEEE, 2016.

Pinele, J., Strapasson, J. E., and Costa, S. I. R. The Fisher–Rao distance between multivariate normal distributions: Special cases, bounds and applications. *Entropy*, 22(4): 404, 2020.

Porikli, F., Tuzel, O., and Meer, P. Covariance tracking using model update based on Lie algebra. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pp. 728–735. IEEE, 2006.

Rao, C. R. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

Rong, Y., Tang, M., and Zhou, J. Intrinsic losses based on information geometry and their applications. *Entropy*, 19 (8):405, 2017.

Siegel, C. L. *Symplectic geometry*. Academic Press, 1964.

Skovgaard, L. T. A Riemannian geometry of the multivariate normal model. *Scandinavian journal of statistics*, pp. 211–223, 1984.

Sra, S. Positive definite matrices and the S-divergence. *Proceedings of the American Mathematical Society*, 144 (7):2787–2797, 2016.

Strapasson, J. E., Pinele, J., and Costa, S. I. Clustering using the Fisher-Rao distance. In *2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 1–5. IEEE, 2016.

Takatsu, A. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005 – 1026, 2011.

Trevisan, L. Lecture notes on graph partitioning, expanders and spectral methods. *University of California, Berkeley*, 2017. URL https://people.eecs.berkeley.edu/luca/books/expanders-2016.pdf.

Tsuda, K., Akaho, S., and Asai, K. The em algorithm for kernel matrix completion with auxiliary data. *The Journal of Machine Learning Research*, 4:67–81, 2003.

Tuzel, O., Porikli, F., and Meer, P. Pedestrian detection via classification on Riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 30 (10):1713–1727, 2008.

Welzl, E. Smallest enclosing disks (balls and ellipsoids). In *New Results and New Trends in Computer Science*, pp. 359–370. Springer, 2005.

Yianilos, P. N. Data structures and algorithms for nearest neighbor. In *Proceedings of the ACM-SIAM Symposium on Discrete algorithms*, volume 66, pp. 311, 1993.

Yoshizawa, S. and Tanabe, K. Dual differential geometry associated with the Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT Journal of Mathematics*, 35(1):113–137, 1999.

Yoshizawa, T. A geometry of parameter space and it's statistical interpretation. *Kokyuroku*, pp. 103–131, 1972.

Zhang, K. and Kwok, J. T. Simplifying mixture models through function approximation. *IEEE Transactions on Neural Networks*, 21(4):644–658, 2010.

A. Fisher-Rao geodesics between MVNs with initial value conditions

The Fisher-Rao geodesics $\gamma(t)$ are smooth curves which are autoparallel with respect to the Levi-Civita connection ∇^g induced by the metric tensor g: $\nabla^g_{\dot\gamma}\dot\gamma = 0$. On the MVN manifold, the system of Riemannian geodesic equations (Skovgaard, 1984) is

$$\begin{cases} \ddot\mu - \dot\Sigma\Sigma^{-1}\dot\mu & = & 0, \\ \ddot\Sigma + \dot\mu\dot\mu^\top - \dot\Sigma\Sigma^{-1}\dot\Sigma & = & 0. \end{cases}$$

We may solve the above differential equation system either using initial value conditions (IVPs) by prescribing $N_0 = (\mu(0), \Sigma(0))$ and a tangent vector $\dot N(0) = (\dot\mu(0), \dot\Sigma(0))$, or with boundary value conditions (BVPs) by prescribing $N_0 = (\mu(0), \Sigma(0))$ and $N_1 = (\mu(1), \Sigma(1))$. Let $\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_0, \dot N_0; t)$ and $\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_0, N_0 1; t)$ denote these two types of geodesics.

