entropy

MDPI

Article

The Fisher–Rao Distance between Multivariate Normal Distributions: Special Cases, Bounds and Applications

Julianna Pinele [1,*], João E. Strapasson [2] and Sueli I. R. Costa [3]

[1] Center of Exact and Technological Sciences, University of Reconcavo of Bahia, Cruz das Almas 44380-000, Brazil
[2] School of Applied Sciences, University of Campinas, Limeira 13484-350, Brazil; joao.strapasson@fca.unicamp.br
[3] Institute of Mathematics, University of Campinas, Campinas 13083-859, Brazil; sueli@ime.unicamp.br
* Correspondence: julianna.pinele@gmail.com

Received: 26 January 2020; Accepted: 11 March 2020; Published: 1 April 2020

Abstract: The Fisher–Rao distance is a measure of dissimilarity between probability distributions, which, under certain regularity conditions of the statistical model, is up to a scaling factor the unique Riemannian metric invariant under Markov morphisms. It is related to the Shannon entropy and has been used to enlarge the perspective of analysis in a wide variety of domains such as image processing, radar systems, and morphological classification. Here, we approach this metric considered in the statistical model of normal multivariate probability distributions, for which there is not an explicit expression in general, by gathering known results (closed forms for submanifolds and bounds) and derive expressions for the distance between distributions with the same covariance matrix and between distributions with mirrored covariance matrices. An application of the Fisher–Rao distance to the simplification of Gaussian mixtures using the hierarchical clustering algorithm is also presented.

Keywords: information geometry; Fisher–Rao distance; multivariate normal distributions; Gaussian mixture simplification

1. Introduction

A proper measure to determine the dissimilarity between probability distributions has been approached in many problems and applications. The Fisher–Rao distance is a very special metric for statistical models of probability distributions. This distance is invariant by reparametrization of the sample space and covariant by reparameterization of the parameter space [1]. Moreover, the Fisher–Rao metric is preserved under Markov morphisms and under centain conditions it is, up to a scaling factor, the unique Riemannian metric satisfying this condition [2,3]. Markov morphisms are associated with the notion of statistical sufficiency which express the criterion of passing from one statistical model to another with no loss of information [4–6]. Therefore it is natural to require the invariance of the geometric structures of statistical models under Markov morphisms. Between finite sample size simplex model $S_{k-1} = \{p \in \mathbb{R}^k; p_i \geq 0 \text{ and } \sum_{i=1}^{k} p_i = 1\}$, a Markov morphism is a linear map $T_Q(x) = xQ$, where $Q \in \mathbb{R}^{n \times l}$, with $n \leq l$, is a matrix with non-negative entries such that every row sums to 1 and every column has precisely one non-zero element. The mapping T_Q corresponds to probabilistic refining of the event space $\{1, \ldots, n\} \to \{1, \ldots, l\}$ where the refinement $i \to j$ occurs with probability Q_{ij} [7]. Chentsov [8,9] has proved the Fisher–Rao uniqueness invariance property under Markov morphisms for the finite sample spaces. The extension of this result to more general statistical models requires careful formulations of statistical sufficiency and Markov morphisms and

has been evolved since then [3,10]. More recently in [5,6] it is shown this uniqueness of the Fisher–Rao metric under an assumption of strong continuity of the information metric.

After previous papers [11–13] connecting geometry and statistics, C. R.Rao in an independent landmark paper [14] considered statistical models with the metric induced by the information matrix defined by R. Fisher in 1921 [15]. This work encouraged several authors to calculate the Fisher–Rao metric distance between other probability distributions [16–18] as well as stimulated approaches to other dissimilarity measures such as Kullback-Leibler divergence [19], total variation and Wasserstein distances [20]. Amari [3,4,21] unified the information geometry theory by organizing and introducing other concepts regarding statistical models [2].

An explicit form for the Fisher–Rao distance in the univariate normal distribution space is known via an association with the classical model of the hyperbolic plane [14,16,18,22]. It was applied to quantization of hyperspectral images [23] and to the space of projected lines in the paracatadioptric images [24]. This Fisher–Rao model was used to simplify Gaussian mixtures through the k-means method [25] and a hierarchical clustering technique [26].

An expression for the geodesic curve (initial value problem) in the multivariate normal distributions space was derived in [27] and in [28]. However, the calculus of the Fisher–Rao distance requires solving non-trivial differential equations under boundary conditions to find the geodesic connecting two distributions and then to calculate the integral along the geodesic. A closed form for this distance in the general case is still an open problem. Expressions for the distance are known only in special cases [16–18].

The Fisher–Rao distance between multivariate normal distributions in specific cases, such as distributions with a common mean, was considered in diffusion tensor image analysis [29–31], in color texture discrimination in several classification experiments [32], in the problem of distributed estimation fusion with unknown correlations [33], and in the machine learning technique [34]. In [35,36], the authors described shapes representing landmarks by a Gaussian model with diagonal covariance matrices and used the Fisher–Rao distance to quantify the difference between two shapes. In [17], this model was applied to statistical inference. Bounds for the Fisher–Rao distance were used to track quality monitoring [37].

This paper is organized as follows. In Section 2, we gather known results (closed forms for special cases and bounds) for the Fisher–Rao distance between multivariate normal distributions. In Section 3, we describe a closed form for the Fisher–Rao distance between distributions with the same covariance matrix and a non-linear system to find the distance between distributions with mirrored covariance matrices. An application of the Fisher–Rao distance to the simplification of Gaussian mixtures using the hierarchical clustering algorithm is presented in Section 4. Some conclusions and perspectives are drawn in Section 5.

2. The Fisher–Rao Distance in the Multivariate Normal Distribution Space: Special Submanifolds and Bounds

In this section, as in [38], we summarize previous results regarding the Fisher–Rao distance in the space of multivariate normal distributions including closed forms for this distance restricted to submanifolds and general bounds.

Given a statistical model $S = \{p_{\boldsymbol{\theta}} = p(\boldsymbol{x};\boldsymbol{\theta}); \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k\}$, a natural Riemannian structure [21] can be provided by the Fisher information matrix $G(\boldsymbol{\theta}) = \left[g_{ij}(\boldsymbol{\theta})\right]$:

$$g_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_i}\log p(\boldsymbol{x};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log p(\boldsymbol{x};\boldsymbol{\theta})\right) \tag{1}$$

$$= \int \frac{\partial}{\partial\theta_i}\log p(\boldsymbol{x};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log p(\boldsymbol{x};\boldsymbol{\theta})p(\boldsymbol{x};\boldsymbol{\theta})dx, \tag{2}$$

where $E_{\boldsymbol{\theta}}$ is the expected value with respect to the distribution $p_{\boldsymbol{\theta}}$. This matrix can also be viewed as the Hessian matrix of the Shannon entropy (concave function) [39],

$$H(p) = -\int p(\boldsymbol{x};\boldsymbol{\theta}) \log p(\boldsymbol{x};\boldsymbol{\theta}) dx, \tag{3}$$

and is used to establish connections between inequalities in information theory and geometrical inequalities.

The Fisher–Rao distance, $d_F(\cdot,\cdot)$, between two distributions $p_{\boldsymbol{\theta}_1}$ and $p_{\boldsymbol{\theta}_2}$ in $\mathcal{S}$, identified with their parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, is given by the shortest length of a curve $\gamma(t)$ in the parameter space Θ connecting these distributions, $d_F(p_{\boldsymbol{\theta}_1}, p_{\boldsymbol{\theta}_2}) \equiv d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \min_{\gamma} \int |\gamma'(t)|_G dt$, where $|\gamma'(t)|_G = \sqrt{\gamma'(t)^t G(\boldsymbol{\theta}) \gamma(t)}$. Note that this is in fact a metric, since for any $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_3$ in Θ, we have: (i) $d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq 0$ and $d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 0$ if only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$; (ii) $d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d_F(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)$; (iii) $d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \leq d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_3) + d_F(\boldsymbol{\theta}_3, \boldsymbol{\theta}_2)$. A curve that provides the shortest length is called a geodesic and is given by the solutions of the differential equations

$$\frac{d^2\theta_m}{dt^2} + \sum_{i,j} \Gamma_{ij}^m \frac{d\theta_i}{dt} \frac{d\theta_j}{dt} = 0, \quad m = 1, \cdots, k, \tag{4}$$

where Γ_{ij}^m are the Christoffel symbols,

$$\Gamma_{ij}^m = \frac{1}{2} \sum_l \left(\frac{\partial g_{jl}}{\partial \theta_i} + \frac{\partial g_{li}}{\partial \theta_j} - \frac{\partial g_{ij}}{\partial \theta_l} \right) g^{lm} \tag{5}$$

and $[g^{ij}]$ is the inverse matrix of the Fisher information matrix.

We consider here the space of the multivariate normal distributions given by:

$$p(\boldsymbol{x};\boldsymbol{\mu},\Sigma) = \frac{(2\pi)^{-\left(\frac{n}{2}\right)}}{\sqrt{Det(\Sigma)}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2} \right), \tag{6}$$

where $\boldsymbol{x}^t = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is the variable vector, $\boldsymbol{\mu}^t = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ is the mean vector, and Σ is the covariance matrix in $P_n(\mathbb{R})$, the space of order n positive definite symmetric matrices.

In this case, the model $\mathcal{S} = \mathcal{M} = \{p_{\boldsymbol{\theta}}; \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^n \times P_n(\mathbb{R})\}$ is a statistical $\left(n + \frac{n(n+1)}{2}\right)$-dimensional manifold.

In this case, the model $\mathcal{S} = \mathcal{M} = \{p_{\boldsymbol{\theta}}; \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^n \times P_n(\mathbb{R})\}$ is a statistical manifold of dimension $k = \left(n + \frac{n(n+1)}{2}\right)$. Considering a parametrization $(\boldsymbol{\mu}, \Sigma) = \phi(\theta_1, \ldots, \theta_k)$ of the model $\mathcal{M}$, the Fisher information matrix is given by [40]

$$g_{ij}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\mu}^t}{\partial \theta_i} \Sigma^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_j} + \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \right). \tag{7}$$

The metric provided by this matrix is invariant with respect to affine transformations. In other words, for any $(\boldsymbol{c}, Q) \in \mathbb{R}^n \times GL_n(\mathbb{R})$, where $GL_n(\mathbb{R})$ is the group of non-singular n-square matrices, the mapping:

$$\begin{array}{rccc} \psi_{(c,Q)}: & \mathcal{M} & \to & \mathcal{M} \\ & (\boldsymbol{\mu}, \Sigma) & \mapsto & (Q\boldsymbol{\mu} + \boldsymbol{c}, Q\Sigma Q^t), \end{array} \tag{8}$$

is an isometry in $\mathcal{M}$ [16]. Consequently, the Fisher–Rao distance between $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_2, \Sigma_2)$ in $\mathcal{M}$ satisfies:

$$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d_F((Q\boldsymbol{\mu}_1 + \boldsymbol{c}, Q\Sigma_1 Q^t), (Q\boldsymbol{\mu}_2 + \boldsymbol{c}, Q\Sigma_2 Q^t)) \tag{9}$$

for any $(\boldsymbol{c}, Q) \in \mathbb{R}^n \times GL_n(\mathbb{R})$. In particular, for $Q = \Sigma_1^{-(1/2)}$ and $\boldsymbol{c} = -\Sigma_1^{(-1/2)}\boldsymbol{\mu}_1$, $\boldsymbol{\theta}_3 = (\boldsymbol{\mu}_3, \Sigma_3) = (\Sigma_1^{-(1/2)}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \Sigma_1^{-(1/2)}\Sigma_2\Sigma_1^{-(1/2)})$, the Fisher–Rao distance admits the form:

$$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d_F(\boldsymbol{\theta}_0, \boldsymbol{\theta}_3), \tag{10}$$

where $\boldsymbol{\theta}_0 = (\mathbf{0}, I_n)$, I_n is the n-order identity matrix, and $\mathbf{0} \in \mathbb{R}^n$ is the null vector.

