HAL open science

The Fisher-Rao geometry of CES distributions

Florent Bouchard, Arnaud Breloy, Antoine Collas, Alexandre Renaux, Guillaume Ginolhac

To cite this version:

Florent Bouchard, Arnaud Breloy, Antoine Collas, Alexandre Renaux, Guillaume Ginolhac. The Fisher-Rao geometry of CES distributions. Springer. Elliptical Distributions in Signal Processing, In press. hal-04225646

HAL Id: hal-04225646

https://hal.science/hal-04225646

Submitted on 3 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Fisher-Rao geometry of CES distributions

Florent Bouchard, Arnaud Breloy, Antoine Collas, Alexandre Renaux, Guillaume Ginolhac

Abstract When dealing with a parametric statistical model, a Riemannian manifold can naturally appear by endowing the parameter space with the Fisher information metric. The geometry induced on the parameters by this metric is then referred to as the Fisher-Rao information geometry. Interestingly, this yields a point of view that allows for leveraging many tools from differential geometry. After a brief introduction about these concepts, we will present some practical uses of these geometric tools in the framework of elliptical distributions. This second part of the exposition is divided into three main axes: Riemannian optimization for covariance matrix estimation, Intrinsic Cramér-Rao bounds, and classification using Riemannian distances.

Florent Bouchard
Université Paris-Saclay, CNRS, CentraleSupélec, L2S,
e-mail: florent.bouchard@centralesupelec.fr

Arnaud Breloy
LEME, Université Paris Nanterre,
e-mail: arnaud.breloy@parisnanterre.fr

Antoine Collas
Université Paris-Saclay, Inria, CEA,
e-mail: antoine.collas@inria.fr

Alexandre Renaux
Université Paris-Saclay, CNRS, CentraleSupélec, L2S,
e-mail: alexandre.renaux@centralesupelec.fr

Guillaume Ginolhac
LISTIC, Université Savoie Mont-Blanc,
e-mail: guillaume.ginolhac@univ-smb.fr

1 Introduction: from CES distributions to information geometry

This section starts with reminders on complex elliptically symmetric distributions (CES)[1]. This part is concluded by introducing the Fisher information matrix of this model, which acts as a transition point to information geometry. Indeed, the Fisher information matrix actually represents a metric that induces an inherent geometry for statistical models, which is referred to as the Fisher-Rao information geometry. In the specific case of CES, this will yield a particular geometry for the space of covariance matrices. After evidencing this transition point, this section concludes by outlining the rest of the chapter.

1.1 Reminders on CES distributions

Circular complex elliptically symmetric (C-CES) distributions [52] refer to a large family of multivariate distributions. Very comprehensive and detailed reviews on the topic can be found in the references [70, 73], and of course, the background chapter of this book. A vector $\mathbf{x} \in \mathbb{C}^p$ follows a centered (zero-mean) C-CES distribution, denoted $\mathbf{x} \sim \text{C-CES}(\mathbf{0}, \Sigma, g)$, if it admits the following stochastic representation:

$$\mathbf{x} =_d \sqrt{Q}\, \Sigma^{\frac{1}{2}}\, \mathbf{u}, \tag{1}$$

where:

- The notation $=_d$ means that random variables on both sides have the same cumulative distribution function.
- The vector $\mathbf{u} \in \mathbb{C}^p$ follows a uniform distribution on the complex unit sphere $\mathbb{C}\mathcal{S}^p = \{\mathbf{u} \in \mathbb{C}^p \mid \|\mathbf{u}\| = 1\}$, denoted $\mathbf{u} \sim \mathcal{U}(\mathbb{C}\mathcal{S}^p)$.
- The scalar $Q \in \mathbb{R}^+$ is non-negative real random variable of probability density function f_Q, independent of $\mathbf{u}$, and called the second-order modular variate (while $\sqrt{Q}$ is called the modular variate).
- The matrix $\Sigma^{\frac{1}{2}} \in \mathbb{C}^{p \times p}$ is a factorization of the scatter matrix $\Sigma = \Sigma^{\frac{1}{2}} \Sigma^{\frac{H}{2}}$. If the covariance matrix of $\mathbf{x}$ exists, it is proportional to the scatter matrix, i.e., $\mathrm{E}\left[\mathbf{x}\mathbf{x}^H\right] \propto \Sigma$. If we then choose the normalization convention $\mathrm{E}[Q] = 1$, these two matrices are equal. Thus we will abusively refer to the scatter matrix Σ as the covariance matrix, as it is a more familiar terminology.

We focus only on the absolutely continuous case where the covariance matrix Σ is full rank (cf. Section 2.3 of the background chapter). In this case, the probability density function of $\mathbf{x}$ is given as

[1] Note that this chapter considers the case where the data and covariance matrix can be complex-valued for the sake of generality. However, we focus solely on the circular case (referred to as C-CES in the background chapter). Hence, most of the presented results can also be obtained in the real-valued case (RES) with proper adjustments.

$$f_{\mathbf{x}}(\mathbf{x}|\Sigma) \propto |\Sigma|^{-1} g\left(\mathbf{x}^H \Sigma^{-1} \mathbf{x}\right), \tag{2}$$

where the function $g : \mathbb{R}^+ \to \mathbb{R}^+$ is called the density generator. The density generator satisfies the finite moment condition $\delta_{p,g} = \int_0^\infty t^{p-1} g(t) dt < \infty$. This function g is directly related to the probability density function of the second-order modular variate by the relation

$$f_Q(Q) = \delta_{p,g}^{-1} Q^{p-1} g(Q). \tag{3}$$

Given a n-sample $\{\mathbf{x}_i\}_{i=1}^n$, assumed to be independent and identically distributed (iid) from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$, its log-likelihood is given as:

$$\mathcal{L}_g\left(\{\mathbf{x}_i\}_{i=1}^n | \Sigma\right) = \sum_{i=1}^n \log\left(g\left(\mathbf{x}_i^H \Sigma^{-1} \mathbf{x}_i\right)\right) - n \log|\Sigma|. \tag{4}$$

The C-CES model being defined, we move to the notion of information brought by the Fisher information matrix: intuitively, the more the sample set $\{\mathbf{x}_i\}_{i=1}^n$ depends on Σ, the more sampling from the likelihood (4) (increasing n) will reveal information about Σ. The score vector is a tool that will help in quantifying this notion of information: to define this quantity, we now consider a parameterization of the covariance matrix through a real-valued vector $\boldsymbol{\nu}$ of appropriate dimension[2], denoted $\Sigma(\boldsymbol{\nu})$. The score vector $\mathbf{s}$ is then defined entry-wise as

$$[\mathbf{s}]_j = \frac{\partial \mathcal{L}_g\left(\{\mathbf{x}_i\}_{i=1}^n | \Sigma(\boldsymbol{\nu})\right)}{\partial \nu_j}, \tag{5}$$

which therefore reflects the variation of the log-likelihood of the sample set $\{\mathbf{x}_i\}_{i=1}^n$ with respect to the parameter ν_j. Under mild regularity conditions (satisfied by $\mathcal{L}_g$ in our case), this vector has zero mean, i.e., $\mathrm{E}[\mathbf{s}] = \mathbf{0}$. However, its covariance matrix is a fundamental quantity referred to as the Fisher information matrix, denoted $\mathbf{F}$, and defined as:

$$[\mathbf{F}]_{j,k} = \mathrm{E}\left[\frac{\partial \mathcal{L}_g\left(\{\mathbf{x}_i\}_{i=1}^n | \Sigma(\boldsymbol{\nu})\right)}{\partial \nu_j} \frac{\partial \mathcal{L}_g\left(\{\mathbf{x}_i\}_{i=1}^n | \Sigma(\boldsymbol{\nu})\right)}{\partial \nu_k}\right]. \tag{6}$$

This matrix quantifies, on average, how much information about the vector $\boldsymbol{\nu}$ we can obtain from a sample set $\{\mathbf{x}_i\}_{i=1}^n$. In practice, the entries of the Fisher information matrix $\mathbf{F}$ for centered C-CES can be obtained thanks to Slepian-Bangs type formula from [16], also presented in the Section 6.5 of the background chapter. The latter

[2] In this chapter, Σ is not assumed to have a specific structure, so $\boldsymbol{\nu}$ is typically of dimension p^2 and stores the entries of the diagonal and upper triangle of the covariance matrix (where the coordinates are split in terms of real and imaginary part). However, the definition extends to any parameterization, e.g., from a choice of decomposition in the case of structured matrices [62].

is briefly recalled below using an alternate expression that is consistent whit the upcoming discussions:

Theorem 1 *(Fisher Information matrix of centered C-CES)*
Let $\Sigma \overset{def}{=} \Sigma(\boldsymbol{\nu})$ be a covariance matrix parameterized by the real-valued vector $\boldsymbol{\nu}$. Let $\{\mathbf{x}_i\}_{i=1}^n$ be a n-sample of iid from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$. The entries of the Fisher information matrix are

$$[\mathbf{F}]_{j,k} = n\alpha_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}_j\Sigma^{-1}\boldsymbol{\xi}_k) + n\beta_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}_j)\operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}_k), \tag{7}$$

with

$$\boldsymbol{\xi}_j = \frac{\partial \Sigma(\boldsymbol{\nu})}{\partial \nu_j}, \tag{8}$$

and where the coefficients α_g and β_g are defined by

$$\alpha_g = 1 - \frac{\mathbb{E}\left[Q^2\phi'(Q)\right]}{p(p+1)} \qquad \textit{and} \qquad \beta_g = \alpha_g - 1, \tag{9}$$

with $\phi(t) = g'(t)/g(t)$.

In the statistical signal processing community [55], the Fisher information matrix has been extensively leveraged thanks to the CramÃľr-Rao inequality:

$$\mathbb{E}\left[(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu})(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu})^\top\right] \succcurlyeq \mathbf{F}^{-1} \quad \Rightarrow \quad \|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_F^2 \geq \operatorname{Tr}(\mathbf{F}^{-1}), \tag{10}$$

that yields a lower bound for the mean squared error of any unbiased estimator $\hat{\boldsymbol{\nu}}$ of $\boldsymbol{\nu}$ (built from a set of observations $\{\mathbf{x}_i\}_{i=1}^n$). On the other hand, the seminal work of Rao [79, 80] also discusses using the Riemannian geometry of the parameter space when the Fisher information matrix is used as a metric tensor. The study of such spaces is now broadly referred to as the Fisher-Rao information geometry, which is introduced in the next section. Before this, we conclude this brief reminder by the example of multivariate (Student's) t distribution (also discussed with more details in Section 5.2 of the background chapter).