Without loss of generality, let us assume $N_0 = N(0, I)$ (standard normal distribution). The task is to perform geodesic shooting, i.e., calculate $N(t) = \gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_0, v_0; t)$ with some prescribed initial condition $(v, S) = \dot\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_0, v_0; 0) \in T_{N\mathrm{std}}\mathcal{M}$ and $t \geq 0$. We report the solution given in (Calvo & Oller, 1991) which relies on the following natural parameterization of the normal distributions

$$\left(\xi = \Sigma^{-1}\mu, \Xi = \Sigma^{-1}\right).$$

The initial conditions are given by $(a = \dot\xi(0), B = \dot\Xi(0)) = \dot\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_0, v_0; 0)$.

The method of (Calvo & Oller, 1991) first calculate those quantities:

$$\begin{aligned} B &= -\Xi(0)^{-\frac{1}{2}}\, \dot\Xi(0)\, \Xi(0)^{-\frac{1}{2}}, \\ a &= \Xi(0)^{-\frac{1}{2}}\dot\xi(0) + B\Xi_0^{-\frac{1}{2}}\xi(0), \\ G &= (B^2 + 2aa^\top)^{\frac{1}{2}}. \end{aligned}$$

Furthermore, let $G^\dagger = G^{-1}$ when G is invertible or $G^\dagger = (G^\top G)^{-1} G^\top$ the Moore-Penrose generalized pseudo-inverse matrix of G otherwise (or any kind of generalized matrix inverse G^- (Calvo & Oller, 1991), see).

Then we have $(\xi(t), \Xi(t)) = \gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_0, v_0; t)$ with

$$\begin{aligned} \Xi(t) &= \Xi(0)^{\frac{1}{2}}\, R(t) R(t)^\top\, \Xi(0)^{\frac{1}{2}}, \\ \xi(t) &= 2\Xi(0)^{\frac{1}{2}}\, R(t) \mathrm{Sinh}\left(\frac{1}{2}Gt\right) G^\dagger a + \Xi(t)\Xi^{-1}(0)\xi(0), \end{aligned}$$

and

$$R(t) = \mathrm{Cosh}\left(\frac{1}{2}Gt\right) - BG^\dagger \mathrm{Sinh}\left(\frac{1}{2}Gt\right).$$

The matrix hyperbolic cosine and sinus functions of M are calculated from the eigen decomposition of $M = O\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\, O^\top$ as follows:

$$\begin{aligned} \mathrm{Sinh}(M) &= O\,\mathrm{diag}(\sinh(\lambda_1), \ldots, \sinh(\lambda_d))\, O^\top, \quad \sinh(u) = \frac{e^u - e^{-u}}{2} = \sum_{i=0}^{\infty} \frac{u^{2i+1}}{(2i+1)!}, \\ \mathrm{Cosh}(M) &= O\,\mathrm{diag}(\cosh(\lambda_1), \ldots, \cosh(\lambda_d))\, O^\top, \quad \cosh(u) = \frac{e^u + e^{-u}}{2} = \sum_{i=0}^{\infty} \frac{u^{2i}}{(2i)!}. \end{aligned}$$

For the general case $\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N, v_0; t)$ with arbitrary $N = (\Sigma, \mu)$, we use the affine equivariance property of the Fisher-Rao geodesics with $P = \Sigma^{-\frac{1}{2}}$:

$$\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N, v_0; t) = (-P\mu, P^{-1}).\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_{\mathrm{std}}, (Pa, -PBP^\top); t). \tag{16}$$

Figure 6 displays several examples of geodesics from the standard normal distribution with various initial value conditions.

