entropy

MDPI

Article

A Simple Approximation Method for the Fisher–Rao Distance between Multivariate Normal Distributions

Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan; frank.nielsen.x@gmail.com

Abstract: We present a simple method to approximate the Fisher–Rao distance between multivariate normal distributions based on discretizing curves joining normal distributions and approximating the Fisher–Rao distances between successive nearby normal distributions on the curves by the square roots of their Jeffreys divergences. We consider experimentally the linear interpolation curves in the ordinary, natural, and expectation parameterizations of the normal distributions, and compare these curves with a curve derived from the Calvo and Oller's isometric embedding of the Fisher–Rao d-variate normal manifold into the cone of $(d+1) \times (d+1)$ symmetric positive–definite matrices. We report on our experiments and assess the quality of our approximation technique by comparing the numerical approximations with both lower and upper bounds. Finally, we present several information–geometric properties of Calvo and Oller's isometric embedding.

Keywords: Fisher–Rao normal manifold; symmetric positive–definite matrix cone; isometric embedding; information geometry; exponential family; elliptical distribution; maximal invariant

Citation: Nielsen, F. A Simple Approximation Method for the Fisher–Rao Distance between Multivariate Normal Distributions. *Entropy* **2023**, *25*, 654. https://doi.org/10.3390/e25040654

Academic Editor: Antonio M. Scarfone

Received: 27 February 2023
Revised: 6 April 2023
Accepted: 12 April 2023
Published: 13 April 2023

1. Introduction

1.1. The Fisher–Rao Normal Manifold

Let $\mathrm{Sym}(d)$ be the set of $d \times d$ symmetric matrices with real entries and $\mathbb{P}(d) \subset \mathrm{Sym}(d)$ denote the set of symmetric positive–definite $d \times d$ matrices that forms a convex regular cone. Let us denote by $\mathcal{N}(d) = \{N(\mu, \Sigma) \ : \ (\mu, \Sigma) \in \Lambda(d) = \mathbb{R}^d \times \mathbb{P}(d)\}$ the set of d-variate normal distributions, MultiVariate Normals or MVNs for short, also called Gaussian distributions. A MVN distribution $N(\mu, \Sigma)$ has probability density function (pdf) on the support $\mathbb{R}^d$:

$$p_{\lambda=(\mu,\Sigma)}(x) = (2\pi)^{\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \quad x \in \mathbb{R}^d,$$

where $|M| = \det(M)$ denotes the determinant of matrix M.

The statistical model $\mathcal{N}(d)$ is of dimension $m = \dim(\Lambda(d)) = d + \frac{d(d+1)}{2} = \frac{d(d+3)}{2}$ since it is identifiable, i.e., there is a one-to-one correspondence $\lambda \leftrightarrow p_\lambda(x)$ between $\lambda \in \Lambda(d)$ and $N(\mu, \Sigma) \in \mathcal{N}(d)$. The statistical model $\mathcal{N}(d)$ is said to be regular since the second-order derivatives $\frac{\partial^2 p_\lambda}{\partial \lambda_i \partial \lambda_j}$ and third-order derivatives $\frac{\partial^3 p_\lambda}{\partial \lambda_i \partial \lambda_j \partial \lambda_k}$ are smooth functions (defining the metric and cubic tensors in information geometry [1]), and the set of first-order partial derivatives $\left\{\frac{\partial p_\lambda}{\partial \lambda_1}, \ldots, \frac{\partial p_\lambda}{\partial \lambda_1}\right\}$ are linearly independent.

Let $\mathrm{Cov}(X)$ denote the covariance of X (variance when X is scalar). A matrix M is a semi-positive–definite if and only if $\forall x \neq 0, x^\top M x \geq 0$. The Fisher information matrix [1,2] (FIM) is the following symmetric semi-positive–definite matrix:

$$I(\lambda) = \mathrm{Cov}[\nabla \log p_\lambda(x)] \succeq 0.$$

For regular statistical models $\{p_\lambda\}$, the FIM is positive–definite: $I(\lambda) \succ 0$, i.e., $\forall x \neq 0$, $x^\top I(\lambda) x > 0$. $M_1 \succ M_2$ denotes Löwner partial ordering, i.e., the fact that $M_1 - M_2$ is positive–definite.

The FIM is covariant under the reparameterization of the statistical model [2]. That is, let $\theta(\lambda)$ be a new parameterization of the MVNs. Then we have:

$$I_\theta(\lambda) = \left[\frac{\partial \lambda}{\partial \theta}\right]^\top \times I_\lambda(\lambda(\theta)) \times \left[\frac{\partial \lambda}{\partial \theta}\right].$$

For example, we may parameterize univariate normal distributions by $\lambda = (\mu, \sigma^2)$ or $\theta = (\mu, \sigma)$. We obtain the following Fisher information matrices for these parameterizations:

$$I_\lambda(\lambda(\mu,\sigma)) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \quad \text{and} \quad I_\theta(\theta(\mu,\sigma)) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^2} \end{bmatrix}.$$

In higher dimensions, parameterization $\lambda = (\mu, \sigma^2)$ corresponds to the parameterization (μ, Σ) while parameterization $\theta = (\mu, L)$ where $\Sigma = LL^\top$ is the unique Cholesky decomposition with $L \in \mathrm{GL}(d)$, the group of invertible $d \times d$ matrices. Another useful parameterization for optimization is the log–Cholesky parameterization [3] ($\eta = (\mu, \log \sigma^2) \in \mathbb{R}^2$ for univariate normal distributions) which ensures that a gradient descent always stays in the domain. The Fisher information matrix with respect to the log–Cholesky parameterization is $I_\eta(\eta(\mu,\sigma)) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 2 \end{bmatrix}$ with $\eta(\mu,\sigma) \in \mathbb{R}^2$.

Since the statistical model $\mathcal{N}(d)$ is identifiable and regular, the Fisher information matrix can be written equivalently as follows [2,4]:

$$\begin{aligned} I(\mu,\Sigma) = \mathrm{Cov}[\nabla \log p_{(\mu,\Sigma)}] &= E_{p_{(\mu,\Sigma)}}\left[\nabla \log p_{(\mu,\Sigma)} \nabla \log p_{(\mu,\Sigma)}^\top\right], &(1)\\ &= -E_{p_{(\mu,\Sigma)}}\left[\nabla^2 \log p_{(\mu,\Sigma)}\right]. &(2)\end{aligned}$$

For multivariate distributions parameterized by a m-dimensional vector (with $m = \frac{d(d+3)}{2}$)

$$\theta = (\theta_1, \ldots, \theta_d, \theta_{d+1}, \ldots, \theta_m) \in \mathbb{R}^m,$$

with $\mu = (\theta_1, \ldots, \theta_d)$ and $\Sigma(\theta) = \mathrm{vech}(\theta_{d+1}, \ldots, \theta_m)$ (inverse half-vectorization of matrices [5]), we have [6–9]:

$$I(\theta) = [I_{ij}(\theta)], \quad \text{with } I_{ij}(\theta) = \left[\frac{\partial \mu}{\partial \theta_i}\right]^\top \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial \mu}{\partial \theta_i}\Sigma^{-1}\frac{\partial \mu}{\partial \theta_j}\right).$$

By equipping the regular statistical model $\mathcal{N}(d)$ with the Fisher information metric

$$g_{\mathcal{N}}^{\mathrm{Fisher}}(\mu,\Sigma) = \mathrm{Cov}[\nabla \log p_{(\mu,\Sigma)}(x)]$$

we obtain a Riemannian manifold $\mathcal{M} = \mathcal{M}_{\mathcal{N}}$ called the Fisher–Rao Gaussian or normal manifold [6,7]. The tangent space $T_N\mathcal{M}$ is identified with the product space $\mathbb{R}^d \times \mathrm{Sym}(d)$. Let $\{\partial\mu, \partial\Sigma\}$ be a natural vector basis in $T_N\mathcal{M}$, and denote by $[v]$ and $[V]$ the vector components in that natural basis. We have

$$\begin{aligned} g_{(\mu,\Sigma)}^{\mathrm{Fisher}}((v_1,V_1),(v_2,V_2)) &= \langle (v_1,V_1),(v_2,V_2)\rangle_{(\mu,\Sigma)},\\ &= [v_1]^\top \Sigma^{-1}[v_2] + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}[V_1]\Sigma^{-1}[V_2]\right).\end{aligned}$$

The induced Riemannian geodesic distance $\rho_{\mathcal{N}}(\cdot,\cdot)$ is called the Rao distance [10] or the Fisher–Rao distance [11,12]:

$$\rho_{\mathcal{N}}(N(\lambda_1), N(\lambda_2)) = \inf_{\substack{c(t) \\ c(0)=p_{\lambda_1} \\ c(1)=p_{\lambda_2}}} \{\mathrm{Length}(c)\}, \tag{3}$$

where the Riemannian length of any smooth curve $c(t) \in \mathcal{M}$ is defined by

$$\mathrm{Length}(c) = \int_0^1 \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}} \mathrm{d}t = \int_0^1 \mathrm{d}s_{\mathcal{N}}(t) \mathrm{d}t = \int_0^1 \|\dot{c}(t)\|_{c(t)} \mathrm{d}t, \tag{4}$$

where $\dot{} = \frac{d}{\mathrm{d}t}$ denotes the derivative with respect to parameter t, $\mathrm{d}s_{\mathcal{N}}(t)$ is the Riemannian length element of $(\mathcal{M}, g_{\mathcal{N}}^{\mathrm{Fisher}})$ and $\| \cdot \|_{c(t)} = \sqrt{\langle \cdot, \cdot \rangle_{c(t)}}$. We also write $\rho_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2})$ for $\rho_{\mathcal{N}}(N(\lambda_1), N(\lambda_2))$.

The minimizing curve $c(t) = \gamma_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)$ of Equation (3) is called the Fisher–Rao geodesic. The Fisher–Rao geodesic is also an autoparallel curve [2] with respect to the Levi–Civita connection $\nabla_{\mathcal{N}}^{\mathrm{Fisher}}$ induced by the Fisher metric $g_{\mathcal{N}}^{\mathrm{Fisher}}$.

Remark 1. *If we consider the Riemannian manifold $(\mathcal{M}, \beta g)$ for $\beta > 0$ instead of $(\mathcal{M}, g)$ then the length element* ds *is scaled by $\sqrt{\beta}$: $\mathrm{ds}_{\beta g} = \beta\, \mathrm{ds}_g$. It follows that the length of a curve c becomes*

$$\mathrm{Length}_{\beta g}(c) = \sqrt{\beta}\, \mathrm{Length}_g(c).$$

However, the geodesics joining any two points p_1 and p_2 of $\mathcal{M}$ are the same: $\gamma_{\beta g}(p_1, p_2; t) = \gamma_g(p_1, p_2; t)$ (with $\gamma_g(p_1, p_2; 0) = p_1$ and $\gamma_g(p_1, p_2; 1) = p_2$).

Historically, Hotelling [13] first used this Fisher Riemannian geodesic distance in the late 1920s. From the viewpoint of information geometry [1], the Fisher metric is the unique Markov invariant metric up to rescaling [14–16]. The counterpart to the Fisher metric on the compact manifold has been reported in [17], proving its uniqueness under the action of the diffeomorphism group. The Fisher–Rao distance has been used to design statistical hypothesis testing [18–21], to measure the distance between the prior and posterior distributions in Bayesian statistics [22], in clustering [23,24], in signal processing [25–28], and in deep learning [29], just to mention a few.

The squared line element induced by the Fisher metric of the multivariate normal family [6,7] is

$$\begin{aligned} \mathrm{ds}^2_{\mathcal{N}}(\mu, \Sigma) &= \begin{bmatrix} \mathrm{d}\mu \\ \mathrm{d}\Sigma \end{bmatrix}^\top I(\mu, \Sigma) \begin{bmatrix} \mathrm{d}\mu \\ \mathrm{d}\Sigma \end{bmatrix}, \\ &= \mathrm{d}\mu^\top \Sigma^{-1} \mathrm{d}\mu + \frac{1}{2} \mathrm{tr}\left(\left(\Sigma^{-1} \mathrm{d}\Sigma \right)^2 \right). \end{aligned} \tag{5}$$

There are many ways to calculate the FIM/length element for multivariate normal distributions [7,9]. Let us give a simple approach based on the fact that the family $\mathcal{N}(d)$ of normal distributions forms a regular exponential family [30]:

$$\mathcal{N}(d) = \left\{ p_{\theta(\lambda)} = \exp\left(\langle \theta_v(\mu), x \rangle + \langle \theta_M(\Sigma), x x^\top \rangle - F_{\mathcal{N}}(\theta_v, \theta_M) \right) \right\},$$

with $\theta(\lambda) = (\theta_v = (\Sigma^{-1}\mu, \theta_M = \frac{1}{2}\Sigma^{-1})$ the natural parameters and log-partition function (also called cumulant function)

$$F_{\mathcal{N}}(\theta) = \frac{1}{2} \left(d \log \pi - \log |\theta_M| + \frac{1}{2} \theta_v^\top \theta_M^{-1} \theta_v \right).$$

The vector inner product is $\langle v_1, v_2 \rangle = v_1^\top v_2$, and the matrix inner product is $\langle M_1, M_2 \rangle = \mathrm{tr}(M_1 M_2^\top)$. The exponential family is said to be regular when the natural parameter space

is open. Using Equation (2), it follows that the MVN FIM is $I_\theta(\theta) = -E[\nabla^2 \log p_\theta] = \nabla^2 F(\theta)$. This proves that the FIM is well-defined, i.e., $(I_\theta(\theta))_{ij} < \infty$. As an exponential family [1], we also have $I_\theta(\theta) = E[t(x)]$, where $t(x) = (x, xx^\top)$ is the sufficient statistic. Thus, the Fisher metric is a Hessian metric [31]. Let $F_{\mathcal{N}}(\theta_v, \theta_M) = F_v(\theta_v) + F_M(\theta_M)$ with $F_v(\theta_v) = \frac{1}{2}\left(d \log \pi + \frac{1}{2}\theta_v^\top \theta_M^{-1} \theta_v\right)$ and $F_M(\theta_M) = -\frac{1}{2}\log|\theta_M|$. We obtain the following block-diagonal expression of the FIM:

$$I(\theta(\lambda)) = \nabla^2 F_{\mathcal{N}}(\theta(\mu, \Sigma)) = \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & \frac{1}{2}\nabla^2_{\theta_M} \log|\frac{1}{2}\Sigma^{-1}| \end{bmatrix}.$$

Therefore $\mathrm{d}s^2_{\mathcal{N}}(\mu, \Sigma) = \mathrm{d}s^2_v + \mathrm{d}s^2_M$ with $\mathrm{d}s^2_v(\mu) = \mathrm{d}\mu^\top \Sigma^{-1} \mathrm{d}\mu$ and $\mathrm{d}s^2_M(\Sigma) = \frac{1}{2}\mathrm{tr}\left((\Sigma^{-1}\mathrm{d}\Sigma)^2\right)$. Let us note in passing that $\nabla^2_{\theta_M} \log|\theta_M|$ is a fourth order tensor [4].

The family $\mathcal{N}(d)$ can also be considered to be an elliptical family [32], thus highlighting the affine-invariance property of the Fisher information metric. That is, the Fisher metric is invariant with respect to affine transformations [33]: Let (a, A) be an element of the affine group $\mathrm{Aff}(d)$ with $a \in \mathbb{R}^d$ and $A \in \mathrm{GL}(d)$. The group identity element of $\mathrm{Aff}(d)$ is $e = (0, I)$ and the group operation is $(a_1, A_1).(a_2, A_2) = (a_1 + A_1 a_2, A_1 A_2)$ with inverse $(a, A)^{-1} = (-A^{-1}a, A^{-1})$). Then we have

Property 1 (Fisher–Rao affine invariance)**.** *For all $A \in \mathrm{GL}(d)$, $a \in \mathbb{R}^d$, we have*

$$\rho_{\mathcal{N}}(N(A\mu_1 + a, A\Sigma_1 A^\top), N(A\mu_2 + a, A\Sigma_2 A^\top)) = \rho_{\mathcal{N}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)). \quad (6)$$

This can be proven by checking that $\mathrm{d}s_{\mathcal{N}}(\mu', \Sigma') = \mathrm{d}s_{\mathcal{N}}(\mu, \Sigma)$ where $\mu' = A\mu + a$ and $\Sigma' = A\Sigma_2 A^\top$. It follows that we can reduce the calculation of the Fisher–Rao distance to a canonical case where one argument is $N_{\mathrm{std}} = N(0, I)$, the standard d-variate distribution:

$$\begin{aligned} \rho_{\mathcal{N}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) &= \rho_{\mathcal{N}}\left(N_{\mathrm{std}}, N\left(\Sigma_1^{-\frac{1}{2}}(\mu_2 - \mu_1), \Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}\right)\right), \\ &= \rho_{\mathcal{N}}\left(N\left(\Sigma_2^{-\frac{1}{2}}(\mu_1 - \mu_2), \Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}\right), N_{\mathrm{std}}\right), \end{aligned}$$

where Σ^p is the fractional matrix power which can be calculated from the Singular Value Decomposition $ODO^\top$ of Σ (where O is an orthogonal matrix and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ a diagonal matrix): $\Sigma^p = OD^pO^\top$ with $D^p = \mathrm{diag}(\lambda_1^p, \ldots, \lambda_d^p)$.

The family of normal elliptical distributions can be obtained from the standard normal distribution by the action of the affine group [12,32] $\mathrm{Aff}(d)$:

$$N(\mu, \Sigma) = (\mu, \Sigma^{\frac{1}{2}}).N_{\mathrm{std}} = N((\mu, \Sigma^{\frac{1}{2}}).(0, I)).$$

1.2. Fisher–Rao Distance between Normal Distributions: Some Subfamilies with Closed-Form Formula

In general, the Fisher–Rao distance $\rho_{\mathcal{N}}(N_1, N_2)$ between two multivariate normal distributions N_1 and N_2 is not known in closed form [34–37], and several lower and upper bounds [38], and numerical techniques such as the geodesic shooting [39–41] have been investigated. See [42] for a recent review. Unfortunately, the geodesic shooting (GS) approach is time-consuming and numerically unstable for large Fisher–Rao distances [21,42]. In 3D Diffusion Tensor Imaging (DTI), 3×3 covariance matrices $\Sigma_{i,j,k}$ are stored a 3D grid locations $\mu_{i,j,k}$ thus generating 3D MVNs $N_{i,j,k} = N(\mu_{i,j,k}, \Sigma_{i,j,k})$ with means $\mu_{i,j,k}$ regularly spaced to each others. The Fisher–Rao distances can be calculated between an MVN $N_{i,j,k}$ and another MVN $N_{i',j',k'}$ in a neighborhood of $N_{i,j,k}$ (using 6- or 26-neighborhood) using geodesic shooting. For larger Fisher–Rao distances between non-neighbors MVNs, we can use the shortest path distance using Dijkstra's algorithm [43] on the graph induced by the MVNs with edges between adjacent MVNs weighted by their Fisher–Rao distances.

The two main difficulties with calculating the Fisher–Rao distance are

1. to know explicitly the expression of the Riemannian Fisher–Rao geodesic $\gamma^{\mathrm{FR}}_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)$ and
2. to integrate, in closed form, the length element $\mathrm{d}s_{\mathcal{N}}$ along this Riemannian geodesic.

Please note that the Fisher–Rao geodesics [1] $\gamma^{\mathrm{FR}}_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)$ are parameterized by constant speed (i.e., $\dot{\mu}(t) = \dot{\mu}(0)$ and $\dot{\Sigma}(t) = \dot{\Sigma}(0)$), or equivalently parametrized using the arc length:

$$\rho_{\mathcal{N}}\left(\gamma^{\mathrm{FR}}_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; s), \gamma^{\mathrm{FR}}_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)\right) = |s-t|\, \rho_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}), \quad \forall s, t \in [0,1].$$

However, in several special cases, the Fisher–Rao distance between normal distributions belonging to restricted subsets of $\mathcal{N}$ is known.

Three such prominent cases are (see [42] for other cases)

1. when the normal distributions are univariate ($d = 1$),
2. when we consider the set $\mathcal{N}_\mu = \{N(\mu, \Sigma) : \Sigma \in \mathcal{P}(d)\} \subset \mathcal{M}_{\mathcal{N}}$ of normal distributions sharing the same mean μ (with the embedded submanifold $\mathcal{S}_\mu \in \mathcal{M}$), and
3. when we consider the set $\mathcal{N}_\Sigma = \{N(\mu, \Sigma) : \Sigma \in \mathcal{P}(d)\} \subset \mathcal{N}$ of normal distributions sharing the same covariance matrix Σ (with the corresponding embedded submanifold $\mathcal{S}_\Sigma \in \mathcal{M}$).

Let us report the formula of the Fisher–Rao distance in these three cases:

- In the univariate case $\mathcal{N}(1)$, the Fisher–Rao distance between $N_1 = N(\mu_1, \sigma_1^2)$ and $N_2 = N(\mu_2, \sigma_2^2)$ can be derived from the hyperbolic distance [44] expressed in the Poincaré upper space since we have

$$\mathrm{d}s^2_{\mathcal{N}} = g_{(\mu,\sigma)}(\mathrm{d}\mu, \mathrm{d}\sigma) = \frac{\mathrm{d}\mu^2 + 2\mathrm{d}\sigma^2}{\sigma^2} = 2\frac{\left(\frac{\mathrm{d}\mu}{\sqrt{2}}\right)^2 + \mathrm{d}\sigma^2}{\sigma^2} = 2\frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2} = \mathrm{d}s^2_{\text{Poincaré}},$$

where $x = \frac{\mu}{\sqrt{2}}$ and $y = \sigma$. It follows that

$$\rho_{\mathcal{N}}(N_1, N_2) = \sqrt{2}\,\rho_{\text{Poincaré}}((x_1, y_1), (x_2, y_2)) = \sqrt{2}\,\rho_{\text{Poincaré}}\left(\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right), \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right)\right).$$

Thus, we have the following expression for the Fisher–Rao distance between univariate normal distributions:

$$\rho_{\mathcal{N}}(N_1, N_2) = \sqrt{2}\log\left(\frac{1 + \Delta(\mu_1, \sigma_1; \mu_2, \sigma_2)}{1 - \Delta(\mu_1, \sigma_1; \mu_2, \sigma_2)}\right), \tag{7}$$

with

$$\Delta(a, b; c, d) = \sqrt{\frac{(c-a)^2 + 2(d-b)^2}{(c-a)^2 + 2(d+b)^2}}, \quad (a, b, c, d) \in \mathbb{R}^4 \backslash \{0\}. \tag{8}$$

In particular, we have

- $\Delta(a, b; a, d) = \left|\frac{d-b}{d+b}\right|$ when $a = c$ (same mean),
- $\Delta(a, b; c, b) = \sqrt{\frac{1}{1+8\frac{b^2}{(c-a)^2}}}$ when $b = d$ (same variance),
- $\Delta(0, 1; c, d) = \sqrt{\frac{c^2+2(d-1)^2}{c^2+2(d+1)^2}}$ when $a = 0$ and $b = 1$ (standard normal).