Example The t-distribution with $d \in \mathbb{N}^*$ degrees of freedom is obtained for the C-CES representation $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g_d)$ with

$$g_d(t) = \left(1 + \frac{t}{d}\right)^{-(d+p)}, \tag{11}$$

and the second-order modular variate is distributed as $Q =_d \mathbb{C}_{\chi_p^2} / \mathbb{C}_{\chi_d^2/d}$ where $\mathbb{C}_{\chi_x^2}$ denotes the Chi-squared distribution with x degrees of freedom. Hence Q follows a scaled $\mathcal{F}$-distribution. We have

$$\phi(t) = -\frac{d+p}{d+t}, \tag{12}$$

and the expectation

$$\mathbb{E}\left[Q^2\phi^2\left(Q\right)\right] = \frac{(d+p)p(p+1)}{d+p+1}, \tag{13}$$

that allows to obtain the coefficients

$$\alpha_g = \frac{d+p}{d+p+1} \qquad \text{and} \qquad \beta_g = \frac{-1}{d+p+1}. \tag{14}$$

for the Fisher information metric as in Theorem 2. The t-distribution also encompasses the well known multivariate Gaussian model $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ (of density generator $g_{\mathcal{N}}(t) = \exp(-t)$) as limit case when $d \to \infty$. The corresponding Fisher information metric coefficient are then $\alpha_g = 1$ and $\beta_g = 0$, which makes Theorem 1 coincide with the classical Slepian-Bangs formula [84, 10].

1.2 From the Fisher information matrix to information geometry

This section aims at linking notions of Riemannian geometry to the classical expression of the Fisher information matrix from Theorem 1. The goal is to shortly build the intuition on why the C-CES statistical model naturally induces a certain geometry for covariance matrices, while the corresponding framework will be presented in details in Section 2. We first need to re-interpret the expression of the Fisher information matrix from two key points:

- **Covariance matrices belong to the smooth manifold $\mathcal{H}_p^{++}$**

  The matrix $\Sigma \stackrel{\text{def}}{=} \Sigma(\boldsymbol{\nu})$ is a point in the space of covariance matrices, i.e., the set of $p \times p$ positive definite Hermitian matrices

  $$\mathcal{H}_p^{++} = \left\{\Sigma \in \mathcal{H}_p : \ \forall\, \mathbf{x} \in \mathbb{C}^p \backslash \{\mathbf{0}\},\ \mathbf{x}^H \Sigma \mathbf{x} > 0\right\}, \tag{15}$$

  where $\mathcal{H}_p$ denotes the set of $p \times p$ Hermitian matrices. As it is an open of the linear space $\mathcal{H}_p$, the space $\mathcal{H}_p^{++}$ is a smooth manifold. This means that it admits a differential structure, and notably, a tangent space at each point Σ, denoted $T_\Sigma \mathcal{H}_p^{++}$. For any point Σ, this tangent space $T_\Sigma \mathcal{H}_p^{++}$ turns out to be identifiable to be $\mathcal{H}_p$ (which again, comes from the fact that $\mathcal{H}_p^{++}$ is an open subspace of $\mathcal{H}_p$). An abstract representation of these spaces is presented in Figure 1.

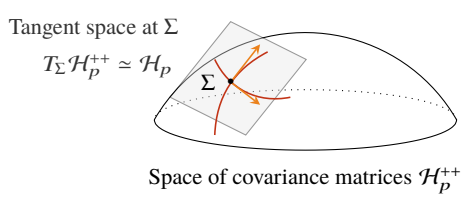

Fig. 1: Space of covariance matrices represented as a smooth manifold.

- **The Fisher information matrix represents an inner product on $T_\Sigma \mathcal{H}_p^{++}$**
  The entries of the Fisher information matrix of Theorem 1 can be compactly denoted as $[\mathbf{F}]_{j,k} = \langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_k \rangle_\Sigma^{\text{FIM}}$, whose expression is identified directly from (7). We then remark that the matrices $\boldsymbol{\xi}_j$ and $\boldsymbol{\xi}_k$ in (8) are, in fact, elements of $T_\Sigma \mathcal{H}_p^{++}$. The expression in (7) can thus be generalized to any pair of matrices $\boldsymbol{\xi}, \boldsymbol{\eta} \in T_\Sigma \mathcal{H}_p^{++}$, which results in a bi-linear form, denoted $\langle \cdot, \cdot \rangle_\Sigma^{\text{FIM}} : T_\Sigma \mathcal{H}_p^{++} \times T_\Sigma \mathcal{H}_p^{++} \to \mathbb{R}$. Because this bi-linear form is positive definite, it defines a metric, i.e., an inner product on the tangent space $T_\Sigma \mathcal{H}_p^{++}$. This inner product on $T_\Sigma \mathcal{H}_p^{++}$ is referred to as the Fisher information metric[3].

These two points being stated, we can last notice that the obtained Fisher information metric $\langle \cdot, \cdot \rangle_\Sigma^{\text{FIM}}$ varies smoothly with Σ. This enables the transition from statistical models to Riemannian geometry: the branch of differential geometry studying smooth manifolds endowed with smooth local inner products (referred to as Riemannian metrics). Such framework indeed applies to parametric statistical models, as it allows us to investigate the geometry of the parameter space equipped with the Fisher information metric. The resulting Riemannian geometry is generally referred to as the Fisher-Rao information geometry. Back to our central example, we have presented enough elements to explicit that the title of this chapter "The Fisher-Rao Geometry of CES distributions" more precisely stands short for "the Riemannian geometry of Hermitian positive definite matrices (covariance matrices) induced by the Fisher information metric of centered circular complex elliptically symmetric distributions", which will be studied in the next sections.

1.3 Outline of the chapter

The previous section showed why an inherent geometry of the parameter space can naturally result from a statistical model. Studying such geometry in detail requires introducing tools from the framework of Riemannian geometry, which is done in section 2. The C-CES distributions will be used as an example throughout the exposition. Hence, we will obtain most tools related to the Fisher-Rao Geometry of C-CES distributions: the Levi-Civita connection, the geodesics (and geodesic distance) between two covariance matrices, as well as the Riemannian exponential and logarithm mappings.

On a larger perspective, the second part of this chapter illustrates where tools obtained from the Fisher-Rao information geometry of C-CES can be leveraged within signal processing and machine learning tasks. In details:

- Section 3 addresses covariance matrix estimation problems, i.e., given a sample set $\{\mathbf{x}_i\}_{i=1}^n$, we infer Σ to perform a task (covariance analysis, filtering, metric learning, etc.). In this setup, we illustrate how the concepts of geodesic convexity

[3] Note that the Fisher information *matrix* being obtained as $[\mathbf{F}]_{j,k} = \langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_k \rangle_\Sigma^{\text{FIM}}$, it is actually a matrix representation (metric tensor) of the Fisher information *metric* $\langle \cdot, \cdot \rangle_\Sigma^{\text{FIM}}$ when the set $\{\boldsymbol{\xi}_j\}$ is chosen as a basis of coordinates for the tangent space $T_\Sigma \mathcal{H}_p^{++}$.

and Riemannian optimization can be helpful in problems related to covariance matrix estimation.
- Still related to covariance matrix estimation problems, Section 4 presents how the statistical performance of an estimator can be evaluated with intrinsic CramÃľr-Rao lower bounds, which generalize the standard CramÃľr-Rao inequality for parameters that lie in a manifold.
- Section 5 discusses how the Fisher-Rao geometry of C-CES provides a measure between the distributions of samples that can be leveraged in classification methodologies. This framework is then applied to Electroencephalography (EEG) signals.

2 An introduction to Riemannian geometry through the Fisher-Rao geometry of CES distributions

This section provides a short introduction to the concepts and tools of Riemannian geometry, while using the Fisher-Rao geometry of C-CES distributions as the main directive example. Some elementary notions are also assumed to be known for the sake of conciseness (e.g., basics matrix differentiation). For more detailed coverages of differential geometry, one can refer to the standard textbooks on the topic [41, 58, 59]. The notations and definitions of this section are mostly inspired from the books [1, 26], which provide very good (optimization-oriented) entry points to smooth manifolds and Riemannian geometry. The Fisher-Rao geometries of multivariate Gaussian and CES models have been studied in, e.g., [8, 15, 28, 63, 64, 85, 83].

2.1 $\mathcal{H}_p^{++}$ as a Riemannian manifold

The set of $p \times p$ Hermitian positive definite matrices $\mathcal{H}_p^{++}$ is an open subspace of the space of $p \times p$ Hermitian matrices $\mathcal{H}_p$. Since $\mathcal{H}_p^{++}$ has the same dimension as its embedding space $\mathcal{H}_p$, it is a smooth manifold of dimension p^2 [26, Definition 3.10]. As every smooth manifolds, $\mathcal{H}_p^{++}$ admits a differential structure, *i.e.*, every point $\Sigma \in \mathcal{H}_p^{++}$ possesses a tangent space $T_\Sigma \mathcal{H}_p^{++}$. The elements of $T_\Sigma \mathcal{H}_p^{++}$ are called tangent vectors, and correspond to the directional derivatives of curves in $\mathcal{H}_p^{++}$ passing through Σ (cf. Figure 1 for an illustration). Since $\mathcal{H}_p^{++}$ is open in $\mathcal{H}_p$, the tangent space $T_\Sigma \mathcal{H}_p^{++}$ at every point Σ can be identified as $\mathcal{H}_p$ [26, Theorem 3.15]. An illustration of the 1-dimensional case $\mathcal{H}_1^{++} = \mathbb{R}_*^+$ is presented in Figure 2.

Remark The space $\mathcal{H}_p^{++}$ is often referred to as the *convex cone* of positive definite matrices. It is indeed a cone because $\Sigma \in \mathcal{H}_p^{++}$ implies that $a\Sigma \in \mathcal{H}_p^{++}$, $\forall\, a \in \mathbb{R}_+^*$. It is furthermore a convex cone because any linear combination $a\Sigma_1 + b\Sigma_2$ is also in $\mathcal{H}_p^{++}$, $\forall\, \Sigma_1, \Sigma_2 \in \mathcal{H}_p^{++}$ and $\forall\, a, b \in \mathbb{R}_+^*$. This cone visually appears in the real 2×2 case of $\mathcal{S}_2^{++}$, which is often used to represent $\mathcal{H}_p^{++}$. Still, this chapter will rely

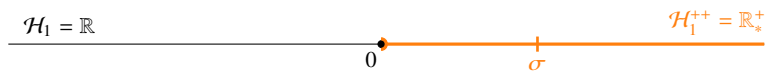

Fig. 2: Illustration of the manifold $\mathcal{H}_1^{++} = \mathbb{R}_*^+$, which is open in $\mathcal{H}_1 = \mathbb{R}$. We observe that the tangent space at every point $\sigma > 0$ simply corresponds to $\mathcal{H}_1 = \mathbb{R}$.

on the representation of Figure 1, which is more convenient to illustrate the generic concepts and tools of Riemannian geometry.