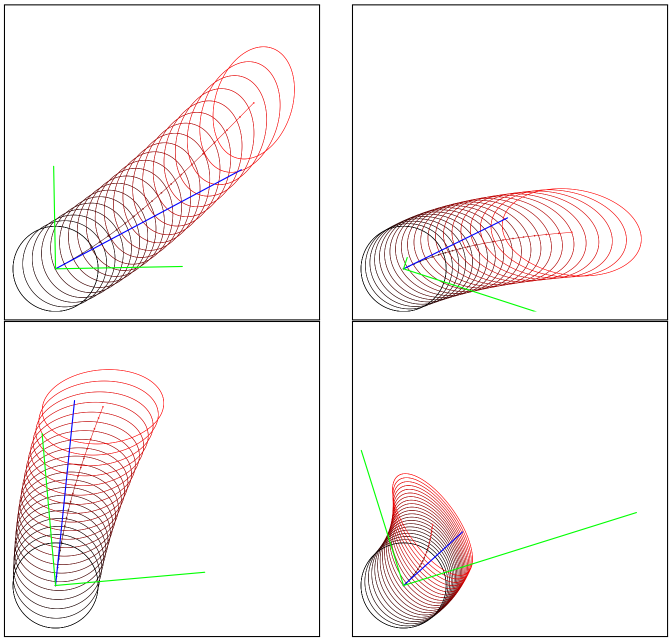


Figure 6. Examples of Fisher-Rao geodesics $(\mu(t), \Sigma(t))$ emanating from the standard bivariate normal distribution $(\mu(0), \Sigma(0)) = N(0, I)$ with initial value conditions $(\dot{\mu}(0), \dot{\Sigma}(0))$. Vectors $\dot{\mu}(0)$ are shown in blue and symmetric matrices $\dot{\Sigma}(0) = \lambda_1 v_1 v_1^\top + \lambda_2 v_2 v_2^\top$ are visualized by their two scaled eigenvectors $\lambda_1 v_1$ and $\lambda_2 v_2$ shown in green.

The geodesics with initial values let us define the Riemannian exponential map $\exp : T_N\mathcal{M} \rightarrow \mathcal{M}$:

$$\exp_N(v) = \gamma_{\mathcal{N}}^{\text{Fisher}}(N, v; 1).$$

The inverse map is the Riemannian logarithm map. In a geodesically complete manifold (e.g., $\mathcal{N}_\mu$), we can express the Fisher-Rao distance as:

$$\rho_{\mathcal{N}}(N_1, N_2) = \|\mathrm{Log}_{N_1}(N_2)\|_{N_1}.$$

Thus computing the Fisher-Rao distance can be done by computing the Riemannian MVN logarithm.

B. Proof of square root of Jeffreys upper bound

Let us prove that the Fisher-Rao distance between normal distributions is upper bounded by the square root of the Jeffreys divergence:

$$\rho_{\mathcal{N}}(N_1, N_2) \leq \sqrt{D_J(N_1, N_2)}.$$

Property 2. We have

$$D_J[p_{\lambda_1}, p_{\lambda_2}] = \int_0^1 \mathrm{d}s^2_{\mathcal{N}}(\gamma^m_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t))\mathrm{d}t = \int_0^1 \mathrm{d}s^2_{\mathcal{N}}(\gamma^e_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t))\mathrm{d}t.$$

Let $S_F(\theta_1; \theta_2) = B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1)$ be a symmetrized Bregman divergence. Let $\mathrm{d}s^2 = \mathrm{d}\theta^\top \nabla^2 F(\theta)\mathrm{d}\theta$ denote the squared length element on the Bregman manifold and denote by $\gamma(t)$ and $\gamma^*(t)$ the dual geodesics connecting θ_1 to θ_2. We can express $S_F(\theta_1; \theta_2)$ as integral energies on dual geodesics:

Property 3. We have $S_F(\theta_1; \theta_2) = \int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = \int_0^1 \mathrm{d}s^2(\gamma^*(t))\mathrm{d}t$.

Proof. The proof that the symmetrized Bregman divergence amount to these energy integrals is based on the first-order and second-order directional derivatives. The first-order directional derivative $\nabla_u F(\theta)$ with respect to vector u is defined by

$$\nabla_u F(\theta) = \lim_{t \to 0} \frac{F(\theta + tv) - F(\theta)}{t} = v^\top \nabla F(\theta).$$

The second-order directional derivatives $\nabla^2_{u,v}F(\theta)$ is

$$\begin{aligned}\nabla^2_{u,v}F(\theta) &= \nabla_u\nabla_v F(\theta),\\ &= \lim_{t\rightarrow 0}\frac{v^\top\nabla F(\theta+tu)-v^\top\nabla F(\theta)}{t},\\ &= u^\top\nabla^2F(\theta)v.\end{aligned}$$