In 1D, the affine-invariance property (Property 1) extends to function Δ as follows:

$$\Delta(\mu_1, \sigma_1; \mu_2, \sigma_2) = \Delta\left(0, 1; \frac{\mu_2 - \mu_1}{\sigma_1}, \frac{\sigma_2}{\sigma_1}\right) = \Delta\left(\frac{\mu_1 - \mu_2}{\sigma_2}, \frac{\sigma_1}{\sigma_2}; 0, 1\right).$$

Using one of the many identities between inverse hyperbolic functions (e.g., arctanh, arccosh, arcsinh), we can obtain an equivalent formula for Equation (7). For example, since $\operatorname{arctanh}(u) := \frac{1}{2}\log\left(\frac{1+u}{1-u}\right)$ for $0 < u < 1$, we have equivalently:

$$\rho_{\mathcal{N}}(N_1, N_2) = 2\sqrt{2}\operatorname{arctanh}(\Delta(\mu_1, \sigma_1; \mu_2, \sigma_2)). \tag{9}$$

The Fisher–Rao geodesics are semi-ellipses with centers located on the x-axis. See Appendix A.1 for the parametric equations of Fisher–Rao geodesics between univariate normal distributions. Figure 1 displays four univariate normal distributions with their pairwise geodesics and Fisher–Rao distances.

Using the identity $\operatorname{arctanh}\left(\frac{u^2-1}{u^2+1}\right) = \operatorname{arccosh}\left(\frac{1+u^2}{2u}\right)$ with $\operatorname{arccosh}(x) := \log(x + \sqrt{x^2-1})$, we also have

$$\rho_{\mathcal{N}}(N_1, N_2) = 2\sqrt{2}\operatorname{arccosh}\left(\frac{1}{\sqrt{(1-\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2))(1+\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2))}}\right),$$

Since the inverse hyperbolic cosecant (CSC) function is defined by $\operatorname{arccsch}(u) := \operatorname{arccosh}(1/u)$, we further obtain

$$\rho_{\mathcal{N}}(N_1, N_2) = 2\sqrt{2}\operatorname{arccsch}\left(\sqrt{(1-\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2))(1+\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2))}\right),$$

We can also write

$$\rho_{\mathcal{N}}(N_1, N_2) = \sqrt{2}\operatorname{arccosh}\left(1 + \frac{(\mu_2-\mu_1)^2 + 2(\sigma_2-\sigma_1)^2}{4\sigma_1\sigma_2}\right)$$

Thus, using the many-conversions formula between inverse hyperbolic functions, we obtain many equivalent different formulas of the Fisher–Rao distance, which are used in the literature.

- In the second case, the Fisher–Rao distance between $N_1 = N(\mu, \Sigma_1)$ and $N_2 = N(\mu, \Sigma_2)$ has been reported in [6,7,45–47]:

$$\rho_{\mathcal{N}_\mu}(N_1, N_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{d}\log^2\lambda_i(\Sigma_1^{-1}\Sigma_2)}, \tag{10}$$

$$= \rho_{\mathcal{N}_\mu}(\Sigma_1, \Sigma_2), \tag{11}$$

where $\lambda_i(M)$ denotes the i-th generalized largest eigenvalue of matrix M, where the generalized eigenvalues are solutions of the equation $|\Sigma_1 - \lambda\Sigma_2| = 0$. Let us notice that $\rho_{\mathcal{N}_\mu}((\mu,\Sigma_1),(\mu,\Sigma_2)) = \rho_{\mathcal{N}_\mu}((\mu,\Sigma_1^{-1}),(\mu,\Sigma_2^{-1}))$ since $\lambda_i(\Sigma_2^{-1}\Sigma_1) = \frac{1}{\lambda_i(\Sigma_1^{-1}\Sigma_2)}$ and $\log^2\lambda_i(\Sigma_2^{-1}\Sigma_1) = (-\log\lambda_i(\Sigma_1^{-1}\Sigma_2))^2 = \log^2\lambda_i(\Sigma_1^{-1}\Sigma_2)$. Matrix $\Sigma_1^{-1}\Sigma_2$ may not be SPD and thus the λ_i's are generalized eigenvalues. We may consider instead the SPD matrix $\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}$ which is SPD and such that $\lambda_i(\Sigma_1^{-1}\Sigma_2) = \lambda_i(\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}})$. The Fisher–Rao distance of Equation (11) can be equivalently written [48] as

$$\rho_{\mathcal{N}_\mu}(N_1, N_2) = \frac{1}{\sqrt{2}}\left\|\operatorname{Log}\left(\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}\right)\right\|_F,$$

where $\operatorname{Log}(M)$ is the matrix logarithm (unique when M is SPD) and $\|M\|_F = \sqrt{\sum_{i,j}M_{i,j}^2} = \sqrt{\operatorname{tr}(MM^\top)}$ is the matrix Fröbenius norm. This metric distance between SPD matrices although first studied by Siegel [45] in 1964 was rediscovered and analyzed recently in [49] (2003). Let $\rho_{\mathrm{SPD}}(P_1,P_2) = \sqrt{\sum_{i=1}^{d}\log^2\lambda_i(P_1^{-1}P_2)}$ so that $\rho_{\mathcal{N}_\mu}(N(\mu,P_1),N(\mu,P_2)) = \frac{1}{\sqrt{2}}\rho_{\mathrm{SPD}}(P_1,P_2)$.

The Riemannian SPD distance ρ_{SPD} enjoys the following well-known invariance properties:

  - Invariance by congruence transformation:

$$\forall X \in \mathrm{GL}(d), \quad \rho_{\mathrm{SPD}}(XP_1X^\top, XP_2X^\top) = \rho_{\mathrm{SPD}}(P_1,P_2), \tag{12}$$

  - Invariance by inversion:

$$\forall P_1, P_2 \in \mathbb{P}(d), \quad \rho(P_1^{-1}, P_2^{-1}) = \rho_{\mathrm{SPD}}(P_1, P_2).$$

Let $P_1 = L_1 L_1^\top$ be the Cholesky decomposition (unique when $P_1 \succ 0$). Then apply the congruence invariance for $X = L_1^{-1}$:

$$\rho_{\mathrm{SPD}}(P_1, P_2) = \rho_{\mathrm{SPD}}(L_1^{-1} P_1 (L_1^{-1})^\top, L_1^{-1} P_2 (L_1^{-1})^\top) = \rho_{\mathrm{SPD}}(I, L_1^{-1} P_2 (L_1^{-1})^\top). \tag{13}$$

We can also consider the factorization $P_1 = S_1 S_1$ where $S_1 = P_1^{\frac{1}{2}}$ is the unique symmetric square root matrix [50]. Then we have

$$\rho_{\mathrm{SPD}}(P_1, P_2) = \rho_{\mathrm{SPD}}(S_1^{-1} P_1 (S_1^{-1})^\top, S_1^{-1} P_2 (S_1^{-1})^\top) = \rho_{\mathrm{SPD}}(I, S_1^{-1} P_2 (S_1^{-1})^\top).$$

- The Fisher–Rao distance between $N_1 = N(\mu_1, \Sigma)$ and $N_2 = N(\mu_2, \Sigma)$ has been reported in closed form [42] (Proposition 3). The method is described with full details in Appendix B. We present a simpler scheme based on the inverse $\Sigma^{-\frac{1}{2}}$ of the symmetric square root factorization [50] of $\Sigma = \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}}$ (ith $(\Sigma^{-\frac{1}{2}})^\top = \Sigma^{-\frac{1}{2}}$). Let us use the affine-invariance property of the Fisher–Rao distance under the affine transformation $\Sigma^{-\frac{1}{2}}$ and then apply affine invariance under translation as follows:

$$\begin{aligned}
\rho_{\mathcal{N}}(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) &= \rho_{\mathcal{N}}(N(\Sigma^{-\frac{1}{2}} \mu_1, \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}}), N(\Sigma^{-\frac{1}{2}} \mu_2, \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}})), \\
&= \rho_{\mathcal{N}}(N(0, I), N(\Sigma^{-\frac{1}{2}} (\mu_2 - \mu_1), I)), \\
&= \rho_{\mathcal{N}}(N(0, 1), N(\|\Sigma^{-\frac{1}{2}} (\mu_2 - \mu_1)\|_2, 1)).
\end{aligned}$$

The right-hand side Fisher–Rao distance is computed from Equation (7) and justified by the method [42] (Proposition 3) described in Appendix B using a rotation matrix R with $R R^\top = I$ so that

$$\begin{aligned}
\rho_{\mathcal{N}}(N(0, I), N(\Sigma^{-\frac{1}{2}} (\mu_2 - \mu_1), I)) &= \rho_{\mathcal{N}}(N(0, I), N(R \Sigma^{-\frac{1}{2}} (\mu_2 - \mu_1), R I R^\top)), \\
&= \rho_{\mathcal{N}}(N(0, I), \|\Sigma^{-\frac{1}{2}} (\mu_2 - \mu_1)\|_2, I)).
\end{aligned}$$

Then we apply the formula of Equation (23) of [42]. Section 1.5 shall report a simpler closed-form formula by proving that the Fisher–Rao distance between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ is a scalar function of their Mahalanobis distance [51] using the algebraic method of maximal invariants [52].

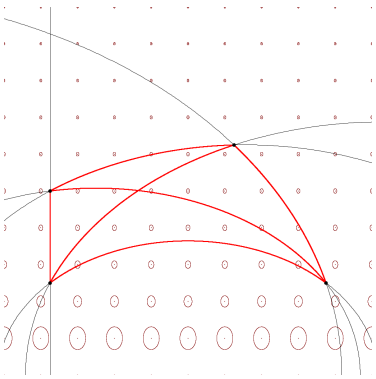


Figure 1. Four univariate normal distributions $N_1 = N(0,1)$, $N_2 = N(3,1)$, $N_3 = N(2,2.5)$, $N_4 = N(0,2)$, and their pairwise full geodesics in gray and geodesics linking them in red. The Fisher–Rao distances are $\rho_{\mathcal{N}}(N_1, N_2) = 2.6124\ldots$, $\rho_{\mathcal{N}}(N_3, N_4) = 0.9317\ldots$, $\rho_{\mathcal{N}}(N_1, N_4) = 0.9803\ldots$, $\rho_{\mathcal{N}}(N_2, N_3) = 1.4225\ldots$ $\rho_{\mathcal{N}}(N_2, N_4) = 2.1362\ldots$, and $\rho_{\mathcal{N}}(N_1, N_3) = 1.7334\ldots$ The ellipses are Tissot indicatrices, which visualize the metric tensor $g_{\mathcal{N}}^{\mathrm{Fisher}}$ at grid positions.

1.3. Fisher–Rao Distance: Totally versus Non-Totally Geodesic Submanifolds

Consider $\mathcal{N}' = \{N(\lambda) : \lambda' \in \Lambda'\} \subset \mathcal{N}$ a statistical submodel of the MVN statistical model $\mathcal{N}$. Using the Fisher information matrix $I_{\lambda'}(\lambda')$, we obtain the intrinsic Fisher–Rao manifold $\mathcal{M}' = \mathcal{M}_{\mathcal{N}'}$. We may also consider $\mathcal{M}'$ to be an embedded submanifold of $\mathcal{M}$. Let us write $\mathcal{S}' = \mathcal{S}_{\mathcal{N}'} \subset \mathcal{M}$ the embedded submanifold.

A totally geodesic submanifold $\mathcal{S}' \subset \mathcal{M}$ is such that the geodesics $\gamma_{\mathcal{M}'}(N_1', N_2'; t)$ fully stay in $\mathcal{M}'$ for any pair of points $N_1', N_2' \in \mathcal{N}'$. For example, the submanifold $\mathcal{M}_\mu = \{N(\mu, \Sigma) : \Sigma \in \mathbb{P}(d)\} \subset \mathcal{M}$ of MVNs with fixed mean μ is a totally geodesic submanifold [53] of $\mathcal{M}$ but the submanifold $\mathcal{M}_\Sigma = \{N(\mu, \Sigma) : \mu \in \mathbb{R}^d\} \subset \mathcal{M}$ of MVNs sharing the same covariance matrix Σ is not totally geodesic. When an embedded submanifold $\mathcal{S} \subset \mathcal{M}$ is totally geodesic, we always have $\rho_{\mathcal{M}}(N_1, N_2) = \rho_{\mathcal{S}}(N_1, N_2)$. Thus, we have $\rho_{\mathcal{N}}(N(\mu, \Sigma_1), N(\mu, \Sigma_2)) = \rho_{\mathrm{SPD}}(\Sigma_1, \Sigma_2)$. However, when an embedded submanifold $\mathcal{S} \subset \mathcal{M}$ is not totally geodesic, we have $\rho_{\mathcal{M}}(N_1, N_2) \leq \rho_{\mathcal{S}}(N_1, N_2)$ because the Riemannian geodesic length in $\mathcal{S}$ is necessarily longer or equal than the Riemannian geodesic length in $\mathcal{M}$. The merit to consider submanifolds is to be able to calculate in closed form the Fisher–Rao distance which may then provide an upper bound on the Fisher–Rao distance for the full statistical model. For example, consider $N_1 = N(\mu_1, \Sigma)$ and $N_2 = N(\mu_2, \Sigma)$ in $\mathcal{M}_\Sigma$, a non-totally geodesic submanifold. The Rao distance between N_1 and N_2 in $\mathcal{M}$ is upper bounded by the Riemannian distance in $\mathcal{M}_\Sigma$ (with line element $\mathrm{d}s_\Sigma^2 = \mathrm{d}\mu^\top \Sigma^{-1} \mathrm{d}\mu$) which corresponds to the Mahalanobis distance [10,51] $\Delta_\Sigma(\mu_1, \mu_2)$:

$$\rho_{\mathcal{M}_\mu}(N_1, N_2) \leq \Delta_\Sigma(\mu_1, \mu_2) := \sqrt{(\mu_2 - \mu_1)^\top \Sigma^{-1} (\mu_2 - \mu_1)}. \tag{14}$$

The Mahalanobis distance can be interpreted as the Euclidean distance $D_E(p, q) = \Delta_I(p, q) = \sqrt{(p-q)^\top (p-q)}$ (where I denotes the identity matrix) after an affine transformation: Let $\Sigma = LL^\top = U^\top U$ be the Cholesky decomposition of $\Sigma \gg 0$ with L a lower triangular matrix or $U = L^\top$ an upper triangular matrix. Then we have

$$\begin{aligned} \Delta_\Sigma(\mu_1, \mu_2) &= \sqrt{(\mu_2 - \mu_1)^\top (L^\top)^{-1} L^{-1} (\mu_2 - \mu_1)}, \\ &= \|\Sigma^{-\frac{1}{2}} (\mu_2 - \mu_1)\|_2, \\ &= \Delta_I(L^{-1}\mu_1, L^{-1}\mu_2) = D_E(L^{-1}\mu_1, L^{-1}\mu_2), \end{aligned}$$

where $\| \cdot \|_2$ denotes the vector ℓ_2-norm.

The Rao distance ρ_Σ of Equation (A1) between two MVNs with fixed covariance matrix emanates from the property that the submanifold $\mathcal{M}_{[v],\Sigma} = \{N(av, \Sigma) : a \in \mathbb{R}\}$ is totally geodesic [54].

Let us emphasize that for a submanifold $\mathcal{S} \subset \mathcal{M}$ to be totally geodesic or not depend on the underlying metric in $\mathcal{M}$. The same subset $\mathcal{N}' \subset \mathcal{N}$ with $\mathcal{N}$ equipped with two different metrics g_1 and g_2 can be totally geodesic regarding g_1 and non-totally geodesic regarding g_2. See Remark 3 for such an example.

In general, using the triangle inequality of the Riemannian metric distance $\rho_{\mathcal{N}}$, we can upper bound $\rho_{\mathcal{N}}(N_1, N_2)$ with $N_1 = (\mu_1, \Sigma_1)$ and $N_1 = (\mu_2, \Sigma_2)$ as follows:

$$\begin{aligned} \rho_{\mathcal{N}}(N_1, N_2) &\leq \rho_{\mathcal{M}_{\mu_1}}(N_1, N_{12}) + \rho_{\mathcal{M}_{\Sigma_2}}(N_{12}, N_2), \\ &\leq \rho_{\mathcal{M}_{\Sigma_1}}(N_1, N_{21}) + \rho_{\mathcal{M}_{\mu_2}}(N_{21}, N_2), \end{aligned}$$

where $N_{12} = (\mu_1, \Sigma_2)$ and $N_{21} = N(\mu_2, \Sigma_1)$. See Figure 2 for an illustration of the Fisher–Rao geodesic triangle $\triangle N_1, N_2, N_{12}$. Furthermore, since $\rho_{\mathcal{N}_{\Sigma_1}}(N_1, N_{21}) \leq \Delta_{\Sigma_1}(\mu_1, \mu_2)$ and $\rho_{\mathcal{N}_{\Sigma_2}}(N_{12}, N_2) \leq \Delta_{\Sigma_2}(\mu_1, \mu_2)$, we obtain the following upper bound on the Rao distance between MVNs:

$$\rho_{\mathcal{N}}(N_1, N_2) \leq \rho_{\mathcal{P}}(\Sigma_1, \Sigma_2) + \min\{\Delta_{\Sigma_1}(\mu_1, \mu_2), \Delta_{\Sigma_2}(\mu_1, \mu_2)\}. \tag{15}$$

See also [55].

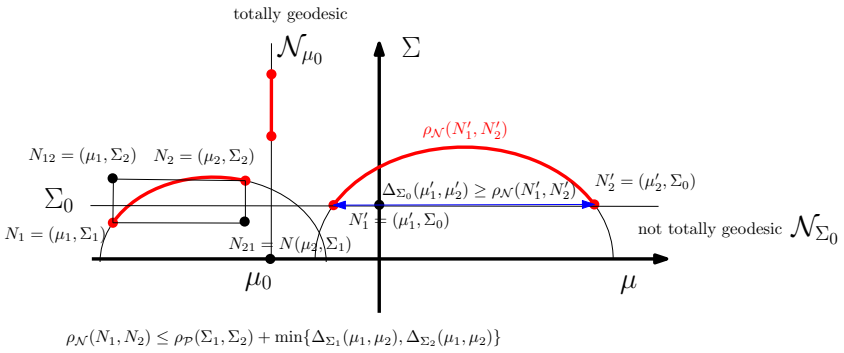


Figure 2. The submanifolds $\mathcal{N}_\Sigma$ are not totally geodesic (i.e., $\rho_{\mathcal{N}}(N_1', N_2')$ is upper bounded by their Mahalanobis distance) but the submanifolds $\mathcal{N}_\mu$ are totally geodesic. Using the triangle inequality of the Riemannian metric distance $\rho_{\mathcal{N}}$, we can upper bound $\rho_{\mathcal{N}}(N_1, N_2)$.

In general, the difficulty with calculating the Fisher–Rao distance comes from the fact that

1. we do not know the Fisher–Rao geodesics with boundary value conditions (BVP) in closed form but the geodesics with initial value conditions [48] (IVP) are known explicitly using the natural parameters $(\Sigma^{-1}\mu, \Sigma^{-1})$ of MVNs,
2. we must integrate the line element $\mathrm{d}s_{\mathcal{N}}$ along the geodesic.

As we shall see in Section 3.1, the above first problem is much harder to solve than the second problem which can be easily approximated by discretizing the curve. The lack of a closed-form formula and fast and good approximations for $\rho_{\mathcal{N}}$ between MVNs is a current limiting factor for its use in applications. Indeed, many applications (e.g., [56,57]) consider the restricted case of the Rao distance between zero-centered MVNs which have closed form (distance of Equation (11) in the SPD cone). The SPD cone is a symmetric Hadamard manifold, and its isometries have been fully studied and classified in [58] (Section 4). The Fisher–Rao geometry of zero-centered generalized MVNs was recently studied in [59].

1.4. Contributions and Paper Outline

The main contribution of this paper is to propose an approximation of $\rho_{\mathcal{N}}$ based on Calvo and Oller's embedding [19] (C&O for short) and report its experimental performance. First, we concisely recall C&O's family of embeddings f_β of $\mathcal{N}(d)$ as submanifolds $\overline{\mathcal{N}}_\beta$ of $\mathcal{P}(d+1)$ in Section 2. Next, we present our approximation technique in Section 3 which differs from the usual geodesic shooting approach [39], and report experimental results. Finally, we study some information–geometric properties [1] of the isometric embedding in Section 5 such as the fact that it preserves mixture geodesics (embedded C&O submanifold is autoparallel with respect to the mixture affine connection) but not exponential geodesics. Moreover, we prove that the Fisher–Rao distance between multivariate normal distributions sharing the same covariance matrix is a scalar function of their Mahalanobis distance in Section 1.5 using the framework of Eaton [52] of maximal invariants.

1.5. A Closed-Form Formula for the Fisher–Rao Distance between Normal Distributions Sharing the Same Covariance Matrix

Consider the Fisher–Rao distance between $N_1 = (\mu_1, \Sigma)$ and $N_1 = (\mu_2, \Sigma)$ for a fixed covariance matrix Σ and the translation action $a.\mu := \mu + a$ of the translation group $\mathbb{R}^d$ (a subgroup of the affine group). Both the Fisher–Rao distance and the Mahalanobis distance are invariant under translations:

$$\rho_{\mathcal{N}}((\mu_1 + a, \Sigma), (\mu_2 + a, \Sigma)) = \rho_{\mathcal{N}}((\mu_1, \Sigma), (\mu_2, \Sigma)), \quad \Delta_\Sigma(\mu_1 + a, \mu_2 + a) = \Delta_\Sigma(\mu_1, \mu_2).$$

To prove that $\rho_{\mathcal{N}}((\mu_1,\Sigma),(\mu_2,\Sigma)) = h_{\mathrm{FR}}(\Delta_\Sigma(\mu_1,\mu_2))$ for a scalar function $h_{\mathrm{FR}}(\cdot)$, we shall prove that the Mahalanobis distance is a maximal invariant, and use the framework of maximal invariants of Eaton [52] (Chapter 2) who proved that any other invariant function is necessarily a function of a maximal invariant, i.e., a function of the Mahalanobis distance in our case.

The Mahalanobis distance is a maximal invariant because we can write $\Delta_\Sigma(\mu_1,\mu_2) = \Delta_1(0,\Delta_\Sigma(\mu_1,\mu_2))$ and when $\Delta_\Sigma(\mu_1,\mu_2) = \Delta_\Sigma(\mu_1',\mu_2')$ in 1D there exists $a \in \mathbb{R}$ such that $(\mu_1+a,\mu_2+a) = (\mu_1',\mu_2')$. We must prove equivalently that when $|m_1-m_2| = |m_1'-m_2'|$ that there exists $a \in \mathbb{R}$ such that $(m_1+a,m_2+a) = (m_1',m_2')$. Assume without loss of generality that $m_1 \geq m_2$. When $m_1-m_2 = m_1'-m_2'$, there exists $a = m_1'-m_1$ so that $m_1' = a.m_1 = m_1+a$ and $m_2' = a.m_2 = m_2+a$ with $m_1'-m_2' = m_1-m_2$. Thus, using Eaton's theorem [52], there exists a scalar function h_{FR} such that $\rho_{\mathcal{N}}((\mu_1,\Sigma),(\mu_2,\Sigma)) = h_{\mathrm{FR}}(\Delta_\Sigma(\mu_1,\mu_2))$.

To find explicitly the scalar function $h_{\mathrm{FR}}(\cdot)$, let us consider the univariate case of normal distributions for which the Fisher–Rao distance is given in closed form in Equation (7). In that case, the univariate Mahalanobis distance is $\Delta_{\sigma^2}(\mu_1,\mu_2) = \sqrt{(\mu_2-\mu_1)(\sigma^2)^{-1}(\mu_2-\mu_1)} = \frac{|\mu_2-\mu_1|}{\sigma}$ and we can write formula of Equation (7) as $h_{\mathrm{FR}}(\Delta_{\sigma^2}(\mu_1,\mu_2))$ with

$$h_{\mathrm{FR}}(u) = \sqrt{2}\log\left(\frac{\sqrt{8+u^2}+u}{\sqrt{8+u^2}-u}\right), \tag{16}$$

$$= \sqrt{2}\,\mathrm{arccosh}\left(1+\frac{1}{4}u^2\right), \tag{17}$$

using the identities

$$\log(x) = \mathrm{arccosh}\left(\frac{1+x^2}{2x}\right) = \mathrm{arctanh}\left(\frac{x^2-1}{1+x^2}\right), \quad x>1,$$

where $\mathrm{arctanh}(u) = \frac{1}{2}\log\frac{1+u}{1-u}$.