The geodesic equations in $\mathcal{M}$ can be expressed as [17]:

$$\begin{cases} \frac{d^2\boldsymbol{\mu}}{dt^2} - \left(\frac{d\Sigma}{dt}\right)\Sigma^{-1}\left(\frac{d\boldsymbol{\mu}}{dt}\right) = 0 \\ \frac{d^2\Sigma}{dt^2} + \left(\frac{d\boldsymbol{\mu}}{dt}\right)\left(\frac{d\boldsymbol{\mu}}{dt}\right)^t - \left(\frac{d\Sigma}{dt}\right)\Sigma^{-1}\left(\frac{d\Sigma}{dt}\right) = 0. \end{cases} \tag{11}$$

and could be partially integrated [27]:

$$\begin{cases} \frac{d\boldsymbol{\mu}}{dt} = \Sigma\boldsymbol{x} \\ \frac{d\Sigma}{dt} = \Sigma(B - \boldsymbol{x}^t\mu), \end{cases} \tag{12}$$

$$\begin{cases} \frac{d\Delta}{dt} = -B\Delta + \boldsymbol{x}\boldsymbol{\delta}^t \\ \frac{d\boldsymbol{\delta}}{dt} = -B\boldsymbol{\delta} + (1 + \boldsymbol{\delta}\Delta^{-1}\boldsymbol{\delta})\boldsymbol{x} \end{cases}. \tag{13}$$

where $(\boldsymbol{\delta}(t), \Delta(t)) = (\Sigma^{-1}(t)\boldsymbol{\mu}(t), \Sigma^{-1}(t))$, $\boldsymbol{x} \in \mathbb{R}^n$, and B is a symmetric matrix. The initial conditions for this problem can be taken as:

$$\begin{cases} (\boldsymbol{\delta}(0), \Delta(0)) = (\mathbf{0}, I_n) \\ \left(\frac{d\boldsymbol{\delta}}{dt}(0), \frac{d\Delta}{dt}(0)\right) = (\boldsymbol{x}, -B). \end{cases} \tag{14}$$

Eriksen [27] and Calvo and Oller [28], in independent works, solved this initial value problem. An explicit solution to the geodesic curve in $\mathcal{M}$ [28] is:

$$\begin{cases} \boldsymbol{\delta}(t) = & -B(\cosh(tG) - I_n)(G^-)^2\boldsymbol{x} + \sinh(tG)G^-\boldsymbol{x} \\ \Delta(t) = & I_n + \frac{1}{2}(\cosh(tG) - I_n) + \frac{1}{2}B(\cosh(tG) - I_n)(G^-)^2B \\ & -\frac{1}{2}\sinh(tG)G^-B - \frac{1}{2}B\sinh(tG)G^- \end{cases}, \tag{15}$$

where I_n is an n-order identity matrix, $G^2 = B^2 + 2\boldsymbol{x}\boldsymbol{x}^t$, and G^- is the generalized inverse square matrix of G, that is $GG^-G = G$.

Due the fact that the geodesic curve has constant velocity at any point, given $(\boldsymbol{x}, B)$ in the tangent space of $\mathcal{M}$, the Fisher–Rao distance between $(\mathbf{0}, I_n)$ and $(\boldsymbol{\delta}(1), \Delta(1))$ is:

$$\int_0^1 \sqrt{\frac{d\boldsymbol{\mu}}{dt}(0)\Sigma^{-1}(0)\frac{d\boldsymbol{\mu}}{dt}(0) + \frac{1}{2}\operatorname{tr}\left[\left(\Sigma^{-1}(0)\frac{\Sigma}{dt}(0)\right)^2\right]}\, dt = \sqrt{\frac{1}{2}\operatorname{tr}(B^2) + |\boldsymbol{x}|^2}, \tag{16}$$

where $|\cdot|$ is the standard Euclidean norm. Note that the above expression provides the Fisher–Rao distance between two distributions only if we can determine the initial value problem from the boundary conditions, which usually is very difficult.

Han and Park in [31] presented a numerical shooting method for computing the minimum geodesic distance between two normal distributions, through parallel transport of a vector field defined along the geodesic curve given in Equation (15).

A closed form for the Fisher–Rao distance between two normal distributions in $\mathcal{M}$ is still an important open question. Next, we present closed forms for this distance in some submanifolds of $\mathcal{M}$.

2.1. Closed Forms for the Fisher–Rao Distance in Submanifolds of $\mathcal{M}$

In this subsection, we consider submanifolds $\mathcal{M}_* \subset \mathcal{M}$ with the distance induced by the Fisher–Rao metric in $\mathcal{M}$. It is important to remark that, in general, given two distributions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\mathcal{M}_*$, the distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ when restricted to a submanifold $\mathcal{M}_*$ is bigger than the distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\mathcal{M}$, that is $d_{\mathcal{M}_*}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq d_{\mathcal{M}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. This is due to the fact that to get $d_{\mathcal{M}_*}$, we consider the minimum length of restricted curves, which are the ones contained in the submanifold $\mathcal{M}_*$. We say that $\mathcal{M}_*$ is totally geodesic if only if $d_{\mathcal{M}_*}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d_{\mathcal{M}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{M}_*$, which means that the geodesic in $\mathcal{M}$ connecting $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is contained in $\mathcal{M}_*$.

2.1.1. The Submanifold $\mathcal{M}_\Sigma$ Where Σ Is Constant

In the n-dimensional manifold composed by multivariate normal distributions with common covariance matrix Σ, $\mathcal{M}_\Sigma = \{p_{\boldsymbol{\theta}}; \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma),\ \Sigma = \Sigma_0 \in P_n(\mathbb{R}) \text{ constant}\}$, the Fisher–Rao distance between two distributions $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1, \Sigma_0)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_2, \Sigma_0)$ is [18]:

$$d_\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma_0^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}. \tag{17}$$

This distance is equal to the Mahalanobis distance [11], which is equal to the Euclidean distance between the image of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ under the transformation $\boldsymbol{\mu} \mapsto P^{-1}\boldsymbol{\mu}$, where $\Sigma_0 = PP^t$ is the Cholesky decomposition [18]. This distance was one of the first dissimilarity measures between datasets with some correlation. Note that this submanifold is not totally geodesic, as it can be seen even in the space of univariate normal distributions [22] and in Example 1 in the next section.

A geodesic curve $\gamma_\Sigma(t)$ in $\mathcal{M}_\Sigma$ connecting $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be provided by:

$$\gamma_\Sigma(t) = ((1-t)\boldsymbol{\mu}_1 - t\boldsymbol{\mu}_2, \Sigma_0). \tag{18}$$

2.1.2. The Submanifold $\mathcal{M}_\mu$ where μ Is Constant

A totally geodesic submanifold of $\mathcal{M}$ is given by $\mathcal{M}_{\boldsymbol{\mu}} = \{p_{\boldsymbol{\theta}}; \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma),\ \boldsymbol{\mu} = \boldsymbol{\mu}_0 \in \mathbb{R}^n \text{ constant}\}$ of dimension $\frac{n(n+1)}{2}$ composed by distributions that have the same mean vector $\boldsymbol{\mu}_0$. The Fisher–Rao distance in $\mathcal{M}_{\boldsymbol{\mu}}$ was studied by several authors in different contexts [16,18,30,41] and for $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_0, \Sigma_1)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_0, \Sigma_2)$ is given by:

$$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{n} [\log(\lambda_i)]^2}, \tag{19}$$

where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$.

An expression for the geodesic curve connecting these two distributions is [30]:

$$\gamma_{\boldsymbol{\mu}}(t) = (\boldsymbol{\mu}_0, \Sigma_1^{1/2} \exp(t \log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}))\Sigma_1^{1/2}). \tag{20}$$

2.1.3. The Submanifold $\mathcal{M}_D$ Where Σ Is Diagonal

Let $\mathcal{M}_D = \{p_{\boldsymbol{\theta}};\ \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma),\ \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2),\ \sigma_i > 0,\ i = 1, \ldots, n\}$, the submanifold of $\mathcal{M}$ composed by distributions with a diagonal covariance matrix. If we consider the parameter $\boldsymbol{\theta} = (\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, \mu_n, \sigma_n)$, it can be shown [22] that the metric in the parametric space of $\mathcal{M}_D$ is equal to the product metric:

$$d_D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\sum_{i=1}^{n} d_{F*}^2((\mu_{1i}, \sigma_{1i}), (\mu_{2i}, \sigma_{2i}))}, \tag{21}$$

where d_{F*} is the Fisher–Rao distance in the univariate case given by [22]:

$$d_{F*}((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \sqrt{2}\log\frac{\left|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right|+\left|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right|}{\left|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right|-\left|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right|}. \tag{22}$$

In this space, a curve $\gamma_D(t) = (\gamma_1(t),\ldots,\gamma_n(t))$ is a geodesic if, and only if, $\gamma_i(t)$ is a geodesic curve in the univariate case, for all $i = 1,\ldots,n$. The geodesic curves in the univariate normal distributions space (upper half plane $\mathbb{R}\times\mathbb{R}^+$) are half-vertical lines and half-ellipses centered at $\sigma = 0$, with eccentricity $\frac{1}{\sqrt{2}}$ [22].

It is important to note that $\mathcal{M}_D \subset \mathcal{M}$ is not totally geodesic. The submanifold of $\mathcal{M}_D$ composed only by normal distributions with covariance matrices which are multiples of the identity (round normals) is totally geodesic [22]. In fact, this submanifold of round normals is also contained in the totally geodesic submanifold described next.

2.1.4. The Submanifold $\mathcal{M}_{D\mu}$ Where Σ Is Diagonal and μ Is an Eigenvector of Σ

Let $\mathcal{M}_{D\boldsymbol{\mu}}$ be the $n+1$-dimensional submanifold composed by distributions with the mean vector $\boldsymbol{\mu} = \mu_1\boldsymbol{e}_i$ for some $\boldsymbol{e}_i \in \{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n\}$ (the canonical basis of $\mathbb{R}^n$) and diagonal covariance matrix Σ, and without loss of generality, we shall assume that $\boldsymbol{e}_i = \boldsymbol{e}_1$. An analytic expression for the distance in $\mathcal{M}_{D\boldsymbol{\mu}}$ is:

$$d^2_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = d^2_{F*}\left((\mu_{11},\sigma_{11}),(\mu_{21},\sigma_{21})\right) + \sum_{i=2}^{n} d^2_{F*}\left((0,\sigma_{1i}),(0,\sigma_{2i})\right). \tag{23}$$

We proved in [42] that this submanifold is totally geodesic.

2.2. Bounds for the Fisher–Rao in $\mathcal{M}$

As mentioned, a closed form for the Fisher–Rao distance between two general normal distributions is not known. In this subsection, we present some bounds for this distance.

2.2.1. A Lower Bound

Calvo and Oller [43] derived a lower bound for the Fisher–Rao distance through an isometric embedding of the parametric space $\mathcal{M}$ into the manifold of the positive definite matrices.

Proposition 1. *[43] Given $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1,\Sigma_1)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_2,\Sigma_2)$, let:*

$$S_i = \begin{pmatrix} \Sigma_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^t & \boldsymbol{\mu}_i^t \\ \boldsymbol{\mu}_i & 1 \end{pmatrix}, \tag{24}$$

$i = 1,2$. A lower bound for the distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is:

$$LB(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{n+1}[\log(\lambda_i)]^2}, \tag{25}$$

where λ_i, $1 \le i \le n+1$, are the eigenvalues of $S_1^{-1/2}S_2S_1^{-1/2}$.

We note that this bound satisfies the distance proprieties in $\mathcal{M}$. In [44], through a similar approach, a lower bound for the Fisher–Rao distance was obtained in the more general space of elliptical distributions, restricted to normal distributions, is the above bound.