Now consider the squared length element $\mathrm{d}s^2(\gamma(t))$ on the primal geodesic $\gamma(t)$ expressed using the primal coordinate system θ: $\mathrm{d}s^2(\gamma(t))=\mathrm{d}\theta(t)^\top\nabla^2F(\theta(t))\mathrm{d}\theta(t)$ with $\theta(\gamma(t))=\theta_1+t(\theta_2-\theta_1)$ and $\mathrm{d}\theta(t)=\theta_2-\theta_1$. Let us express the $\mathrm{d}s^2(\gamma(t))$ using the second-order directional derivative:

$$\mathrm{d}s^2(\gamma(t))=\nabla^2_{\theta_2-\theta_1}F(\theta(t)).$$

Thus we have $\int_0^1\mathrm{d}s^2(\gamma(t))\mathrm{d}t=[\nabla_{\theta_2-\theta_1}F(\theta(t))]_0^1$, where the first-order directional derivative is $\nabla_{\theta_2-\theta_1}F(\theta(t))=(\theta_2-\theta_1)^\top\nabla F(\theta(t))$. Therefore we get $\int_0^1\mathrm{d}s^2(\gamma(t))\mathrm{d}t=(\theta_2-\theta_1)^\top(\nabla F(\theta_2)-\nabla F(\theta_1))=S_F(\theta_1;\theta_2)$.

Similarly, we express the squared length element $\mathrm{d}s^2(\gamma^*(t))$ using the dual coordinate system η as the second-order directional derivative of $F^*(\eta(t))$ with $\eta(\gamma^*(t))=\eta_1+t(\eta_2-\eta_1)$:

$$\mathrm{d}s^2(\gamma^*(t))=\nabla^2_{\eta_2-\eta_1}F^*(\eta(t)).$$

Therefore, we have $\int_0^1\mathrm{d}s^2(\gamma^*(t))\mathrm{d}t=[\nabla_{\eta_2-\eta_1}F^*(\eta(t))]_0^1=S_{F^*}(\eta_1;\eta 2)$. Since $S_{F^*}(\eta_1;\eta_2)=S_F(\theta_1;\theta_2)$, we conclude that

$$S_F(\theta_1;\theta_2)=\int_0^1\mathrm{d}s^2(\gamma(t))\mathrm{d}t=\int_0^1\mathrm{d}s^2(\gamma^*(t))\mathrm{d}t$$

Note that in 1D, both pregeodesics $\gamma(t)$ and $\gamma^*(t)$ coincide. We have $\mathrm{d}s^2(t)=(\theta_2-\theta_1)^2f''(\theta(t))=(\eta_2-\eta_1)f^{*''}(\eta(t))$ so that we check that $S_F(\theta_1;\theta_2)=\int_0^1\mathrm{d}s^2(\gamma(t))\mathrm{d}t=(\theta_2-\theta_1)[f'(\theta(t))]_0^1=(\eta_2-\eta_1)[f^{*'}(\eta(t))]_0^1=(\eta_2-\eta_1)(\theta_2-\theta_2)$. □

Property 4 ((Amari, 2016)). We have

$$D_J[p_{\lambda_1},p_{\lambda_2}]=\int_0^1\mathrm{d}s^2_{\mathcal{N}}(\gamma^m_{\mathcal{N}}(p_{\lambda_1},p_{\lambda_2};t))\mathrm{d}t=\int_0^1\mathrm{d}s^2_{\mathcal{N}}(\gamma^e_{\mathcal{N}}(p_{\lambda_1},p_{\lambda_2};t))\mathrm{d}t.$$