Proposition 1. *The Fisher–Rao distance $\rho_{\mathcal{N}}((\mu_1,\Sigma),(\mu_2,\Sigma))$ between two MVNs with same covariance matrix is*

$$\rho_{\mathcal{N}}((\mu_1,\Sigma),(\mu_2,\Sigma)) = \rho_{\mathcal{N}}((0,1),(\Delta_\Sigma(\mu_1,\mu_2),1)), \tag{18}$$

$$= \sqrt{2}\log\left(\frac{\sqrt{8+\Delta_\Sigma^2(\mu_1,\mu_2)}+\Delta_\Sigma(\mu_1,\mu_2)}{\sqrt{8+\Delta_\Sigma^2(\mu_1,\mu_2)}-\Delta_\Sigma(\mu_1,\mu_2)}\right), \tag{19}$$

$$= \sqrt{2}\,\mathrm{arccosh}\left(1+\frac{1}{4}\Delta_\Sigma^2(\mu_1,\mu_2)\right), \tag{20}$$

where $\Delta_\Sigma(\mu_1,\mu_2) = \sqrt{(\mu_2-\mu_1)^\top\Sigma^{-1}(\mu_2-\mu_1)}$ is the Mahalanobis distance.

Indeed, notice that the d-variate Mahalanobis distance $\Delta_\Sigma(\mu_1,\mu_2)$ can be interpreted as a univariate Mahalanobis distance between the standard normal distribution $N(0,1)$ and $N(\Delta_\Sigma(\mu_1,\mu_2),1)$:

$$\Delta_\Sigma(\mu_1,\mu_2) = \Delta_1(0,\Delta_\Sigma(\mu_1,\mu_2)).$$

Thus, we have $\rho_{\mathcal{N}}((\mu_1,\Sigma),(\mu_2,\Sigma)) = \rho_{\mathcal{N}}((0,1),(\Delta_\Sigma(\mu_1,\mu_2),1))$, where the right-hand-side term is the univariate Fisher–Rao distance of Equation (7). Let us notice that the square length element on $\mathcal{M}_\Sigma$ is $\mathrm{d}s^2 = \mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu = \Delta_\Sigma^2(\mu,\mu+\mathrm{d}\mu)$. This result can be extended to elliptical distributions [12] (Theorem 1).

Let us corroborate this result by checking the formula of Equation (1) with two examples in the literature: In [38] (Figure 4), we Fisher–Rao distance between $N_1 = (0,I)$

and $N_2 = \left(\begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, I \right)$ is studied. We find $\rho_{\mathcal{N}}(N_1, N_2) = 0.69994085$ in accordance with their result shown in Figure 4. The second example is Example 1 of [42] (p. 11) with $N_1 = \left(\begin{bmatrix} -1 \\ 0 \end{bmatrix}, \Sigma \right)$ and $N_2 = \left(\begin{bmatrix} 6 \\ 3 \end{bmatrix}, \Sigma \right)$ for $\Sigma = \begin{bmatrix} 1.1 & 0.9 \\ 0.9 & 1.1 \end{bmatrix}$. Formula of Equation (18) yields the Fisher–Rao distance 5.006483034546878 in accordance with [42] which reports 5.00648.

Similarly, the statistical Ali–Silvey–Csiszár f-divergences [60,61]

$$I_f[p_{(\mu_1,\Sigma)} : p_{(\mu_2,\Sigma)}] = \int_{\mathbb{R}^d} p_{(\mu_1,\Sigma)}(x) \, f\left(\frac{p_{(\mu_2,\Sigma)}}{p_{(\mu_1,\Sigma)}} \right) \mathrm{d}x,$$

between two MVNs sharing the same covariance matrix are increasing functions of the Mahalanobis distance because the f-divergences between two MVNs sharing the same covariance matrix are invariant under the action of the translation group [62]. Thus, we have $I_f[p_{(\mu_1,\Sigma} : p_{(\mu_2,\Sigma)}] = h_f(\Delta_\Sigma(\mu_1, \mu_2))$. Since $\Delta_\Sigma(\mu_1, \mu_2) = \Delta_1(0, \Delta_\Sigma(\mu_1, \mu_2))$, we thus have

$$I_f[p_{(\mu_1,\Sigma} : p_{(\mu_2,\Sigma)}] = h_f(\Delta_1(0, \Delta_\Sigma(\mu_1, \mu_2)) = I_f[p_{(0,1} : p_{(\Delta_\Sigma(\mu_1,\mu_2),1)}],$$

where the right-hand side f-divergence is between univariate normal distributions. See Table 2 of [62] for some explicit functions h_f.

2. Calvo and Oller's Family of Diffeomorphic Embeddings

Calvo and Oller [19,32] noticed that we can embed the space of normal distributions in $\mathcal{P}(d+1)$ by using the following mapping:

$$f_\beta(N) = f_\beta(\mu, \Sigma) = \begin{bmatrix} \Sigma + \beta \mu \mu^\top & \beta \mu \\ \beta \mu^\top & \beta \end{bmatrix} \in \mathcal{P}(d+1), \tag{21}$$

where $\beta \in \mathbb{R}_{>0}$ and $N = N(\mu, \Sigma)$. Notice that since the dimension of $\mathcal{P}(d+1)$ is $\frac{(d+1)(d+2)}{2}$, we only use $\frac{(d+1)(d+2)}{2} - \frac{d(d+3)}{2} = 1$ extra dimension for embedding $\mathcal{N}(d)$ into $\mathcal{P}(d+1)$. By foliating $\mathbb{P} = \mathbb{R}_{>0} \times \mathbb{P}_c$ where $\mathbb{P}_c = \{P \in \mathbb{P} \ : \ |P| = c\}$ denotes the subsets of $\mathbb{P}$ with determinant c, we obtain the following Riemannian Calvo and Oller metric on the SPD cone:

$$\begin{aligned} \mathrm{d}s^2_{\mathrm{CO}} &= \frac{1}{2} \mathrm{tr}\left(\left(f^{-1}(\mu, \Sigma) \mathrm{d}f(\mu, \Sigma) \right)^2 \right), \\ &= \frac{1}{2} \left(\frac{\mathrm{d}\beta}{\beta} \right)^2 + \beta \mathrm{d}\mu^\top \Sigma^{-1} \mathrm{d}\mu + \frac{1}{2} \mathrm{tr}\left(\left(\Sigma^{-1} \mathrm{d}\Sigma \right)^2 \right). \end{aligned}$$

Let

$$\overline{\mathcal{N}}_\beta(d) = \left\{ \bar{P} = f_\beta(\mu, \Sigma) \ : \ (\mu, \Sigma) \in \mathcal{N}(d) = \mathbb{R}^d \times \mathcal{P}(d) \right\}$$

denote the submanifold of $\mathcal{P}(d+1)$ of codimension 1, and $\overline{\mathcal{N}} = \overline{\mathcal{N}}_1$ (i.e., $\beta = 1$). The family of mappings f_β provides diffeomorphisms between $\mathcal{N}(d)$ and $\overline{\mathcal{N}}_\beta(d)$. Let $f_\beta^{-1}(\bar{P}) = (\mu_{\bar{P}}, \Sigma_{\bar{P}})$ denote the inverse mapping for $\bar{P} \in \overline{\mathcal{N}}_\beta(d)$, and let $f = f_1$ (i.e., $\beta = 1$):

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

By equipping the cone $\mathcal{P}(d+1)$ by the trace metric [63,64] (also called the affine invariant Riemannian metric, AIRM) scaled by $\frac{1}{2}$:

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

(yielding the squared line element $\mathrm{d}s^2_{\mathcal{P}} = \frac{1}{2}\mathrm{tr}((P\,\mathrm{d}P)^2)$), Calvo and Oller [19] proved that $\overline{\mathcal{N}}(d)$ is isometric to $\mathcal{N}(d)$ (i.e., the Riemannian metric of $\mathcal{P}(d+1)$ restricted to $\mathcal{N}(d)$ coincides with the Riemannian metric of $\mathcal{N}(d)$ induced by f) but $\overline{\mathcal{N}}(d)$ is not totally geodesic (i.e., the geodesics $\gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)$ for $\bar{P}_1 = f(N_1), \bar{P}_2 = f(N_2) \in \overline{\mathcal{N}}(d)$ leaves the embedded normal submanifold $\overline{\mathcal{N}}(d)$). Please note that g^{trace}_P can be interpreted as the Fisher metric for the family $\mathcal{N}_0$ of 0-centered normal distributions. Thus, we have $(\mathcal{N}(d), g^{\mathrm{Fisher}}) \hookrightarrow (\mathcal{P}(d+1), g^{\mathrm{trace}})$, and the following diagram between parameter spaces and corresponding distributions:

$$\begin{array}{ccc} \mathcal{N}(d) & \hookrightarrow & \mathcal{N}_0(d+1) \\ \updownarrow & & \updownarrow \\ \Lambda(d) & \hookrightarrow & \mathbb{P}(d+1) \end{array}$$

Remark 2. *The trace metric was first studied by Siegel [45,65] using the wider scope of complex symmetric matrices with positive–definite imaginary parts generalizing the Poincaré upper half-plane (see Appendix D).*

We omit to specify the dimensions and write for short $\mathcal{N}$, $\overline{\mathcal{N}}$, and $\mathcal{P}$ when clear from the context. Thus, C&O proposed to use the embedding $f = f_1$ to give a lower bound ρ_{CO} of the Fisher–Rao distance $\rho_{\mathcal{N}}$ between normals:

$$\mathrm{LC}_{\mathrm{CO}}: \quad \rho_{\mathcal{N}}(N_1, N_2) \geq \rho_{\mathrm{CO}}(\underbrace{f(\mu_1, \Sigma_1)}_{\bar{P}_1}, \underbrace{f(\mu_2, \Sigma_2)}_{\bar{P}_2}) = \sqrt{\frac{1}{2}\sum_{i=1}^{d+1} \log^2 \lambda_i(\bar{P}_1^{-1}\bar{P}_2)}. \tag{22}$$

We let $\rho_{\mathrm{CO}}(N_1, N_2) = \rho_{\mathrm{CO}}(f(N_1), f(N_2))$. The ρ_{CO} distance is invariant under affine transformations such as the Fisher–Rao distance of Property 1:

Property 2 (affine invariance of C&O distance [19])**.** *For all $A \in \mathrm{GL}(d), a \in \mathbb{R}^d$, we have* $\rho_{\mathrm{CO}}((A\mu_1 + a, A\Sigma_1 A^\top), (A\mu_2 + a, A\Sigma_2 A^\top)) = \rho_{\mathrm{CO}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2))$.

When $\Sigma_1 = \Sigma_2 = \Sigma$, we have $|\bar{P}_1| = |\bar{P}_2| = |\Sigma|$. Since the Riemannian geodesics $\gamma_{\mathbb{P}}(P_1, P_2; t)$ in the SPD cone are given by $\gamma_{\mathbb{P}}(P_1, P_2; t) = P_1^{\frac{1}{2}}(P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}})^t P_1^{\frac{1}{2}}$ [66] (also written $\gamma_{\mathrm{SPD}}(P_1, P_2; t)$), we have $|\gamma_{\mathbb{P}}(P_1, P_2; t)| = |\Sigma|$. Although the submanifold $\mathbb{P}_c = \{P \in \mathbb{P} : |P| = c\}$ is totally geodesic with respect to the trace metric, it is not totally geodesic with respect to $\frac{1}{2}\mathrm{tr}((\bar{P}\mathrm{d}\bar{P})^2)$. Thus, although $\gamma_{\mathbb{P}}(P_1, P_2) \in \overline{\mathcal{N}}$, it does not correspond to the embedded MVN geodesics with respect to the Fisher metric. The C&O distance between two MVNs $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ sharing the same covariance matrix [19] is

$$\rho_{\mathrm{CO}}(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) = \mathrm{arccosh}\left(1 + \frac{1}{2}\Delta^2_{\Sigma}(\mu_1, \mu_2)\right), \tag{23}$$

where $\mathrm{arccosh}(x) := \log(x + \sqrt{x^2 - 1})$ for $x \geq 1$ and $\Delta_{\Sigma}(\mu_1, \mu_2)$ is the Mahalanobis distance between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. In that case, we thus have $\rho_{\mathrm{CO}}(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) = h_{\mathrm{CO}}(\Delta_{\Sigma}(\mu_1, \mu_2))$ where $h_{\mathrm{CO}}(u) = \mathrm{arccosh}\left(1 + \frac{1}{2}u^2\right)$ is a strictly monotone increasing function. Let us note in passing that in [19] (Corollary, page 230) there is a confusing or typographic error since the distance is reported as $\mathrm{arccosh}\left(1 + \frac{1}{2}d_M(\mu_1, \mu_2)\right)$ where d_M denotes "Mahalanobis distance" [51]. Therefore, either $d_M = \Delta^2_{\Sigma}$, Mahalanobis D^2-distance, or there is a missing square in the equation of the Corollary page 230. To obtain a flavor of how good is the approximation of the C&O distance, we may consider the same covariance case where we have both closed-form solutions for $\rho_{\mathcal{N}}$ (Equation (20)) and ρ_{CO} (Equation (23)). Figure 3 plots the two functions h_{CO} and h_{FR} (with $h_{\mathrm{CO}}(u) \leq h_{\mathrm{FR}}(u) \leq u$ for $u \in [0, \infty)$).

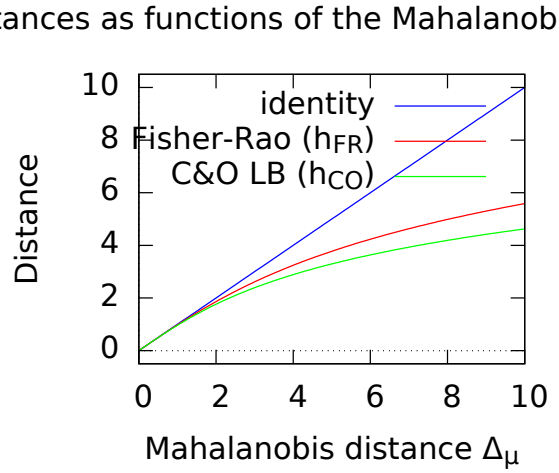


Figure 3. Quality of the C&O lower bound compared to the exact Fisher–Rao distance in the case of $N_1, N_2 \in \mathcal{M}_\Sigma$ (MVNs sharing the same covariance matrix Σ). We have $\rho_{\mathrm{CO}} \leq \rho_{\mathcal{N}} \leq \Delta_\Sigma$.

Let us remark that similarly all f-divergences between $N_1 = (\mu_1, \Sigma)$ and $N_2 = (\mu_2, \Sigma)$ are scalar functions of their Mahalanobis distance $\Delta_\Sigma(\mu_1, \mu_2)$ too, see [62].

The C&O distance ρ_{CO} is a metric distance that has been used in many applications ranging from computer vision [57,67–69] to signal/sensor processing, statistics [70,71], machine learning [29,72–76] and analogical reasoning [77].

Remark 3. *In a second paper, Calvo and Oller [32] noticed that we can embed normal distributions in $\mathcal{P}(d+1)$ by the following more general mapping (Lemma 3.1 [32]):*

$$g_{\alpha,\beta,\gamma}(\mu,\Sigma) = |\Sigma|^\alpha \begin{bmatrix} \Sigma + \beta\gamma^2\mu\mu^\top & \beta\gamma\mu \\ \beta\gamma\mu^\top & \beta \end{bmatrix} \in \mathcal{P}(d+1), \tag{24}$$

where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}_{>0}$ and $\gamma \in \mathbb{R}$. It is show in [32] that the induced length element is

$$\begin{aligned} \mathrm{ds}^2_{\alpha,\beta,\gamma} &= \frac{1}{2}\Big(\alpha((d+1)+2\alpha)\mathrm{tr}^2(\Sigma^{-1}\mathrm{d}\Sigma) + \mathrm{tr}((\Sigma^{-1}\mathrm{d}\Sigma)^2) \\ &\quad + 2\beta\gamma^2\mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu + 2\alpha\mathrm{tr}(\Sigma^{-1}\mathrm{d}\Sigma)\frac{\mathrm{d}\beta}{\beta} + \left(\frac{\mathrm{d}\beta}{\beta}\right)^2\Big). \end{aligned}$$

When $\gamma = \beta = 1$, we have

$$\mathrm{ds}^2_\alpha = \frac{1}{2}\Big(\alpha((d+1)+2\alpha)\mathrm{tr}^2(\Sigma^{-1}\mathrm{d}\Sigma) + \mathrm{tr}((\Sigma^{-1}\mathrm{d}\Sigma)^2) + 2\beta\gamma^2\mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu\Big).$$

Thus, to cancel the term $\mathrm{tr}^2(\Sigma^{-1}\mathrm{d}\Sigma)$, we may either choose $\alpha = 0$ or $\alpha = -\frac{2}{1+d}$.

In some applications [78], the embedding

$$g_{-\frac{1}{d+1},1,1}(\mu,\Sigma) = |\Sigma|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix} := \hat{f}(\mu,\Sigma), \tag{25}$$

is used to ensure that $\left|g_{-\frac{1}{d+1},1,1}(\mu,\Sigma)\right| = 1$. That is normal distributions are embedded diffeomorphically into the submanifold of positive–definite matrices with a unit determinant (also called SSPD, acronym of Special SPD). In [32], C&O showed that there exists a second isometric embedding of the Fisher–Rao Gaussian manifold $\mathcal{N}(d)$ into a submanifold of the cone $\mathcal{P}(d+1)$: $f_{\mathrm{SSPD}}(\mu,\Sigma) = |\Sigma|^{-\frac{2}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix}$. Let $\hat{P} = f_{\mathrm{SSPD}}(\mu,\Sigma)$. This mapping can be understood as taking the elliptic isometry $P \mapsto |P|^{-\frac{2}{d+1}}P$ of $P \in \mathcal{P}(d+1)$ [64] since $|\Sigma| = |\bar{P}(\mu,\Sigma)|$ (see proof in Proposition 3). It follows that

$$\rho_{\mathrm{CO}}(N_1, N_2) = \rho_{\mathcal{P}}(\bar{P}_1, \bar{P}_2) = \rho_{\mathcal{P}}(\hat{P}_1, \hat{P}_2) \leq \rho_{\mathcal{N}}(N_1, N_2).$$

Similarly, we could have mapped $P \mapsto P^{-1}$ to obtain another isometric embedding. See the four types of elliptic isometric of the SPD cone described in [64]. Finally, let us remark that the SSPD submanifold is totally geodesic with respect to the trace metric but not with respect to the C&O metric.

Interestingly, Calvo and Oller [48] (p. 131) proved that $((\bar{\mu}_1, \ldots, \bar{\mu}_d), \mathrm{diag}(\bar{\sigma}_1^2, \ldots, \bar{\sigma}_d^2))$ is a maximal invariant for the action of the affine group $\mathrm{Aff}(d)$, where $\bar{\mu} = Q^{-1}(\mu_2 - \mu_1)$ and $\Sigma_2\Sigma_1^{-1} = Q\,\mathrm{diag}(\bar{\sigma}_1^2, \ldots, \bar{\sigma}_d^2)\,Q^{-1}$ (in [48], the authors considered $\Sigma_1\Sigma_1^{-2}$). Thus, we consider the following dissimilarity

$$D_{\mathrm{CO}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) = \sqrt{2}\sqrt{\sum_{i=1}^{d} \log^2\left(\frac{1 + \Delta(0, 1; \bar{\mu}_i, \bar{\sigma}_i)}{1 - \Delta(0, 1; \bar{\mu}_i, \bar{\sigma}_i)}\right)}. \tag{26}$$

Dissimilarity D_{CO} is symmetric (i.e., $D_{\mathrm{CO}}(N_1, N_2) = D_{\mathrm{CO}}(N_2, N_1)$) and $D_{\mathrm{CO}}(N_1, N_2) = 0$ if and only if $N_1 = N_2$. Please note that when $d = 1$, D_{CO} is different from the Fisher–Rao distance of Equation (7).

3. Approximating the Fisher–Rao Distance

3.1. Approximating Length of Curves

Recall that the Fisher–Rao's distance [79] is the Riemannian geodesic distance

$$\rho_{\mathcal{N}}(N(\lambda_1), N(\lambda_2)) = \inf_{\substack{c(t) \\ c(0)=p_{\lambda_1} \\ c(1)=p_{\lambda_2}}} \{\mathrm{Length}(c)\},$$

where

$$\mathrm{Length}(c) = \int_0^1 \underbrace{\sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}}}_{\mathrm{d}s_{\mathcal{N}}(t)} \mathrm{d}t.$$

We can approximate the Rao distance $\rho_{\mathcal{N}}(N_1, N_2)$ by discretizing regularly any smooth curve $c(t)$ joining $N_1 = c(0)$ to $N_2 = c(1)$ (Figure 4):

$$\rho_{\mathcal{N}}(N_1, N_2) \leq \frac{1}{T} \sum_{i=1}^{T-1} \rho_{\mathcal{N}}\left(c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right),$$

with equality holding iff $c(t) = \gamma_{\mathcal{N}}(N_1, N_2; t)$ is the Riemannian geodesic defined by the Levi–Civita metric connection induced by the Fisher information metric.

When the number of discretization steps T is sufficiently large, the normal distributions $c\left(\frac{i}{T}\right)$ and $c\left(\frac{i+1}{T}\right)$ are close to each other, and we can approximate $\rho_{\mathcal{N}}\left(c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right)$ by $\sqrt{D_J\left[c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right]}$, where $D_J[N_1, N_2] = D_{\mathrm{KL}}[N_1, N_2] + D_{\mathrm{KL}}[N_2, N_1]$ is Jeffreys divergence, and D_{KL} is the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}[p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}] = \frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \Delta\mu^\top \Sigma_2^{-1}\Delta\mu - d + \log\frac{|\Sigma_2|}{|\Sigma_1|}\right).$$

Thus, the costly determinant computations cancel each other in Jeffreys divergence (i.e., $\log\frac{|\Sigma_2|}{|\Sigma_1|} + \log\frac{|\Sigma_1|}{|\Sigma_2|} = 0$) and we have:

$$D_J[p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}] = \mathrm{tr}\left(\frac{\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2}{2} - I\right) + \Delta\mu^\top \frac{\Sigma_1^{-1} + \Sigma_2^{-1}}{2}\Delta\mu.$$

Figure 4 summarizes our method to approximate the Fisher–Rao geodesic distance.