2.2.2. The Upper Bound UB_1

In [45], we proposed an upper bound based on an isometry (8) in the manifold $\mathcal{M}$ and on the distance in the non-totally geodesic submanifold $\mathcal{M}_D$ (21), as follows:

Proposition 2. *[45] The Fisher–Rao distance between two multivariate normal distributions $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_2, \Sigma_2)$ is upper bounded by,*

$$UB_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\sum_{i=1}^{n} d_{F*}^2((0,1),(\mu_i, \lambda_i))}, \tag{26}$$

where λ_i are the diagonal terms of the matrix Λ given by the eigenvalues of $A = \Sigma_1^{-(1/2)}\Sigma_2\Sigma_1^{-(1/2)} = Q\Lambda Q^t$, μ_i are the coordinates of $\boldsymbol{\mu} = Q^t\Sigma_1^{-(1/2)}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, Q is the orthogonal matrix whose columns are the eigenvectors of A and d_{F*} is the Fisher–Rao distance between univariate normal distributions given in Equation (22).

2.2.3. The Upper Bounds UB_2 and UB_3

Considering the Fisher–Rao distance in the totally geodesic submanifold $\mathcal{M}_{D\boldsymbol{\mu}}$ and the triangular inequality, we propose another upper bound [42].

Given $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_2, \Sigma_2)$, we consider the Fisher–Rao distance between $\boldsymbol{\theta}_0 = (\mathbf{0}, I_n)$ and $\boldsymbol{\theta}_3 = (\boldsymbol{\mu}_3, \Sigma_3)$ as in Equation (10). Let $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\mu}}, \bar{\Sigma})$; by the triangular inequality, it follows that:

$$d_F(\boldsymbol{\theta}_0, \boldsymbol{\theta}_3) \leq d_F(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) + d_F(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_3). \tag{27}$$

To calculate this bound, we choose $\bar{\boldsymbol{\theta}}$ appropriately. For $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_3$, note that $d_F(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_3) = d_{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_3)$. Let P be an orthogonal matrix such that $P\boldsymbol{\mu} = (|\boldsymbol{\mu}_3|, 0, \ldots, 0)$ and $D = \text{diag}(d_1^2, d_2^2, \ldots, d_n^2)$ a diagonal matrix. We will consider $\bar{\Sigma} = P^{-1}DP^{-t}$ and $\boldsymbol{\theta}_P = (P\boldsymbol{\mu}, D)$. By the isometry $\psi_{(c,Q)}$, given in Equation (9), for $Q = P^{-1}$ and $c = \mathbf{0}$, it follows:

$$d_F(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) = d_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_P). \tag{28}$$

Then, combining Inequality (27) and Equation (28), the left side of the equation below is an upper bound for the Fisher–Rao distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$,

$$d_F(\boldsymbol{\theta}_0, \boldsymbol{\theta}_3) \leq d_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_P) + d_{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_3). \tag{29}$$

In [42], we derived the upper bound:

$$UB_2 = d_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_P) + d_{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_3). \tag{30}$$

through a numerical minimization process by considering the diagonal elements of D as a vector that minimizes $d_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_P) + d_{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_3)$,

$$(\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_n) = \min_{(d_1, d_2, \ldots, d_n)} \{d_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_P) + d_{\boldsymbol{\mu}}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_3)\}. \tag{31}$$

We also derive an analytic upper bound UB_3 by minimizing of the distance $d_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_P)$. By expressing this distance in terms of the parameters $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_P) = ((\mathbf{0}, I_n), (P\boldsymbol{\mu}, D))$, we can show that it reaches the minimum at:

$$D = \text{diag}\left(\sqrt{\frac{|\boldsymbol{\mu}_3| + 2}{2}}, 1, \cdots, 1\right). \tag{32}$$

The lower bound of Section 2.2.1 and the upper bounds of Sections 2.2.2 and 2.2.3 are summarized in Table 1.

Table 1. The lower bound $LB(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and the upper bounds $UB_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $UB_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $UB_3(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for the Fisher–Rao distance, $d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, between distributions $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_2, \Sigma_2)$ in $\mathcal{M}$. d_{F*} is the distance between univariate normal distributions given in Equation (22).

$LB(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{n+1}[\log(\lambda_i)]^2}$	– $S_i = \begin{pmatrix} \Sigma_i + \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i & \boldsymbol{\mu}_i^t \\ \boldsymbol{\mu}_i & 1 \end{pmatrix}$; – λ_i are the eigenvalues of $S_1^{-1/2} S_2 S_1^{-1/2}$.
$UB_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\sum_{i=1}^{n} d_{F*}^2((0,1),(\mu_i, \lambda_i))}$	– $\Sigma_1^{-(1/2)} \Sigma_2 \Sigma_1^{-(1/2)} = Q \Lambda Q^t$; – λ_i are the diagonal terms of the matrix Λ; – μ_i are the coordinates of $\boldsymbol{\mu} = Q^t \Sigma_1^{-(1/2)} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$.
$UB_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}[\log(\lambda_i)]^2 + \sqrt{d_{F*}^2((0,1),(\lvert\boldsymbol{\mu}_3\rvert, \bar{d}_1)) + \sum_{i=2}^{m} d_{F*}^2((0,1),(0,\bar{d}_i))}}$	– $\boldsymbol{\mu}_3 = \Sigma_1^{-(1/2)}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$; – $\bar{d}_i$ is given by (31); – $\Sigma_3 = \Sigma_1^{-(1/2)} \Sigma_2 \Sigma_1^{-(1/2)}$; – P is an orthogonal matrix such that $P\boldsymbol{\mu} = (\lvert\boldsymbol{\mu}_3\rvert, 0, \ldots, 0)$; – $\bar{\Sigma} = P^{-1} \Sigma_3 P^{-t}$; – λ_i are the eigenvalues of $\bar{\Sigma}^{-(1/2)} \Sigma_3 \bar{\Sigma}^{-(1/2)}$.
$UB_3(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}[\log(\lambda_i)]^2 + d_{F*}\left((0,1),\left(\lvert\boldsymbol{\mu}_3\rvert, \sqrt{\frac{\lvert\boldsymbol{\mu}_3\rvert + 2}{2}}\right)\right)}$	– $\boldsymbol{\mu}_3 = \Sigma_1^{-(1/2)}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$; – $\Sigma_3 = \Sigma_1^{-(1/2)} \Sigma_2 \Sigma_1^{-(1/2)}$; – P is an orthogonal matrix such that $P\boldsymbol{\mu} = (\lvert\boldsymbol{\mu}_3\rvert, 0, \ldots, 0)$; – $\bar{\Sigma} = P^{-1} \Sigma_3 P^{-t}$; – λ_i are the eigenvalues of $\bar{\Sigma}^{-(1/2)} \Sigma_3 \bar{\Sigma}^{-(1/2)}$.

Upper and lower bounds have been used to estimate the Fisher–Rao distance in applications such as [37].

2.2.4. Comparisons of the Bounds

In this section, as in [42], we illustrate comparisons between the bounds presented previously.

We consider the bivariate normal distributions model ($n = 2$) and distributions $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\Sigma})$, where:

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \right). \tag{33}$$

From (10), we can see that there always exists an isometry that converts any two pairs of bivariate distributions into a pair of distributions as above.

We present next a comparison between the lower bound "LB" (25), the upper bounds UB_1 (26), UB_2 (31), and UB_3 (32), and the numerical solution given by the geodesic shooting algorithm (GS) [31] in specific situations.

In Figure 1, we consider the eigenvalues $\lambda_1 = 2$, $\lambda_2 = 0.5$, and $\mu = 1$ to be fixed and α varying from zero to $\frac{\pi}{2}$. We note that the upper bound UB_1 is very near the lower bound LB and to the numerical solution GS. The other upper bounds are bigger than the bound UB_1. In Figure 2, it is considered $\mu = 10$ and the previous eigenvalues. Now, the best performance is of bounds UB_2 and UB_3, which are similar. In Figure 3a, we again keep the eigenvalues; the rotation angle is fixed $\alpha = \frac{\pi}{4}$; and μ varies from zero to 10. We can see similar performances of UB_2 and UB_3, which are better than UB_1 for larger values of μ.

We may also consider the upper bound:

$$UB_{123}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \min\{UB_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), UB_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), UB_3(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}. \tag{34}$$

Figure 3b displays the comparison between LB, UB_{123}, and GS for the same data of the comparison in Figure 3a.

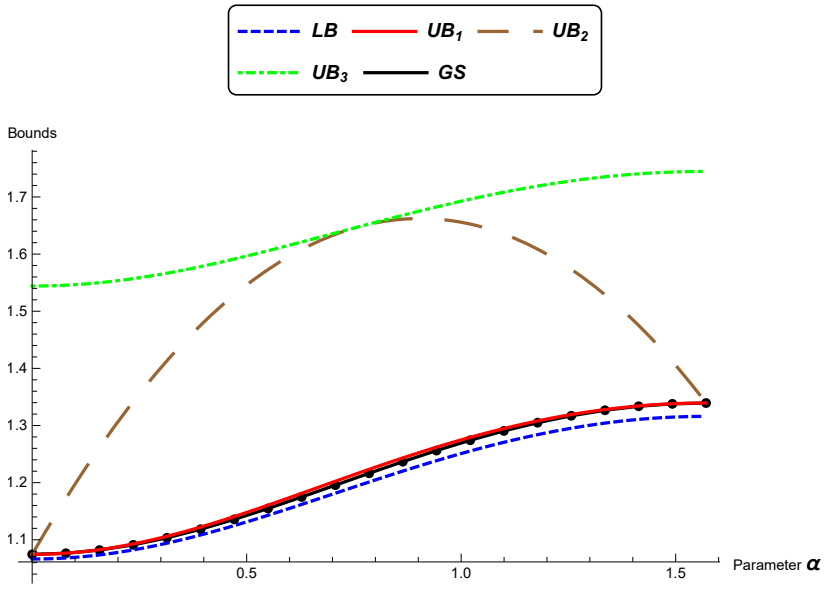


Figure 1. A comparison between the bounds LB, UB_1, UB_2, UB_3, and GS. ($\lambda_1 = 2$, $\lambda_2 = 0.5$, and $\mu = 1$ are fixed, and α varies from zero to $\frac{\pi}{2}$).

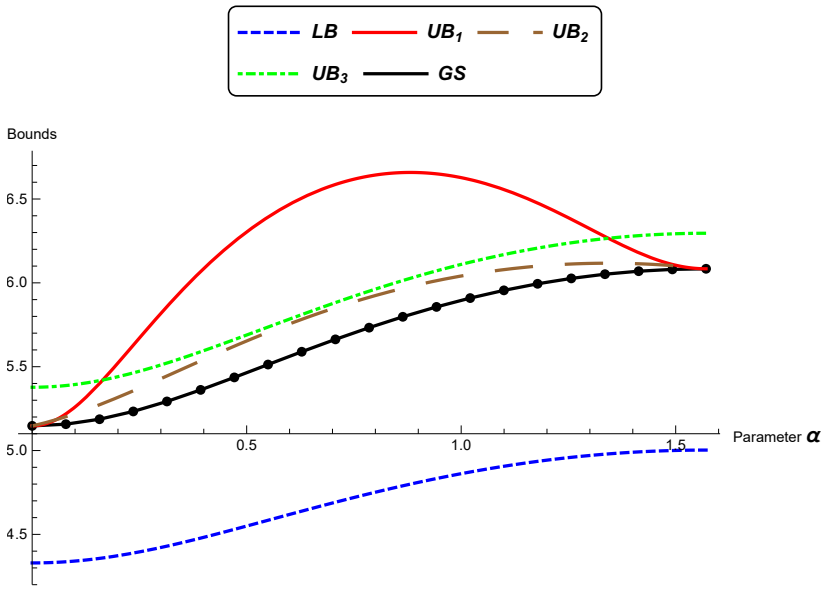
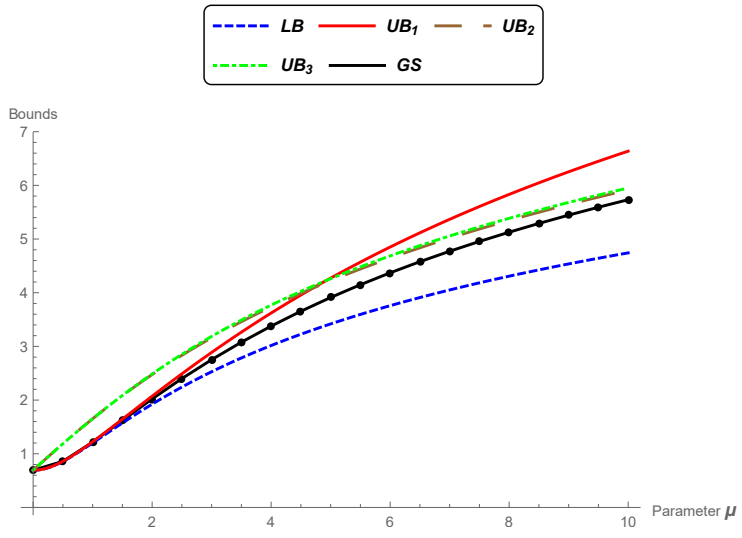
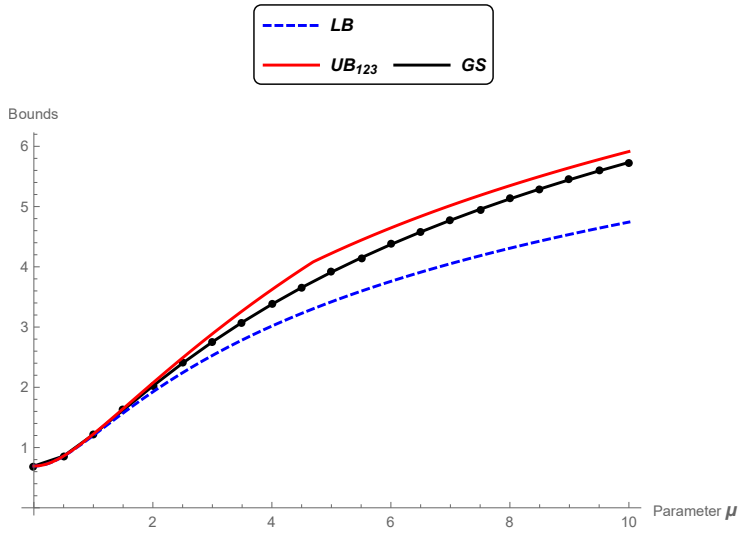