In order to further harness the differential structure of $\mathcal{H}_p^{++}$, we endow it with a Riemannian metric. This consists in a mapping that equips every tangent space $T_\Sigma \mathcal{H}_p^{++}$ with an inner product[4] $\langle \cdot, \cdot \rangle_\Sigma$ that varies smoothly with respect to the point Σ. This allows notably for locally defining the notion of angle and length for vectors in $T_\Sigma \mathcal{H}_p^{++}$. A smooth manifold equipped with such Riemmanian metric is then referred to as a Riemannian manifold. Notice that the definition of the metric is a choice that induces a corresponding geometry. In particular, if $\mathcal{H}_p^{++}$ is endowed with the Euclidean metric $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_\Sigma^{\mathcal{E}} = \Re\mathfrak{e}(\mathrm{Tr}(\boldsymbol{\xi}\boldsymbol{\eta}))$, where $\Re\mathfrak{e}(\cdot)$ returns the real part of its argument, all the geometrical objects of the manifold $\mathcal{H}_p^{++}$ are exactly the same as those of the space $\mathcal{H}_p$. In this case, there is no distinction between $\mathcal{H}_p^{++}$ and $\mathcal{H}_p$ from a geometrical point of view, and the true structure of $\mathcal{H}_p^{++}$ cannot be exploited. This motivates the use of other metrics, that induce a more meaningful geometry on $\mathcal{H}_p^{++}$ (e.g., ensuring that the boundaries of the space are not reachable). In this scope, various options have been considered, such as the affine invariant metric [17, 65], the log-Euclidean metric [7], or the Bures-Wasserstein one [18, 43]. Overviews of the different metrics and their corresponding geometries can be found in [92, 93].

When dealing with a statistical model the Fisher information metric is generally to be favored, as it is naturally suited to the underlying geometry of the data. Without resorting to the tedious parameterization and identification of Section 1.2, a general expression of this metric can directly be obtained following [85, Theorem 1] as:

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_\Sigma^{\mathrm{FIM}} = \mathrm{E}\left[\mathrm{D}\,\mathcal{L}_g(\mathbf{x}|\Sigma)[\boldsymbol{\xi}] \cdot \mathrm{D}\,\mathcal{L}_g(\mathbf{x}|\Sigma)[\boldsymbol{\eta}]\right] = -\mathrm{E}\left[\mathrm{D}^2\,\mathcal{L}_g(\mathbf{x}|\Sigma)[\boldsymbol{\xi}, \boldsymbol{\eta}]\right], \quad (16)$$

where $\mathrm{D}\,\mathcal{L}_g$ and $\mathrm{D}^2\,\mathcal{L}_g$ are the first and second order directional derivatives of the log-likelihood $\mathcal{L}_g$ of the distribution with respect to Σ. Recall from [44] that the first and second derivatives of a function $L : \mathcal{H}_p^{++} \to \mathbb{R}$ at $\Sigma \in \mathcal{H}_p^{++}$ in directions $\boldsymbol{\xi}$ and $\boldsymbol{\eta} \in T_\Sigma \mathcal{H}_p^{++}$ are defined as

$$\begin{aligned} \mathrm{D}\,L(\Sigma)[\boldsymbol{\xi}] &= L(\Sigma + \boldsymbol{\xi}) - L(\Sigma) + o(\|\boldsymbol{\xi}\|) \\ \mathrm{D}^2\,L(\Sigma)[\boldsymbol{\xi}, \boldsymbol{\eta}] &= \mathrm{D}\,L(\Sigma + \boldsymbol{\eta})[\boldsymbol{\xi}] - \mathrm{D}\,L(\Sigma)[\boldsymbol{\xi}] + o(\|\boldsymbol{\xi}\|). \end{aligned} \quad (17)$$

Notice that $\mathrm{D}^2\,L(\Sigma)[\boldsymbol{\xi}, \boldsymbol{\eta}]$ is symmetrical with respect to $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. In the case of the CES distributions, the Fisher information metric was studied in [8, 15, 28, 63, 64], and its derivation is reported in the following Theorem:

[4] An inner product is a bilinear, symmetric, positive definite function.

Theorem 2 (Fisher Information metric of centered CES) *Let* $\Sigma \in \mathcal{H}_p^{++}$. *Let* $\{\mathbf{x}_i\}_{i=1}^n$ *be a n-sample of iid from* $\mathbf{x} \sim$ *C-CES*$(\mathbf{0}, \Sigma, g)$. *The Fisher information metric is obtained* $\forall\ \boldsymbol{\xi}, \boldsymbol{\eta} \in T_\Sigma \mathcal{H}_p^{++}$ *as*

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_\Sigma^{\mathrm{FIM}} = n\alpha_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}) + n\beta_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\eta}),$$

with α_g *and* β_g *defined in* (9).

Proof The first things to compute to obtain the Fisher information metric are the derivatives $\mathrm{D}\,\mathcal{L}_g(\{\mathbf{x}_i\}_{i=1}^n|\Sigma)[\boldsymbol{\xi}]$ and $\mathrm{D}^2\,\mathcal{L}_g(\{\mathbf{x}_i\}_{i=1}^n|\Sigma)[\boldsymbol{\xi}, \boldsymbol{\eta}]$ at $\Sigma \in \mathcal{H}_p^{++}$ in directions $\boldsymbol{\xi}$ and $\boldsymbol{\eta} \in T_\Sigma \mathcal{H}_p^{++}$. To do so, recall that $\mathrm{D}\log\det(\Sigma)[\boldsymbol{\xi}] = \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi})$ and $\mathrm{D}(\Sigma^{-1})[\boldsymbol{\xi}] = -\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}$. It follows that

$$\mathrm{D}\,\mathcal{L}_g(\{\mathbf{x}_i\}_{i=1}^n|\Sigma)[\boldsymbol{\xi}] = -n \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}) - \sum_{i=1}^n \phi(\mathbf{x}_i^H\Sigma^{-1}\mathbf{x}_i) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H).$$

Moreover, also recall that the trace is invariant to any permutation of the product of three Hermitian matrices. Thus,

$$\begin{aligned}\mathrm{D}^2\,\mathcal{L}_g(\{\mathbf{x}_i\}_{i=1}^n|\Sigma)[\boldsymbol{\xi}, \boldsymbol{\eta}] &= \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}) \\ &\quad + 2\sum_{i=1}^n \phi(\operatorname{Tr}(\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H)) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H) \\ &\quad + \sum_{i=1}^n \phi'(\operatorname{Tr}(\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H)) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\eta}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H).\end{aligned}$$

We now need to compute the expectation. To do so, we exploit the stochastic representation $\mathbf{x}_i =_d \sqrt{Q_i}\Sigma^{1/2}\mathbf{u}_i$. Recall that Q_i and $\mathbf{u}_i$ are independent, $\mathbf{u}_i^H\mathbf{u}_i = 1$ and $\mathrm{E}\left[\mathbf{u}_i\mathbf{u}_i^H\right] = \frac{1}{p}\mathbf{I}_p$ (since $\mathbf{u}_i \sim \mathcal{U}(\mathbb{C}\mathcal{S}^p)$). Furthermore, from (3), $\mathrm{E}\left[Q_i\phi(Q_i)\right] = -p$. It follows that

$$\begin{aligned}&\mathrm{E}\left[\phi(\operatorname{Tr}(\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H)) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H)\right] \\ &\quad = \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}\mathrm{E}\left[\phi(Q_i)Q_i\mathbf{u}_i\mathbf{u}_i^H\right]) = \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}\mathrm{E}\left[\phi(Q_i)Q_i\right]\mathrm{E}\left[\mathbf{u}_i\mathbf{u}_i^H\right]) \\ &\quad = -\operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}).\end{aligned}$$

For the second expectation, from [16], we need

$$\mathrm{E}\left[(\mathbf{u}_i^H\boldsymbol{A}\mathbf{u}_i)^2\right] = \frac{\operatorname{Tr}(\boldsymbol{A}^2) + (\operatorname{Tr}(\boldsymbol{A}))^2}{p(p+1)}.$$

Applying the polarization formula $\frac{1}{4}[(\mathbf{u}_i^H(\boldsymbol{A}+\boldsymbol{B})\mathbf{u}_i)^2 - (\mathbf{u}_i^H(\boldsymbol{A}-\boldsymbol{B})\mathbf{u}_i)^2]$, we get

$$\mathrm{E}\left[(\mathbf{u}_i^H\boldsymbol{A}\mathbf{u}_i)(\mathbf{u}_i^H\boldsymbol{B}\mathbf{u}_i)\right] = \frac{\operatorname{Tr}(\boldsymbol{A}\boldsymbol{B}) + \operatorname{Tr}(\boldsymbol{A})\operatorname{Tr}(\boldsymbol{B})}{p(p+1)}.$$

Therefore,

$$\begin{aligned}
&\mathrm{E}\left[\phi'(\mathrm{Tr}(\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H))\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H)\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H)\right]\\
&\quad= \mathrm{E}\left[Q_i^2\phi'(Q_i)(\mathbf{u}_i^H\Sigma^{-1}\boldsymbol{\xi}\mathbf{u}_i)(\mathbf{u}_i^H\Sigma^{-1}\boldsymbol{\eta}\mathbf{u}_i)\right]\\
&\quad= \mathrm{E}\left[Q_i^2\phi'(Q_i)\right]\mathrm{E}\left[(\mathbf{u}_i^H\Sigma^{-1}\boldsymbol{\xi}\mathbf{u}_i)(\mathbf{u}_i^H\Sigma^{-1}\boldsymbol{\eta}\mathbf{u}_i)\right]\\
&\quad= \frac{\mathrm{E}\left[Q_i^2\phi'(Q_i)\right]}{p(p+1)}\left(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}) + \mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta})\right)
\end{aligned}$$

From there, basic manipulations yield the result with coefficients α_g and β_g defined in (9). Notice that the dependency on i in Q is omitted since these parameters are assumed iid. □

The Fisher information metric of C-CES thus corresponds to a general form of the well known affine invariant metric on $\mathcal{H}_p^{++}$ [17, 83]. Hence, if not specified otherwise the remainder of this chapter will use the more common generic denotation:

$$\langle\boldsymbol{\xi},\boldsymbol{\eta}\rangle_\Sigma \overset{\text{def}}{=} \mathfrak{g}_\Sigma(\boldsymbol{\xi},\boldsymbol{\eta}) = \Re\mathfrak{e}(\alpha\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}) + \beta\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta})) \tag{18}$$

and study the corresponding Riemannian geometry of $\mathcal{H}_p^{++}$ for any $\alpha \in \mathbb{R}_*^+$ and $\beta > -\alpha/p$ (necessary conditions so that $\langle\cdot,\cdot\rangle_\Sigma$ is positive definite). The Fisher-Rao information geometry of the considered C-CES model is then recovered by fixing α and β according to Theorem 2. We can also point out that the most studied case corresponds to $\alpha = 1$ and $\beta = 0$, which coincide with the Fisher information metric of the Gaussian distribution, as $\alpha_g = 1$ and $\beta_g = 0$ in this case [28, 85].