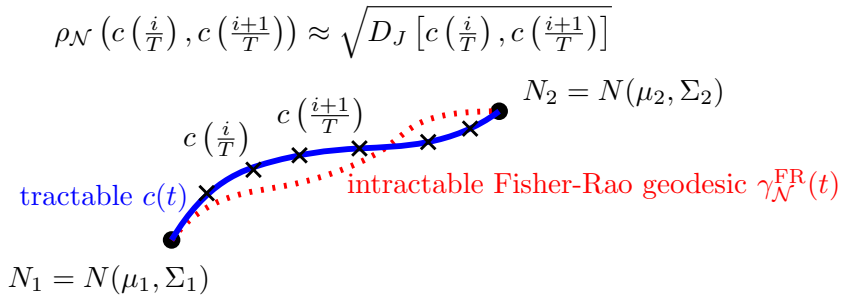


Figure 4. Approximating the Fisher–Rao geodesic distance $\rho_{\mathcal{N}}(N_1, N_2)$: The Fisher–Rao geodesic $\gamma^{\mathrm{FR}}_{\mathcal{N}}$ is not known in closed form. We consider a tractable curve $c(t)$, discretize $c(t)$ at $T+1$ points $c(\frac{i}{T})$ with $c(0) = N_1$ and $c(1) = N_2$, and approximate $\rho_{\mathcal{N}}\left(c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right)$ by $\sqrt{D_J\left[c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right]}$, considering that different tractable curves $c(t)$ yield different approximations.

In general, it holds that

$$I_f[p:q] \approx \frac{f''(1)}{2} \mathrm{ds}^2_{\mathrm{Fisher}},$$

between infinitesimally close distributions p and q ($\mathrm{d}s \approx \sqrt{\frac{2\, I_f[p:q]}{f''(1)}}$), where $I_f[\cdot:\cdot]$ denotes a f-divergence [1]. The Jeffreys divergence is a f-divergence obtained for $f_J(u) = -\log u + u \log u$ with $f''_J(1) = 2$. It is thus interesting to find low computational cost f-divergences between multivariate normal distributions to approximate the infinitesimal length element $\mathrm{d}s$. Please note that f-divergences between MVNs are also invariant under the action of the affine group [62]. Thus, for infinitesimally close distributions p and q, this informally explains that $\mathrm{d}s_{\mathrm{Fisher}}$ is invariant under the action of the affine group (see Proposition 1).

Although the definite integral of the length element along the Fisher–Rao geodesic $\gamma^{\mathrm{FR}}_{\mathcal{N}}$ is not known in closed form (i.e., Fisher–Rao distance), the integral of the squared length element along the mixture geodesic $\gamma^{m}_{\mathcal{N}}(N_1, N_2)$ and exponential geodesic $\gamma^{e}_{\mathcal{N}}(N_1, N_2)$ coincide with Jeffreys divergence $D_J[N_1, N_2]$ between N_1 and N_2 [1]:

Property 3 ([1])**.** *We have*

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

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 $F^*(\eta) = -h(p_{\mu,\Sigma})$ [1], i.e., the convex conjugate function is Shannon negentropy.

Then we conclude using the fact that $S_F(\theta_1; \theta_2) = \int_0^1 \mathrm{ds}^2(\gamma(t))\mathrm{d}t = \int_0^1 \mathrm{ds}^2(\gamma^*(t))\mathrm{d}t$, 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 E. □

It follows the following upper bound on the Fisher–Rao distance:

Property 4 (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 obtain:

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

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

Please 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 $L(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|\mathrm{d}t$ and $\|\dot{\gamma}(t)\|$ is constant. Using Cauchy-Schwartz inequality, we can show that $L(\gamma) \leq E(\gamma)$. □

This upper bound is tight at infinitesimal scale (i.e., when $N_2 = N_1 + \mathrm{d}N$) since $\rho_{\mathcal{N}}(N_1, N_2) \approx \mathrm{d}s_{\mathcal{N}}(N_1) \approx \sqrt{\frac{2\, I_f[N_1:N_2]}{f''(1)}}$ and the f-divergence in right-hand side of the identity can be chosen as Jeffreys divergence. To appreciate the quality of the square root of Jeffreys divergence upper bound of Property 4, consider the case where $N_1, N_2 \in \mathcal{M}_\Sigma$. In that case, we have $\rho_{\mathcal{N}}(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) = \sqrt{2}\,\mathrm{arccosh}(1 + \frac{1}{4}\Delta^2_\Sigma(\mu_1, \mu_2))$ and $\sqrt{D_J[N(\mu_1, \Sigma), N(\mu_2, \Sigma)]} = \Delta_\Sigma(\mu_1, \mu_2)$ (since $D_{\mathrm{KL}}[N(\mu_1, \Sigma), N(\mu_2, \Sigma)] = \frac{1}{2}\Delta^2_\Sigma(\mu_1, \mu_2)$). The upper bound can thus be checked since we have $\sqrt{2}\,\mathrm{arccosh}(1 + \frac{1}{4}x^2) \leq x$ for $x \geq 0$. The plots of Figure 5 shows visually the quality of the $\sqrt{D_J}$ upper bound.

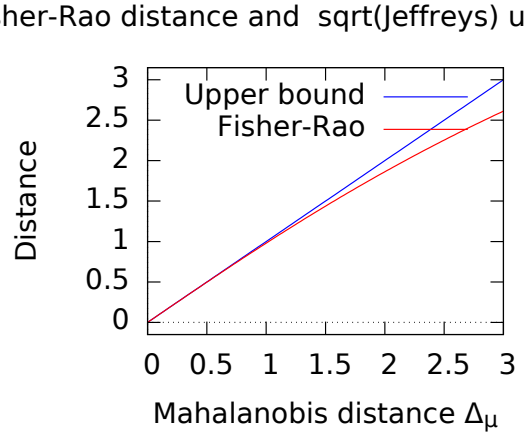


Figure 5. Quality of the $\sqrt{D_J}$ upper bound on the Fisher–Rao distance $\rho_{\mathcal{N}}$ when normal distributions have the same covariance matrix.

For any smooth curve $c(t)$, we can thus approximate $\rho_{\mathcal{N}}$ for large T by

$$\tilde{\rho}^{c}_{\mathcal{N}}(N_1, N_2) := \frac{1}{T}\sum_{i=1}^{T-1}\sqrt{D_J\left[c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right]}. \tag{27}$$

For example, we may consider the following curves on $\mathcal{M}_{\mathcal{N}}$ which admit closed-form parameterizations in $t \in [0,1]$:

- linear interpolation (LERP, Linear intERPolation) $c_\lambda(t) = t(\mu_1, \Sigma_1) + (1-t)(\mu_2, \Sigma_2)$ between (μ_1, Σ_1) and (μ_2, Σ_2),
- the mixture geodesic [80] $c_m(t) = \gamma^m_{\mathcal{N}}(N_1, N_2; t) = (\mu^m_t, \Sigma^m_t)$ with $\mu^m_t = \bar{\mu}_t$ and $\Sigma^m_t = \bar{\Sigma}_t + t\mu_1\mu_1^\top + (1-t)\mu_2\mu_2^\top - \bar{\mu}_t\bar{\mu}_t^\top$ where $\bar{\mu}_t = t\mu_1 + (1-t)\mu_2$ and $\bar{\Sigma}_t = t\Sigma_1 + (1-t)\Sigma_2$,
- the exponential geodesic [80] $c_e(t) = \gamma^e_{\mathcal{N}}(N_1, N_2; t) = (\mu^e_t, \Sigma^e_t)$ with $\mu^e_t = \bar{\Sigma}^H_t(t\Sigma_1^{-1}\mu_1 + (1-t)\Sigma_2^{-1}\mu_2)$ and $\Sigma^e_t = \bar{\Sigma}^H_t$ where $\bar{\Sigma}^H_t = (t\Sigma_1^{-1} + (1-t)\Sigma_2^{-1})^{-1}$ is the matrix harmonic mean,
- the curve $c_{em}(t) = \frac{1}{2}\left(\gamma^e_{\mathcal{N}}(N_1, N_2; t) + \gamma^m_{\mathcal{N}}(N_1, N_2; t)\right)$ which is obtained by averaging the mixture geodesic with the exponential geodesic.

Figure 6 visualizes the exponential and mixture geodesics between two bivariate normal distributions.

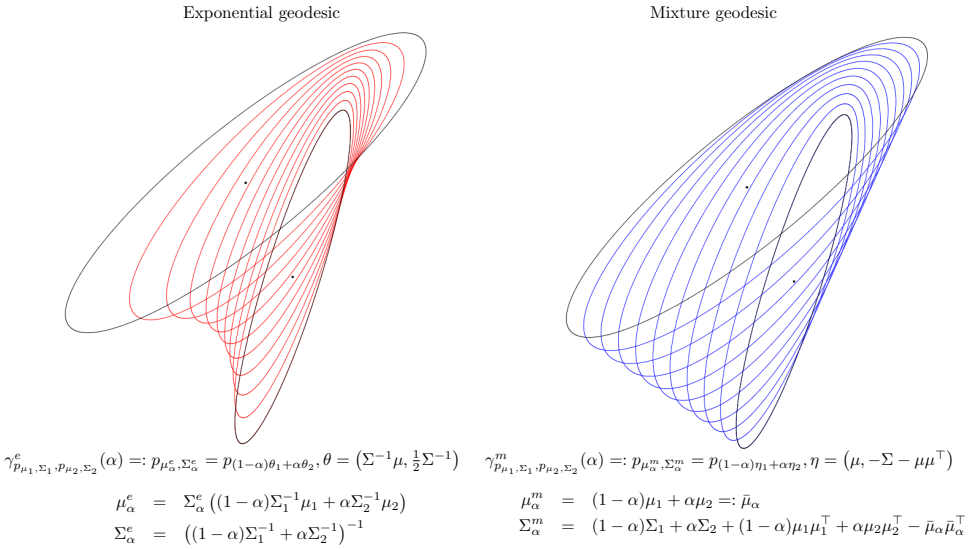


Figure 6. Visualizing the exponential and mixture geodesics between two bivariate normal distributions.

Let us denote by $\tilde{\rho}^{\lambda}_{\mathcal{N}} = \tilde{\rho}^{c_\lambda}_{\mathcal{N}}$, $\tilde{\rho}^{m}_{\mathcal{N}} = \tilde{\rho}^{c_m}_{\mathcal{N}}$, $\tilde{\rho}^{e}_{\mathcal{N}} = \tilde{\rho}^{c_e}_{\mathcal{N}}$ and $\tilde{\rho}^{em}_{\mathcal{N}} = \tilde{\rho}^{c_{em}}_{\mathcal{N}}$ the approximations obtained by these curves following from Equation (27). When T is sufficiently large, the approximated distances $\tilde{\rho}^x$ are close to the length of curve x, and we may thus consider a set of several curves $\{c_i\}_{i\in I}$ and report the smallest Fisher–Rao distance approximations obtained among these curves: $\rho_{\mathcal{N}}(N_1, N_2) \approx \min_{i\in I}\tilde{\rho}^{c_i}_{\mathcal{N}}(N_1, N_2)$.

Please note that we consider the regular spacing for approximating a curve length and do not optimize the position of the sample points on the curve. Indeed, as $T \to \infty$, the curve length approximation tends to the Riemannian curve length. In other words, we can measure approximately finely the length of any curve available with closed-form reparameterization by increasing T. Thus, the key question of our method is how to best approximate the Fisher–Rao geodesic by a curve that can be parametrized by a closed-form formula and is close enough to the Fisher–Rao geodesic.

Next, we introduce our approximation curve $c_{\mathrm{CO}}(t)$ derived from Calvo and Oller isometric mapping f which experimentally behaves better when normals are not *too far* from each other.

3.2. A Curve Derived from Calvo and Oller's Embedding

This approximation consists of leveraging the closed-form expression of the SPD geodesics [63,66]:

$$\gamma_{\mathcal{P}}(P,Q;t) = P^{\frac{1}{2}}\left(P^{-\frac{1}{2}}Q^{\frac{1}{2}}P^{-\frac{1}{2}}\right)^{t}P^{\frac{1}{2}}, \quad t\in[0,1]$$

to approximate the Fisher–Rao normal geodesic $\gamma_{\mathcal{N}}^{\mathrm{Fisher}}(N_1,N_2;t)$ as follows: Let $\bar{P}_1 = f(N_1), \bar{P}_2 = f(N_2) \in \overline{\mathcal{N}}$, and consider the smooth curve

$$\bar{c}_{\mathrm{CO}}(\bar{P}_1,\bar{P}_2;t) = \mathrm{proj}_{\overline{\mathcal{N}}}(\gamma_{\mathcal{P}}(\bar{P}_1,\bar{P}_2;t)), \tag{28}$$

where $\mathrm{proj}_{\overline{\mathcal{N}}}(P)$ denotes the orthogonal projection of $P\in\mathcal{P}(d+1)$ onto $\overline{\mathcal{N}}$ (Figure 7). Thus, curve $c_{\mathrm{CO}}(t)$ ($t\in[0,1]$) is then defined by taking the inverse mapping $f^{-1}(\bar{c}_{\mathrm{CO}})$ (Figure 8):

$$c_{\mathrm{CO}}(t) = f^{-1}\left(\mathrm{proj}_{\overline{\mathcal{N}}}(\gamma_{\mathcal{P}}(\bar{P}_1,\bar{P}_2;t))\right). \tag{29}$$

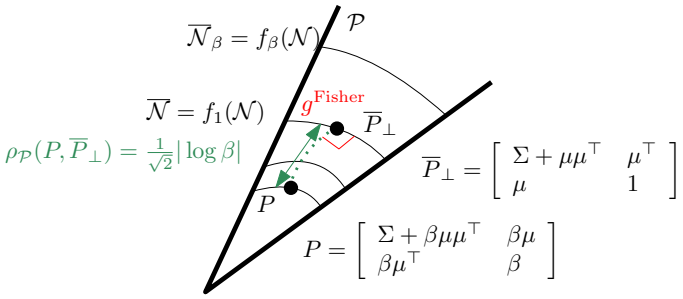


Figure 7. Projecting an SPD matrix $P\in\mathcal{P}$ onto $\overline{\mathcal{N}} = f(\mathcal{N})$: $\gamma_{\mathcal{P}}(P,\bar{P}_\perp)$ is orthogonal to $\overline{\mathcal{N}}$ with respect to the trace metric.

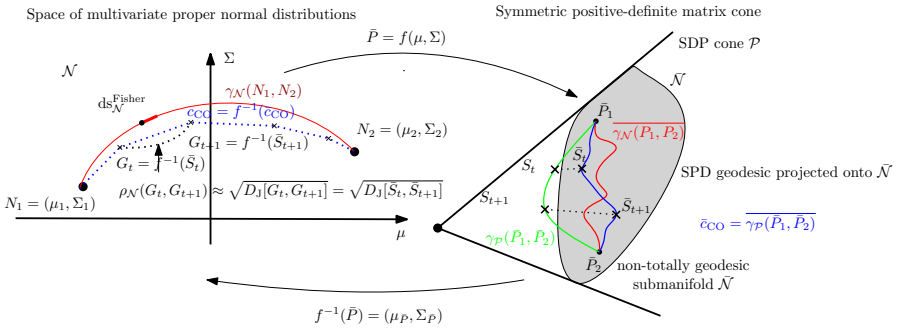


Figure 8. Illustration of the approximation of the Fisher–Rao distance between two multivariate normals N_1 and N_2 (red geodesic length $\gamma_{\mathcal{N}}(N_1,N_2)$ by discretizing curve $\bar{c}_{\mathrm{CO}}\in\overline{\mathcal{N}}$ or equivalently curve $c_{\mathrm{CO}}\in\mathcal{N}$.

Please note that the matrix power P^t can be computed as $P^t = U\,\mathrm{diag}(\lambda_1^t,\ldots,\lambda_d^t)\,V^\top$ where $P = U\,\mathrm{diag}(\lambda_1^t,\ldots,\lambda_d^t)\,V^\top$ is the eigenvalue decomposition of P.

Let us now explain how to project $P=[P_{i,j}]\in\mathcal{P}(d+1)$ onto $\overline{\mathcal{N}}$ based on the analysis of the Appendix of [19] (p. 239):

Proposition 2 (Projection of an SPD matrix onto the embedded normal submanifold $\overline{\mathcal{N}}$)**.** *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:*

$$\bar{P}_\perp := \mathrm{proj}_{\overline{\mathcal{N}}}(P) = \begin{bmatrix}\Sigma+\mu\mu^\top & \mu^\top\\ \mu & 1\end{bmatrix}, \tag{30}$$

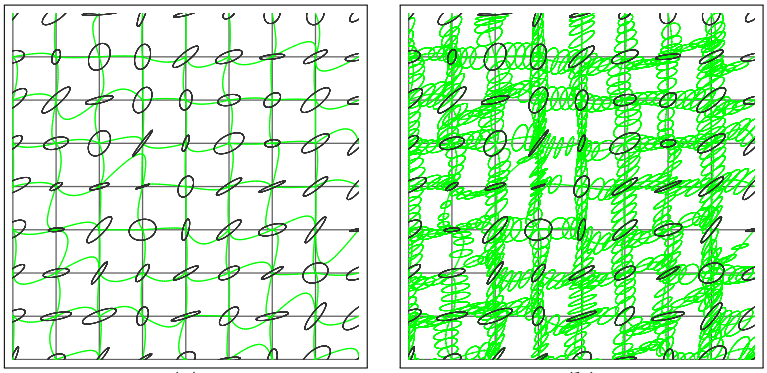
and the SPD distance between P and $\bar{P}_\perp$ is

$$\rho_{\mathcal{P}}(P,\bar{P}_\perp) = \frac{1}{\sqrt{2}}|\log\beta|. \tag{31}$$

Notice that the projection of P is easily computed since $\beta = P_{d+1,d+1}$.

$$\mathrm{proj}_{\overline{\mathcal{N}}}\left(\begin{bmatrix} \Sigma + \beta\mu\mu^\top & \beta\mu \\ \beta\mu^\top & \beta \end{bmatrix}\right) = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu^\top \\ \mu & 1 \end{bmatrix}$$

Remark 4. *In Diffusion Tensor Imaging [39] (DTI), the Fisher–Rao distance can be used to evaluate the distance between three-dimensional normal distributions with means located at a 3D grid position. We may consider* $3 \times 3 \times 3 - 1 = 26$ *neighbor graphs induced by the grid, and for each normal* N *of the grid, calculate the approximations of the Fisher–Rao distance of* N *with its neighbors* N' *as depicted in Figure 9. Then the distance between two tensors* N_1 *and* N_2 *of the 3D grid is calculated as the shortest path on the weighted graph using Dijkstra's algorithm [39].*



(**a**) (**b**)

Figure 9. Diffusion tensor imaging (DTI) on a 2D grid: (**a**) Ellipsoids shown at the 8×8 grid locations with C&O curves in green, and (**b**) some interpolated ellipsoids are further shown along the C&O curves.

Please note that the Fisher–Rao projection of $N_1 = (\mu_1, \Sigma_1)$ onto a submanifold $\mathcal{M}_{\mu_2}$ with fixed mean μ_2 was recently reported in closed form in [72] (Equation (21)):

$$N^* = N\left(\mu_2, \Sigma_1 + \frac{1}{2}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^\top\right),$$

with

$$\rho_{\mathcal{N}}(N_1, N^*) = \frac{1}{\sqrt{2}} \operatorname{arccosh}\left(d + (\mu_2 - \mu_1)^\top \Sigma_1^{-1}(\mu_2 - \mu_1)\right),$$

and the Fisher–Rao projection of $N_1 = (\mu_1, \Sigma_1)$ onto submanifold $\mathcal{M}_{\Sigma_2}$ is the "vertical projection" $N^* = (\mu_1, \Sigma_2)$ (Figure 10) with

$$\rho_{\mathcal{N}}(N_1, N^*) = \rho_{\mathcal{N}_\mu}(\Sigma_1, \Sigma_2).$$

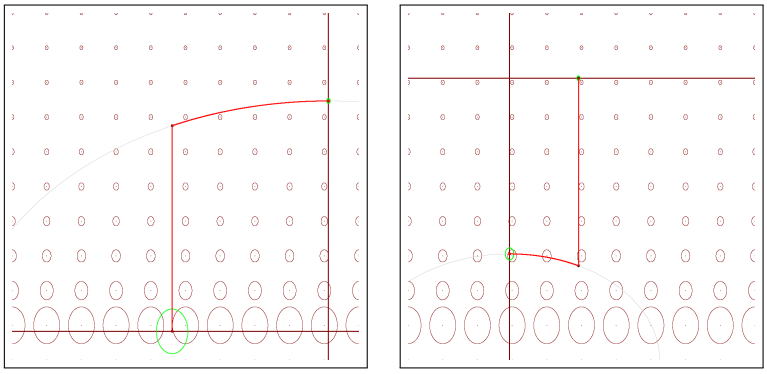


Figure 10. Examples of projection of $N(\mu, \Sigma)$ onto the submanifolds $\mathcal{M}_{\mu_0}$ and $\mathcal{M}_{\Sigma_0}$. Tissot indicatrices are rendered in green at the projected normal distributions $\left(\mu_0, \Sigma + \frac{1}{2}(\mu_0 - \mu)(\mu_0 - \mu)^\top\right)$ and (μ, Σ_0), respectively.

We can upper bound the Fisher–Rao distance $\rho_{\mathcal{N}}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))$ by projecting Σ_1 onto $\mathcal{M}_{\mu_2}$ and projecting Σ_2 onto $\mathcal{M}_{\mu_1}$. Let $\Sigma_{12} \in \mathcal{M}_{\mu_2}$ and $\Sigma_{21} \in \mathcal{M}_{\mu_1}$ denote those Fisher–Rao orthogonal projections. Using the triangular inequality property of the Fisher–Rao distance, we obtain the following upper bounds:

$$\begin{aligned}\rho_{\mathcal{N}}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)) &\leq \rho_{\mathcal{N}}((\mu_1,\Sigma_1),(\mu_2,\Sigma_{12})) + \rho_{\mathcal{N}}(\mu_2,\Sigma_{12},(\mu_2,\Sigma_2)), &(32)\\ &\leq \rho_{\mathcal{N}}((\mu_2,\Sigma_2),(\mu_1,\Sigma_{21})) + \rho_{\mathcal{N}}((\mu_1,\Sigma_{21}),(\mu_1,\Sigma_1)). &(33)\end{aligned}$$

See Figure 11 for an illustration.

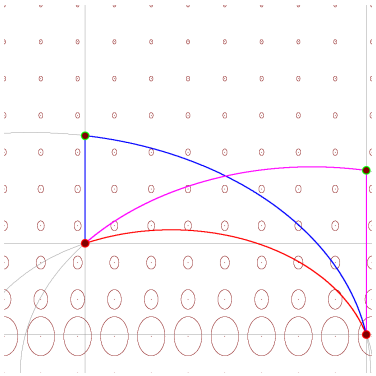


Figure 11. Upper bounding the Fisher–Rao's distance $\rho_{\mathcal{N}}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))$ (red points) using projections (green points) onto submanifolds with fixed means.

Let $\bar{c}_{\mathrm{CO}}(t) = \bar{S}_t$ and $c_{\mathrm{CO}}(t) = f^{-1}(c_{\mathrm{CO}}(t)) =: G_t$. The following proposition shows that we have $D_J[\bar{S}_t,\bar{S}_{t+1}] = D_J[G_t,G_{t+1}]$.

Proposition 3. *The Kullback–Leibler divergence between p_{μ_1,Σ_1} and p_{μ_2,Σ_2} amounts to the KLD 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 $\bar{P}_i = f(\mu_i,\Sigma_i)$:*

$$D_{\mathrm{KL}}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] = D_{\mathrm{KL}}[q_{\bar{P}_1} : q_{\bar{P}_2}].$$

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 [81]. The SPD cone equipped with $\frac{1}{2}$ of the trace metric can be interpreted as Fisher–Rao centered normal manifolds: $(\mathcal{N}_\mu, g^{\mathrm{Fisher}}_{\mathcal{N}_\mu}) = (\mathcal{P}, \frac{1}{2}g^{\mathrm{trace}})$.