Figure 2. A comparison between the bounds LB, UB_1, UB_2, UB_3, and GS. ($\lambda_1 = 2$, $\lambda_2 = 0.5$, and $\mu = 10$ are fixed, and α varies from zero to $\frac{\pi}{2}$).



(**a**)



(**b**)

Figure 3. (**a**) A comparison between the bounds LB, UB_1, UB_2, UB_3, and GS. ($\lambda_1 = 2$, $\lambda_2 = 0.5$, and the rotation angle $\alpha = \pi/4$ are fixed, and μ varies from zero to 10). (**b**) A comparison between the bounds LB, UB_{123}, and GS. ($\lambda_1 = 2$, $\lambda_2 = 0.5$, and the rotation angle $\alpha = \pi/4$ are fixed, and μ varies from zero to 10).

3. Fisher–Rao Distance Between Special Distributions

In this section, we describe the Fisher–Rao distance in the full space $\mathcal{M}$ between special kinds of distributions.

3.1. The Fisher–Rao Distance Between Distributions with Common Covariance Matrices

The Fisher–Rao distance between distributions with common covariance matrices given in Section 2.1.1 was restricted to non-totally geodesic submanifold $\mathcal{M}_\Sigma$. We show next that using the isometry given in (8) and the distance in the submanifold $\mathcal{M}_{D\mu}$, it is possible to find a closed form for the distance between two distributions with the same covariance matrix, in the full manifold $\mathcal{M}$.

Proposition 3. *Given two distributions $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}_2, \Sigma)$ in $\mathcal{M}$, let P be an orthogonal matrix such that $P(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = |\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{e}_1$, and consider the decomposition UDU^t of the matrix $P\Sigma P^t$,*

$$P\Sigma P^t = UDU^t, \tag{35}$$

where U is an upper triangular matrix with all diagonal entries equal to one and D is a diagonal matrix. The Fisher–Rao distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is given by:

$$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d_{D\boldsymbol{\mu}}((\mathbf{0}, D), (|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{e}_1, D)). \tag{36}$$

Proof. By considering the isometries $\psi = \psi_{(-P\boldsymbol{\mu}_1, P)}$ and $\hat{\psi} = \psi_{(\mathbf{0}, U^{-1})}$ and the decomposition given by Equation (35), it follows from Equation (9) that:

$$\begin{aligned}
d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= d_F(\psi_{(-P\boldsymbol{\mu}_1, P)}(\boldsymbol{\theta}_1), \psi_{(-P\boldsymbol{\mu}_1, P)}(\boldsymbol{\theta}_2)) \\
&= d_F((P\boldsymbol{\mu}_1 - P\boldsymbol{\mu}_1, P\Sigma P^t), (P\boldsymbol{\mu}_2 - P\boldsymbol{\mu}_1, P\Sigma P^t)) \\
&= d_F((\mathbf{0}, P\Sigma P^t), (|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{e}_1, P\Sigma P^t)). \\
&= d_F(\hat{\psi}(\mathbf{0}, P\Sigma P^t), \hat{\psi}(|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{e}_1, P\Sigma P^t)) \\
&= d_F((U^{-1}\mathbf{0}, U^{-1}P\Sigma P^t U^{-t}), (|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|U^{-1}\boldsymbol{e}_1, U^{-1}P\Sigma P^t U^{-t})) \\
&= d_F((\mathbf{0}, D), (|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{e}_1, D)).
\end{aligned} \tag{37}$$

Since the distributions $(\mathbf{0}, D)$ and $(|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{e}_1, D)$ belong to the submanifold $\mathcal{M}_{D\boldsymbol{\mu}}$, we conclude that:

$$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d_{D\boldsymbol{\mu}}((\mathbf{0}, D), (|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{e}_1, D)). \tag{38}$$

□

Example 1. *Consider two bivariate normal distributions $\boldsymbol{\theta}_1 = ((-1, 0)^t, \Sigma)$ and $\boldsymbol{\theta}_2 = ((6, 3)^t, \Sigma)$ with the same covariance matrix:*

$$\Sigma = \begin{pmatrix} 1.1 & 0.9 \\ 0.9 & 1.1 \end{pmatrix}.$$

Figure 4a illustrates the normal distributions in the geodesic curve connecting connecting $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\mathcal{M}$, and Figure 4b illustrates the geodesic in the submanifold $\mathcal{M}_\Sigma$. We observe that in $\mathcal{M}$, the shape of the ellipses (contour curves) changes along the path. Furthermore, the Fisher–Rao distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is $d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 5.00648$, which is less than the Mahalanobis distance given in Equation (17), $d_\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 8.06226$, as expected, since the submanifold $\mathcal{M}_\Sigma$ is not totally geodesic.

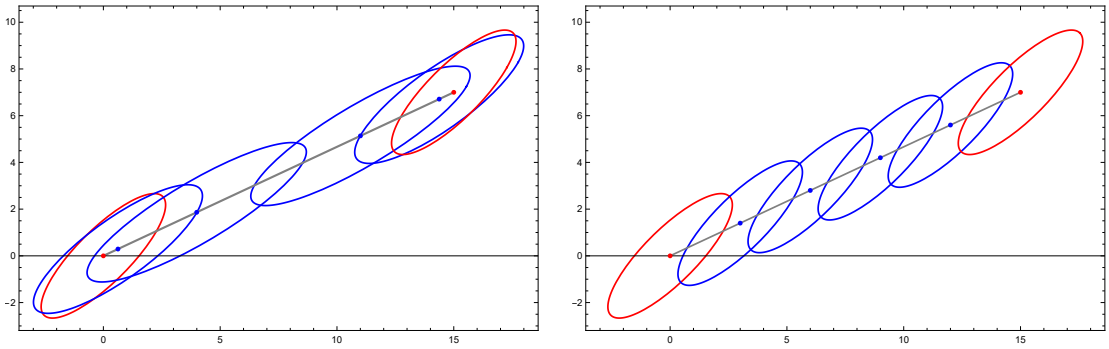


Figure 4. (**a**) Level curves of the distributions in the geodesic curve connecting the bivariate normal distributions $\boldsymbol{\theta}_1 = ((-1,0)^t, \Sigma)$ and $\boldsymbol{\theta}_2 = ((6,3)^t, \Sigma)$ in $\mathcal{M}$. (**b**) Level curves of the distributions in the geodesic curve connecting the bivariate normal distributions $\boldsymbol{\theta}_1 = ((-1,0)^t, \Sigma)$ and $\boldsymbol{\theta}_2 = ((6,3)^t, \Sigma)$ in $\mathcal{M}_\Sigma$.

3.2. The Fisher–Rao Distance between Mirrored Distributions

We consider here two mirrored normal distributions; that is, without loss of generality, if we consider up rotation, the line connecting $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as parallel to the $\boldsymbol{e}_1$-axis, and the covariance matrices Σ_1 and Σ_2 satisfying:

$$\Sigma_2 = M_1 \, \Sigma_1 \, M_1, \text{ where } M_1 = \begin{pmatrix} -1 & 0 \\ 0 & I_{n-1} \end{pmatrix}. \tag{39}$$

This condition implies also the same eigenvalues for both matrices.

For bivariate normal distributions, we then should have:

$$\boldsymbol{\theta}_1 = \left(\begin{pmatrix} \mu_1 \\ \mu_0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right) \quad \text{e} \quad \boldsymbol{\theta}_2 = \left(\begin{pmatrix} \mu_2 \\ \mu_0 \end{pmatrix} \begin{pmatrix} \sigma_{11} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{22} \end{pmatrix} \right); \tag{40}$$

see Figure 5.

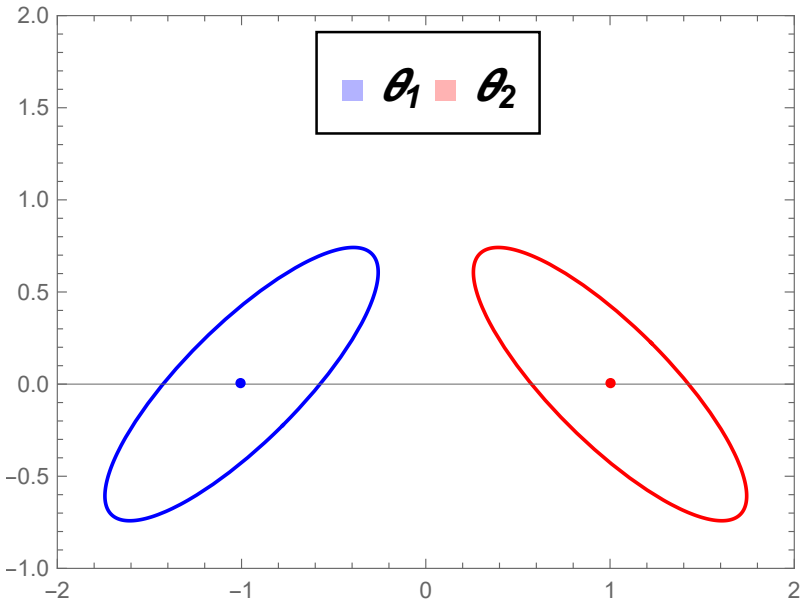


Figure 5. Example of level curves of mirrored distributions where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are given by Equation (40).

After several experiments using the algorithm *geodesic shooting* for the $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we have observed that for $t = 0$ the geodesic curve connecting these distributions ($\gamma(t) = (\boldsymbol{\mu}(t), \Sigma(t))$, with $\gamma(-1) = \boldsymbol{\theta}_1$ and $\gamma(1) = \boldsymbol{\theta}_2$), satisfies

$$\gamma(0) \approx \boldsymbol{\theta}_{1/2} = (\boldsymbol{\mu}_{1/2}, \Sigma_{1/2}) = \left(\begin{pmatrix} \frac{\mu_1+\mu_2}{2} \\ \eta \end{pmatrix}, \begin{pmatrix} d_{11}^2 & 0 \\ 0 & d_{22}^2 \end{pmatrix} \right), \tag{41}$$

$$\gamma'(0) \approx \hat{\boldsymbol{\theta}}_{1/2} = (\hat{\boldsymbol{\mu}}_{1/2}, \hat{\Sigma}_{1/2}) = \left(\begin{pmatrix} \hat{\mu}_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & 0 \end{pmatrix} \right). \tag{42}$$

where η, d_{11}, and d_{22} are real values; see Figure 6.