Remark Taking the real part in the metric (18) defines a proper inner product on $T_\Sigma\mathcal{H}_p^{++}$ from the original Hermitian inner product. This way, we implicitly identify the complex space as its underlying real vector space ($\mathbb{C} \sim \mathbb{R}^2$), so that we can use the usual derivatives (defined as those used on $\mathbb{R}$). As a direct consequence, in this chapter, both $\mathcal{H}_p$ and $\mathcal{H}_p^{++}$ are of dimension p^2. Notice that, even though it is not always stated, most works that deal with complex-valued matrices (e.g., [85]) also implicitly use the real part of the Fisher information metric.

Remark Among many other properties, the Fisher information metric from Theorem 2 has a notable quadratic dependence on Σ^{-1}. This makes the norm of tangent vectors $\|\boldsymbol{\xi}\|_\Sigma^2 = \langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_\Sigma$ tend to infinity when the point Σ tends to the boundaries of the manifold (i.e, when any number of its eigenvalues tend 0). This Riemannian metric thus allows to actually perceive the boundary of $\mathcal{H}_p^{++}$ as being infinitely far, which was not the case for the Euclidean metric. An illustration of the effect of the metric is displayed for $\mathcal{H}_1^{++} = \mathbb{R}_*^+$ in Figure 3.

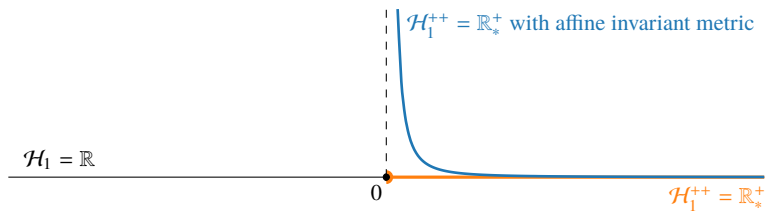

Fig. 3: Illustration of the effect of the affine invariant metric on $\mathcal{H}_1^{++} = \mathbb{R}_*^+$. Thanks to the metric, the excluded point 0 becomes truly unreachable.

2.2 Levi-Civita connection

One of the most – if not the most – important tools of Riemannian geometry is the Levi-Civita connection, which generalizes the notion of directional derivatives of vector fields on manifolds. A vector field is a function that associates a unique tangent vector $\boldsymbol{\xi}_\Sigma \in T_\Sigma \mathcal{H}_p^{++}$ to every point $\Sigma \in \mathcal{H}_p^{++}$, which is illustrated in Figure 4. An example of a vector field that will be involved in Section 3 is the gradient of a cost function. The set of vector fields on $\mathcal{H}_p^{++}$ is denoted $\mathfrak{X}(\mathcal{H}_p^{++})$.

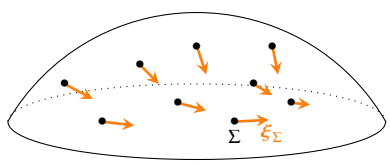

Fig. 4: Illustration of a vector field on $\mathcal{H}_p^{++}$.

To differentiate a vector field on a manifold, one needs to resort to an affine connection. This is an application from $\mathfrak{X}(\mathcal{H}_p^{++}) \times \mathfrak{X}(\mathcal{H}_p^{++})$ onto $\mathfrak{X}(\mathcal{H}_p^{++})$. The connection of $\boldsymbol{\eta}_\Sigma$ in the direction $\boldsymbol{\xi}_\Sigma$ is denoted $\nabla_{\boldsymbol{\xi}_\Sigma} \boldsymbol{\eta}_\Sigma$ and generalizes the directional derivative of $\boldsymbol{\eta}_\Sigma$ in the direction $\boldsymbol{\xi}_\Sigma$ (i.e., $\mathrm{D}\,\boldsymbol{\eta}_\Sigma[\boldsymbol{\xi}_\Sigma]$). Such generalization is needed because the tangent space changes when one moves from one point to another on a manifold. Thus, the usual directional derivative might not be properly defined, as it does not account for the structure of the manifold (constraints, Riemannian metric, etc.). This specificity is illustrated in Figure 5.
Many affine connections can be defined on a manifold. However, there is a unique one that is in accordance with the chosen Riemannian metric, which is referred to as the Levi-Civita connection. This Levi-Civita connection, denoted $\nabla_{\boldsymbol{\xi}_\Sigma} \boldsymbol{\eta}_\Sigma$, is the unique solution in the tangent space $T_\Sigma \mathcal{H}_p^{++}$ to the Koszul formula

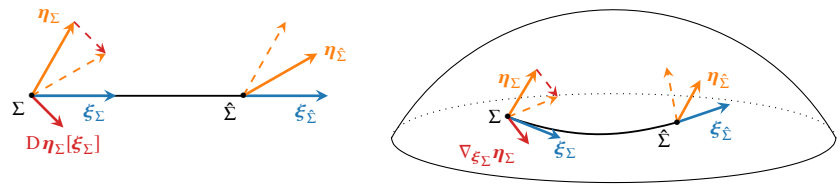

Fig. 5: Illustration of directional derivative $\mathrm{D}\,\boldsymbol{\eta}_\Sigma[\boldsymbol{\xi}_\Sigma]$ (left) and affine connection $\nabla_{\boldsymbol{\xi}_\Sigma}\boldsymbol{\eta}_\Sigma$ (right) of a vector field $\boldsymbol{\eta}$ in the direction $\boldsymbol{\xi}$ at Σ. As the directional derivative, an affine connection describes how the vector field $\boldsymbol{\eta}$ evolves in a given direction $\boldsymbol{\xi}$. In addition, the affine connection takes into account the structure of the manifold (curvature, and non-constant metric).

$$\begin{aligned}2\mathfrak{g}_\Sigma(\nabla_{\boldsymbol{\xi}_\Sigma}\boldsymbol{\eta}_\Sigma, \boldsymbol{\nu}_\Sigma) &= 2\mathfrak{g}_\Sigma(\mathrm{D}\,\boldsymbol{\eta}_\Sigma[\boldsymbol{\xi}_\Sigma], \boldsymbol{\nu}_\Sigma)\\ &\quad + \mathrm{D}\,\mathfrak{g}_\Sigma[\boldsymbol{\xi}_\Sigma](\boldsymbol{\eta}_\Sigma, \boldsymbol{\nu}_\Sigma) + \mathrm{D}\,\mathfrak{g}_\Sigma[\boldsymbol{\eta}_\Sigma](\boldsymbol{\xi}_\Sigma, \boldsymbol{\nu}_\Sigma) - \mathrm{D}\,\mathfrak{g}_\Sigma[\boldsymbol{\nu}_\Sigma](\boldsymbol{\eta}_\Sigma, \boldsymbol{\xi}_\Sigma), \quad (19)\end{aligned}$$

where we use the alternate notation of the metric, i.e., $\mathfrak{g}_\Sigma(\cdot,\cdot) = \langle\cdot,\cdot\rangle_\Sigma$. Notice that the presented formula is simpler than the general case [1]. It is because the Lie bracket is $[\boldsymbol{\xi}_\Sigma, \boldsymbol{\eta}_\Sigma] = \mathrm{D}\,\boldsymbol{\eta}_\Sigma[\boldsymbol{\eta}_\Sigma] - \mathrm{D}\,\boldsymbol{\xi}_\Sigma[\boldsymbol{\eta}_\Sigma]$ since $\mathcal{H}_p^{++}$ is an open subset of a vector space, i.e., $\mathcal{H}_p$. The Levi-Civita connection of $\mathcal{H}_p^{++}$ associated with the Riemannian metric (18) is provided in Theorem 3.

Theorem 3 (Levi-Civita connection) *The Levi-Civita connection on $\mathcal{H}_p^{++}$ associated with the affine invariant metric* (18) *is defined for $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathfrak{X}(\mathcal{H}_p^{++})$ and $\Sigma \in \mathcal{H}_p^{++}$, as*

$$\nabla_{\boldsymbol{\xi}_\Sigma}\boldsymbol{\eta}_\Sigma = \mathrm{D}\,\boldsymbol{\eta}_\Sigma[\boldsymbol{\xi}_\Sigma] - \mathrm{Herm}(\boldsymbol{\eta}_\Sigma\Sigma^{-1}\boldsymbol{\xi}_\Sigma),$$

where $\mathrm{Herm}(\cdot)$ *returns the Hermitian part of its argument.*

Proof First recall that for $\boldsymbol{A} \in \mathcal{H}_p$ and $\boldsymbol{B} \in \mathbb{R}^{p\times p}$, $\mathrm{Tr}(\boldsymbol{AB}) = \mathrm{Tr}(\boldsymbol{A}\,\mathrm{Herm}(\boldsymbol{B}))$. Further recall that the trace is invariant to any permutation of the product of three Hermitian matrices. Since $\mathrm{D}(\Sigma^{-1})[\boldsymbol{\xi}] = -\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}$, we have

$$\mathrm{D}\,\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}))[\boldsymbol{\nu}] = -2\,\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}\Sigma^{-1}\boldsymbol{\nu}))$$

and

$$\begin{aligned}&\mathrm{D}\,\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta}\Sigma^{-1}\boldsymbol{\nu}))[\boldsymbol{\xi}] + \mathrm{D}\,\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\nu}))[\boldsymbol{\eta}] - \mathrm{D}\,\mathfrak{Re}\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}))[\boldsymbol{\nu}]\\ &\quad = -2\,\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}\Sigma^{-1}\boldsymbol{\nu})) = -2\,\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\,\mathrm{Herm}(\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta})\Sigma^{-1}\boldsymbol{\nu})).\end{aligned}$$

We also have

$$\begin{aligned}\mathrm{D}\left(\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta}))\right)[\boldsymbol{\nu}] &= -\,\mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\nu}\Sigma^{-1}\boldsymbol{\xi})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta}))\\ &\quad - \mathfrak{Re}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\nu}\Sigma^{-1}\boldsymbol{\eta})).\end{aligned}$$

It follows that

$$\begin{aligned}\mathrm{D}\left(\Re\mathrm{e}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\nu}))\right)[\boldsymbol{\xi}] + \mathrm{D}\left(\Re\mathrm{e}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\nu}))\right)[\boldsymbol{\eta}]\\ - \mathrm{D}\left(\Re\mathrm{e}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\eta}))\right)[\boldsymbol{\nu}] = -2\,\Re\mathrm{e}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\nu})\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}))\\ = -2\,\Re\mathrm{e}(\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\nu})\,\mathrm{Tr}(\Sigma^{-1}\,\mathrm{Herm}(\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}))).\end{aligned}$$

From there, we can deduce that

$$\begin{aligned}\mathrm{D}\,\mathfrak{g}_\Sigma[\boldsymbol{\xi}_\Sigma](\boldsymbol{\eta}_\Sigma,\boldsymbol{\nu}_\Sigma) + \mathrm{D}\,\mathfrak{g}_\Sigma[\boldsymbol{\eta}_\Sigma](\boldsymbol{\xi}_\Sigma,\boldsymbol{\nu}_\Sigma) - \mathrm{D}\,\mathfrak{g}_\Sigma[\boldsymbol{\nu}_\Sigma](\boldsymbol{\eta}_\Sigma,\boldsymbol{\xi}_\Sigma)\\ = -2\mathfrak{g}_\Sigma(\mathrm{Herm}(\boldsymbol{\xi}_\Sigma\Sigma^{-1}\boldsymbol{\eta}_\Sigma),\boldsymbol{\nu}_\Sigma).\end{aligned}$$

Injecting this into the Koszul formula yields the result. □

Remark Notice that the Levi-Civita connection of $\mathcal{H}_p^{++}$ associated with the Riemannian metric of the metric in (18) does not depend on α and β. Hence it remains the same for any underlying C-CES distribution.