Since the determinant of a block matrix is

$$\left|\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right| = \left|A - BD^{-1}C\right|,$$

we obtain with $D = 1$: $|f(\mu,\Sigma)| = |\Sigma + \mu\mu^\top - \mu\mu^\top| = |\Sigma|$.

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

$$\begin{aligned}\mathrm{tr}(\bar{P}_2^{-1}\bar{P}_1) &= \mathrm{tr}\left(\begin{bmatrix} \Sigma_2^{-1} & -\Sigma_2^{-1}\mu_2 \\ -\mu_2^\top\Sigma_2^{-1} & 1+\mu_2^\top\Sigma_2^{-1}\mu_2 \end{bmatrix}\begin{bmatrix} \Sigma_1+\mu_1\mu_1^\top & \mu_1 \\ \mu_1^\top & 1 \end{bmatrix}\right), \\ &= 1+\mathrm{tr}(\Sigma_2^{-1}\Sigma_1+\Delta_\mu^\top\Sigma_2^{-1}\Delta_\mu).\end{aligned}$$

Thus, even if the dimension of the sample spaces of $p_{\mu,\Sigma}$ and $q_{\bar{P}=f(\mu,\Sigma)}$ differs by one, we obtain the same KLD by Calvo and Oller's isometric mapping f.

This property holds for the KLD/Jeffreys divergence D_J but not for all f-divergences [1] I_f in general (e.g., it fails for the Hellinger divergence).

Figure 12 shows the various geodesics and curves used to approximate the Fisher–Rao distance with the Fisher metric shown using Tissot indicatrices.

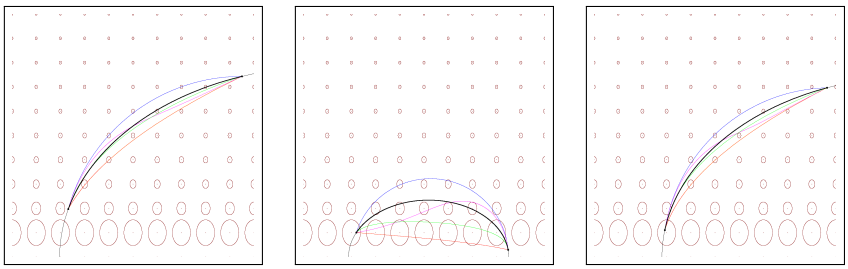


Figure 12. Geodesics and curves used to approximate the Fisher–Rao distance with the Fisher metric shown using Tissot's indicatrices: exponential geodesic (red), mixture geodesic (blue), mid-exponential-mixture curve (purple), projected CO curve (green), and target Fisher–Rao geodesic (black). (Visualization in the parameter space of normal distributions).

Please note that the introduction of parameter β is related to the foliation of the SPD cone $\mathcal{P}$ by $\{f_\beta(\mathcal{N}) : \beta > 0\}$: $\mathcal{P}(d+1) = \mathbb{R}_{>0} \times f_\beta(\mathcal{N})$. See Figure 7. Thus, we may define how good the projected C&O curve is to the Fisher–Rao geodesic by measuring the average distance between points on $\gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)$ and their projections $\overline{\gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)}^{\perp}$ onto $\overline{\mathcal{N}}$:

$$\delta^{\mathrm{CO}}(N_1, N_2) = \delta^{\mathrm{CO}}(\bar{P}_1, \bar{P}_2) = \int_0^1 \rho_{\mathcal{P}}(\gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t), \overline{\gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)}^{\perp})\,\mathrm{d}t.$$

In practice, we evaluate this integral at the sampling points S_t:

$$\delta^{\mathrm{CO}}(P_1, P_2) \approx \delta_T^{\mathrm{CO}}(P_1, P_2) := \frac{1}{T}\sum_{i=1}^{T} \rho_{\mathcal{P}}(S_t, \bar{S}_t), \tag{34}$$

where $S_t = \gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)$ and $\bar{S}_t = \gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)^{\perp}$. We checked experimentally (see Section 3.3) that for close by normals N_1 and N_1, we have $\delta^{\mathrm{CO}}(\bar{N}_1, \bar{N}_2)$ small, and that when N_1 becomes further separated from N_2, the average projection error $\delta^{\mathrm{CO}}(\bar{N}_1, \bar{N}_2)$ increases. Thus, $\delta_T^{\mathrm{CO}}(P_1, P_2)$ is a good measure of the precision of our Fisher–Rao distance approximation.

Lemma 1. *We have* $\rho_{\overline{\mathcal{N}}}(\bar{S}_t, \bar{S}_{t+1}) \leq \rho_{\mathcal{P}}(\bar{S}_t, S_t) + \rho_{\mathcal{P}}(S_t, S_{t+1}) + \rho_{\mathcal{P}}(S_{t+1}, \bar{S}_{t+1})$.

Proof. The proof consists of applying twice the triangle inequality of metric distance $\rho_{\mathcal{P}}$:

$$\begin{aligned}\rho_{\overline{\mathcal{N}}}(\bar{S}_t, \bar{S}_{t+1}) &\leq \rho_{\mathcal{P}}(\bar{S}_t, S_{t+1}) + \rho_{\mathcal{P}}(S_{t+1}, \bar{S}_{t+1}), \\ &\leq \rho_{\mathcal{P}}(\bar{S}_t, S_t) + \rho_{\mathcal{P}}(S_t, S_{t+1}) + \rho_{\mathcal{P}}(S_{t+1}, \bar{S}_{t+1}).\end{aligned}$$

See Figure 13 where the left-hand-side geodesic length is shown in blue and the right-hand-side upper bound is visualized in red. □

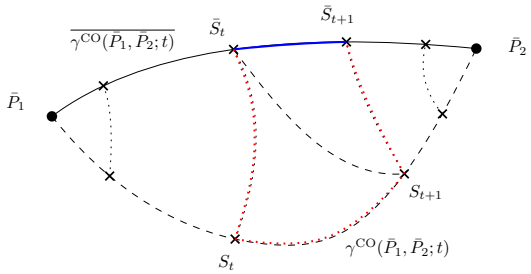


Figure 13. Bounding $\rho_{\mathcal{N}}(\bar{S}_t, \bar{S}_{t+1})$ using the triangular inequality of $\rho_{\mathcal{P}}$ in the SPD cone $\mathcal{P}(d+1)$.

Property 5. *We have* $\rho_{\mathcal{N}}(N_1, N_2) \leq \rho_{\mathcal{N}}^{\mathrm{CO}}(N_1, N_2) \leq \rho_{\mathcal{N}}(N_1, N_2) + 2\,\delta_T^{\mathrm{CO}}(\bar{P}_1, \bar{P}_2)$.

Proof. At infinitesimal scale when $S_{t+1} \approx S_t$, using Lemma 1 and $\rho_{\mathcal{P}}(S_{t+1}, \bar{S}_{t+1}) \approx \rho_{\mathcal{P}}(\bar{S}_t, S_t)$ we have

$$\mathrm{d}s_{\mathcal{N}}(\bar{S}_t) \leq \mathrm{d}s_{\mathcal{P}}(S_t) + 2\rho_{\mathcal{P}}(S_t, \bar{S}_t).$$

Taking the integral along the curve $c^{\mathrm{CO}}(t) = \overline{\gamma^{\mathrm{CO}}(\bar{P}_1, \bar{P}_2; t)}$, we obtain

$$\rho_{\mathcal{N}}^{\mathrm{CO}}(N_1, N_2) \leq \rho_{\mathcal{P}}(\bar{P}_1, \bar{P}_2) + 2\delta_T^{\mathrm{CO}}(\bar{P}_1, \bar{P}_2)$$

Since $\rho_{\mathcal{P}}(\bar{P}_1, \bar{P}_2) \leq \rho_{\mathcal{N}}(N_1, N_2)$, we have

$$\rho_{\mathcal{N}}(N_1, N_2) \leq \rho_{\mathcal{N}}^{\mathrm{CO}}(N_1, N_2) \leq \rho_{\mathcal{N}}(N_1, N_2) + 2\delta_T^{\mathrm{CO}}(\bar{P}_1, \bar{P}_2).$$

□

Notice that $\sum_{i=0}^{T-1} \rho_{\mathcal{P}}(S_t, S_{t+1}) = \rho_{\mathcal{P}}(\bar{P}_1, \bar{P}_2)$.

Example 1. *Let us consider Example 1 of [42] (p. 11):*

$$N_1 = \left(\begin{bmatrix} -1 \\ 0 \end{bmatrix}, \Sigma \right), \quad N_2 = \left(\begin{bmatrix} 6 \\ 3 \end{bmatrix}, \Sigma \right), \Sigma = \begin{bmatrix} 1.1 & 0.9 \\ 0.9 & 1.1 \end{bmatrix}.$$

The Fisher–Rao distance is evaluated numerically in [42] as 5.00648. *We have the lower bound* $\rho_{\mathcal{N}}^{\mathrm{CO}}(N_1, N_2) = 4.20447$, *and the Mahalanobis distance* 8.06226 *upper bounds the Fisher–Rao distance (not totally geodesic submanifold* $\mathcal{N}_\Sigma$*). Our projected C&O curve discretized with* $T = 1000$ *yields an approximation* $\tilde{\rho}_{\mathcal{N}}^{\mathrm{CO}}(N_1, N_2) = 5.31667$. *The average projection distance* $\rho_{\mathcal{P}}(S_t, \bar{S}_t)$ *is* $\delta_T^{\mathrm{CO}}(N_1, N_2) = 0.61791$, *and the maximum projected distance is* 1.00685. *We check that*

$$5.00648 \approx \rho_{\mathcal{N}}(N_1, N_2) \leq \tilde{\rho}_{\mathcal{N}}^{\mathrm{CO}}(N_1, N_2) \approx 5.31667 \leq \rho_{\mathcal{N}}(N_1, N_2) + 2\delta_T^{\mathrm{CO}}(\bar{P}_1, \bar{P}_2) \approx 5.44028.$$

The Killing distance [82] obtained for $\kappa_{\mathrm{Killing}} = 2$ *is* $\rho_{\mathrm{Killing}}(N_1, N_2) \approx 6.82028$ *(see Appendix C). Notice that geodesic shooting is time-consuming compared to our approximation technique.*

3.3. Some Experiments

The KLD D_{KL} and Jeffreys divergence D_J, the Fisher–Rao distance $\rho_{\mathcal{N}}$ and the Calvo and Oller distance ρ_{CO} are all invariant under the congruence action of the affine group $\mathrm{Aff}(d) = \mathbb{R}^d \rtimes \mathrm{GL}(d)$ with the group operation

$$(a_1, A_1)(a_2, A_2) = (a_1 + A_1 a_2, A_1 A_2).$$

Let $(A, a) \in \mathrm{Aff}(d)$, and define the action on the normal space $\mathcal{N}$ as follows:

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

Then we have:

$$\rho_{\mathcal{N}}((A, a).N_1, (A, a).N_2) \quad = \quad \rho_{\mathcal{N}}(N_1, N_2),$$

$$\begin{aligned}\rho_{\mathrm{CO}}((A,a).N_1,(A,a).N_2) &= \rho_{\mathrm{CO}}(N_1,N_2),\\ D_{\mathrm{KL}}[(A,a).N_1:(A,a).N_2] &= D_{\mathrm{KL}}[N_1:N_2].\end{aligned}$$

This invariance extends to our approximations $\tilde{\rho}^c_{\mathcal{N}}$ (see Equation (27)).

Since we have

$$\tilde{\rho}^c_{\mathcal{N}}(N_1,N_2) \approx \rho_{\mathcal{N}}(N_1,N_2) \geq \rho_{\mathrm{CO}}(N_1,N_2),$$

the ratio $\kappa_c = \frac{\tilde{\rho}^c_{\mathcal{N}}}{\rho_{\mathrm{CO}}} \geq \kappa = \frac{\tilde{\rho}^c_{\mathcal{N}}}{\rho_{\mathcal{N}}}$ gives an upper bound on the approximation factor of $\tilde{\rho}^c_{\mathcal{N}}$ compared to the true Fisher–Rao distance $\rho_{\mathcal{N}}$:

$$\kappa_c \rho_{\mathcal{N}}(N_1,N_2) \geq \kappa \rho_{\mathcal{N}}(N_1,N_2) \geq \tilde{\rho}^c_{\mathcal{N}}(N_1,N_2) \approx \rho_{\mathcal{N}}(N_1,N_2) \geq \rho_{\mathrm{CO}}(N_1,N_2).$$

Let us now report some numerical experiments of our approximated Fisher–Rao distances $\tilde{\rho}^x_{\mathcal{N}}$ with $x \in \{l,m,e,\mathrm{em},\mathrm{CO}\}$. Although that dissimilarity $\tilde{\rho}_{\mathcal{N}}$ is positive–definite, it does not satisfy the triangular inequality of metric distances (e.g., Riemannian distances $\rho_{\mathcal{N}}$ and ρ_{CO}).

First, we draw multivariate normals by sampling means $\mu \sim \mathrm{Unif}(0,1)$ and sample covariance matrices Σ as follows: We draw a lower triangular matrix L with entries L_{ij} iid sampled from $\mathrm{Unif}(0,1)$, and take $\Sigma = LL^\top$. We use $T = 1000$ samples on curves and repeat the experiment 1000 times to gather average statistics on κ_c's of curves. Results are summarized in Table 1.

Table 1. First set of experiments demonstrates the advantage of the $c_{\mathrm{CO}}(t)$ curve.

d	κ_{CO}	κ_l	κ_e	κ_m	κ_{em}
1	**1.0025**	1.0414	1.1521	1.0236	1.0154
2	**1.0167**	1.0841	1.1923	1.0631	1.0416
3	**1.0182**	1.8997	2.6072	1.9965	1.07988
4	**1.0207**	2.0793	1.8080	2.1687	1.1873
5	**1.0324**	4.1207	12.3804	5.6170	4.2349

For that scenario that the C&O curve (either $\bar{c}_{\mathrm{CO}} \in \overline{\mathcal{N}}$ or $c_{\mathrm{CO}} \in \mathcal{N}$) performs best compared to the linear interpolation curves with respect to source parameter (l), mixture geodesic (m), exponential geodesic (e), or exponential-mixture mid-curve (em). Let us point out that we sample $\gamma_{\mathcal{P}}(\bar{P}_1,\bar{P}_2;\frac{i}{T})$ for $i \in \{0,\ldots,T\}$.

Strapasson, Porto, and Costa [38] (SPC)reported the following upper bound on the Fisher–Rao distance between multivariate normals

$$\rho_{\mathrm{CO}}(N_1,N_2) \leq \rho_{\mathcal{N}}(N_1,N_2) \leq U_{\mathrm{SPC}}(N_1,N_2),$$

with:

$$U_{\mathrm{SPC}}(N_1,N_2) = \sqrt{2\sum_{i=1}^{d}\log^2\left(\frac{\sqrt{(1+D_{ii})^2+\mu_i^2}+\sqrt{(1-D_{ii})^2+\mu_i^2}}{\sqrt{(1+D_{ii})^2+\mu_i^2}-\sqrt{(1-D_{ii})^2+\mu_i^2}}\right)}, \tag{35}$$

where $\Sigma = \Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}$, $\Sigma = \Omega D \Omega^\top$ is the eigen decomposition, and $\mu = \Omega^\top \Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1)$. This upper bound performs better when the normals are well-separated and worse than the $\sqrt{D_J}$-upper bound when the normals are close to each other.

Let us compare $\rho_{\mathrm{CO}}(N_1,N_2)$ with $\rho_{\mathcal{N}}(N_1,N_2) \approx \tilde{\rho}^{c_{\mathrm{CO}}}(N_1,N_2)$ and the upper bound $U(N_1,N_2)$ by averaging over 1000 trials with N_1 and N_2 chosen randomly as before and $T = 1000$. We have $\rho_{\mathrm{CO}}(N_1,N_2) \leq \rho_{\mathcal{N}}(N_1,N_2) \approx \tilde{\rho}^{c_{\mathrm{CO}}}(N_1,N_2) \leq U(N_1,N_2)$. Table 2 shows that our Fisher–Rao approximation is close to the lower bound (and hence to the underlying true Fisher–Rao distance) and that the upper bound is about twice the lower bound for that particular scenario.

Second, since the distances are invariant under the action of the affine group, we can set wlog. $N_1 = (0,I)$ (standard normal distribution) and let $N_2 = \mathrm{diag}(u_1,\ldots,u_d)$ where

$u_i \sim \mathrm{Unif}(0, a)$. As normals N_1 and N_2 separate from each other, we notice experimentally that the performance of the c_{CO} curve degrades in the second experiment with $a = 5$ (see Table 3): Indeed, the mixture geodesic works experimentally better than the C&O curve when $d \geq 11$.

Table 2. Comparing our Fisher–Rao approximation with the Calvo and Oller lower bound and the upper bound of [38].

d	$\rho_{\mathrm{CO}}(N_1, N_2)$	$\tilde{\rho}^{c_{\mathrm{CO}}}(N_1, N_2)$	$U(N_1, N_2)$
1	1.7563	1.8020	3.1654
2	3.2213	3.3194	6.012
3	4.6022	4.7642	8.7204
4	5.9517	6.1927	11.3990
5	7.156	7.3866	13.8774

Table 3. Second set of experiments shows limitations of the $c_{\mathrm{CO}}(t)$ curve.

d	κ_{CO}	κ_l	κ_e	κ_m
1	**1.0569**	1.1405	1.139	1.0734
5	**1.1599**	1.4696	1.5201	1.1819
10	**1.2180**	1.6963	1.7887	1.2184
11	1.2260	1.7333	1.8285	**1.2235**
12	1.2301	1.7568	1.8539	**1.2282**
15	1.2484	1.8403	1.9557	**1.2367**
20	1.2707	1.9519	2.0851	**1.2466**

Figure 14 display the various curves considered for approximating the Fisher–Rao distance between bivariate normal distributions: For a curve $c(t)$, we visualize its corresponding bivariate normal distributions $(\mu_{c(t)}, \Sigma_{c(t)})$ at several increment steps $t \in [0, 1]$ by plotting the ellipsoid

$$E_{c(t)} = \mu_{c(t)} + \left\{ L^\top x, x = (\cos\theta, \sin\theta), \theta \in [0, 2\pi) \right\},$$

where $\Sigma_{c(t)} = L_{c(t)} L_{c(t)}^\top$.

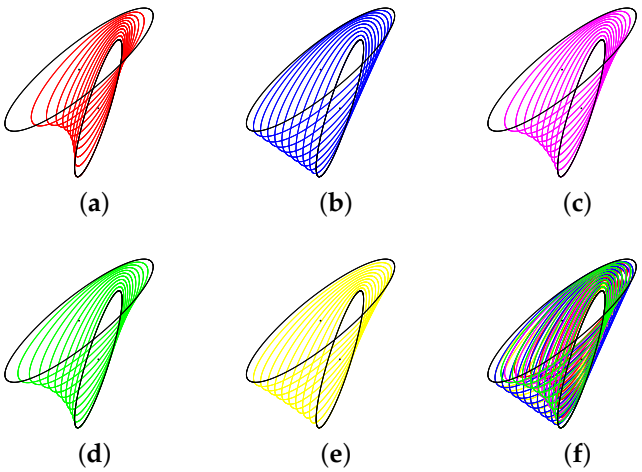


Figure 14. Visualizing at discrete positions (10 increment steps between 0 and 1) some curves used to approximate the Fisher–Rao distance between two bivariate normal distributions: (**a**) exponential geodesic $c^e = \gamma^e_{\mathcal{N}}$ (red), (**b**) mixture geodesic $c^m = \gamma^m_{\mathcal{N}}$ (blue), (**c**) mid-mixture-exponential curve c^{em} (purple), (**d**) projected Calvo and Oller curve c^{CO} (green), (**e**) c^λ: ordinary linear interpolation in λ (yellow), and (**f**) All superposed curves at once.

Example 2. *Let us report some numerical results for bivariate normals with* $T = 1000$*:*

- *We use the following example of Han and Park [39] (Equation (26)):*

$$N_1 = \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix} \right), \quad N_2 = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

  Their geodesic shooting algorithm [39] evaluates the Fisher–Rao distance to $\rho_{\mathcal{N}}(N_1, N_2) \approx$ **3.1329** *(precision* 10^{-5}*).*
  We obtain:
  - *Calvo and Oller lower bound:* $\rho_{\mathrm{CO}}(N_1, N_2) \approx$ **3.0470**,
  - *Upper bound using Equation (15):* 7.92179,
  - *SPC upper bound (Equation (35)):* $U_{\mathrm{SPC}}(N_1, N_2) \approx 5.4302$,
  - $\sqrt{D_J}$ *upper bound:* $U_{\sqrt{J}}(N_1, N_2) \approx$ **4.3704**,
  - $\tilde{\rho}^{\lambda}_{\mathcal{N}}(N_1, N_2) \approx 3.4496$,
  - $\tilde{\rho}^{m}_{\mathcal{N}}(N_1, N_2) \approx 3.5775$,
  - $\tilde{\rho}^{e}_{\mathcal{N}}(N_1, N_2) \approx 3.7314$,
  - $\tilde{\rho}^{\mathrm{em}}_{\mathcal{N}}(N_1, N_2) \approx 3.1672$,
  - $\tilde{\rho}^{\mathrm{CO}}_{\mathcal{N}}(N_1, N_2) \approx$ **3.1391**.

  In that setting, the $\sqrt{D_J}$ *upper bound is better than the upper bound of Equation (35), and the projected Calvo and Oller geodesic yields the best approximation of the Fisher–Rao distance (Figure 15) with an absolute error of* 0.0062 *(about* 0.2% *relative error). When* $T = 10$*, we have* $\tilde{\rho}^{\mathrm{CO}}_{\mathcal{N}}(N_1, N_2) \approx 3.1530$*, when* $T = 100$*, we obtain* $\tilde{\rho}^{\mathrm{CO}}_{\mathcal{N}}(N_1, N_2) \approx 3.1136$*, and when* $T = 500$ *we obtain* $\tilde{\rho}^{\mathrm{CO}}_{\mathcal{N}}(N_1, N_2) \approx 3.1362$ *(which is better than the approximation obtained for* $T = 1000$*). Figure 16 shows the fluctuations of the approximation of the Fisher–Rao distance by the projected C&O curve when T ranges from* 3 *to* 100*.*
- *Bivariate normal* $N_1 = (0, I)$ *and bivariate normal* $N_2 = (\mu_2, \Sigma_2)$ *with* $\mu_2 = [1\ 0]^\top$ *and* $\Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$*. We obtain*
  - *Calvo and Oller lower bound:* 1.4498
  - *Upper bound of Equation (35):* 2.6072
  - $\sqrt{D_J}$ *upper bound:* 1.5811
  - $\tilde{\rho}^{\lambda}$*:* 1.5068
  - $\tilde{\rho}^{m}$*:* 1.5320
  - $\tilde{\rho}^{e}$*:* 1.5456
  - $\tilde{\rho}^{\mathrm{em}}$*:* 1.4681
  - $\tilde{\rho}^{\mathrm{co}}$*:* 1.4673
- *Bivariate normal* $N_1 = (0, I)$ *and bivariate normal* $N_2 = (\mu_2, \Sigma_2)$ *with* $\mu_2 = [5\ 0]^\top$ *and* $\Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$*. We get:*
  - *Calvo and Oller lower bound:* 3.6852
  - *Upper bound of Equation (35):* 6.0392
  - $\sqrt{D_J}$ *upper bound:* 6.2048
  - $\tilde{\rho}^{\lambda}$*:* 5.7319
  - $\tilde{\rho}^{m}$*:* 4.4039
  - $\tilde{\rho}^{e}$*:* 5.9205
  - $\tilde{\rho}^{\mathrm{em}}$*:* 4.2901
  - $\tilde{\rho}^{\mathrm{co}}$*:* 4.3786

See Supplementary Materials for further experiments.