The focus here is the "shape" of these distributions. Note that at $t = 0$, the distribution $\gamma(0)$ appears as $\boldsymbol{\theta}_{1/2}$, which has a diagonal covariance matrix, and the tangent vector $\gamma'(0)$ appears as $\hat{\boldsymbol{\theta}}_{1/2}$, which is composed by a mean vector with the second entry equal to zero and by a symmetric covariance matrix with a null diagonal.

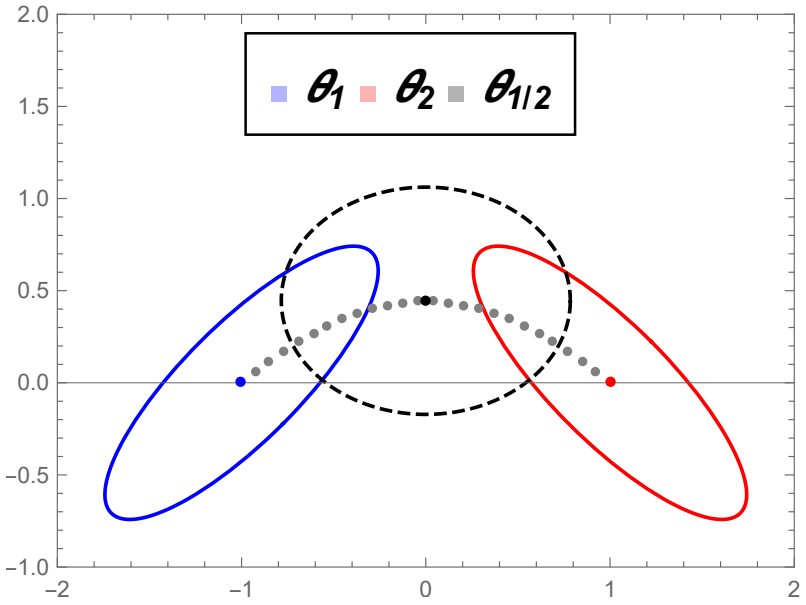


Figure 6. Approximation of the geodesic curve connecting $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ via the geodesic shooting algorithm. The level curve of $\boldsymbol{\theta}_{1/2}$ is the dashed one.

This observation inspired us to get an explicit expression for the geodesic connecting two mirrored distributions. Starting with the bi-dimensional case again, we will prove that in fact we have equality in Expressions (41) and (42).

Let $\gamma(t) = (\boldsymbol{\mu}(t), \Sigma(t))$, $-1 \leq t \leq 1$, and the geodesic curve in $\mathcal{M}$ connecting $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and consider that $\gamma(0) = \boldsymbol{\theta}_{1/2}$ and $\gamma'(0) = \hat{\boldsymbol{\theta}}_{1/2}$. Given the isometry $\psi = \psi_{\left(-\Sigma_{1/2}^{-1/2}\boldsymbol{\mu}_{1/2}, \Sigma_{1/2}^{-1/2}\right)}$, we define:

$$\bar{\gamma}(t) = (\bar{\boldsymbol{\mu}}(t), \bar{\Sigma}(t)) := \psi(\gamma(t)) = \left(\Sigma_{1/2}^{-1/2}(\boldsymbol{\mu}(t) - \boldsymbol{\mu}_{1/2}), \Sigma_{1/2}^{-1/2}\, \Sigma(t)\, \Sigma_{1/2}^{-1/2} \right). \tag{43}$$

Then:

$$\bar{\gamma}'(t) = \left(\frac{d\bar{\boldsymbol{\mu}}(t)}{dt}, \frac{\bar{\Sigma}(t)}{dt} \right) = \left(\Sigma_{1/2}^{-1/2} \left(\frac{d\boldsymbol{\mu}(t)}{dt} \right), \Sigma_{1/2}^{-1/2} \left(\frac{\Sigma(t)}{dt} \right) \Sigma_{1/2}^{-1/2} \right), \tag{44}$$

$$\begin{aligned} \bar{\gamma}(0) &= \left(\Sigma_{1/2}^{-1/2}(\boldsymbol{\mu}_{1/2} - \boldsymbol{\mu}_{1/2}), \Sigma_{1/2}^{-1/2}\, \Sigma_{1/2}\, \Sigma_{1/2}^{-1/2} \right) \\ &= (\mathbf{0}, I_2) =: \boldsymbol{\theta}_0 \end{aligned} \tag{45}$$

and:

$$\begin{aligned}\bar{\gamma}'(0) &= \left(\Sigma_{1/2}^{-1/2}\,\boldsymbol{\mu}_{1/2}', \Sigma_{1/2}^{-1/2}\,\Sigma_{1/2}'\,\Sigma_{1/2}^{-1/2}\right)\\ &= \left(\begin{pmatrix}\frac{\mu'(0)}{d_{11}}\\ 0\end{pmatrix}, \begin{pmatrix}0 & \frac{\sigma_{12}'(0)}{d_{11}d_{22}}\\ \frac{\sigma_{12}'(0)}{d_{11}d_{22}} & 0\end{pmatrix}\right).\end{aligned} \tag{46}$$

Applying the natural changing of parameters:

$$(\delta(t), \Delta(t)) = \varphi(\bar{\boldsymbol{\mu}}(t), \bar{\Sigma}(t)) = (\bar{\Sigma}(t)^{-1}\bar{\boldsymbol{\mu}}(t), \bar{\Sigma}(t)^{-1}), \tag{47}$$

it follows that:

$$\begin{cases}\dfrac{d\Delta}{dt}(t) = -\,\Delta(t)\left(\dfrac{d\bar{\Sigma}}{dt}(t)\right)\Delta(t)\\ \dfrac{d\boldsymbol{\delta}}{dt}(t) = \left(\dfrac{d\Delta}{dt}(t)\right)\bar{\boldsymbol{\mu}}(t) + \Delta(t)\left(\dfrac{d\bar{\boldsymbol{\mu}}}{dt}(t)\right)\end{cases}. \tag{48}$$

Then, given that $(\delta(0), \Delta(0)) = (\bar{\boldsymbol{\mu}}(0), \bar{\Sigma}(0)) = (\mathbf{0}, I_2)$,

$$\begin{cases}\dfrac{d\Delta}{dt}(0) = -\,\Delta(0)\left(\dfrac{d\bar{\Sigma}}{dt}(0)\right)\Delta(0) = -\dfrac{d\bar{\Sigma}}{dt}(0)\\ \dfrac{d\boldsymbol{\delta}}{dt}(0) = \left(\dfrac{d\Delta}{dt}(0)\right)\bar{\boldsymbol{\mu}}(0) + \Delta(0)\left(\dfrac{d\bar{\boldsymbol{\mu}}}{dt}(0)\right) = \dfrac{d\bar{\boldsymbol{\mu}}}{dt}(0)\end{cases}. \tag{49}$$

That is, at $t = 0$, the tangent vector $\left(\frac{d\delta}{dt}(0), \frac{d\Delta}{dt}(0)\right)$ is equal to the tangent vector in (46). Furthermore, the distributions $\boldsymbol{\vartheta}_1 = \varphi(\bar{\boldsymbol{\theta}}_1) = (\bar{\Sigma}_1^{-1}\bar{\boldsymbol{\mu}}_1, \bar{\Sigma}_1^{-1})$ and $\boldsymbol{\vartheta}_2 = \varphi(\bar{\boldsymbol{\theta}}_2) = (\bar{\Sigma}_2^{-1}\bar{\boldsymbol{\mu}}_2, \bar{\Sigma}_2^{-1})$ are also mirrored $\bar{\Sigma}_2^{-1} = M_1\,\bar{\Sigma}_1^{-1}\,M_1$. In fact,

$$\begin{aligned}\boldsymbol{\vartheta}_1 =& (\bar{\Sigma}_1^{-1}\bar{\boldsymbol{\mu}}_1, \bar{\Sigma}_1^{-1})\\ =& ((\Sigma_{1/2}^{-1/2}\,\Sigma_1\,\Sigma_{1/2}^{-1/2})^{-1}\Sigma_{1/2}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{1/2}), (\Sigma_{1/2}^{-1/2}\,\Sigma_1\,\Sigma_{1/2}^{-1/2})^{-1})\\ =& \left(\frac{1}{\det(\Sigma_1)}\begin{pmatrix}\sigma_{22}d_{11}\frac{\mu_1-\mu_2}{2} - \sigma_{12}d_{11}(\mu_0-\eta)\\ \sigma_{11}d_{22}(\mu_0-\eta) - \sigma_{12}d_{22}\frac{\mu_1-\mu_2}{2}\end{pmatrix}, \frac{1}{\det(\Sigma_1)}\begin{pmatrix}\sigma_{22}d_{11}^2 & -\sigma_{12}d_{11}d_{22}\\ -\sigma_{12}d_{11}d_{22} & \sigma_{11}d_{22}^2\end{pmatrix}\right),\end{aligned} \tag{50}$$

and by similar arguments, we obtain:

$$\boldsymbol{\vartheta}_2 = \left(\frac{1}{\det(\Sigma_1)}\begin{pmatrix}\sigma_{22}d_{11}\frac{\mu_2-\mu_1}{2} + \sigma_{12}d_{11}(\mu_0-\eta)\\ \sigma_{11}d_{22}(\mu_0-\eta) - \sigma_{12}d_{22}\frac{\mu_1-\mu_2}{2}\end{pmatrix}, \frac{1}{\det(\Sigma_1)}\begin{pmatrix}\sigma_{22}d_{11}^2 & \sigma_{12}d_{11}d_{22}\\ \sigma_{12}d_{11}d_{22} & \sigma_{11}d_{22}^2\end{pmatrix}\right). \tag{51}$$

Figure 7 illustrates the distributions $\boldsymbol{\theta}_0$, $\boldsymbol{\vartheta}_1$, and $\boldsymbol{\vartheta}_2$.

Conversely, by considering:

$$(\boldsymbol{x}, B) = \left(\begin{pmatrix}x\\ 0\end{pmatrix}, \begin{pmatrix}0 & b\\ b & 0\end{pmatrix}\right) \tag{52}$$

in the initial value problem given in Equations (13) and (14), it follows that the matrix $G^2 = B^2 + \boldsymbol{x}\boldsymbol{x}^t$ is diagonal. Therefore, the geodesic curve $(\delta(t), \Delta(t))$ with initial value $(\delta(0), \Delta(0)) = \boldsymbol{\theta}_0$ and tangent vector $(\boldsymbol{x}, B)$ given in Equation (15) can be simplified as follows:

$$\begin{cases} \delta(t) = \begin{pmatrix} \frac{x\sinh(t\sqrt{b^2+2x^2})}{\sqrt{b^2+2x^2}} \\ -\frac{bx(\cosh(t\sqrt{b^2+2x^2})-1)}{b^2+2x^2} \end{pmatrix} \\ \Delta(t) = \begin{pmatrix} \frac{1}{2}(\cosh(bt)+\cosh(t\sqrt{b^2+2x^2})) & -\frac{1}{2}\left(\sinh(bt)+\frac{b\sinh(t\sqrt{b^2+2x^2})}{\sqrt{b^2+2x^2}}\right) \\ -\frac{1}{2}\left(\sinh(bt)+\frac{b\sinh(t\sqrt{b^2+2x^2})}{\sqrt{b^2+2x^2}}\right) & \frac{1}{2}\left(\cosh(bt)+\frac{2x^2+b^2\cosh(t\sqrt{b^2+2x^2})}{b^2+2x^2}\right) \end{pmatrix} \end{cases}. \tag{53}$$

From the parity of the functions $\sinh(t)$ and $\cosh(t)$, it is possible to show that, given $t_0 \in \mathbb{R}$, the distributions:

$$(\delta(-t_0), \Delta(-t_0)) \quad \text{and} \quad (\delta(t_0), \Delta(t_0))$$

are also mirrored.