Proof. Let us report a proof of this remarkable fact in the general setting of Bregman manifolds. Indeed, since

$$D_J[p_{\lambda_1},p_{\lambda_2}]=D_{\mathrm{KL}}[p_{\lambda_1}:p_{\lambda_2}]+D_{\mathrm{KL}}[p_{\lambda_2}:p_{\lambda_1}],$$

and $D_{\mathrm{KL}}[p_{\lambda_1}:p_{\lambda_2}]=B_F(\theta(\lambda_2):\theta(\lambda_1))$, where B_F denotes the Bregman divergence induced by the cumulant function of the multivariate normals and $\theta(\lambda)$ is the natural parameter corresponding to λ, we have

$$\begin{aligned}D_J[p_{\lambda_1},p_{\lambda_2}] &= B_F(\theta_1:\theta_2)+B_F(\theta_2:\theta_1),\\ &= S_F(\theta_1;\theta_2)=(\theta_2-\theta_1)^\top(\eta_2-\eta_1)=S_{F^*}(\eta_1;\eta_2),\end{aligned}$$

where $\eta=\nabla F(\theta)$ and $\theta=\nabla F^*(\eta)$ denote the dual parameterizations obtained by the Legendre-Fenchel convex conjugate $F^*(\eta)$ of $F(\theta)$. Moreover, we have $F^*(\eta)=-h(p_{\mu,\Sigma})$ (Amari, 2016), i.e., the convex conjugate function is Shannon negentropy.

Then we conclude by using the fact that $S_F(\theta_1;\theta_2)=\int_0^1\mathrm{d}s^2(\gamma(t))\mathrm{d}t=\int_0^1\mathrm{d}s^2(\gamma^*(t))\mathrm{d}t$, i.e., the symmetrized Bregman divergence amounts to integral energies on dual geodesics on a Bregman manifold. The proof of this general property is reported in Appendix B. □

Property 5 (Fisher–Rao upper bound). The Fisher-Rao distance between normal distributions is upper bounded by the square root of the Jeffreys divergence: $\rho_{\mathcal{N}}(N_1,N_2)\leq\sqrt{D_J(N_1,N_2)}$.

Proof. Consider the Cauchy-Schwarz inequality for positive functions $f(t)$ and $g(t)$: $\int_0^1 f(t)g(t)\mathrm{d}t \leq \sqrt{(\int_0^1 f(t)^2\mathrm{d}t)(\int_0^1 g(t)^2\mathrm{d}t)}$, and let $f(t) = \mathrm{d}s_{\mathcal{N}}(\gamma^c_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)$ and $g(t) = 1$. Then we get:

$$\left(\int_0^1 \mathrm{d}s_{\mathcal{N}}(\gamma^c_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)\mathrm{d}t\right)^2 \leq \left(\int_0^1 \mathrm{d}s^2_{\mathcal{N}}(\gamma^c_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)\mathrm{d}t\right)\left(\underbrace{\int_0^1 1^2\mathrm{d}t}_{=1}\right).$$

Furthermore since by definition of $\gamma^{\mathrm{FR}}_{\mathcal{N}}$, we have

$$\int_0^1 \mathrm{d}s_{\mathcal{N}}(\gamma^c_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)\mathrm{d}t \geq \int_0^1 \mathrm{d}s_{\mathcal{N}}(\gamma^{\mathrm{FR}}_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)\mathrm{d}t =: \rho_{\mathcal{N}}(N_1, N_2).$$

It follows for $c = \gamma^e_{\mathcal{N}}$ (i.e., e-geodesic) using Property 4 that we have:

$$\rho_{\mathcal{N}}(N_1, N_2)^2 \leq \int_0^1 \mathrm{d}s^2_{\mathcal{N}}(\gamma^e_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)\mathrm{d}t = D_J(N_1, N_2).$$

Thus we conclude that $\rho_{\mathcal{N}}(N_1, N_2) \leq \sqrt{D_J(N_1, N_2)}$.