2.3 Geodesics, Riemannian exponential, logarithm and distance

One of the main reasons why the Levi-Civita connection is so crucial is because it allows to define geodesics. The geodesics generalize the concept of straight lines on a manifold. These are curves $\gamma : [0,1] \to \mathcal{H}_p^{++}$ with no acceleration, where acceleration is defined thanks to the Levi-Civita connection. They are parameterized by the choice of starting point $\gamma(0) = \Sigma \in \mathcal{H}_p^{++}$ and either initial direction $\dot\gamma(0) = \boldsymbol{\xi} \in \mathcal{H}_p$ or ending point $\gamma(1) = \hat\Sigma \in \mathcal{H}_p^{++}$. An illustration of geodesics is provided in Figure 6. Formally, the geodesic $\gamma : [0,1] \to \mathcal{H}_p^{++}$ is the solution to the differential equation

$$\nabla_{\dot\gamma(t)}\,\dot\gamma(t) = \mathbf{0}. \tag{20}$$

The geodesics on $\mathcal{H}_p^{++}$ according to the Levi-Civita connection of Theorem 3 are given in Theorem 4 along with the proof.

Theorem 4 (Geodesics) *The geodesic $\gamma : [0,1] \to \mathcal{H}_p^{++}$ such that $\gamma(0) = \Sigma \in \mathcal{H}_p^{++}$ and $\dot\gamma(0) = \boldsymbol{\xi} \in \mathcal{H}_p$ is defined as*

$$\gamma(t) \;=\; \Sigma\exp(t\Sigma^{-1}\boldsymbol{\xi}) \;=\; \exp(t\boldsymbol{\xi}\Sigma^{-1})\Sigma \;=\; \Sigma^{1/2}\exp(t\Sigma^{-1/2}\boldsymbol{\xi}\Sigma^{-1/2})\Sigma^{1/2},$$

where $\exp(\cdot)$ *denotes the matrix exponential. Equivalently, one can define the geodesic $\gamma : [0,1] \to \mathcal{H}_p^{++}$ such that $\gamma(0) = \Sigma \in \mathcal{H}_p^{++}$ and $\gamma(1) = \hat\Sigma \in \mathcal{H}_p^{++}$ by*

$$\gamma(t) = \Sigma^{1/2}\left(\Sigma^{-1/2}\hat\Sigma\Sigma^{-1/2}\right)^t\Sigma^{1/2},$$

where $(\cdot)^t = \exp(t\log(\cdot))$, $\log(\cdot)$ *denoting the matrix logarithm.*

Proof We only provide the proof for $\gamma(0) = \Sigma$ and $\dot{\gamma}(0) = \boldsymbol{\xi}$. The result for $\gamma(1) = \hat{\Sigma}$ is obtained by choosing

$$\boldsymbol{\xi} \;=\; \Sigma \log(\Sigma^{-1}\hat{\Sigma}) \;=\; \log(\hat{\Sigma}\Sigma^{-1})\Sigma \;=\; \Sigma^{1/2}\log(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2})\Sigma^{1/2}.$$

Notice that the equality between the three versions of $\gamma(t)$ given in the theorem above and of $\boldsymbol{\xi}$ given here solely rely on the fact that $\boldsymbol{A}\exp(\boldsymbol{B})\boldsymbol{A}^{-1} = \exp(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{-1})$ and $\boldsymbol{A}\log(\boldsymbol{B})\boldsymbol{A}^{-1} = \log(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{-1})$.

The differential equation (20) for the Levi-Civita connection defined in Theorem 3 is

$$\ddot{\gamma}(t) - \dot{\gamma}(t)\gamma(t)^{-1}\dot{\gamma}(t) = \mathbf{0}.$$

Recall that $\frac{\mathrm{d}}{\mathrm{d}t}\exp(t\boldsymbol{A}) = \boldsymbol{A}\exp(t\boldsymbol{A})$. Thus, with $\gamma(t) = \Sigma\exp(t\Sigma^{-1}\boldsymbol{\xi})$, we have $\dot{\gamma}(t) = \boldsymbol{\xi}\exp(t\Sigma^{-1}\boldsymbol{\xi})$ and $\ddot{\gamma}(t) = \boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\xi}\exp(t\Sigma^{-1}\boldsymbol{\xi})$. From there, we easily obtain $\dot{\gamma}(t)\gamma(t)^{-1} = \boldsymbol{\xi}\Sigma^{-1}$. Simple computations show that $\gamma(t)$ satisfies the differential equation above, which is enough to conclude. □

Remark Since the Levi-Civita connection does not depend on α and β, neither does geodesics. Hence, the Fisher-Rao geometries of C-CES models share the same geodesics whatever the underlying distribution.

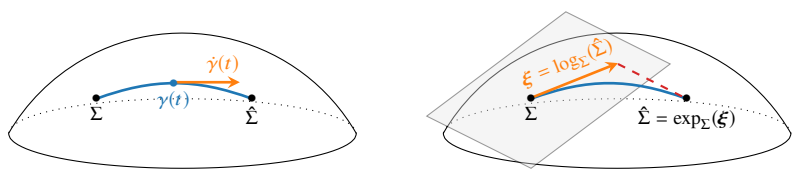

Fig. 6: Illustration of geodesics (left), Riemannian exponential and logarithm mappings (right). The Riemannian distance $\delta(\Sigma, \hat{\Sigma})$ is the length of the geodesic joining Σ and $\hat{\Sigma}$.

Geodesics allow to define the Riemannian exponential mapping. By definition, for all $\Sigma \in \mathcal{H}_p^{++}$, this is the mapping from $T_\Sigma\mathcal{H}_p^{++} \simeq \mathcal{H}_p$ onto $\mathcal{H}_p^{++}$ such that, for all $\boldsymbol{\xi} \in \mathcal{H}_p$, $\exp_\Sigma(\boldsymbol{\xi}) = \gamma(1)$, where γ is the geodesic such that $\gamma(0) = \Sigma$ and $\dot{\gamma}(t) = \boldsymbol{\xi}$. Thus, for all $\Sigma \in \mathcal{H}_p^{++}$ and $\boldsymbol{\xi} \in \mathcal{H}_p$, we have

$$\exp_\Sigma(\boldsymbol{\xi}) \;=\; \Sigma\exp(\Sigma^{-1}\boldsymbol{\xi}) \;=\; \exp(\boldsymbol{\xi}\Sigma^{-1})\Sigma \;=\; \Sigma^{1/2}\exp(\Sigma^{-1/2}\boldsymbol{\xi}\Sigma^{-1/2})\Sigma^{1/2}. \tag{21}$$

From there we can define the Riemannian logarithm mapping, which is the inverse of the Riemannian exponential mapping. Given $\Sigma \in \mathcal{H}_p^{++}$, it is the mapping from $\mathcal{H}_p^{++}$ onto $T_\Sigma\mathcal{H}_p^{++} \simeq \mathcal{H}_p$ such that, for $\hat{\Sigma} \in \mathcal{H}_p^{++}$, $\log_\Sigma(\hat{\Sigma})$ is the solution to equation $\exp_\Sigma(\log_\Sigma(\hat{\Sigma})) = \hat{\Sigma}$. In our case, we have

$$\log_\Sigma(\hat{\Sigma}) \;=\; \Sigma\log(\Sigma^{-1}\hat{\Sigma}) \;=\; \log(\hat{\Sigma}\Sigma^{-1})\Sigma \;=\; \Sigma^{1/2}\log(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2})\Sigma^{1/2}. \tag{22}$$

Illustrations of Riemannian exponential and logarithm mappings are given in Figure 6.

The last object from Riemannian geometry presented in this chapter is the Riemannian distance. The distance between two points corresponds to the length of the geodesic joining them. Formally, it is defined as

$$\delta(\Sigma, \hat{\Sigma}) = \int_0^1 \langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{\gamma(t)}^{1/2} \mathrm{d}t, \tag{23}$$

where γ is the geodesic such that $\gamma(0) = \Sigma$ and $\gamma(1) = \hat{\Sigma}$. The Riemannian distance on $\mathcal{H}_p^{++}$ associated to the metric of Theorem 2 is given in Theorem 5 along with the proof. It was derived in [28].

Theorem 5 (Fisher-Rao distance of C-CES distributions) *The square of the Fisher distance of C-CES distributions over $\mathcal{H}_p^{++}$ is defined, for all Σ and $\hat{\Sigma} \in \mathcal{H}_p^{++}$, by*

$$\delta^2(\Sigma, \hat{\Sigma}) = \alpha \| \log(\Sigma^{-1}\hat{\Sigma}) \|_F^2 + \beta (\log\det(\Sigma^{-1}\hat{\Sigma}))^2.$$

Proof From the proof of Theorem 4, $\dot{\gamma}(t)\gamma(t)^{-1} = \boldsymbol{\xi}\Sigma^{-1}$ for all $t \in [0, 1]$. Thus, we can deduce that $\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{\gamma(t)} = \langle \dot{\gamma}(0), \dot{\gamma}(0) \rangle_{\gamma(0)}$ for all $t \in [0, 1]$. Therefore, $\delta^2(\Sigma, \hat{\Sigma}) = \langle \dot{\gamma}(0), \dot{\gamma}(0) \rangle_{\gamma(0)}$, with $\gamma(0) = \Sigma$ and $\dot{\gamma}(0) = \Sigma \log(\Sigma^{-1}\hat{\Sigma})$. It follows that

$$\delta^2(\Sigma, \hat{\Sigma}) = \alpha \operatorname{Tr}((\log(\Sigma^{-1}\hat{\Sigma}))^2) + \beta (\operatorname{Tr}(\log(\Sigma^{-1}\hat{\Sigma})))^2.$$

To conclude, it is enough to recall that $\operatorname{Tr}(\log(\boldsymbol{A})) = \log\det(\boldsymbol{A})$. □

Remark We previously noticed that the Levi-Civita connection and geodesics do not depend on the coefficients α and β of the metric. However, since the Riemannian distance integrates the metric along the geodesics, it does well depend on these factors. This means that the Fisher-Rao distance (Riemannian distance according to the Fisher in formation geometry) actually depends on the underlying C-CES distribution.