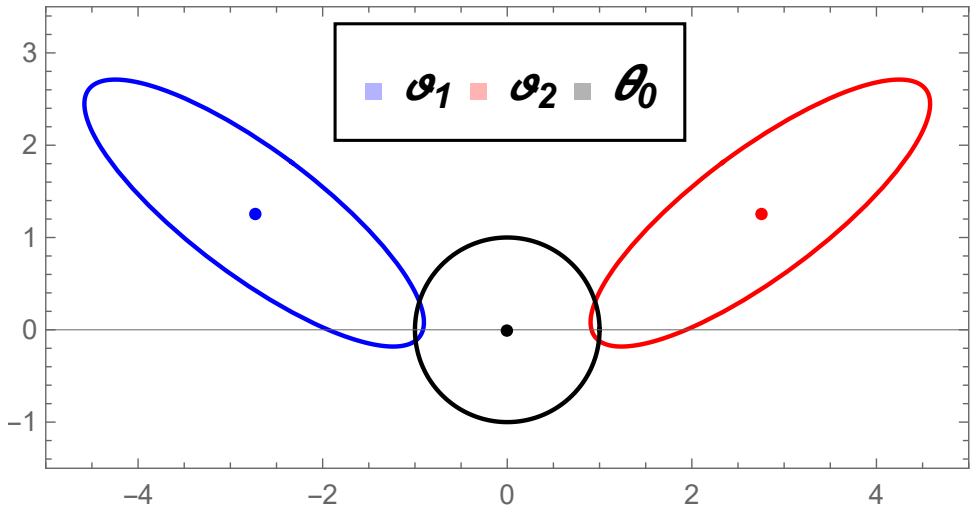


Figure 7. Contour curves of distributions $\vartheta_1 = \varphi(\bar{\theta}_1)$ and $\vartheta_2 = \varphi(\bar{\theta}_2)$.

By the above discussion, we conclude that it is possible to calculate the geodesic curve connecting θ_1 and θ_2 making $\psi^{-1}(\varphi^{-1}(\delta(-1), \Delta(-1))) = \theta_1$ and $\psi{-}1(\varphi^{-1}(\delta(1), \Delta(1))) = \theta_2$. That is, we need to find the values of η, d_{11}, and d_{22} of the isometry ψ and the values of (x, B) such that:

$$\begin{cases} \varphi(\psi(\theta_1)) = (\delta(-1), \Delta(-1)) \\ \varphi(\psi(\theta_2)) = (\delta(1), \Delta(1)) \end{cases}. \tag{54}$$

Since the two equations above are equivalent, it is enough to solve the equation:

$$(\delta(1), \Delta(1)) = \varphi(\psi(\mu_2, \Sigma_2)). \tag{55}$$

This is equivalent to solving the system:

$$\begin{cases} \begin{pmatrix} \frac{1}{d_{11}} & 0 \\ 0 & \frac{1}{d_{22}} \end{pmatrix} \Delta(1) \begin{pmatrix} \frac{1}{d_{11}} & 0 \\ 0 & \frac{1}{d_{22}} \end{pmatrix} = \Delta_2 \\ \begin{pmatrix} \frac{1}{d_{11}} & 0 \\ 0 & \frac{1}{d_{22}} \end{pmatrix} \delta(1) + \Delta_2 \begin{pmatrix} \frac{\mu_1+\mu_2}{2} \\ \eta \end{pmatrix} = \delta_2 \end{cases}, \tag{56}$$

where $(\delta_2, \Delta_2) = \varphi(\mu_2, \Sigma_2)$.

The above non-linear system has five equations and five variables $(d_{11}, d_{22}, \eta, x, b)$ and can be solved by an iterative method. With the solution of this system, we can determine the geodesic curve connecting the distributions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Moreover, by Equation (16), the Fisher–Rao distance is:

$$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 2d_F(\boldsymbol{\theta}_0, \boldsymbol{\theta}_2) = 2d_F((\mathbf{0}, I_n), (\delta(1), \Delta(1))) = 2\sqrt{\frac{1}{2}\operatorname{tr}(B^2) + \boldsymbol{x}^t\boldsymbol{x}} = 2\sqrt{b^2 + x^2}. \tag{57}$$

We also remark that the curve of the means $\delta(t)$ (and therefore, $\boldsymbol{\mu}(t)$) satisfies the equation of a hyperbola; in fact:

$$\frac{\left(-\frac{bx(\cosh(t\sqrt{b^2+2x^2})-1)}{b^2+2x^2} - \frac{bx}{b^2+2x^2}\right)^2}{\left(\frac{bx}{b^2+2x^2}\right)^2} - \frac{\left(\frac{x\sinh(t\sqrt{b^2+2x^2})}{\sqrt{b^2+2x^2}}\right)^2}{\left(\frac{x}{\sqrt{b^2+2x^2}}\right)^2} = 1. \tag{58}$$

Summarizing the above discussion, we have:

Proposition 4.

(i) *Expression (57) provides a closed form for the Fisher–Rao distance between two mirrored bivariate normal distributions, based on the solutions of the non-linear system (56).*

(ii) *The plane curve given by the coordinates of the mean vector in the geodesic connecting two of these distributions is a hyperbola.*

Table 2 shows a time comparison between the numerical method proposed here and the geodesic shooting to obtain the Fisher–Rao distance. The distributions used in this experiment were:

$$\boldsymbol{\theta}_1 = \left(\begin{pmatrix} -\mu \\ 0 \end{pmatrix}, \begin{pmatrix} 0.55 & -0.45 \\ -0.45 & 0.55 \end{pmatrix}\right) \quad \text{and} \quad \boldsymbol{\theta}_2 = \left(\begin{pmatrix} \mu \\ 0 \end{pmatrix} \begin{pmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{pmatrix}\right). \tag{59}$$

for different values of μ.

Table 2. A time comparison between the numerical method proposed here and the geodesic shooting to calculate the distance between two mirrored distributions.

μ	$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$	Time Systems (s)	Time G.Shooting (s)
1	2.77395	0.046875	4.70313
2	3.67027	0.046875	5.60938
3	4.52933	0.0625	7.10938
4	5.26093	0.078125	9.17188
5	5.87480	0.046875	12.5313
6	6.39439	0.0625	18.4219
7	6.84043	0.078125	492.563
8	7.22903	0.0625	574.422
9	7.57221	0.046875	917.859
10	7.87896	0.046875	1007.13

The method proposed here uses a non-linear system for the calculus of the Fisher–Rao distance, so it is faster the geodesic shooting algorithm. Furthermore, we remark that for $\mu \geq 7$, the geodesic shooting requires additional adaptation to convergence.

Next, we generalize the results of Proposition 4 to pairs of general multivariate normal mirrored distributions. Without loss of generality, we may assume:

$$\boldsymbol{\theta}_1 = (\mu_1 \boldsymbol{e}_1, \Sigma_1) \quad \text{and} \quad \boldsymbol{\theta}_2 = (\mu_2 \boldsymbol{e}_1, \Sigma_2), \tag{60}$$

with $\Sigma_2 = M_1 \Sigma_1 M_1$ as in (39), that is:

$$\Sigma_2 = \begin{cases} \hat{\sigma}_{1j} = \hat{\sigma}_{j1} = -\sigma_{1j}, & j = 2, \ldots, n \\ \hat{\sigma}_{ij} = \sigma_{ij}, & \text{otherwise.} \end{cases}.$$

Proposition 5. *The Fisher–Rao distance between a pair of multivariate mirrored normal distributions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ (65) is:*

$$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 2\sqrt{\sum_{l=1}^{n-1} b_l^2 + x^2}, \tag{61}$$

where:

$$(\boldsymbol{x}, B) = \left(\begin{pmatrix} x \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & b_1 & \cdots & b_{n-1} \\ b_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{n-1} & 0 & \cdots & 0 \end{pmatrix} \right). \tag{62}$$

The values x and b_l, the non-zero entries of $(\boldsymbol{x}, B)$, are obtained by the solution of the $\left(n + \frac{n(n+1)}{2}\right)$ order non-linear system:

$$\begin{cases} L^{-1}\Delta(1)L^{-1} = \Delta_2 \\ L^{-1}\boldsymbol{\delta}(1) + \Delta_2 \boldsymbol{\mu}_{1/2} = \boldsymbol{\delta}_2 \end{cases}, \tag{63}$$

where $\boldsymbol{\mu}_{1/2} = \left(\frac{\mu_1+\mu_2}{2}, \eta_1, \ldots, \eta_{n-1}\right)^t$, L is the Cholesky factor of the matrix $\Sigma_{1/2} = \begin{pmatrix} d_{11} & \mathbf{0}^t \\ \mathbf{0} & D \end{pmatrix}$, with D a symmetric $(n-1)$ order matrix, $(\boldsymbol{\delta}_2, \Delta_2) = \varphi(\boldsymbol{\mu}_2, \Sigma_2)$, and $(\boldsymbol{\delta}(t), \Delta(t))$ is the geodesic curve with initial value $(\boldsymbol{\delta}(0), \Delta(0)) = \boldsymbol{\theta}_0$ and tangent vector $(\boldsymbol{x}, B)$ given in Equation (15).

Let $\gamma(t) = (\boldsymbol{\mu}(t), \Sigma(t))$, $-1 \leq t \leq 1$, be the geodesic curve in $\mathcal{M}$ connecting $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The proof is similar to the bivariate case, by considering $\gamma(0) = (\boldsymbol{\mu}_{1/2}, \Sigma_{1/2})$,

$$\gamma'(0) = \hat{\boldsymbol{\theta}}_{1/2} = (\hat{\boldsymbol{\mu}}_{1/2}, \hat{\Sigma}_{1/2})) = \left(\begin{pmatrix} \hat{\mu} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1n} \\ \hat{\sigma}_{12} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{1n} & 0 & \cdots & 0 \end{pmatrix} \right) \tag{64}$$

and $\bar{\gamma}(t) = \psi(\gamma(t))$ where $\psi = \psi_{\left(-L^{-1}\mu_{1/2}L^{-1}\right)}$.

$$\boldsymbol{\theta}_1 = (\mu_1 \boldsymbol{e}_1, \Sigma_1) \quad \text{and} \quad \boldsymbol{\theta}_2 = (\mu_2 \boldsymbol{e}_1, \Sigma_2), \tag{65}$$

with $\Sigma_2 = M_1 \Sigma_1 M_1$.

Table 3 collects the results in Section 2.1 and the new results of this section.

Table 3. Closed forms for the Fisher–Rao distance in submanifolds of $\mathcal{M}$ and the distance in $\mathcal{M}$ between pairs of special distributions.