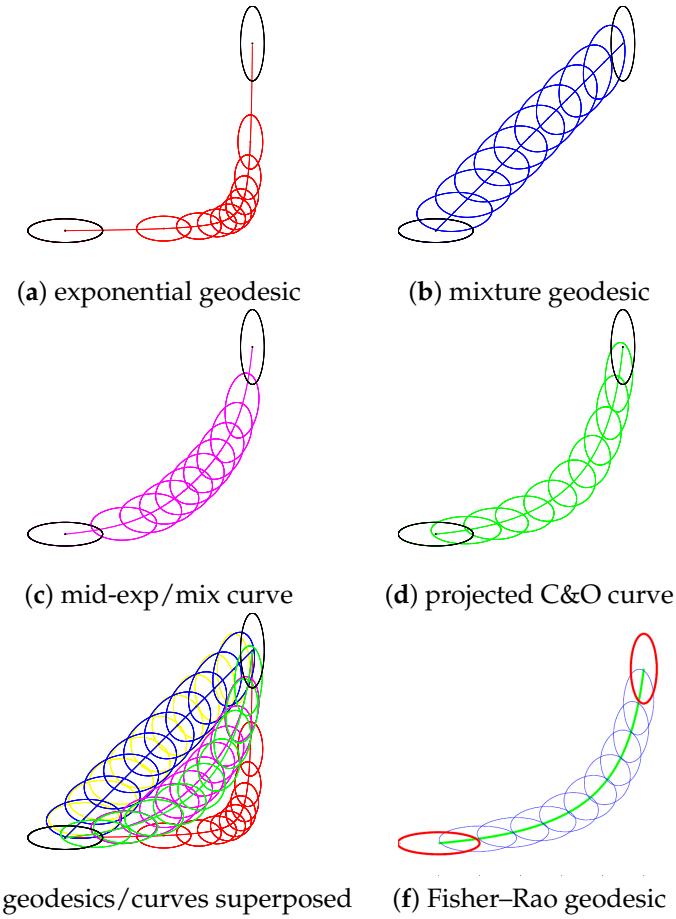


Figure 15. Comparison of our approximation curves with the Fisher–Rao geodesic (**f**) obtained by geodesic shooting (Figure 5 of [39]). Exponential (**a**) and mixture (**b**) geodesics with the mid-exponential-mixture curve (**c**), and the projected C&O curve (**d**). Superposed curves (**e**) and comparison with geodesic shooting (Figure 5 of [39]). Beware that color coding is not related between (**a**) and (**f**), and scale for depicting ellipsoids are different.

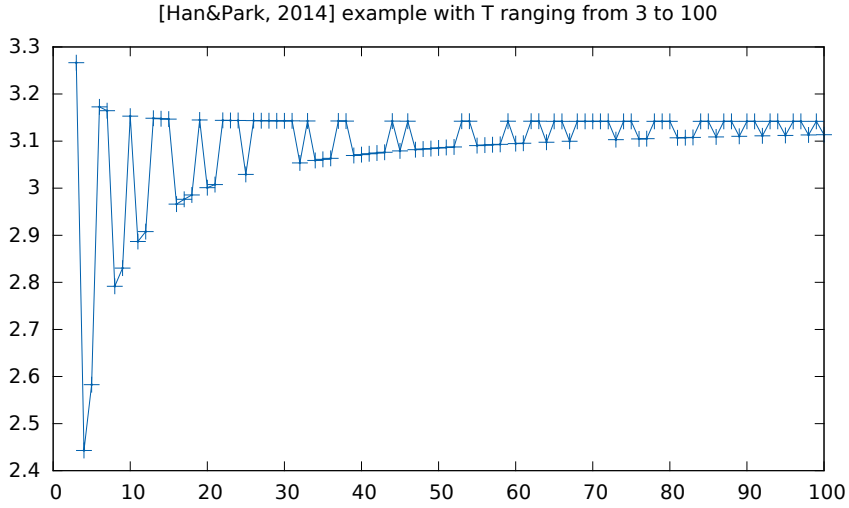


Figure 16. Approximation of the Fisher–Rao distance obtained using the projected C&O curve when T ranges from 3 to 100 [39].

4. Approximating the Smallest Enclosing Fisher–Rao Ball of MVNs

We may use these closed-form distance $\rho_{\mathrm{CO}}(N, N')$ between N and N' to compute an approximation (of the center) of the smallest enclosing Fisher–Rao ball $B^* = \mathrm{ball}(C^*, r^*)$ of a set $\mathcal{G} = \{N_1 = (\mu_1, \Sigma_1), \ldots, N_n = (\mu_n, \Sigma_n)\}$ of n d-variate normal distributions:

$$C^* = \arg\min_{C \in \mathcal{N}} \max_{i \in \{1,\ldots,n\}} \rho_{\mathcal{N}}(C, N_i)$$

where $\mathrm{ball}(C, r) = \{N \in \mathcal{N} : \rho_{\mathcal{N}}(C, N) \leq r\}$.

The method proceeds as follows:

- First, we convert MVN set $\mathcal{G}$ into the equivalent set of $(d+1)$-dimensional SPD matrices $\bar{\mathcal{G}} = \{\bar{P}_i = f(N_i)\}$ using the C&O embedding. We relax the problem of approximating the circumcenter C^* of the smallest enclosing Fisher–Rao ball by

$$P^* = \arg\min_{P \in \mathbb{P}(d+1)} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{CO}}(P, \bar{P}_i).$$

- Second, we approximate the center of the smallest enclosing Riemannian ball of $\bar{\mathcal{G}}$ using the iterative smallest enclosing Riemannian ball algorithm in [66] with say $T = 1000$ iterations. Let $\tilde{P} \in \mathbb{P}(d+1)$ denote this approximation center: $P_T = \mathrm{RieSEB}_{\mathrm{SPD}}(\bar{\mathcal{G}}, T)$.
- Finally, we project back P_T onto $\overline{\mathcal{N}}$: $\bar{P}_T = \mathrm{proj}_{\overline{\mathcal{N}}}(P_T)$. We return $\bar{P}_T$ as the approximation of C^*.

Algorithm [66] $\mathrm{RieSEB}_{\mathrm{SPD}}(\{P_1, \ldots, P_n\}, T)$ is described for a set of SPD matrices $\{P_1, \ldots, P_n\}$ as follows:

- Let $C_1 \leftarrow P_1$
- For $t = 1$ to T
  - Compute the index of the SPD matrix which is farthest from the current circumcenter C_t:

$$f_t = \arg\max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{SPD}}(C_t, P_i)$$

  - Update the circumcenter by walking along the geodesic linking C_t to P_{f_t}:

$$C_{t+1} = \gamma_{\mathrm{SPD}}\left(C_t, P_{f_t}; \frac{1}{t+1}\right) = C_t^{\frac{1}{2}} (C_t^{-\frac{1}{2}} P_{f_t} C_t^{-\frac{1}{2}})^{\frac{1}{t+1}} C_t^{\frac{1}{2}}$$

- Return C_T

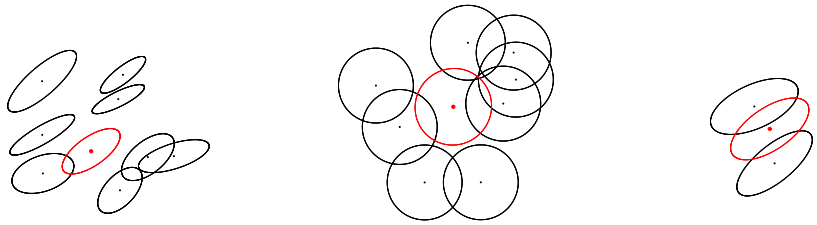
The convergence of the algorithm $\mathrm{RieSEB}_{\mathrm{SPD}}$ follows from the fact that the SPD trace manifold is a Hadamard manifold (with negative sectional curvatures). See [66] for proof of convergence.

The SPD distance $\rho_{\mathcal{P}}(C_T, \bar{C}_T)$ indicates the quality of the approximation. Figure 17 shows the result of implementing this heuristic.

Let us notice that when all MVNs share the same covariance matrix Σ, we have from Equation (18) or Equation (23) that $\rho_{\mathcal{N}}(\mu_1, \Sigma), N(\mu_2, \Sigma)$ and $\rho_{\mathrm{CO}}(N(\mu_1, \Sigma), N(\mu_2, \Sigma))$ are strictly increasing function of their Mahalanobis distance. Using the Cholesky decomposition $\Sigma^{-1} = LL^\top$, we deduce that the smallest Fisher–Rao enclosing ball coincides with the smallest Calvo and Oller enclosing ball, and the circumcenter of that ball can be found as an ordinary Euclidean circumcenter [83] (Figure 17b). Please note that in 1D, we can find the exact smallest enclosing Fisher–Rao ball as an equivalent smallest enclosing ball in hyperbolic geometry.

Furthermore, we may extend the computation of the approximated circumcenter to k-center clustering [84] of n multivariate normal distributions. Since the circumcenter of the clusters is approximated and not exact, we extend straightforwardly the variational approach of k-means described in [85] to k-center clustering. An application of k-center clustering of MVNs is to simplify a Gaussian mixture model [42] (GMM).

Similarly, we can consider other Riemannian distances with closed-form formulas between MVNs such as the Killing distance in the symmetric space [82] (see Appendix C) or the Siegel-based distance proposed in Appendix D.



(**a**) (**b**) (**c**)

Figure 17. Approximation of the smallest enclosing Riemannian ball of a set of n bivariate normals $N_i = N(\mu_i, \Sigma_i)$ with respect to C&O distance ρ_{CO} (the approximate circumcenter $\bar{C}_T$ is depicted as a red ellipse): (**a**) $n = 8$ with different covariance matrices, (**b**) $n = 8$ with identical covariance matrices amount to the smallest enclosing ball of a set of n points $\{\mu_i\}$, (**c**) $n = 2$ displays the midpoint of the C&O geodesic visualized as an equivalent bivariate normal distribution in the sample space.

5. Some Information–Geometric Properties of the C&O Embedding

In information geometry [1], the manifold $\mathcal{N}$ admits a dual structure denoted by the quadruple

$$(\mathcal{N}, g_{\mathcal{N}}^{\mathrm{Fisher}}, \nabla_{\mathcal{N}}^{e}, \nabla_{\mathcal{N}}^{m}),$$

when equipped with the exponential connection $\nabla_{\mathcal{N}}^{e}$ and the mixture connection $\nabla_{\mathcal{N}}^{m}$. The connections $\nabla_{\mathcal{N}}^{e}$ and $\nabla_{\mathcal{N}}^{m}$ are said to be dual since $\frac{\nabla_{\mathcal{N}}^{e}+\nabla_{\mathcal{N}}^{m}}{2} = \bar{\nabla}_{\mathcal{N}}$, the Levi–Civita connection induced by $g_{\mathcal{N}}^{\mathrm{Fisher}}$. Furthermore, by viewing $\mathcal{N}$ as an exponential family $\{p_\theta\}$ with natural parameter $\theta = (\theta_v, \theta_M)$ (using the sufficient statistics [80] $(x, -xx^\top)$), and taking the convex log-normalizer function $F_{\mathcal{N}}(\theta)$ of the normals, we can build a dually flat space [1] where the canonical divergence amounts to a Bregman divergence which coincides with the reverse Kullback–Leibler divergence [30,86] (KLD). The Legendre duality

$$F^*(\eta) = \langle \nabla F(\theta), \eta \rangle - F(\nabla F(\theta))$$

(with $\langle (v_1, M_1), (v_2, M_2) \rangle = \mathrm{tr}(v_1 v_2^\top + M_1 M_2^\top) = v_1 \cdot v_2 + \mathrm{tr}(M_1 M_2^\top)$) yields: $\theta = (\theta_v, \theta_M) = \left(\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}\right)$,

$$F_{\mathcal{N}}(\theta) = \frac{1}{2}\left(d \log \pi - \log |\theta_M| + \frac{1}{2}\theta_v^\top \theta_M^{-1} \theta_v\right),$$

$\eta = (\eta_v, \eta_M) = \nabla F_{\mathcal{N}}(\theta) = \left(\frac{1}{2}\theta_M^{-1}\theta_v, \theta_M^{-1}\right)$,

$$F_{\mathcal{N}}^*(\eta) = -\frac{1}{2}\left(\log(1 + \eta_v^\top \eta_M^{-1} \eta_v) + \log |-\eta_M| + d(\log 2\pi e)\right),$$

and we have

$$B_{F_{\mathcal{N}}}(\theta_1, \theta_2) = D_{\mathrm{KL}}^*(p_{\lambda_1} : p_{\lambda_2}) = D_{\mathrm{KL}}(p_{\lambda_2} : p_{\lambda_1}) = B_{F_{\mathcal{N}}^*}(\eta_2 : \eta_1),$$

where $D_{\mathrm{KL}}^*[p : q] = D_{\mathrm{KL}}[q : p]$ is the reverse KLD.

In a dually flat space, we can express the canonical divergence as a Fenchel–Young divergence using the mixed coordinate systems $B_{F_{\mathcal{N}}}(\theta_1 : \theta_2) = Y_{F_{\mathcal{N}}}(\theta_1 : \eta_2)$ where $\eta_i = \nabla F_{\mathcal{N}}(\theta_i)$ and

$$Y_{F_{\mathcal{N}}}(\theta_1 : \eta_2) := F_{\mathcal{N}}(\theta_1) + F_{\mathcal{N}}^*(\eta_2) - \langle \theta_1, \eta_2 \rangle.$$

The moment η-parameterization of a normal is $(\eta = \mu, H = -\Sigma - \mu\mu^\top)$ with its reciprocal function $(\lambda = \eta, \Lambda = -H - \eta\eta^\top)$.

Let $F_{\mathcal{P}}(P) = F_{\mathcal{N}}(0, P)$, $\bar{\theta} = \frac{1}{2}\bar{P}^{-1}$, $\bar{\eta} = \nabla F_{\mathcal{P}}(\bar{\theta})$. Then we have the following proposition which proves that the Fenchel–Young divergences in $\mathcal{N}$ and $\overline{\mathcal{N}}$ (as a submanifold of $\mathcal{P}$) coincide:

Proposition 4. *We have*

$$\begin{aligned} D_{\mathrm{KL}}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] &= B_{F_{\mathcal{N}}}(\theta_2 : \theta_1) = Y_{F_{\mathcal{N}}}(\theta_2 : \eta_1) = Y_{F_{\mathcal{P}}}(\bar{\theta}_2 : \bar{\eta}_1) \\ &= B_{F_{\mathcal{P}}}(\bar{\theta}_2 : \bar{\theta}_1) = D_{\mathrm{KL}}[p_{0,\bar{P}_1=f(\mu_1,\Sigma_2)} : p_{0,\bar{P}_2=f(\mu_2,\Sigma_2)}]. \end{aligned}$$

Consider now the ∇^e-geodesics and ∇^m-geodesics on $\mathcal{N}$ (linear interpolation with respect to natural and dual moment parameterizations, respectively): $\gamma^e_{\mathcal{N}}(N_1, N_2; t) = (\mu^e_t, \Sigma^e_t)$ and $\gamma^m_{\mathcal{N}}(N_1, N_2; t) = (\mu^m_t, \Sigma^m_t)$.

Proposition 5 (Mixture geodesics preserved)**.** *The mixture geodesics are preserved by the embedding f: $f(\gamma^m_{\mathcal{N}}(N_1, N_2; t)) = \gamma^m_{\mathcal{P}}(f(N_1), f(N_2); t)$. The exponential geodesics are preserved for the subspace of $\mathcal{N}$ with fixed mean μ: $\mathcal{N}_\mu$.*

Proof. For the m-geodesics, let us check that

$$f(\mu^m_t, \Sigma^m_t) = \begin{bmatrix} \Sigma^m_t + \mu^m_t {\mu^m_t}^\top & \mu^m_t \\ (\mu^m_t)^\top & 1 \end{bmatrix} = t \underbrace{f(\mu_1, \Sigma_1)}_{\bar{P}_1} + (1-t) \underbrace{f(\mu_2, \Sigma_2)}_{\bar{P}_2},$$

since $\Sigma^m_t + \mu_t {\mu^m_t}^\top = \bar{\Sigma}_t + t\mu_1\mu_1^\top + (1-t)\mu_2\mu_2^\top = t(\Sigma_1 + \mu_1\mu_1^\top) + (1-t)(\Sigma_2 + \mu_2\mu_2^\top)$. Thus, we have $f(\gamma^m_{\mathcal{N}}(N_1, N_2; t)) = \gamma^m_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)$. □

Therefore, all algorithms on $\mathcal{N}$ which only require m-geodesics or m-projections [1] by minimizing the right-hand side of the KLD can be implemented by algorithms on $\mathcal{P}$. See, for example, the minimum enclosing ball approximation algorithm called BBC in [87]. Notice that $\overline{\mathcal{N}}_\mu$ (fixed mean normal submanifolds) preserve both mixture and exponential geodesics: The submanifolds $\overline{\mathcal{N}}_\mu$ are said to be doubly autoparallel [88].

Remark 5. *In [2] (p. 355), exercises 13.8 and 13.9 ask to prove the equivalence of the following statements for $\mathcal{S}$ a submanifold of $\mathcal{M}$:*

- *$\mathcal{S}$ is an exponential family $\Leftrightarrow$ $\mathcal{S}$ is ∇^1-autoparallel in $\mathcal{M}$ (exercise 13.8),*
- *$\mathcal{S}$ is a mixture family $\Leftrightarrow$ $\mathcal{S}$ is ∇^{-1}-autoparallel in $\mathcal{M}$ (exercise 13.9).*

Let $\bar{P} = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix}$ (with $|\bar{P}| = |\Sigma|$), $\bar{P}^{-1} = \begin{bmatrix} \Sigma^{-1} & -\Sigma^{-1}\mu \\ -\mu^\top\Sigma^{-1} & 1 + \mu^\top\Sigma^{-1}\mu \end{bmatrix}$, and $y = (x, 1)$. Then we have

$$\begin{aligned} q_{\bar{P}}(y) &= \frac{1}{(2\pi)^{\frac{d+1}{2}}\sqrt{|\bar{P}|}} \exp\left(-\frac{1}{2} y^\top \bar{P}^{-1} y\right), \\ &= \frac{1}{(2\pi)^{\frac{d+1}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} y^\top \bar{P}^{-1} y\right), \\ &= \frac{1}{(2\pi)^{\frac{d+1}{2}}\sqrt{|\Sigma|}} \exp\left([x^\top\ 1] \begin{bmatrix} \Sigma^{-1} & -\Sigma^{-1}\mu \\ -\mu^\top\Sigma^{-1} & 1 + \mu^\top\Sigma^{-1}\mu \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}\right). \end{aligned}$$

Thus, $\overline{\mathcal{N}} = \{q_{\bar{P}}(x, 1)\}$ is an exponential family. Therefore, we deduce that $\mathcal{P}$ is ∇^e-autoparallel in $\mathcal{P}$. However, $\overline{\mathcal{N}}$ is not a mixture family and thus $\mathcal{P}$ is not ∇^m-autoparallel in $\mathcal{P}$.

6. Conclusions and Discussion

In general, the Fisher–Rao distance between multivariate normals (MVNs) is not known in closed form. In practice, the Fisher–Rao distance is usually approximated by costly geodesic shooting techniques [39–41] which requires time-consuming computations of the Riemannian exponential map and are nevertheless limited to normals within a

short range of each other. In this work, we consider a simple alternative approach for approximating the Fisher–Rao distance by approximating the Riemannian lengths of curves, which admits closed-form parameterizations. In particular, we considered the mixed exponential-mixture curved and the projected symmetric positive–definite matrix geodesic obtained from Calvo and Oller isometric submanifold embedding into the SPD cone [19]. We summarize our method to approximate $\rho_{\mathcal{N}}(N_1, N_2)$ between $N_1 = N(\mu_1, \Sigma_1)$ and $N_2 = N(\mu_2, \Sigma_2)$ as follows:

$$\tilde{\rho}_T^{\mathrm{CO}}(N_1, N_2) := \frac{1}{T} \sum_{i=1}^{T-1} \sqrt{D_J[\bar{S}_t, \bar{S}_{t+1}]},$$

where

$$\bar{S}_t = \mathrm{proj}_{\overline{\mathcal{N}}}(S_t)), \quad \mathrm{proj}_{\overline{\mathcal{N}}}\left(\begin{bmatrix} \Sigma + \beta\mu\mu^\top & \beta\mu \\ \beta\mu^\top & \beta \end{bmatrix} \right) = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu^\top \\ \mu & 1 \end{bmatrix}$$

and

$$S_t = \bar{P}_1^{\frac{1}{2}} \left(\bar{P}_1^{-\frac{1}{2}} \bar{P}_2^{\frac{1}{2}} \bar{P}_1^{-\frac{1}{2}} \right)^{\frac{t}{T}} \bar{P}_1^{\frac{1}{2}}$$

with

$$\bar{P}_1 = f(N_1) = \begin{bmatrix} \Sigma_1 + \mu_1\mu_1^\top & \mu_1 \\ \mu_1^\top & 1 \end{bmatrix}, \quad \bar{P}_2 = f(N_2) = \begin{bmatrix} \Sigma_2 + \mu_2\mu_2^\top & \mu_2 \\ \mu_2^\top & 1 \end{bmatrix}.$$

We proved the following sandwich bounds of our approximation

$$\rho_{\mathcal{N}}(N_1, N_2) \leq \tilde{\rho}_T^{\mathrm{CO}}(N_1, N_2) \leq \rho_{\mathcal{N}}(N_1, N_2) + 2\delta_T^{\mathrm{CO}}(\bar{P}_1, \bar{P}_2),$$

where

$$\delta_T^{\mathrm{CO}}(P_1, P_2) := \frac{1}{T} \sum_{i=1}^{T} \rho_{\mathcal{P}}(S_t, \bar{S}_t).$$

Notice that we may calculate equivalently $D_J[\bar{S}_t, \bar{S}_{t+1}]$ as $D_J[G_t, G_{t+1}]$ where $G_i = f^{-1}(\bar{S}_i) = N(m_i, C_i)$ for $i \in \{0, \ldots, T\}$ (see Proposition 3).