3 Covariance matrix estimation with Riemannian optimization

The estimation of the covariance matrix of a set of observations is a ubiquitous problem in signal processing and machine learning. Among many applications involving this quantity, we can mention: adaptive filtering and detection, metric learning in classification, data analysis (e.g., graph learning), and dimension reduction. This section discusses covariance matrix estimation within the class of C-CES, and illustrates how the concepts related to Fisher-Rao information geometry can be leveraged in this context. First, Section 3.1 provides some reminders on covariance matrix estimation in the C-CES framework (cf. Section 6 of the background chapter for more details). Second, section 3.2 presents an introduction to Riemannian optimization, where maximum likelihood estimation of C-CES models is used as a driving

example. Finally, Section 3.3 shortly presents how this framework can be leveraged to more general regularized covariance matrix estimation problems and points to references on the matter.

3.1 Reminders on covariance matrix estimation within CES

Given a n-sample $\{\mathbf{x}_i\}_{i=1}^n$ assumed to be iid from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$, with unknown covariance matrix Σ, we consider inferring this matrix. The most common approach to tackle this problem consists in maximizing the log-likelihood function in (4). The maximum likelihood estimator is thus obtained as a solution to the optimization problem

$$\underset{\Sigma \in \mathcal{H}_p^{++}}{\text{minimize}} \quad L(\Sigma) \tag{24}$$

where L denotes in short the negative log-likelihood of the sample set $\{\mathbf{x}_i\}_{i=1}^n$, i.e.:

$$L(\Sigma) = -\mathcal{L}_g\left(\{\mathbf{x}_i\}_{i=1}^n | \Sigma\right), \tag{25}$$

with $\mathcal{L}_g$ defined in (4). The solution of (24) yields the MLE in the form of a fixed point equation

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \psi\left(\mathbf{x}_i^H \hat{\Sigma}^{-1} \mathbf{x}_i\right) \mathbf{x}_i \mathbf{x}_i^H =_d \mathcal{T}_\psi\left(\hat{\Sigma}\right), \tag{26}$$

where $\psi(t) = -g'(t)/g(t)$. This solution is most commonly evaluated thanks to a fixed-point algorithm

$$\Sigma_{(k+1)} = \mathcal{T}_\psi\left(\Sigma_{(k)}\right) = \frac{1}{n}\sum_{i=1}^n \psi(\mathbf{x}_i^H \Sigma_{(k)}^{-1} \mathbf{x}_i)\, \mathbf{x}_i \mathbf{x}_i^H. \tag{27}$$

The existence and uniqueness of the fixed-point solution (26), as well as the convergence of the fixed-point algorithm (27) is subject subject to conditions on the function ψ (resp. the density generator g) and the sample set $\{\mathbf{x}_i\}_{i=1}^n$, e.g., obtained in [73, Theorems 6 and 7]. A notable condition in the absolutely continuous case is that the sample size is required to be larger than the dimension, i.e., $n > p$.

Remark In practice, the true density generator g may not be known or accurately specified. In the robust estimation theory, an M-estimator of the scatter matrix [60, 95] refers to an estimator built from (26)-(27) using a function $\psi(t)$ that is not necessarily linked to the density generator g (cf. Section 6.3 of the background chapter). In this chapter, we focus on the example of the MLE, but the tools that will be presented apply to any generic cost function L.

3.2 Computing MLEs with Riemannian optimization

Riemannian optimization [1, 26] is a general framework to solve optimization problems on manifolds. This extends to Riemannian manifolds classical Euclidean optimization methods such as steepest gradient descent, conjugate gradient, Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, Newton method, trust region, *etc*. This section introduces Riemannian optimization on $\mathcal{H}_p^{++}$ as a framework to solve (24) that can leverage tools from the Fisher-Rao information geometry. At the end of this section, we will see that this framework actually yields the fixed point algorithm (27) as a special case (specifically, a Riemannian steepest gradient descent with a specific choice of metric, retraction, and step-size).

We consider an optimization problem of the from (24) that has no obvious closed-form solution on $\mathcal{H}_p^{++}$. In order to evaluate this solution, we resort to iterative methods, i.e., methods that yield a sequence of iterates $\{\Sigma_{(k)}\}$ in $\mathcal{H}_p^{++}$ from a starting point $\Sigma_{(0)} \in \mathcal{H}_p^{++}$. This sequence is constructed so that it eventually converges to a critical point of the objective in (24). When the variable is constrained to lie in the manifold $\mathcal{H}_p^{++}$, a generic first-order Riemannian optimization method operates as follows:

1. At iterate $\Sigma_{(k)} \in \mathcal{H}_p^{++}$, a descent direction in the tangent space, denoted $\boldsymbol{\xi}_{(k)} \in T_{\Sigma_{(k)}}\mathcal{H}_p^{++} \simeq \mathcal{H}_p$, is computed by leveraging the Riemannian gradient.
2. The direction descent $\boldsymbol{\xi}_{(k)}$ is used to obtain the next iterate $\Sigma_{(k+1)}$ on $\mathcal{H}_p^{++}$. This is achieved through a retraction on $\mathcal{H}_p^{++}$, which is an operator that maps tangent vectors back onto the manifold. □

An illustration of such an optimization process is presented in Figure 7, while the design of these two steps is discussed to solve (24) on $\mathcal{H}_p^{++}$ in the following.

For the first step, the steepest descent direction is given by the gradient, which is defined through the metric in the Riemannian setting. The Riemannian gradient of the negative log-likelihood L at $\Sigma \in \mathcal{H}_p^{++}$ according to the metric of Theorem 2 is the unique tangent vector $\operatorname{grad} L(\Sigma) \in T_\Sigma\mathcal{H}_p^{++} \simeq \mathcal{H}_p$ such that, for all $\boldsymbol{\xi} \in \mathcal{H}_p$, we have

$$\langle \operatorname{grad} L(\Sigma), \boldsymbol{\xi} \rangle_\Sigma = \mathrm{D}\, L(\Sigma)[\boldsymbol{\xi}]. \tag{28}$$

This Riemannian gradient is provided in Proposition 1.

Proposition 1 (Riemannian gradient of L) *The Riemannian gradient* $\operatorname{grad} L(\Sigma)$ *of the negative log-likelihood L defined in* (25) *at* $\Sigma \in \mathcal{H}_p^{++}$ *according to metric* (18) *is*

$$\operatorname{grad} L(\Sigma) = \left(\frac{n}{\alpha + p\beta} + \frac{\beta}{\alpha(\alpha + p\beta)} \sum_{i=1}^{n} \psi(\mathbf{x}_i^H \Sigma^{-1} \mathbf{x}_i)\, \mathbf{x}_i^H \Sigma^{-1} \mathbf{x}_i \right) \Sigma - \frac{1}{\alpha} \sum_{i=1}^{n} \psi(\mathbf{x}_i^H \Sigma^{-1} \mathbf{x}_i)\, \mathbf{x}_i \mathbf{x}_i^H .$$

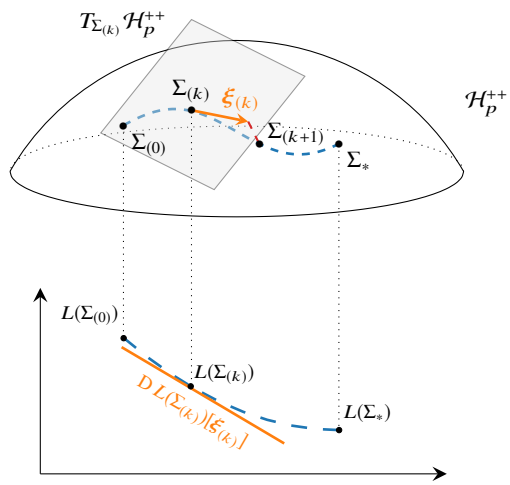

Fig. 7: Illustration of Riemannian optimization. Given some initialization $\Sigma_{(0)}$, the goal is to reach the minimum Σ_*. At $\Sigma_{(k)}$, the descent direction $\boldsymbol{\xi}_{(k)}$ is such that it induces a decrease in L, *i.e.*, $\mathrm{D}\,L(\Sigma_{(k)})[\boldsymbol{\xi}_{(k)}] < 0$ (slope of the orange line).

Proof From the beginning of the proof of Theorem 2, we get that the directional derivative of L at $\Sigma \in \mathcal{H}_p^{++}$ in direction $\boldsymbol{\xi} \in \mathcal{H}_p$ is

$$\begin{aligned}\mathrm{D}\,L(\Sigma)[\boldsymbol{\xi}] &= n\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}) + \textstyle\sum_{i=1}^n \frac{g'}{g}(\mathbf{x}_i^H\Sigma^{-1}\mathbf{x}_i)\,\mathrm{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\mathbf{x}_i\mathbf{x}_i^H)\\ &= \mathrm{Tr}\left(\Sigma^{-1}\left(n\Sigma - \textstyle\sum_{i=1}^n \psi(\mathbf{x}_i^H\Sigma^{-1}\mathbf{x}_i)\,\mathbf{x}_i\mathbf{x}_i^H\right)\Sigma^{-1}\boldsymbol{\xi}\right).\end{aligned}$$

Thus, for $\alpha = 1$ and $\beta = 0$, we immediately get the result by identification. To obtain the result in the general case, notice that, given $\boldsymbol{A}$ and $\boldsymbol{B}$ in $\mathcal{H}_p$, if we set $\tilde{\boldsymbol{A}} = \frac{1}{\alpha}\boldsymbol{A} - \frac{\beta}{\alpha(\alpha+p\beta)}\mathrm{Tr}(\boldsymbol{A})\boldsymbol{I}_p$, then we have $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \alpha\,\mathrm{Tr}(\tilde{\boldsymbol{A}}\boldsymbol{B}) + \beta\,\mathrm{Tr}(\tilde{\boldsymbol{A}})\,\mathrm{Tr}(\boldsymbol{B})$. Taking $\boldsymbol{A} = \Sigma^{-1/2}(n\Sigma - \sum_{i=1}^n \psi(\mathbf{x}_i^H\Sigma^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^H)\Sigma^{-1/2}$ and $\boldsymbol{B} = \Sigma^{-1/2}\boldsymbol{\xi}\Sigma^{-1/2}$, and basic calculations allow to conclude. □

To perform the second step, it remains to define a retraction that maps tangent vectors back onto the manifold. Formally, given $\Sigma \in \mathcal{H}_p^{++}$, a retraction is a mapping $R_\Sigma : T_\Sigma\mathcal{H}_p^{++} \simeq \mathcal{H}_p \to \mathcal{H}_p^{++}$ such that, for all $\boldsymbol{\xi} \in \mathcal{H}_p$,

$$R_\Sigma(\boldsymbol{\xi}) = \Sigma + \boldsymbol{\xi} + o(\|\boldsymbol{\xi}\|). \tag{29}$$

From a geometric point of view, the Riemannian exponential mapping provides the ideal retraction for a manifold equipped with a Riemannian metric (in the sense that it is the most reflective of the considered geometry). In our case, it is defined in (21) and illustrated in Figure 6. However, this retraction involves computing the matrix exponential of some Hermitian matrix, which can be computationally costly and/or numerically unstable, as the exponential tends quickly to infinity or zero. From a practical point of view, it might thus be more advantageous to employ alternate

retractions. Notice that (29) means that a proper retraction is (at least) a first-order approximation of the Riemannian exponential mapping. Since $\mathcal{H}_p^{++}$ is open in $\mathcal{H}_p$, a proper first order approximation is simply obtained as

$$R_\Sigma^{(1)}(\boldsymbol{\xi}) = \Sigma + \boldsymbol{\xi}. \tag{30}$$

The main limitation of $R^{(1)}$ is that, given $\Sigma \in \mathcal{H}_p^{++}$, there are many $\boldsymbol{\xi} \in \mathcal{H}_p$ such that $R_\Sigma^{(1)}(\boldsymbol{\xi}) \notin \mathcal{H}_p^{++}$. This means that the iterative algorithms that employ this retraction are not guaranteed to be numerically stable. To overcome this issue, Proposition 2 provides a retraction that is a second-order approximation of the Riemannian exponential (21) (initially proposed in [51]), that does not suffer the same limitation as $R^{(1)}$.