Submanifold	Distance
Distance in Non-totally Geodesic Submanifolds	
$\mathcal{M}_\Sigma = \left\{ \begin{array}{c} p_{\boldsymbol{\theta}};\ \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma), \\ \Sigma = \Sigma_0 \in P_n(\mathbb{R}) \text{ constant} \end{array} \right\}$, $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_0)$	$d_\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma_0^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$
$\mathcal{M}_D = \left\{ \begin{array}{c} p_{\boldsymbol{\theta}};\ \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma), \\ \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) \end{array} \right\}$, $\boldsymbol{\theta}_i = (\mu_{1i}, \sigma_{1i}, \mu_{2i}, \sigma_{2i}, \ldots, \mu_{ni}, \sigma_{ni})$	$d_D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\sum_{i=1}^{n} d_F^2((\mu_{1i}, \sigma_{1i}), (\mu_{2i}, \sigma_{2i}))}$
Distance in Totally Geodesic Submanifolds	
$\mathcal{M}_{\boldsymbol{\mu}} = \left\{ \begin{array}{c} p_{\boldsymbol{\theta}};\ \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma), \\ \boldsymbol{\mu} = \boldsymbol{\mu}_0 \in \mathbb{R}^n \text{ constant} \end{array} \right\}$, $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_0, \Sigma_i)$	$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\frac{1}{2} \sum_{i=1}^{n} [\log(\lambda_i)]^2}$, where λ_i are the eigenvalues of $\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}$
$\mathcal{M}_{D\boldsymbol{\mu}} = \left\{ \begin{array}{c} \{p_{\boldsymbol{\theta}};\ \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma), \\ \boldsymbol{\mu} \text{ is an eigenvector of} \\ \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)\} \end{array} \right\}$, $\boldsymbol{\theta}_i = (\mu_{1i}, \sigma_{1i}, \sigma_{2i}, \ldots, \sigma_{ni})$	$d_{D\boldsymbol{\mu}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \left(d_F^2((\mu_{11}, \sigma_{11}), (\mu_{21}, \sigma_{21})) + \sum_{i=2}^{n} d_F^2((0, \sigma_{1i}), (0, \sigma_{2i})) \right)^{1/2}$
Distance Between Special Distributions in $\mathcal{M}$	
Distributions with Common Covariance Matrices, $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_0)$	$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d_{D\boldsymbol{\mu}}((\mathbf{0}, D), (\lvert\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\rvert \mathbf{e}_1, D))$, where P is an orthogonal matrix such that $P(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \lvert\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\rvert \mathbf{e}_1$ and $P \Sigma P^t = U D U^t$
Mirrored Distributions, $\boldsymbol{\theta}_1 = (\mu_1 \mathbf{e}_1, \Sigma_1)$ and $\boldsymbol{\theta}_2 = (\mu_2 \mathbf{e}_1, \Sigma_2)$, with $\Sigma_2 = M_1 \Sigma_1 M_1$	$d_F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 2 \sqrt{\sum_{l=1}^{n-1} b_l^2 + x^2}$, where x and b_i are obtained by the solution of Equation (63)

4. Hierarchical Clustering for Diagonal Gaussian Mixture Simplification

A parameterized Gaussian mixture model f is a weighted sum of m multivariate normal distributions, that is,

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} w_i p_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i),$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $p_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, \cdots, m$, are normal distributions and w_i, $i = 1, \cdots, m$, are mixture, $\sum_{i=1}^{m} w_i = 1$. In this paper, we call the diagonal Gaussian mixture model (DGMM) the mixture composed only by distributions with diagonal covariance matrices.

Gaussian mixture models (GMM) are used in modeling datasets: image processing, signal processing, and density estimation problems [46–48]. In many applications involving mixture models, the computational requirements are of a very high level due to the large number of mixture components. This can be handled if we reduce the number of components of the mixture: given a mixture f of m components, we want to find a mixture g of l components, $1 \leq l < m$, such that g is a good approximation of f with respect to a similarity measure [49]. Gaussian mixture simplification was considered in statistical inference in [50] and to decode low-density lattice codes [51].

In [49] was proposed a hierarchical clustering algorithm to simplify an exponential family mixture model based on Bregman divergences. This section describes an agglomerative hierarchical clustering

method based on the Fisher–Rao distance in the submanifold $\mathcal{M}_D$ (21) to simplify DGMM, and we present an application to image segmentation, complementing what was developed in [52]. We start by introducing the concept of the centroid for a set of distributions in $\mathcal{M}_D$.

4.1. Centroids in the Submanifold $\mathcal{M}_D$

In [53], Galperin described centroids in the two-dimensional Minkowski model, which can be translated also to the Klein disk and Poincare half-plane models. Given a set of points $\boldsymbol{q}_i = (x_{\boldsymbol{q}_i}, y_{\boldsymbol{q}_i}, z_{\boldsymbol{q}_i})$ in the Minkowski model, with associated weights u_i, the centroid is computed and normalized as:

$$\boldsymbol{c}' = \sum_i u_i \boldsymbol{q}_i \quad \text{and} \quad \boldsymbol{c} = \frac{\boldsymbol{c}'}{-x_{\boldsymbol{c}'}^2 - y_{\boldsymbol{c}'}^2 + z_{\boldsymbol{c}'}^2}. \tag{66}$$

To calculate the centroid $\boldsymbol{c}$ of a subset of points $\mathcal{C} = \{(w_i, \boldsymbol{\theta}_i)\}$, $\boldsymbol{\theta}_i = (\mu_i, \sigma_i)$, the isometries presented in [25] and the relation between the media $\times$ standard deviation plane of parameters of univariate normal distributions and the Poincare half-plane given in [22] are used.

Given a dataset $\mathcal{C} = \{(w_i, \boldsymbol{\theta}_i)\}$, where $\boldsymbol{\theta}_i = (\mu_{1i}, \sigma_{1i}, \cdots, \mu_{ni}, \sigma_{ni})$ are distributions in $\mathcal{M}_D$, the centroid of $\mathcal{C}$ is:

$$\boldsymbol{c} := (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_n), \tag{67}$$

where $\boldsymbol{c}_j$, $j = 1, \cdots, n$, is the centroid of $\mathcal{C}_j = \{(w_j, (\mu_{ji}, \sigma_{ji}))\}$ given in Equation (66).

4.2. Hierarchical Clustering Algorithm

Let a DGMM f with parameters $\mathcal{C} = \{(w_1, \boldsymbol{\theta}_1), \cdots, (w_m, \boldsymbol{\theta}_m)\}$.

In order to apply the hierarchical clustering algorithm, we need to consider the distance between two subsets A and B. The three most common distances are called linkage criteria and are given by [54]:

- Single linkage:
$$D(A, B) = \min\{d_D(a, b); a \in A, b \in B\}; \tag{68}$$
- Complete linkage:
$$D(A, B) = \max\{d_D(a, b); a \in A, b \in B\}; \tag{69}$$
- Group average linkage:
$$D(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d_D(a, b), \tag{70}$$

where d_D is the distance in the submanifold $\mathcal{M}_D$ and $|X|$ is the number of elements of a set X.

A summary of the hierarchical clustering algorithm (Algorithm 1) [49] using one of these distances is given next.

Algorithm 1: Hierarchical Clustering Algorithm

1: Form m clusters $\mathcal{C}_j = \{(w_j, \boldsymbol{\theta}_j)\}$ with one element.
2: Find the two closest clusters, $\mathcal{C}_i$ and $\mathcal{C}_j$, with respect to a distance D, and merge them into a single cluster $\mathcal{C}_i \cup \mathcal{C}_j$.
3: Compute distances between the new cluster and each of the old clusters.
4: Repeat Steps 2 and 3 until all items are clustered into a single cluster of size n.

The simplified DGMM:

$$g = \sum_{j=1}^{l} \beta_j g_j$$

of l components is built from the l subsets $\mathcal{C}_1, \ldots, \mathcal{C}_l$ remaining after the iteration $n - l$ of the hierarchical clustering algorithm. In this work, we choose the parameters of g_j in two ways: as the centroid in the submanifold $\mathcal{M}_D$ (Fisher–Rao hierarchical clustering) and as the Bregman left-sided centroid [49] (Bregman–Fisher–Rao hierarchical clustering) of the subset $\mathcal{C}_j$ with weights $\beta_j = \sum_{(w_i, \boldsymbol{\theta}_i) \in \mathcal{C}_j} w_i$.

As remarked in [49], the hierarchical clustering algorithm allows introducing a method to learn the optimal number of components in the simplified mixture g. Thus, g must be as compact as possible and reach a minimum prescribed quality $d_{KL}(f||g) \leq \tau$, where $d_{KL}(f||g)$ is the Kullback–Leibler divergence.

4.3. Experiments in Image Segmentation

We can apply the Fisher–Rao and the Bregman–Fisher–Rao hierarchical clusterings to simplify a mixture of exponential families in the context of clustering-based image segmentation as was done in [49] for the Bregman hierarchical clustering. Given an input color image I, we adapt the Bregman soft clustering algorithm to generate a DGMM f of 32 components, which models the image pixels. We point out that the restriction considered in this paper (only DGMM) is also used in many applications due its much lower computational cost. We consider here a pixel $\boldsymbol{\rho} = (\rho_R, \rho_G, \rho_B)$ as a point in $\mathbb{R}^3$, where ρ_R, ρ_G, and ρ_B are the RGB color information. For image segmentation, we can say that the image pixel ρ belongs to the class $\mathcal{C}_j$ when:

$$p_j(\boldsymbol{\rho}; \mu_j, \Sigma_j) > p_i(\boldsymbol{\rho}; \mu_i, \Sigma_i), \ \forall i \in \{1, \cdots, m\} \setminus \{j\}.$$

Thus, the segmented image is illustrated by replacing the color value of the pixel $\boldsymbol{\rho}$ by the mean μ_j of the Gaussian p_j.

Using the the Fisher–Rao and the Bregman–Fisher–Rao hierarchical clusterings, we simplify the mixture f into mixtures g of l components with $l = \{2, 4, 8, 16\}$. Each mixture gives one image segmentation. The linkage criterion used here was the complete linkage (68), which has presented better results in our simulations. Figure 8 shows the segmentation of the Baboon, Lena, and Clown input images given by the Bregman–Fisher–Rao hierarchical clustering. The number of colors in each image is equal to the number of components in the simplified mixture g.

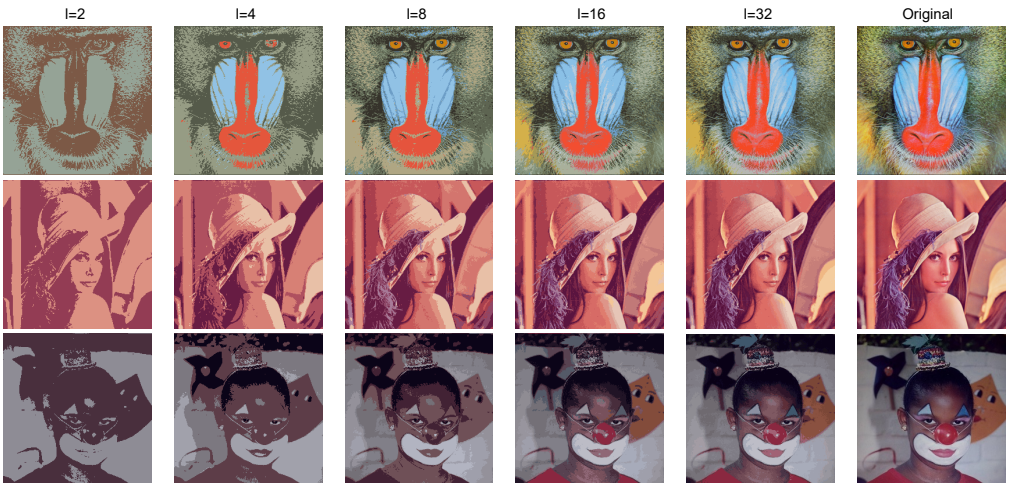


Figure 8. Illustration of the mixture simplification using the Fisher–Rao clustering, where l is the number of components of the mixture (the last column is the original figure).

The quality of the segmentation was analyzed as a function of l through the Kullback–Leibler divergence estimated by the Monte Carlo method, since there was no closed form for this measure (five thousand points were randomly drawn to estimate $d_{KL}(f||g)$). Figures 9–11 show the evolution of the simplification quality as a function of the number of components l for the Baboon, Lena, and Clown

images, using the Bregman, the Fisher–Rao, and the Bregman–Fisher–Rao hierarchical clustering algorithms. We observed that the image quality increased ($d_{KL}(f||g)$ decreased) with l, as expected, and the behavior was similar in all clustering algorithms. In general, the Bregman–Fisher–Rao hierarchical clustering algorithm presented better results. Considering the constraint $\tau = 0.2$, the learning process provided, for the Bregman–Fisher–Rao hierarchical clustering, mixtures of 19, 21, and 21 as optimal simplifications for the images of the Baboon, Lena, and Clown, respectively.