We also reported a fast way to upper bound the Fisher–Rao distance by the square root of Jeffreys' divergence: $\rho_{\mathcal{N}}(N_1, N_2) \leq \sqrt{D_J[N_1, N_2]}$ which is tight at infinitesimal scale. In practice, this upper bound beats the upper bound of [38] when normal distributions are not too far from each other. Finally, we show that not only is Calvo and Oller SPD submanifold embedding [19] isometric, but it also preserves the Kullback–Leibler divergence, the Fenchel–Young divergence, and the mixture geodesics. Our approximation technique extends to elliptical distribution, which generalizes multivariate normal distributions [32,55]. Moreover, we obtained a closed form for the Fisher–Rao distance between normals sharing the same covariance matrix using the technique of maximal invariance under the action of the affine group in Section 1.5. We may also consider other distances different from the Fisher–Rao distance, which admits a closed-form formula: For example, the Calvo and Oller metric distance [19] (a lower bound on the Fisher–Rao distance) or the metric distance proposed in [82] (see Appendix C) whose geodesics enjoys the asymptotic property of the Fisher–Rao geodesics [89]). The C&O distance is very well-suited for short Fisher–Rao distances while the symmetric space distance is well-tailored for large Fisher–Rao distances. The calculations of these closed-form distances rely on generalized eigenvalues. We also propose an embedding of normals into the Siegel upper space in Appendix D. To conclude, let us propose yet another alternative distance, The Hilbert projective distance on the SPD cone [90], which only needs to calculate the minimal and maximal eigenvalues (say, using the power iteration method [91]):

$$\rho_{\mathrm{Hilbert}}(P_1, P_2) = \log\left(\frac{\lambda_{\max}(P_1^{-1}P_2)}{\lambda_{\min}(P_1^{-1}P_2)} \right). \tag{36}$$

The dissimilarity is said projective on the SPD cone because $\rho_{\mathrm{Hilbert}}(P_1, P_2) = 0$ if and only if $P_1 = \lambda P_2$ for some $\lambda > 0$. However, let us notice that it yields a proper metric distance on $\overline{\mathcal{N}}$:

$$\rho_{\mathrm{Hilbert}}(N_1, N_2) := \rho_{\mathrm{Hilbert}}(\bar{P}_1, \bar{P}_2),$$

since $\bar{P}_1 = \lambda \bar{P}_2$ if and only if $\lambda = 1$ because the array element $(P_1)_{d+1,d+1} = (P_2)_{d+1,d+1} = 1$, i.e., $\bar{P}_1 = \bar{P}_2$ implying $P_1 = P_2$ by the isometric diffeomorphism f.

Notice that since $\lambda_{\max}(P) = \lambda_{\min}(P^{-1})$, $\lambda_{\min}(P) = \lambda_{\max}(P^{-1})$, $\lambda_{\max}(P_1 P_2) \leq \lambda_{\max}(P_1)\lambda_{\max}(P_2)$, and $\lambda_{\min}(P_1 P_2) \geq \lambda_{\min}(P_1)\lambda_{\min}(P_2)$, we have the following upper bound on Hilbert distance: $\rho_{\mathrm{Hilbert}}(P_1, P_2) \leq \log \frac{\lambda_{\max}(P_1)}{\lambda_{\min}(P_1)} + \log \frac{\lambda_{\max}(P_2)}{\lambda_{\min}(P_2)}$.

Supplementary Materials: The following supporting information can be downloaded at: https://franknielsen.github.io/FisherRaoMVN.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Acknowledgments: I warmly thank Frédéric Barbaresco (Thales) and Mohammad Emtiyaz Khan (Riken AIP) for fruitful discussions about this work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Entities	
$N(\mu, \Sigma)$	d-variate normal distribution (mean μ, covariance matrix Σ)
$p_{(\mu,\Sigma)}(x)$	Probability density function of $N(\mu, \Sigma)$
$q_\Sigma(y) = p_{(0,\Sigma)}(y)$	Probability density function of $N(0, \Sigma)$
$P = (P_{ij})$	Positive–definite matrix with matrix entries P_{ij}
Mappings	
$\bar{P} = f_1(N)$	Calvo and Oller mapping [19] (1990)
$\hat{P} = f_{-\frac{1}{d+1},1}(N) = \hat{f}(N)$	Calvo and Oller mapping [32] (2002) or [82]
Groups	
$\mathrm{GL}(d)$	Group of linear transformations (invertible $d \times d$ matrices)
$\mathrm{SL}(d)$	Special linear group ($d \times d$ matrices with unit determinant)
$\mathrm{Aff}(d)$	Affine group of dimension d
Sets	
$\mathcal{N}$	Set of multivariate normal distributions $N(\mu, \Sigma)$ (MVNs)
$\mathrm{Sym}(d)$	Set of symmetric $d \times d$ real matrices
$\mathbb{P}$	Symmetric positive–definite matrix cone (SPD matrix cone)
$\mathbb{P}_c$	Set of SPD matrices with fixed determinant c ($\mathbb{P} = \mathbb{R}_{>0} \times \mathbb{P}_c$)
SSPD, $\mathbb{P}_1$	Set of SPD matrices with unit determinant
Λ	Parameter space of $N(\mu, \Sigma)$: $\mathbb{R}^d \times \mathbb{P}(d)$
$\mathcal{N}_0$, $\mathcal{P}$	Set of zero-centered normal distributions $N(0, \Sigma)$
$\mathcal{N}_\Sigma$	Set of normal distributions $N(\mu, \Sigma)$ with fixed Σ
$\mathcal{N}_\mu$	Set of normal distributions $N(\mu, \Sigma)$ with fixed μ
$\overline{\mathcal{N}}$	Set of SPD matrices $f(N)$
Riemannian length elements	
MVN Fisher	$\mathrm{d}s^2_{\mathrm{Fisher},\mathcal{N}} = \mathrm{d}\mu^\top \Sigma^{-1} \mathrm{d}\mu + \frac{1}{2}\mathrm{tr}\left((\Sigma^{-1}\mathrm{d}\Sigma)^2 \right)$
0-MVN Fisher	$\mathrm{d}s^2_{\mathrm{Fisher},\mathcal{N}_0} = \frac{1}{2}\mathrm{tr}\left((\Sigma^{-1}\mathrm{d}\Sigma)^2 \right)$
SPD trace	$\mathrm{d}s^2_{\beta,\mathrm{trace}} = \beta \mathrm{tr}((P\,\mathrm{d}P)^2)$ (when $\beta = \frac{1}{2}$, $\mathrm{d}s_{\mathrm{trace}} = \mathrm{d}s_{\mathrm{Fisher},\mathcal{N}_0}$)
SPD Calvo and Oller metric	$\mathrm{d}s^2_{\mathrm{CO}} = \frac{1}{2}\left(\frac{\mathrm{d}\beta}{\beta}\right)^2 + \beta \mathrm{d}\mu^\top \Sigma^{-1}\mathrm{d}\mu + \frac{1}{2}\mathrm{tr}\left((\Sigma^{-1}\mathrm{d}\Sigma)^2 \right)$ (with $\mathrm{d}s_{\mathrm{CO}} = \mathrm{d}s_{\mathcal{P}}(f(\mu, \Sigma))$) when $\beta = 1$, $\mathrm{d}s_{\mathrm{CO}} = \mathrm{d}s_{\mathrm{Fisher},\mathcal{N}}$ in $\overline{\mathcal{N}}$
SPD symmetric space	$\mathrm{d}s^2_{\mathrm{SS}} = \frac{1}{2}\mathrm{d}\mu^\top \Sigma^{-1}\mathrm{d}\mu + \mathrm{tr}\left((\Sigma^{-1}\mathrm{d}\Sigma)^2 \right) - \frac{1}{2}\mathrm{tr}^2(\Sigma^{-1}\mathrm{d}\Sigma)$
Siegel upper space	$\mathrm{d}s^2_{\mathbb{SH}}(Z) = 2\mathrm{tr}(Y^{-1}\mathrm{d}Z\, Y^{-1}\mathrm{d}\bar{Z})$ ($\mathrm{d}s_{\mathbb{SH}}(iY) = 2\mathrm{d}s_{\mathrm{Fisher},\mathcal{N}_0}$)

Manifolds and submanifolds	
$\mathcal{M}$ $(=\mathcal{M}_{\mathcal{N}})$	Manifold of multivariate normal distributions
$T_p\mathcal{M}$	Tangent space at $p \in \mathcal{M}$
$\mathcal{S}_\mu \subset \mathcal{M}$	Submanifold of MVNs with μ prescribed
$\mathcal{S}_\Sigma \subset \mathcal{M}$	Submanifold of MVNs with Σ prescribed
$\mathcal{M}_\Sigma$	manifold of $\mathcal{N}_\Sigma$ (non-embedded in $\mathcal{M}$)
$\mathcal{M}_\mu$	manifold of $\mathcal{N}_\mu$ (non-embedded in $\mathcal{M}$)
$\mathcal{S}_{[v],\Sigma}$	Submanifold of MVN set $\{N(\lambda v, \Sigma) \ : \ \lambda > 0\}$ where v is an eigenvector of Σ
$\mathcal{P}$	manifold of symmetric positive–definite matrices
Distances	
$\rho_N(N_1, N_2)$	Fisher–Rao distance between normal distributions N_1 and N_2
$\rho_{\mathrm{SPD}}(P_1, P_2)$	Riemannian SPD distance between P_1 and P_2
$\rho_{\mathrm{CO}}(N_1, N_2)$	Calvo and Oller distance from embedding N to $\bar{P} = f(N)$
$\rho_{\mathrm{SS}}(N_1, N_2)$	Symmetric space distance from embedding N to $\hat{P} = \hat{f}(N)$
$\rho_{\mathrm{Hilbert}}(N_1, N_2)$	Hilbert distance $\rho_{\mathrm{Hilbert}}(\bar{P}_1, \bar{P}_2)$
$D_{\mathrm{KL}}(N_1, N_2)$	Kullback–Leibler divergence between MVNs N_1 and N_2
$D_J(N_1, N_2)$	Jeffreys divergence between MVNs N_1 and N_2
$D_{\mathrm{CO}}(N_1, N_2)$	Calvo and Oller dissimilarity measure of Equation (26)
Geodesics and curves	
$\gamma^{\mathrm{FR}}_{\mathcal{N}}(N_1, N_2; t)$	Fisher–Rao geodesic between MVNs N_1 and N_2
$\gamma^{\mathrm{FR}}_{\mathcal{P}}(P_1, P_2; t)$	Fisher–Rao geodesic between SPD P_1 and P_2
$\gamma^{e}_{\mathcal{N}}(N_1, N_2; t)$	exponential geodesic between MVNs N_1 and N_2
$\gamma^{m}_{\mathcal{N}}(N_1, N_2; t)$	mixture geodesic between MVNs N_1 and N_2
$\gamma^{\mathrm{CO}}_{\mathcal{N}}(N_1, N_2; t)$	projection curve (not geodesic) of $\gamma_{\mathcal{P}}(\bar{P}_1, \bar{P}_2; t)$ onto $\overline{\mathcal{N}}$
Metrics and connections	
$g^{\mathrm{Fisher}}_{\mathcal{N}}$	Fisher information metric of MVNs
$g_{\mathbb{P}}$	trace metric
$g_{\mathcal{P}}$	Fisher information metric of centered MVNs
g^{Killing}	Killing metric studied in [82]
$\nabla^{\mathrm{Fisher}}_{\mathcal{N}}$	Levi–Civita metric connection
$\nabla^{e}_{\mathcal{N}}$	exponential connection
$\nabla^{m}_{\mathcal{N}}$	mixture connection

Appendix A. Geodesics on the Fisher–Rao Normal Manifold

Appendix A.1. Parametric Equations of the Fisher–Rao Geodesics between Univariate Normal Distributions

The Fisher–Rao geodesics $\gamma^{\mathrm{FR}}_{\mathcal{N}}(N_1, N_2)$ on the Fisher–Rao univariate normal manifolds are either vertical line segments when $\mu_1 = \mu_2$, or semi-circle with origin on the x-axis and x-axis stretched by $\sqrt{2}$ [92] (Figure A1):

$$\gamma^{\mathrm{FR}}_{\mathcal{N}}(\mu_1, \sigma_1; \mu_2, \sigma_2) = \begin{cases} (\mu, (1-t)\sigma_1 + t\sigma_2), & \mu_1 = \mu_2 = \mu \\ (\sqrt{2}(c + r\cos t, r \sin t), t \in [\min\{\theta_1, \theta_2\}, \max\{\theta_1, \theta_2\}], & \mu_1 \neq \mu_2, \end{cases},$$

where

$$c = \frac{\frac{1}{2}(\mu_2^2 - \mu_1^2) + \sigma_2^2 - \sigma_1^2}{\sqrt{2}(\mu_1 - \mu_2)}, \quad r = \sqrt{\left(\frac{\mu_i}{\sqrt{2}} - c\right)^2 + \sigma_i^2}, i \in \{1, 2\},$$

and

$$\theta_i = \arctan\left(\frac{\sigma_i}{\frac{\mu_i}{\sqrt{2}} - c}\right), i \in \{1, 2\},$$

provided that $\theta_i \geq 0$ for $i \in \{1, 2\}$ (otherwise, we let $\theta_i \leftarrow \theta_i + \pi$).

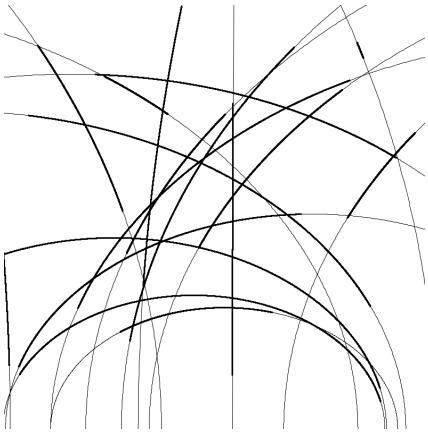


Figure A1. Visualizing some Fisher–Rao geodesics of univariate normal distributions on the stretched Poincaré upper plane (semi-circles with origin on the x-axis and stretched by $\sqrt{2}$ on the x-axis). Full geodesics are plotted with a thin gray style and geodesic arcs are plotted with a thick black style.

Notice that it is remarkable that the Fisher–Rao distance between normal distributions is available in closed form: Indeed, the Euclidean length (with respect to the Euclidean metric) of semi-ellipse curves (perimeters) is not known in closed form but can be expressed using the so-called complete elliptic integral of the second kind [93].

Appendix A.2. Geodesics with Initial Values on the Multivariate Fisher–Rao Normal Manifold

The geodesic equation is given by

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

We concisely report the parametric geodesics using another variant of the natural parameters of the normal distributions (slightly differing from the θ-coordinate system since natural parameters can be chosen up to a fixed affine transformation by changing accordingly the sufficient statistics by the inverse affine transformation) viewed as an exponential family:

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

In general, the geodesics with boundary values $\gamma_{\mathcal{N}}^{\text{Fisher}}(N_1, N_2; t)$ are not known in closed form. However, Calvo and Oller [48] (Theorem 3.1 and Corollary 1) reported the explicit equations of the geodesics when the initial values are given, i.e., $\gamma_{\mathcal{N}}^{\text{Fisher}}(N_0, v_0; t)$ where $v_0 = \dot{\gamma}_{\mathcal{N}}^{\text{Fisher}}(N_0, v_0; 0) = (\dot{\xi}(0), \dot{\Xi}(0))$ is in $T_{N_0}\mathcal{M}$ and $\gamma_{\mathcal{N}}^{\text{Fisher}}(N_0, v_0; 0) = N_0$.

Let

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

and $G^{\dagger}$ be the Moore–Penrose generalized inverse matrix of G: $G^{\dagger} = (G^{\top}G)^{-1}G^{\top}$ or $G^{\dagger} = G^{\top}(GG^{\top})^{-1}$. The Moore–Penrose pseudo-inverse matrix can be replaced by any other pseudo-inverse matrix G^{-} [48].

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

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

where the Cosh and Sinh functions of a matrix M are defined by the following absolutely convergent series [48] (Equation (9), p. 122):

$$\begin{aligned}\mathrm{Sinh}(M) &= M + \sum_{i=1}^{\infty} \frac{M^{2i+1}}{(2i+1)!},\\ \mathrm{Cosh}(M) &= I + \sum_{i=1}^{\infty} \frac{M^{2i}}{(2i)!},\end{aligned}$$

and satisfies the identity $\mathrm{Sinh}^2(M) + \mathrm{Cosh}^2(M) = I$. The matrix Cosh and Sinh functions can be calculated from the eigendecomposition of $M = O\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\,O^\top$ as follows:

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

When we restrict the manifold to a totally geodesic submanifold $\mathcal{M}_\mu = \{P \succ 0\}$, the geodesic equation becomes $\ddot{P} - \dot{P}P^{-1}\dot{P} = 0$, and the geodesic with initial values $P(0) = P$ and $\dot{P}(0) = S \in \mathrm{Sym}$ is:

$$P(t) = P^{\frac{1}{2}}\,\exp\left(tP^{-\frac{1}{2}}SP^{-\frac{1}{2}}\right)P^{\frac{1}{2}}.$$

The geodesic with boundary values $P(0) = P_1$ and $P(1) = P_2$ is

$$P(t) = P_1^{\frac{1}{2}}\,\exp\left(t\mathrm{Log}(P_1^{-\frac{1}{2}}P_2P_1^{-\frac{1}{2}})\right)P_1^{\frac{1}{2}}.$$

Furthermore, we can convert a geodesic with boundary values $\gamma_{\mathbb{P}}(P_1, P_2; t)$ to an equivalent geodesic with initial values $\gamma_{\mathbb{P}}(P, S; t)$ by letting

$$S = P_1^{\frac{1}{2}}\mathrm{Log}(P_1^{-\frac{1}{2}}P_2P_1^{-\frac{1}{2}})P_1^{\frac{1}{2}}.$$

Appendix B. Fisher–Rao Distance between Normal Distributions Sharing the Same Covariance Matrix

The Rao distance between $N_1 = N(\mu_1, \Sigma)$ and $N_2 = N(\mu_2, \Sigma)$ has been reported in closed form [42] (Proposition 3). We shall explain the geometric method in full as follows: Let $(e_1, \ldots, e_d)$ be the standard frame of $\mathbb{R}^d$ (ordered basis): The e_i's are the unit vectors of the axis x_i's. Let P be an orthogonal matrix such that $P(\mu_2 - \mu_1) = \|\mu_2 - \mu_1\|_2\, e_1$ (i.e., matrix P aligns vector $\mu_2 - \mu_1$ to the first axis x_1). Let $\Delta_{12} = \|\mu_2 - \mu_1\|_2$ be the Euclidean distance between μ_1 and μ_2. Furthermore, factorize matrix $P\Sigma P^\top$ using the LDL decomposition (a variant of the Cholesky decomposition) as $P\Sigma P^\top = LDL^\top$ where L is a lower triangular matrix with all diagonal entries equal to one (lower unitriangular matrix of unit determinant) and D a diagonal matrix. Let $\sigma_{12} = \sqrt{D_{11}}$. Then we have [42]:

$$\rho_\Sigma(\mu_1, \mu_2) = \rho_{\mathcal{N}}(N(\mu_1, \Sigma), N(\mu_2, \Sigma)) = \rho_{\mathcal{N}}(N(0, \sigma), N(\Delta_{12}e_1, \sigma_{12})). \tag{A1}$$

Please note that the right-hand side term is the Fisher–Rao distance between univariate normal distributions of Equation (7).

To find matrix P, we proceed as follows: Let $u = \frac{\mu_2 - \mu_1}{\|\mu_2 - \mu_1\|_2}$ be the normalized vector to align on axis x_1. Let $v = u - e_1$. Consider the Householder reflection matrix [94] $M = I - \frac{2vv^\top}{\|v\|_2^2}$, where $vv^\top$ is an outer product matrix. Since Householder reflection matrices have determinant -1, we let P be a copy of M with the last row multiplied by -1 so that we obtain $\det(P) = 1$. By construction, we have $Pu = \|\mu_2 - \mu_1\|_2\, e_1$. We then use the affine-invariance property of the Fisher–Rao distance as follows:

$$\begin{aligned}
\rho_{\mathcal{N}}(N(\mu_1,\Sigma),N(\mu_2,\Sigma)) &= \rho_{\mathcal{N}}(N(0,\Sigma),N(\mu_2-\mu_1,\Sigma)),\\
&= \rho_{\mathcal{N}}(N(0,P\Sigma P^\top),N(P(\mu_2-\mu_1),P\Sigma P^\top)),\\
&= \rho_{\mathcal{N}}(N(0,P\Sigma P^\top),N(\Delta_{12}\,e_1,P\Sigma P^\top)),\\
&= \rho_{\mathcal{N}}(N(0,LDL^\top),N(\Delta_{12}\,e_1,LDL^\top)),\\
&= \rho_{\mathcal{N}}(N(0,D),N(\Delta_{12}\,e_1,D)).
\end{aligned}$$

The last row follows from the fact that $L^{-1}e_1 = e_1$ since L^{-1} is an upper unitriangular matrix, and $L^\top(L^{-1})^\top = (L^{-1}L)^\top = I$. The right-hand side Fisher–Rao distance is computed from Equation (7).

Appendix C. Embedding the Set of Multivariate Normal Distributions in a Riemannian Symmetric Space

The multivariate Gaussian manifold $\mathcal{N}(d)$ can also be embedded into the SPD cone $\mathcal{P}(d+1)$ as a Riemannian symmetric space [82,89] by f_{SSPD}: $\hat{\mathcal{P}} = \{f_{\mathrm{SSPD}}(N) \subset \mathcal{P}(d+1) : N \in \mathcal{N}(d)\}$. We have $\hat{\mathcal{P}} \cong \mathrm{SL}(d+1)/\mathrm{SO}(d+1)$ [82,95,96] (and textbook [97], Part II Chapter 10), and the symmetric space $\mathrm{SL}(d+1)/\mathrm{SO}(d+1)$ can be embedded with the Killing Riemannian metric instead of the Fisher information metric:

$$g^{\mathrm{Killing}}(N_1,N_2) = \kappa_{\mathrm{Killing}}\left(\mu_1^\top\Sigma^{-1}\mu_2 + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma_2\right) - \frac{1}{2(d+1)}\mathrm{tr}\left(\Sigma^{-1}\Sigma_1\right)\mathrm{tr}\left(\Sigma^{-1}\Sigma_2\right)\right),$$

where $\kappa_{\mathrm{Killing}} > 0$ is a predetermined constant (e.g., 1). The length element of the Killing metric is

$$\mathrm{d}s^2_{\mathrm{SS}} = \kappa_{\mathrm{Killing}}\left(\frac{1}{2}\mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu + \mathrm{tr}\left(\left(\Sigma^{-1}\mathrm{d}\Sigma\right)^2\right) - \frac{1}{2}\mathrm{tr}^2\left(\Sigma^{-1}\mathrm{d}\Sigma\right)\right).$$

When we consider $\mathcal{N}_\Sigma$, we may choose $\kappa_{\mathrm{Killing}} = 2$ so that the Killing metric coincides with the Fisher information metric. The induced Killing distance [82] is available in closed form:

$$\rho_{\mathrm{Killing}}(N_1,N_2) = \sqrt{\kappa_{\mathrm{Killing}}\sum_{i=1}^{d+1}\log^2\lambda_i\left(\hat{L}_1^{-1}\hat{P}_2\left(\hat{L}_1^{-1}\right)^\top\right)}, \tag{A2}$$

where $\hat{L}_1$ is the unique lower triangular matrix obtained from the Cholesky decomposition of $\hat{P}_1 = f_{\mathrm{SSPD}}(N_1) = \hat{L}_1\hat{L}_1^\top$. Please note that $\hat{L}_1^{-1}\hat{P}_2\left(\hat{L}_1^{-1}\right)^\top \in \mathcal{P}(d+1)$ and $|\hat{L}_1|$, i.e., $\hat{L}_1 \in \mathrm{SL}(d+1)$.

When $N_1 = (\mu_1,\Sigma)$ and $N_2 = (\mu_2,\Sigma)$ ($N_1, N_2 \in \mathcal{N}_\Sigma$), we have [82]

$$\rho_{\mathrm{Killing}}(N_1,N_2) = \sqrt{2\kappa_{\mathrm{Killing}}}\,\mathrm{arccosh}\left(1+\frac{1}{2}\Delta^2_\Sigma(\mu_1,\mu_2)\right),$$

where Δ^2_Σ is the squared Mahalanobis distance. Thus, $\rho_{\mathrm{Killing}}(N_1,N_2) = h_{\mathrm{Killing}}(\Delta_\Sigma(\mu_1,\mu_2))$ where $h_{\mathrm{Killing}}(u) = \sqrt{2\kappa_{\mathrm{Killing}}}\,\mathrm{arccosh}\left(1+\frac{1}{2}u^2\right)$.