Proposition 2 (Second order retraction) *The retraction $R^{(2)}$ such that, for all $\Sigma \in \mathcal{H}_p^{++}$ and $\boldsymbol{\xi} \in \mathcal{H}_p$,*

$$R_\Sigma^{(2)}(\boldsymbol{\xi}) = \Sigma + \boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\xi}$$

is a second order approximation of the Riemannian exponential mapping (21). *Furthermore, for all $\Sigma \in \mathcal{H}_p^{++}$ and $\boldsymbol{\xi} \in \mathcal{H}_p$, $R_\Sigma^{(2)}(\boldsymbol{\xi})$ belongs to $\mathcal{H}_p^{++}$.*

Proof Recall that the matrix exponential of $\boldsymbol{A}$ is $\exp(\boldsymbol{A}) = \sum_{k=0}^{\infty} \frac{\boldsymbol{A}^k}{k!}$. Hence the second order approximation is $\exp(\boldsymbol{A}) = \boldsymbol{I}_p + \boldsymbol{A} + \frac{1}{2}\boldsymbol{A}^2 + o(\|\boldsymbol{A}\|^2)$. Applying this to (21), we obtain $\exp_\Sigma(\boldsymbol{\xi}) = \Sigma(\boldsymbol{I}_p + \Sigma^{-1}\boldsymbol{\xi} + \frac{1}{2}\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\xi}) + o(\|\boldsymbol{\xi}\|^2)$. Basic calculations yield the result. Moreover, it is obviously a proper retraction. It remains to show that we always get a matrix in $\mathcal{H}_p^{++}$. To do so, notice that

$$R_\Sigma^{(2)}(\boldsymbol{\xi}) = \Sigma^{1/2}(\boldsymbol{I}_p + \Sigma^{-1/2}\boldsymbol{\xi}\Sigma^{-1/2} + \frac{1}{2}(\Sigma^{-1/2}\boldsymbol{\xi}\Sigma^{-1/2})^2)\Sigma^{1/2}.$$

Let the eigenvalue decomposition $\Sigma^{-1/2}\boldsymbol{\xi}\Sigma^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^H$. Then

$$R_\Sigma^{(2)}(\boldsymbol{\xi}) = \Sigma^{1/2}\boldsymbol{U}(\boldsymbol{I}_p + \boldsymbol{\Lambda} + \frac{1}{2}\boldsymbol{\Lambda}^2)\boldsymbol{U}^H\Sigma^{1/2}.$$

The result follows from the fact that the second order polynomial $\lambda \mapsto 1 + \lambda + \frac{1}{2}\lambda^2$ is strictly positive for all values of λ. □

We now have everything needed to define an iterative algorithm that solves the MLE optimization problem (24). Given the retraction R, we can, for instance, define the Riemannian gradient descent that yields the sequence of iterates

$$\Sigma_{(k+1)} = R_{\Sigma_{(k)}}(-\lambda_k \operatorname{grad} L(\Sigma_{(k)})), \tag{31}$$

where λ_k is the step size, which can be set by the user or computed through a line search; see e.g. [1, 26].

Our final point in this section is to show that the fixed point algorithm (27) is, in fact, a particular case of (31). Indeed, if we choose $\alpha = 1$ and $\beta = 0$, the Riemannian gradient of Proposition 1 is

$$\operatorname{grad} L(\Sigma) = n\Sigma - \sum_{i=1}^{n} \psi(\mathbf{x}_i^H \Sigma^{-1} \mathbf{x}_i)\, \mathbf{x}_i \mathbf{x}_i^H .$$

Algorithm (27) is then obtained from (31) by choosing the first order retraction (30) and constant step size $\lambda_k = \frac{1}{n}$. Notice that in this particular case, the choice of the first order retraction (30) is a valid choice because the particular structure of the gradient ensures that all iterates remain in $\mathcal{H}_p^{++}$. Though alternate choices of α and β in the metric, step size, and retraction could improve the convergence speed in some cases, this fixed-point is generally a good all-purpose candidate to compute MLEs as in (26). However, having recast it from the prism of Riemannian geometry opens many perspectives, which are discussed in the next section.

3.3 Beyond MLE and fixed-point algorithms

The MLEs (and M-estimators) are known for their good asymptotic performance in terms of estimation accuracy [37, 38, 39, 70, 100]. Still, they suffer from two limitations: i) they do not exist when the sample set is lower than the dimension ($n < p$); ii) they can be inaccurate when $n \simeq p$, as they do not leverage any bias-variance trade-off improvement. These limitations motivated the development of generalized estimation procedures by expressing new estimators as solutions to penalized optimization problems of the form:

$$\underset{\Sigma \in \mathcal{H}_p^{++}}{\text{minimize}} \quad L(\Sigma) + \lambda h(\Sigma) \tag{32}$$

where L is the negative log-likelihood as in (25), $\lambda \in \mathbb{R}^+$ is a regularization parameter, and h is a penalty function that promotes some form of regularization. Among many options considered in the literature for h, we can mention shrinkage to a target matrix [72, 74, 86], shrinkage of the eigenvalues [97, 29], promoting a sparse graphical structure [45, 101], or pooling from groups of observations [35, 71]. For appropriate choices of regularization penalty and parameters, the regularized estimators, as formulated in (32), can overcome the aforementioned issues of their non-regularized counterparts. In this scope, the Riemannian geometry provides useful tools to address and study (32), which is discussed next.

3.3.1 Riemannian options for computing solutions of (32)

The optimization problems expressed in (32) generally do not exhibit closed-form or fixed-point solutions and, thus, require the use of iterative algorithms to be evaluated.

In this setup, the Riemannian optimization framework is a good candidate in order to ensure that the variable remains in $\mathcal{H}_p^{++}$ along the iterations. Beyond the introduction of the Riemannian gradient descent presented in Section 3.2, this flexible framework extends to many other algorithms:

- Conjugate gradient, or BFGS-type algorithms, require the notion of Riemannian vector transport operator [26, Section 10.3], which allows to transport tangent vectors between tangent spaces at different points.
- Second-order methods, such as trust region or Newton methods require the definition of the Riemannian Hessian [26, Section 5.5].
- For large dimensional datasets, stochastic optimization methods can also be extended to the Riemannian setting [19, 99, 23].

A last remark is that in these algorithms, the metric is left as a choice that conditions the gradient and possibly the retraction. There are various options for $\mathcal{H}_p^{++}$ (cf. Section 2.1), with their respective pros and cons. It is still noticed that the gradient obtained from the Fisher information metric, also referred to as the natural gradient [2], is generally experienced to yield a faster convergence when dealing with a cost function related to the statistical model of the data (cf. examples in [43, 36]).

3.3.2 Geodesic convexity on $\mathcal{H}_p^{++}$

The classical results on the existence and uniqueness of the MLEs [73, Theorems 6 and 7] do not directly extend to the formulation in (32), so one might inquire about the optimally of the solution obtained by reaching a local minimum of this problem. In this scope, the Riemannian perspective offers some answers by generalizing the property of convexity. First, we recall that the geometry induced by the Fisher information metric (18) yields geodesic curves $\gamma(t)$ as defined in Theorem 4 between any two points $\Sigma_0, \Sigma_1 \in \mathcal{H}_p^{++}$. A function f is then said to be geodesically convex (g-convex) on $\mathcal{H}_p^{++}$ if $\forall \Sigma_0, \Sigma_1 \in \mathcal{H}_p^{++}$, it satisfies the inequality

$$f(\gamma(t)) \leq (1-t)f(\Sigma_0) + tf(\Sigma_1), \ \forall t \in [0,1]. \tag{33}$$

If the above inequality is strict, the function is then said to be strictly g-convex. The g-convexity enjoys properties similar to those of the convexity in the standard Euclidean case, in particular:

Theorem 6 *(Global minimizer of g-convex functions on $\mathcal{H}_p^{++}$).*
Let $f : \mathcal{H}_p^{++} \to \mathbb{R}$ be g-convex as defined in (33)*, then any local minimum of f over $\mathcal{H}_p^{++}$ is a global minimum. Furthermore, if f is strictly g-convex, this global minimum is unique.* □

This property offers an alternate proof for the uniqueness of MLEs as in (24) [72], and had practical impacts for the design of regularized covariance matrix estimators as in (32): many examples of penalty functions (with various regularization effects) can be found in the overviews in [98, 40], and the references [9, 97, 96, 98, 72, 40].

4 Intrinsic CramÃľr-Rao Bound for covariance matrix estimation

The CramÃľr-Rao inequality is a staple tool in statistics that characterizes the optimal mean-squared error an unbiased estimator can reach given a model and setup [55]. This tool can either be used to validate estimation procedures, or to design systems so that a certain level of accuracy is guaranteed to be theoretically reachable. While the Euclidean formulation of this inequality was briefly introduced in Section 1.1, the so-called *intrinsic* Cramér-Rao bounds extend it to parameters living in a manifold, and for any chosen Riemannian metric. This perspective is especially interesting as: *i*) some metrics can be more meaningful to assess the estimation performance in a given application; *ii*) a suitable Riemannian geometry (as opposed to the Euclidean one) can reveal hidden properties that make the bound more informative (such as curvature terms, intrinsic biases, etc.). First, Section 4.1 introduces the background on intrinsic CramÃľr-Rao bound from [85], where the C-CES model is used as a driving example. We also refer the reader to [25, Chapter 6] and the reference [14], for more details on the topic. Then, CramÃľr-Rao bounds are derived for various distances in the context of covariance matrix estimation within C-CES distributions [28] in Section 4.2.