Note that in Riemannian geometry, a curve γ minimizes the energy $E(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|^2\mathrm{d}t$ if it minimizes the length $\mathrm{Len}(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|\mathrm{d}t$ and $\|\dot{\gamma}(t)\|$ is constant. Using Cauchy-Schwartz inequality, we can show that $\mathrm{Len}(\gamma) \leq E(\gamma)$. □

C. Riemannian SPD geodesic and the arithmetic-harmonic inductive mean

The Riemannian SPD geodesic $\gamma(X, Y; t)$ joining two SPD matrices X and Y with respect to the trace metric can be expressed using the weighted matrix geometric mean:

$$\gamma(X, Y; t) = X \#_t Y = X^{\frac{1}{2}} \left(X^{-\frac{1}{2}}\, Y\, X^{-\frac{1}{2}}\right)^t X^{\frac{1}{2}}. \tag{17}$$

We denote by $X \# Y = X \#_{\frac{1}{2}} Y =$.

The matrix geometric mean can be computed inductively using the following arithmetic-harmonic sequence:

$$\begin{aligned} A_{t+1} &= A(A_t, H_t), \\ H_{t+1} &= H(A_t, H_t), \end{aligned}$$

where the matrix arithmetic mean is $A(X, Y) = \frac{X+Y}{2}$ and the matrix harmonic mean is $H(X, Y) = 2(X^{-1} + Y^{-1})^{-1}$. The sequence is initialized with $A_0 = X$ and $H_0 = Y$. We have $\mathrm{AHM}(X, Y) = \lim_{t\to\infty} A_t = \lim_{t\to\infty} H_t = X \#_{\frac{1}{2}} Y$, and the convergence is of quadratic order (Nakamura, 2001). This iterative method converges to $X \# Y$, the non-weighted matrix geometric mean. In general, taking weighted arithmetic and harmonic means $A(X, Y) = (1-\alpha)X + \alpha Y$ and $H(X, Y) = ((1-\alpha)X^{-1} + \alpha Y^{-1})^{-1}$ yields convergence to a matrix which is not the weighted geometric mean $X \#_\alpha Y$ (except when $\alpha = \frac{1}{2}$. The method requires to compute the matrix harmonic mean which requires to inverse matrices. The closed-form formula of the matrix weighted geometric mean of Eq. 17 requires to compute a matrix fractional power which can be done from a matrix eigen decomposition.

D. Power method to approximate the largest and smallest eigenvalues

We concisely recall the power method and its computational complexity to approximate the largest eigenvalue λ_1 of a $d \times d$ symmetric positive-definite matrix P following (Trevisan, 2017):

- Pick uniformly at random $x^{(0)} \in \{-1, 1\}^d$
- For $t \in (1, \ldots, T)$ do $x^{(t)} \leftarrow P\, x^{(t-1)}$

- Return $\tilde{\lambda}_1 = \frac{\langle x^{(T)}, Px^{(T)} \rangle}{\langle x^{(T)}, x^{(T)} \rangle}$, where $\langle x, y \rangle = x^\top y$ denotes the dot product.

The complexity of the power method with T iterations is $O(T(d + m))$ where $m = O(d^2)$ is the number of non-zero entries of P. Furthermore, with probability $\geq \frac{3}{16}$, the iterative power method with $T = O\left(\frac{d}{\epsilon}d\right)$ iterations yields $\tilde{\lambda}_1 \geq (1 - \epsilon)\lambda_1$ for any $\epsilon > 0$ (Trevisan, 2017). Due to its vector-matrix product operations, the power method can be efficiently implemented on GPU (Ballard et al., 2011).

To compute an approximation $\tilde{\lambda}_d$ of the smallest eigenvalues λ_d of P, we first compute the matrix inverse P^{-1} and then compute the approximation of the largest eigenvalue of P^{-1}. We report $\tilde{\lambda}_d(P) = \tilde{\lambda}_1(P^{-1})$. The complexity of computing a matrix inverse is as hard as computing the matrix product (Cormen et al., 2022). The current best algorithm requires $O(d^\omega)$ operations with $\omega = 2.373$ (Alman & Williams, 2021).