When $N_1 = (\mu,\Sigma_1)$ and $N_2 = (\mu,\Sigma_2)$ ($N_1, N_2 \in \mathcal{N}_\mu$), we have [82]:

$$\rho_{\mathrm{Killing}}(N_1,N_2) = \sqrt{\kappa_{\mathrm{Killing}}\left(\sum_{i=1}^{d}\log^2\lambda_i\left(L_1^{-1}P_2\left(L_1^{-1}\right)^\top\right) - \frac{1}{(d+1)^2}\left(\sum_{i=1}^{d}\log\lambda_i\left(L_1^{-1}P_2\left(L_1^{-1}\right)^\top\right)\right)\right)}.$$

See Example 1. Let us emphasize that the Killing distance is not the Fisher–Rao distance but is available in closed form as an alternative metric distance between MVNs.

A Fisher geodesic defect measure of a curve c is defined in [89] by

$$\delta(c) = \lim_{s\to\infty} \frac{1}{s} \int_0^s \|\nabla^{g^{\text{Fisher}}}_{\dot{c}} \dot{c}\|^{\text{Fisher}}_{c(t)} \mathrm{d}t,$$

where $\nabla^{g^{\text{Fisher}}}$ denotes the Levi–Civita connection induced by the Fisher metric. When $\delta(c) = 0$ the curve is said to be an asymptotic geodesic of the Fisher geodesic. It is proven that Killing geodesics at (μ, Σ) are asymptotic Fisher geodesics when the initial condition $c'(0)$ is orthogonal to $\mathcal{N}_\mu$.

Appendix D. Embedding the Set of Multivariate Normal Distributions in the Siegel Upper Space

The Siegel upper space is the space of symmetric complex matrices $Z = X + iY = Z^\top$ with imaginary positive–definite matrices $Y \succ 0$ [45,65] (so-called Riemann matrices [98]):

$$\mathbb{SH}(d) := \{Z = X + iY \ : \ X \in \mathrm{Sym}(d), Y \in \mathcal{P}(d)\}, \tag{A3}$$

where $\mathrm{Sym}(d)$ is the space of symmetric real $d \times d$ matrices. $\mathbb{SH}(1)$ corresponds to the Poincaré upper plane. See Figure A2 for an illustration.

The Siegel infinitesimal square line element is

$$\mathrm{d}s^2_{\mathbb{SH}}(Z) = 2\mathrm{tr}\Big(Y^{-1}\mathrm{d}Z\, Y^{-1}\mathrm{d}\bar{Z}\Big). \tag{A4}$$

When $X = 0$ and $Z = iY$, we have $\mathrm{d}Z = i\mathrm{d}Y$, $\mathrm{d}\bar{Z} = -i\mathrm{d}Y$, and it follows that

$$\mathrm{d}s^2_{\mathbb{SH}}(iY) = 2\mathrm{tr}\Big((Y^{-1}\mathrm{d}Y)^2\Big).$$

That is, four times the square length of the Fisher matrix of centered normal distributions $\mathrm{d}s^2_{\mathcal{N}_0} = \frac{1}{2}\mathrm{tr}\big((P^{-1}\mathrm{d}P)^2\big)$.

The Siegel distance [45] between Z_1 and $Z_2 \in \mathbb{SH}(d)$ is

$$\rho_{\mathbb{SH}}(Z_1, Z_2) = \sqrt{\sum_{i=1}^{d} \log^2\left(\frac{1+\sqrt{r_i}}{1-\sqrt{r_i}}\right)}, \tag{A5}$$

where

$$r_i = \lambda_i(R(Z_1, Z_2)), \tag{A6}$$

with $R(Z_1, Z_2)$ denoting the matrix generalization of the cross-ratio

$$R(Z_1, Z_2) := (Z_1 - Z_2)(Z_1 - \bar{Z}_2)^{-1}(\bar{Z}_1 - \bar{Z}_2)(\bar{Z}_1 - Z_2)^{-1}, \tag{A7}$$

and $\lambda_i(M)$ denoting the i-th largest (real) eigenvalue of (complex) matrix M. (In practice, we numerically must round off the tiny imaginary parts to obtain proper real eigenvalues [65].) The Siegel upper half space is a homogeneous space where the Lie Group $\mathrm{SU}(d,d)/S(U(d) \times U(d))$ acts transitively on it.

We can embed a multivariate normal distribution $N = (\mu, \Sigma)$ into $\mathbb{SH}(d)$ as follows:

$$N(\mu, \Sigma) \to Z(N) := \Big(\mu\mu^\top + i\Sigma\Big),$$

and consider the Siegel distance on the embedded normal distributions as another potential metric distance between multivariate normal distributions:

$$\rho_{\mathbb{SH}}(N_1, N_2) = \rho_{\mathbb{SH}}(Z(N_1), Z(N_2)). \tag{A8}$$

Notice that the real matrix part of the $Z(N)$'s are all of rank one by construction.

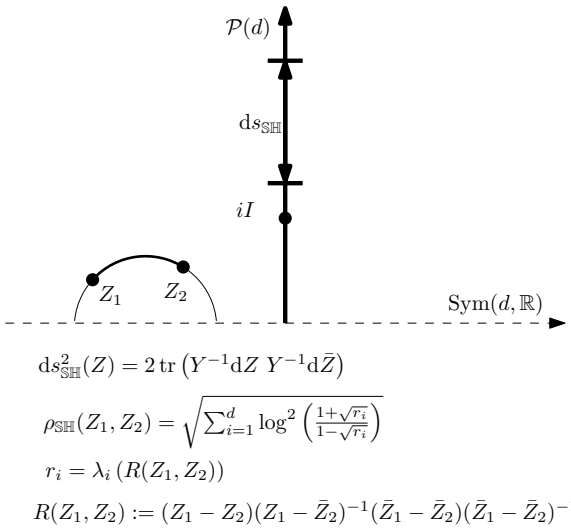


Figure A2. Siegel upper space generalizes the Poincaré hyperbolic upper plane.

Appendix E. The Symmetrized Bregman Divergence Expressed as Integral Energies on Dual Geodesics

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 A1. *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

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

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: $\mathrm{d}s^2(\gamma(t)) = \mathrm{d}\theta(t)^\top\nabla^2 F(\theta(t))\mathrm{d}\theta(t)$ 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 obtain $\int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = (\theta_2-\theta_1)^\top(\nabla F(\theta_2)-\nabla F(\theta_1)) = S_F(\theta_1;\theta_2)$.

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

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

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

References

1. Amari, S.I. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016.
2. Calin, O.; Udrişte, C. *Geometric Modeling in Probability and Statistics*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 121.
3. Lin, Z. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM J. Matrix Anal. Appl.* **2019**, *40*, 1353–1370. [CrossRef]
4. Soen, A.; Sun, K. On the variance of the Fisher information for deep learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 5708–5719.
5. Barachant, A.; Bonnet, S.; Congedo, M.; Jutten, C. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing* **2013**, *112*, 172–178. [CrossRef]
6. Skovgaard, L.T. *A Riemannian Geometry of the Multivariate Normal Model*; Technical Report 81/3; Statistical Research Unit, Danish Medical Research Council, Danish Social Science Research Council: Copenhagen, Denmark, 1981.
7. Skovgaard, L.T. A Riemannian geometry of the multivariate normal model. *Scand. J. Stat.* **1984**, *11* , 211–223.
8. Malagò, L.; Pistone, G. Information geometry of the Gaussian distribution in view of stochastic optimization. In Proceedings of the ACM Conference on Foundations of Genetic Algorithms XIII, Aberystwyth, UK, 17–22 January 2015; pp. 150–162.
9. Herntier, T.; Peter, A.M. Transversality Conditions for Geodesics on the Statistical Manifold of Multivariate Gaussian Distributions. *Entropy* **2022**, *24*, 1698. [CrossRef] [PubMed]
10. Atkinson, C.; Mitchell, A.F. Rao's distance measure. *SankhyĀ Indian J. Stat. Ser.* **1981**, *43*, 345–365.
11. Radhakrishna Rao, C. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
12. Chen, X.; Zhou, J.; Hu, S. Upper bounds for Rao distance on the manifold of multivariate elliptical distributions. *Automatica* **2021**, *129*, 109604. [CrossRef]
13. Hotelling, H. Spaces of statistical parameters. *Bull. Am. Math. Soc.* **1930**, *36*, 191.
14. Cencov, N.N. *Statistical Decision Rules and Optimal Inference*; American Mathematical Soc.: Providence, RI, USA, 2000; Volume 53 .
15. Bauer, M.; Bruveris, M.; Michor, P.W. Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bull. Lond. Math. Soc.* **2016**, *48*, 499–506. [CrossRef]
16. Fujiwara, A. Hommage to Chentsov's theorem. *Inf. Geom.* **2022**, 1–20. [CrossRef]
17. Bruveris, M.; Michor, P.W. Geometry of the Fisher–Rao metric on the space of smooth densities on a compact manifold. *Math. Nachrichten* **2019**, *292*, 511–523. [CrossRef]
18. Burbea, J.; Oller i Sala, J.M. *On Rao Distance Asymptotic Distribution*; Technical Report Mathematics Preprint Series No. 67; Universitat de Barcelona: Barcelona, Spain, 1989.
19. Calvo, M.; Oller, J.M. A distance between multivariate normal distributions based in an embedding into the Siegel group. *J. Multivar. Anal.* **1990**, *35*, 223–242. [CrossRef]
20. Rios, M.; Villarroya, A.; Oller, J.M. Rao distance between multivariate linear normal models and their application to the classification of response curves. *Comput. Stat. Data Anal.* **1992**, *13*, 431–445. [CrossRef]
21. Park, P.S.; Kshirsagar, A.M. Distances between normal populations when covariance matrices are unequal. *Commun. Stat. Theory Methods* **1994**, *23*, 3549–3556. [CrossRef]
22. Gruber, M.H. Some applications of the Rao distance to shrinkage estimators. *Commun. Stat. Methods* **2008**, *37*, 180–193. [CrossRef]
23. Strapasson, J.E.; Pinele, J.; Costa, S.I. Clustering using the Fisher-Rao distance. In Proceedings of the 2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), Rio de Janeiro, Brazil, 10–13 July 2016; pp. 1–5.
24. Le Brigant, A.; Puechmorel, S. Quantization and clustering on Riemannian manifolds with an application to air traffic analysis. *J. Multivar. Anal.* **2019**, *173*, 685–703. [CrossRef]
25. Said, S.; Bombrun, L.; Berthoumieu, Y. Texture classification using Rao's distance on the space of covariance matrices. In Proceedings of the Geometric Science of Information: Second International Conference, GSI 2015, Proceedings 2, Palaiseau, France, 28–30 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 371–378.
26. Legrand, L.; Grivel, E. Evaluating dissimilarities between two moving-average models: A comparative study between Jeffrey's divergence and Rao distance. In Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 8 August–2 September 2016; pp. 205–209.
27. Halder, A.; Georgiou, T.T. Gradient flows in filtering and Fisher-Rao geometry. In Proceedings of the 2018 Annual American Control Conference (ACC), Milwaukee, WI, USA, 27–29 June 2018; pp. 4281–4286.

28. Collas, A.; Breloy, A.; Ren, C.; Ginolhac, G.; Ovarlez, J.P. Riemannian optimization for non-centered mixture of scaled Gaussian distributions. *arXiv* **2022**, arXiv:2209.03315.
29. Liang, T.; Poggio, T.; Rakhlin, A.; Stokes, J. Fisher-Rao metric, geometry, and complexity of neural networks. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, PMLR, Naha, Japan, 16–18 April 2019; pp. 888–896.
30. Yoshizawa, S.; Tanabe, K. Dual differential geometry associated with the Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT J. Math.* **1999**, *35*, 113–137. [CrossRef]
31. Shima, H. *The Geometry of Hessian Structures*; World Scientific: Singapore, 2007.
32. Calvo, M.; Oller, J.M. A distance between elliptical distributions based in an embedding into the Siegel group. *J. Comput. Appl. Math.* **2002**, *145*, 319–334. [CrossRef]
33. Burbea, J. *Informative Geometry of Probability Spaces*; Technical Report; Pittsburgh Univ. PA Center for Multivariate Analysis: Pittsburgh, PA, USA, 1984.
34. Eriksen, P.S. *Geodesics Connected with the Fischer Metric on the Multivariate Normal Manifold*; Institute of Electronic Systems, Aalborg University Centre: Aalborg, Denmark, 1986.
35. Berkane, M.; Oden, K.; Bentler, P.M. Geodesic estimation in elliptical distributions. *J. Multivar. Anal.* **1997**, *63*, 35–46. [CrossRef]
36. Imai, T.; Takaesu, A.; Wakayama, M. *Remarks on Geodesics for Multivariate Normal Models*; Technical Report; Faculty of Mathematics, Kyushu University: Fukuoka, Japan, 2011.
37. Inoue, H. Group theoretical study on geodesics for the elliptical models. In Proceedings of the Geometric Science of Information: Second International Conference, GSI 2015, Proceedings 2, Palaiseau, France, 28–30 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 605–614.
38. Strapasson, J.E.; Porto, J.P.; Costa, S.I. On bounds for the Fisher-Rao distance between multivariate normal distributions. *AIP Conf. Proc.* **2015**, *1641*, 313–320.
39. Han, M.; Park, F.C. DTI segmentation and fiber tracking using metrics on multivariate normal distributions. *J. Math. Imaging Vis.* **2014**, *49*, 317–334. [CrossRef]
40. Pilté, M.; Barbaresco, F. Tracking quality monitoring based on information geometry and geodesic shooting. In Proceedings of the 2016 17th International Radar Symposium (IRS), Krakow, Poland, 10–12 May 2016; pp. 1–6.
41. Barbaresco, F. Souriau exponential map algorithm for machine learning on matrix Lie groups. In Proceedings of the Geometric Science of Information: 4th International Conference, GSI 2019, Proceedings 4, Toulouse, France, 27–29 August 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 85–95.
42. Pinele, J.; Strapasson, J.E.; Costa, S.I. The Fisher–Rao distance between multivariate normal distributions: Special cases, bounds and applications. *Entropy* **2020**, *22*, 404. [CrossRef]
43. Dijkstra, E.W. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 287–290.
44. Anderson, J.W. *Hyperbolic Geometry*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
45. Siegel, C.L. *Symplectic Geometry*; First Printed in 1964; Elsevier: Amsterdam, The Netherlands, 2014.
46. James, A.T. The variance information manifold and the functions on it. In *Multivariate Analysis–III*; Elsevier: Amsterdam, The Netherlands, 1973; pp. 157–169.
47. Wells, J.; Cook, M.; Pine, K.; Robinson, B.D. Fisher-Rao distance on the covariance cone. *arXiv* **2020**, arXiv:2010.15861.
48. Calvo, M.; Oller, J.M. An explicit solution of information geodesic equations for the multivariate normal model. *Stat. Risk Model.* **1991**, *9*, 119–138. [CrossRef]
49. Förstner, W.; Moonen, B. A metric for covariance matrices. In *Geodesy-the Challenge of the 3rd Millennium*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 299–309.
50. Dolcetti, A.; Pertici, D. Real square roots of matrices: Differential properties in semi-simple, symmetric and orthogonal cases. *arXiv*, **2020**, arXiv:2010.15609.
51. Mahalanobis, P.C. On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India*; Springer: New Delhi, India, 1936; Volume 12, pp. 49–55.
52. Eaton, M.L. *Group Invariance Applications in Statistics*; Institute of Mathematical Statistics: Beachwood, OH, USA, 1989.
53. Godinho, L.; Natário, J. An introduction to Riemannian geometry: With Applications to Mechanics and Relativity. In *Universitext*; Springer International Publishing: Cham, Switzerland, 2014.
54. Strapasson, J.E.; Pinele, J.; Costa, S.I. A totally geodesic submanifold of the multivariate normal distributions and bounds for the Fisher-Rao distance. In Proceedings of the IEEE Information Theory Workshop (ITW), Cambridge, UK, 1–11 September 2016; pp. 61–65.
55. Chen, X.; Zhou, J. Multisensor Estimation Fusion on Statistical Manifold. *Entropy* **2022**, *24*, 1802. [CrossRef]
56. Cherian, A.; Sra, S. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2859–2871. [CrossRef]
57. Nguyen, X.S. Geomnet: A neural network based on Riemannian geometries of SPD matrix space and Cholesky space for 3d skeleton-based interaction recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13379–13389.
58. Dolcetti, A.; Pertici, D. Differential properties of spaces of symmetric real matrices. *arXiv* **2018**, arXiv:1807.01113.

59. Verdoolaege, G.; Scheunders, P. On the geometry of multivariate generalized Gaussian models. *J. Math. Imaging Vis.* **2012**, *43*, 180–193. [CrossRef]
60. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B* **1966**, *28*, 131–142. [CrossRef]
61. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
62. Nielsen, F.; Okamura, K. A note on the f-divergences between multivariate location-scale families with either prescribed scale matrices or location parameters. *arXiv* **2022**, arXiv:2204.10952.
63. Moakher, M.; Zéraï, M. The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *J. Math. Imaging Vis.* **2011**, *40*, 171–187. [CrossRef]
64. Dolcetti, A.; Pertici, D. Elliptic isometries of the manifold of positive definite real matrices with the trace metric. *Rend. Circ. Mat. Palermo Ser. 2* **2021**, *70*, 575–592. [CrossRef]
65. Nielsen, F. The Siegel–Klein Disk: Hilbert Geometry of the Siegel Disk Domain. *Entropy* **2020**, *22*, 1019. [CrossRef]
66. Arnaudon, M.; Nielsen, F. On approximating the Riemannian 1-center. *Comput. Geom.* **2013**, *46*, 93–104. [CrossRef]
67. Ceolin, S.R.; Hancock, E.R. Computing gender difference using Fisher-Rao metric from facial surface normals. In Proceedings of the 25th SIBGRAPI Conference on Graphics, Patterns and Images, Ouro Preto, Brazil, 22–25 August 2012; pp. 336–343.
68. Wang, Q.; Li, P.; Zhang, L. G2DeNet: Global Gaussian distribution embedding network and its application to visual recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2730–2739.
69. Miyamoto, H.K.; Meneghetti, F.C.; Costa, S.I. The Fisher–Rao loss for learning under label noise. *Inf. Geom.* **2022**, 1–20. [CrossRef]
70. Kurtek, S.; Bharath, K. Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika* **2015**, *102*, 601–616. [CrossRef]
71. Marti, G.; Andler, S.; Nielsen, F.; Donnat, P. Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series. In Proceedings of the 2016 IEEE Statistical Signal Processing Workshop (SSP), Palma de Mallorca, Spain, 26–29 June 2016; pp. 1–5.
72. Tang, M.; Rong, Y.; Zhou, J.; Li, X.R. Information geometric approach to multisensor estimation fusion. *IEEE Trans. Signal Process.* **2018**, *67*, 279–292. [CrossRef]
73. Wang, W.; Wang, R.; Huang, Z.; Shan, S.; Chen, X. Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2048–2057.
74. Li, P.; Wang, Q.; Zeng, H.; Zhang, L. Local log-Euclidean multivariate Gaussian descriptor and its application to image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 803–817. [CrossRef]
75. Picot, M.; Messina, F.; Boudiaf, M.; Labeau, F.; Ayed, I.B.; Piantanida, P. Adversarial robustness via Fisher-Rao regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2698–2710 [CrossRef] [PubMed]
76. Collas, A.; Bouchard, F.; Ginolhac, G.; Breloy, A.; Ren, C.; Ovarlez, J.P. On the Use of Geodesic Triangles between Gaussian Distributions for Classification Problems. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 5697–5701.
77. Murena, P.A.; Cornuéjols, A.; Dessalles, J.L. Opening the parallelogram: Considerations on non-Euclidean analogies. In Proceedings of the Case-Based Reasoning Research and Development: 26th International Conference, ICCBR 2018, Proceedings 26, Stockholm, Sweden, 9–12 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 597–611.
78. Popović, B.; Janev, M.; Krstanović, L.; Simić, N.; Delić, V. Measure of Similarity between GMMs Based on Geometry-Aware Dimensionality Reduction. *Mathematics* **2022**, *11*, 175. [CrossRef]
79. Micchelli, C.A.; Noakes, L. Rao distances. *J. Multivar. Anal.* **2005**, *92*, 97–115. [CrossRef]
80. Nielsen, F. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef]
81. Davis, J.; Dhillon, I. Differential entropic clustering of multivariate Gaussians. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 337–344.
82. Lovrić, M.; Min-Oo, M.; Ruh, E.A. Multivariate normal distributions parametrized as a Riemannian symmetric space. *J. Multivar. Anal.* **2000**, *74*, 36–48. [CrossRef]
83. Welzl, E. Smallest enclosing disks (balls and ellipsoids). In *Proceedings of the New Results and New Trends in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 359–370.
84. Gonzalez, T.F. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.* **1985**, *38*, 293–306. [CrossRef]
85. Acharyya, S.; Banerjee, A.; Boley, D. Bregman divergences and triangle inequality. In Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM, Austin, TX, USA, 2–4 May 2013; pp. 476–484.
86. Ohara, A.; Suda, N.; Amari, S.i. Dualistic differential geometry of positive definite matrices and its applications to related problems. *Linear Algebra Appl.* **1996**, *247*, 31–53. [CrossRef]
87. Nock, R.; Nielsen, F. Fitting the smallest enclosing Bregman ball. In Proceedings of the Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Proceedings 16, Porto, Portugal, 3–7 October 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 649–656.
88. Ohara, A. Doubly autoparallel structure on positive definite matrices and its applications. In Proceedings of the International Conference on Geometric Science of Information, Toulouse, France, 27–29 August 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 251–260.

89. Globke, W.; Quiroga-Barranco, R. Information geometry and asymptotic geodesics on the space of normal distributions. *Inf. Geom.* **2021**, *4*, 131–153. [CrossRef]
90. Nielsen, F.; Sun, K. Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices. In *Geometric Structures of Information*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 297–331.
91. Journée, M.; Nesterov, Y.; Richtárik, P.; Sepulchre, R. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.* **2010**, *11*, 517–553.
92. Verdoolaege, G. A new robust regression method based on minimization of geodesic distances on a probabilistic manifold: Application to power laws. *Entropy* **2015**, *17*, 4602–4626. [CrossRef]
93. Chandrupatla, T.R.; Osler, T.J. The perimeter of an ellipse. *Math. Sci.* **2010**, *35*.
94. Householder, A.S. Unitary triangularization of a nonsymmetric matrix. *J. ACM* **1958**, *5*, 339–342. [CrossRef]
95. Fernandes, M.A.; San Martin, L.A. Fisher information and α-connections for a class of transformational models. *Differ. Geom. Appl.* **2000**, *12*, 165–184. [CrossRef]
96. Fernandes, M.A.; San Martin, L.A. Geometric proprieties of invariant connections on $\mathrm{SL}(n,R)/\mathrm{SO}(n)$. *J. Geom. Phys.* **2003**, *47*, 369–377. [CrossRef]
97. Bridson, M.R.; Haefliger, A. *Metric Spaces of Non-Positive Curvature*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 319.
98. Frauendiener, J.; Jaber, C.; Klein, C. Efficient computation of multidimensional theta functions. *J. Geom. Phys.* **2019**, *141*, 147–158. [CrossRef]