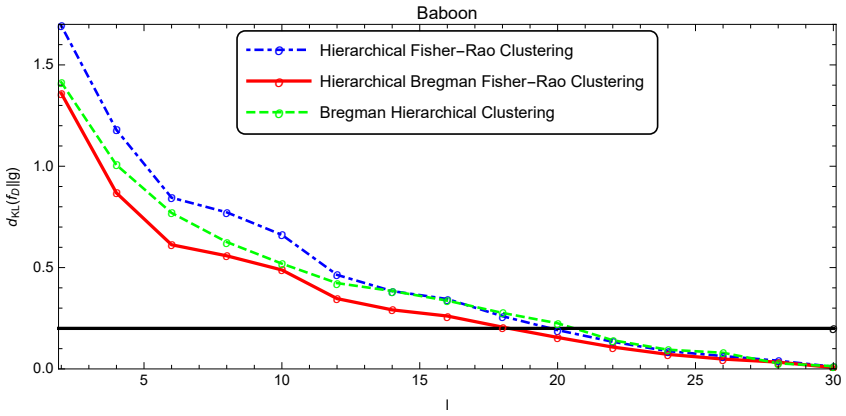


Figure 9. Illustration of the simplification quality of the mixture modeling Baboon image.

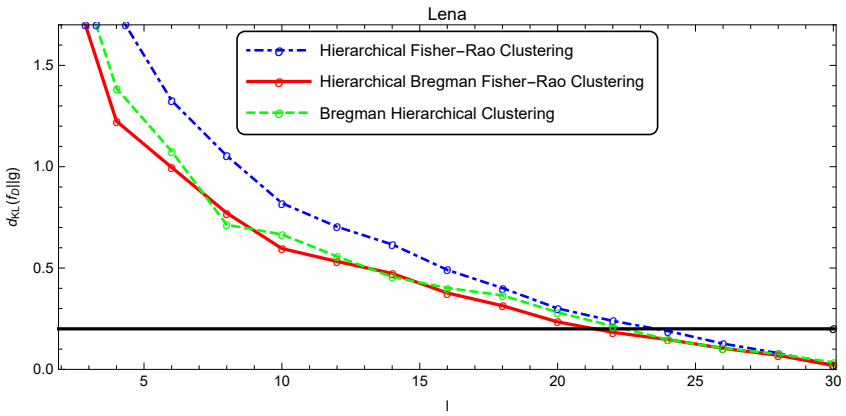


Figure 10. Illustration of the simplification quality of the mixture modeling Lena image.

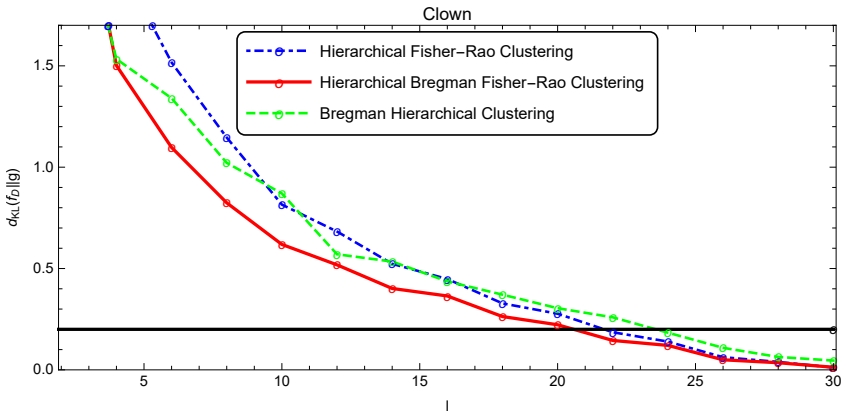


Figure 11. Illustration of the simplification quality of the mixture modeling Clown image.

5. Concluding Remarks

The Fisher–Rao distance was approached here in the space of multivariate normal distributions. Initially, as in [38], we summarized some known closed forms for this distance in submanifolds of this model and some bounds for the general case. A closed form for the Fisher–Rao distance between distributions with the same covariance matrix was obtained in Proposition 3, and we also have derived a non-linear system characterizing the distance between two distributions with mirrored covariance matrices in Proposition 5. Some perspectives for future research related to this topic include deriving new bounds for the Fisher–Rao distance in the general case, by using these special distributions, to characterize as non-linear systems the distances between other types of distributions and to extend the closed forms and bounds presented here to the space of elliptical distributions. Finally, we have extended the analysis of the Bregman–Fisher–Rao hierarchical clustering algorithm to simplify Gaussian mixtures in the context of clustering-based image segmentation given in [52] with comparative results that encourage the use of the Fisher–Rao distance in other clustering or classification algorithms.

Author Contributions: All authors contributed equally to the research and the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgments: The authors are thankful to the referees, as their comments and suggestions have contributed to improve the presentation of the text. The authors were partially supported by grants FAPESP (13/25977-7) and CNPq (313326/2017-7) foundations.

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

References

1. Calin, O.; Udriste, C. Geometric Modeling in Probability and Statistics. In *Mathematics and Statistics*; Springer International: Cham, Switzerland, 2014.
2. Nielsen, F. An elementary introduction to information geometry. *arXiv* **2018**, arXiv:1808.08271.
3. Amari, S.; Nagaoka, H. Methods of Information Geometry. In *Translations of Mathematical Monographs*; Oxford University Press: Oxford, UK, 2000; Volume 191.
4. Amari, S. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016.
5. Ay, N.; Jost, J.; Vân Lê, H.; Schwachhöfer, L. Information geometry and sufficient statistics. *Probab. Theory Relat. Fields* **2015**, *162*, 327–364. [CrossRef]
6. Vân Lê, H. The uniqueness of the Fisher metric as information metric. *Ann. Inst. Stat. Math.* **2017**, *69*, 879–896.
7. Gibilisco, P.; Riccomagno, E.; Rogantin, M.P.; Wynn, H.P. *Algebraic and Geometric Methods in Statistics;* Cambridge University Press: New York, NY, USA, 2010.
8. Chentsov, N.N. *Statistical Decision Rules and Optimal Inference*; AMS Bookstore: Providence, RI, USA, 1982; Volume 53.
9. Campbell, L.L. An extended Cencov characterization of the information metric. *Proc. Am. Math. Soc.* **1986**, *98*, 135–141.
10. Vân Lê, H. Statistical manifolds are statistical models. *J. Geom.* **2006**, *84*, 83–93.
11. Mahalanobis, P.C. On the generalized distance in statistics. *Proc. Natl. Inst. Sci.* **1936**, *2*, 49–55.
12. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–110.
13. Hotelling, H. Spaces of statistical parameters. *Bull. Am. Math. Soc. (AMS)* **1930**, *36*, 191.
14. Rao, C.R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
15. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond.* **1921**, *222*, 309–368.
16. Burbea, J. Informative geometry of probability spaces. *Expo. Math.* **1986**, *4*, 347–378.
17. Skovgaard, L.T. A Riemannian geometry of the multivariate normal model. *Scand. J. Stat.* **1984**, *11*, 211–223.
18. Atkinson, C.; Mitchell, A.F.S. Rao's Distance Measure. *Sankhyã Indian J. Stat.* **1981**, *43*, 345–365.

19. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
20. Villani, C. Optimal Transport, Old and New. In *Grundlehren der Mathematischen Wissenschaften*; Springer: Berlin/Heidelberg, Germany, 2009.
21. Amari, S. *Differential Geometrical Methods in Statistics*; Springer: Berlin, Germany, 1985.
22. Costa, S.I.R.; Santos, S.A.; Strapasson, J.E. Fisher information distance: A geometrical reading. *Discret. Appl. Math.* **2015**, *197*, 59–69. [CrossRef]
23. Angulo, J.; Velasco-Forero, S. Morphological processing of univariate Gaussian distribution-valued images based on Poincaré upper-half plane representation. In *Geometric Theory of Information*; Springer International Publishing: Cham, Switzerland, 2014; pp. 331–366.
24. Maybank, S.J.; Ieng, S.; Benosman, R. A Fisher–Rao metric for paracatadioptric images of lines. *Int. J. Comput. Vis.* **2012**, *99*, 147–165. [CrossRef]
25. Schwander, O.; Nielsen, F. Model centroids for the simplification of kernel density estimators. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012.
26. Taylor, S. Clustering Financial Return Distributions Using the Fisher Information Metric. *Entropy* **2019**, *21*, 110. [CrossRef]
27. Eriksen, P.S. *Geodesics Connected with the Fischer Metric on the Multivariate Normal Manifold*; Institute of Electronic Systems, Aalborg University Centre: Aalborg, Denmark, 1986.
28. Calvo, M.; Oller, J.M. An explicit solution of information geodesic equations for the multivariate normal model. *Stat. Decis.* **1991**, *9*, 119–138. [CrossRef]
29. Lenglet, C.; Rousson, M.; Deriche, R.; Faugeras, O. Statistics on the manifold of multivariate normal distributions. Theory and application to diffusion tensor MRI processing. *J. Math. Imaging Vis.* **2006**, *25*, 423–444. [CrossRef]
30. Moakher, M.; Mourad, Z. 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]
31. 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]
32. Verdoolaege, G.; Scheunders, P. Geodesics on the manifold of multivariate generalized Gaussian distributions with an application to multicomponent texture discrimination. *Int. J. Comput. Vis.* **2011**, *95*, 265. [CrossRef]
33. 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]
34. Poon, C.; Keriven, N.; Peyré, G. Support Localization and the Fisher Metric for off-the-grid Sparse Regularization. *arXiv* **2018**, arXiv:1810.03340.
35. Gattone, S.A.; De Sanctis, A.; Puechmorel, S.; Nicol, F. On the geodesic distance in shapes K-means clustering. *Entropy* **2018**, *20*, 647. [CrossRef]
36. Gattone, S.A.; De Sanctis, A.; Russo, T.; Pulcini, D. A shape distance based on the Fisher–Rao metric and its application for shapes clustering. *Phys. A Stat. Mech. Appl.* **2017**, **487**, 93–102. [CrossRef]
37. 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.
38. Pinele, J.; Costa, S.I.; Strapasson, J.E. On the Fisher–Rao Information Metric in the Space of Normal Distributions. In *International Conference on Geometric Science of Information*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; pp. 676–684.
39. Burbea, J.; Rao, C.R. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multivar. Anal.* **1982**, *12*, 575–596. [CrossRef]
40. Porat, B.; Benjamin, F. Computation of the exact information matrix of Gaussian time series with stationary random components. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 118–130. [CrossRef]
41. Siegel, C.L. Symplectic geometry. *Am. J. Math.* **1943**, *65*, 1–86. [CrossRef]
42. Strapasson, J.E.; Pinele, J.; Costa, S.I.R. 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, 11–14 September 2016; pp. 61–65.
43. 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]

44. 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]
45. Strapasson, J.E.; Porto, J.; Costa, S.I.R. On bounds for the Fisher–Rao distance between multivariate normal distributions. *Aip Conf. Proc.* **2015**, *1641*, 313–320.
46. Zhang, K.; Kwok, J.T. Simplifying mixture models through function approximation. *IEEE Trans. Neural Netw.* **2010**, *21*, 644–658. [CrossRef] [PubMed]
47. Davis, J.V.; Dhillon, I.S. Differential entropic clustering of multivariate gaussians. In Proceedings of the 2006 Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006.
48. Goldberger, J.; Greenspan, H.K.; Dreyfuss, J. Simplifying mixture models using the unscented transform. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1496–1502. [CrossRef]
49. Garcia, V.; Nielsen, F. Simplification and hierarchical representations of mixtures of exponential families. *Signal Process.* **2010**, *90*, 3197–3212. [CrossRef]
50. Bar-Shalom, Y.; Li, X. *Estimation and Tracking: Principles, Techniques and Software*; Artech House: Norwood, MA, USA, 1993.
51. Kurkoski, B.; Dauwels, J. Message-passing decoding of lattices using Gaussian mixtures. In Proceedings of the 2008 IEEE International Symposium on Information Theory, Toronto, ON, Canada, 6–11 July 2008.
52. Strapasson, J.E.; Pinele, J.; Costa, S.I.R. Clustering using the Fisher–Rao distance. In Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop, Rio de Janerio, Brazil, 10–13 July 2016.
53. Galperin, G.A. A concept of the mass center of a system of material points in the constant curvature spaces. *Commun. Math. Phys.* **1993**, *154.1*, 63–84. [CrossRef]
54. Nielsen, F. Introduction to HPC with MPI for Data Science. In *Undergraduate Topics in Computer Science*; Springer: Cham, Switzerland, 2016.