4.1 Introduction to intrinsic CramÃľr-Rao bounds

This subsection will present tools that can be applied to any chosen Riemannian geometry on $\mathcal{H}_p^{++}$. The needed objects are the Riemannian metric, logarithm mapping and square of the distance, which are denoted $\prec \cdot, \cdot \succ_{\cdot}$, $\log_{\cdot}(\cdot)$ and $\mathfrak{d}^2(\cdot,\cdot)$, respectively. As in Section 3, we consider the problem of estimating the matrix Σ from a given n-sample $\{\mathbf{x}_i\}_{i=1}^n$ assumed to be iid from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$. We denote $\hat{\Sigma}$ an estimator of this parameter; e.g. the MLE presented in Section 3. We then consider the evaluation of the performance of such estimator $\hat{\Sigma}$. To do so, we exploit the chosen Riemannian metric $\prec \cdot, \cdot \succ_{\Sigma}$. Such metric can, for example, be the Fisher information one (18), or one of the many other options from the literature [93]. The performance criterion is the resulting square of the Riemannian distance, i.e., the error is measured through $\mathfrak{d}^2(\Sigma, \hat{\Sigma})$. The intrinsic CramÃľr-Rao theory from [85] then allows us to obtain a lower bound on the expectation of this error for any unbiased estimator $\hat{\Sigma}$. Eventually, this retrieves the well-known inequality " $\mathbf{C} \geq \mathbf{F}^{-1}$," with $\mathbf{C} \in \mathbb{R}^{p^2 \times p^2}$ being the covariance matrix of the estimation error and $\mathbf{F} \in \mathbb{R}^{p^2 \times p^2}$ being the Fisher information matrix, where $p^2 = \dim(\mathcal{H}_p^{++})$. However, these parameters have different definitions due to the specific nature of the considered objects. The point of this section is to briefly present the key ingredients to obtain such inequality and the corresponding main theorem.

First, we need to generalize the notion of estimation error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^{p^2}$ to the Riemannian context. Notice that, in the Euclidean case, such vector is generally con-

structed by vectorizing the entry-wise subtraction of the covariance matrix Σ to its estimate $\hat{\Sigma}$, i.e., $\boldsymbol{\varepsilon}^{\mathcal{E}} = \text{vech}(\hat{\Sigma} - \Sigma)$, where $\text{vech}(\cdot)$ denotes the half-vectorization operator. As it happens, from a Riemannian geometry point of view, $\hat{\Sigma} - \Sigma$ corresponds to the Euclidean logarithm mapping at Σ. Therefore, the Riemannian logarithm $\mathfrak{log}_{\Sigma}(\hat{\Sigma})$ provides a natural way to extend the error to any geometry. It is indeed an element of the tangent space $T_{\Sigma}\mathcal{H}_p^{++}$ of Σ that "points towards" $\hat{\Sigma}$, and whose norm corresponds to the Riemannian distance. It remains to actually get an error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^{p^2}$ from $\mathfrak{log}_{\Sigma}(\hat{\Sigma})$. To do so, we leverage a basis $\{\boldsymbol{\xi}_q\}_{q=1}^{p^2}$ of $T_{\Sigma}\mathcal{H}_p^{++} \simeq \mathcal{H}_p$ that is orthonormal with respect to the chosen metric $\prec \cdot, \cdot \succ_{\Sigma}$. In practice, such a basis can be obtained either analytically from mathematical calculations or numerically, thanks to the Gram-Schmidt orthonormalization process. This basis yields the decomposition

$$\mathfrak{log}_{\Sigma}(\hat{\Sigma}) = \sum_{q=1}^{p^2} \varepsilon_q \boldsymbol{\xi}_q, \tag{34}$$

and we denote $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_{p^2}] \in \mathbb{R}^{p^2}$ the corresponding coordinates error vector, obtained as

$$\varepsilon_q = \prec \mathfrak{log}_{\Sigma}(\hat{\Sigma}), \boldsymbol{\xi}_q \succ_{\Sigma} . \tag{35}$$

Moreover, the norm of this vector corresponds to the Riemannian distance between Σ and $\hat{\Sigma}$, i.e.,

$$\mathfrak{d}^2(\Sigma, \hat{\Sigma}) = \prec \mathfrak{log}_{\Sigma}(\hat{\Sigma}), \mathfrak{log}_{\Sigma}(\hat{\Sigma}) \succ_{\Sigma} = \|\boldsymbol{\varepsilon}\|_2^2, \tag{36}$$

which will be instrumental in the next derivations. The basis $\{\boldsymbol{\xi}_q\}_{q=1}^{p^2}$ also yields a Fisher information matrix $\mathbf{F}$, with entries

$$\mathbf{F}_{q\ell} = \langle \boldsymbol{\xi}_q, \boldsymbol{\xi}_\ell \rangle_{\Sigma}^{\text{FIM}}. \tag{37}$$

The matrix $\mathbf{F}$ represents the Fisher information metric of Theorem 2 according to this system of coordinates. Then, from [85, Corrolary 2], we obtain Theorem 7.

Theorem 7 (Intrinsic CramÃľr-Rao bound) *Let $\Sigma \in \mathcal{H}_p^{++}$. Let $\{\mathbf{x}_i\}_{i=1}^n$ a iid n-sample from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$. Let $\hat{\Sigma}$ an unbiased estimator of Σ with corresponding error vector $\boldsymbol{\varepsilon}$ defined in* (35)*. Then*

$$\mathbf{C} = E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] \geq \mathbf{F}^{-1} + \text{ curvature terms},$$

where $\mathbf{F}$ is the Fisher information matrix in (37) *and the curvature terms – which are not detailed here – depend on the Riemannian curvature tensor corresponding to the chosen geometry and on $\mathbf{F}$; see [85, 24, 25] for further details.*

In practice, the curvature terms can usually be neglected in Theorem 7. Furthermore, taking the trace of the inequality yields the desired result, i.e.,

$$\text{E}\left[\mathfrak{d}^2(\Sigma, \hat{\Sigma})\right] \geq \text{Tr}(\mathbf{F}^{-1}). \tag{38}$$

It offers a bound that can be derived for any chosen Riemannian distance $\mathfrak{d}^2(\cdot,\cdot)$ (and corresponding metric $\prec \cdot, \cdot \succ_{\cdot}$).

Remark The inequality in Theorem 7 interestingly takes into account the curvature of the manifold, which, for $\mathcal{H}_p^{++}$, only depends on the chosen metric. In the Euclidean case, such curvature term is null, and we recover the standard CramÃľr-Rao inequality. We also notice that the theorem in [85] also incorporates an intrinsic bias terms, which was excluded here for the sake of conciseness. This intrinsic bias (expectation of the Riemannian logarithm) depends on the estimator and the chosen metric, and can reveal unexpected properties. A main example is that the MLE of the covariance matrix of the Gaussian model appears unbiased in the Euclidean setting, but is, in fact, biased when using the Fisher information metric [85]. Such analysis thus opens prospects for improved estimation from the intrinsic perspective.

4.2 Bounds for various matrix distances in C-CES distributions

This Section presents the derivation of special cases of Theorem 7 when considering various usual metrics. Hence, it yields intrinsic CramÃľr-Rao bounds for the problem of covariance matrix estimation in C-CES distributions for the corresponding Riemannian distances. Since the Fisher information metric is already obtained in Theorem 2, the derivation boils down to the following steps:

a) Selecting the performance metric $\prec \cdot, \cdot \succ_\Sigma$ and computing $\{\boldsymbol{\xi}_q\}_{q=1}^{p^2}$, a corresponding orthonormal basis of $T_\Sigma \mathcal{H}_p^{++}$;
b) Computing the elements of the Fisher information matrix with this basis, according to (37);
c) Inverting the Fisher information matrix, applying Theorem 7, then (38).

These operations are conducted in the following for the Euclidean metric, the so-called natural Riemannian metric (the affine invariant metric (18) with $\alpha = 1$ and $\beta = 0$), and the Fisher-Rao metric of the assumed model (i.e., the metric of Theorem 2: (18) with $\alpha = \alpha_g$ and $\beta = \beta_g$, where α_g and β_g are defined in (9)). In order for the chosen values of α and β to be clear, in this subsection, the metric (18) is denoted $\langle \cdot, \cdot \rangle_{\cdot}^{(\alpha,\beta)}$ and the distance of Theorem 5 is denoted $\delta^2_{(\alpha,\beta)}(\cdot,\cdot)$.

4.2.1 Euclidean distance

We first recall the elementary tools of the Euclidean metric for $\mathcal{H}_p^{++}$:

$$\begin{aligned} &\text{Metric:} && \langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_\Sigma^{\mathcal{E}} = \mathfrak{Re}(\mathrm{Tr}(\boldsymbol{\xi}\boldsymbol{\eta})) \\ &\text{Logarithm:} && \log_\Sigma^{\mathcal{E}}(\hat{\Sigma}) = \hat{\Sigma} - \Sigma \\ &\text{Distance:} && \delta^2_{\mathcal{E}}(\Sigma, \hat{\Sigma}) = \|\hat{\Sigma} - \Sigma\|_2^2. \end{aligned} \tag{39}$$

A basis of the tangent space $T_\Sigma \mathcal{H}_p^{++}$ that is orthonormal with respect to the metric in (39) can be obtained as follows:

1. For $1 \leq i \leq p$, $\boldsymbol{\xi}_{ii}^{\mathcal{E}}$ is a $p \times p$ symmetric matrix whose i^{th} diagonal element is one, zeros elsewhere
2. For $1 \leq i < j \leq p$, $\boldsymbol{\xi}_{ij}^{\mathcal{E}}$ is a $p \times p$ symmetric matrix whose ij^{th} and ji^{th} elements are both $1/\sqrt{2}$, zeros elsewhere.
3. For $1 \leq i < j \leq p$, $\bar{\boldsymbol{\xi}}_{ij}^{\mathcal{E}}$ is a $p \times p$ Hermitian matrix whose ij^{th} and ji^{th} elements are $\sqrt{-1}/\sqrt{2}$ and $-\sqrt{-1}/\sqrt{2}$, respectively, zeros elsewhere. □

To shorten notations, we simply denote this basis $\{\boldsymbol{\xi}_q^{\mathcal{E}}\}_{q=1}^{p^2}$, where the p^2 elements are ordered following items 1), 2), and 3). The squared Euclidean distance between an estimator $\hat{\Sigma}$ and the true value Σ also corresponds to the summed squared errors on the coordinates in this basis. We then have the following result:

Theorem 8 (Cramér-Rao bound on Euclidean distance) *Let $\hat{\Sigma}$ an unbiased estimator of Σ built from iid data $\{\mathbf{x}_i\}_{i=1}^n$ drawn from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$. The Euclidean distance between $\hat{\Sigma}$ and Σ is bounded in expectation as*

$$\mathbf{C}_{\mathcal{E}} = E\left[\delta_{\mathcal{E}}^2(\hat{\Sigma}, \Sigma)\right] \geq \operatorname{Tr}(\mathbf{F}_{\mathcal{E}}^{-1}),$$

where

$$[\mathbf{F}_{\mathcal{E}}]_{q\ell} = \Re\mathfrak{e}(n\alpha_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}_q^{\mathcal{E}}\Sigma^{-1}\boldsymbol{\xi}_\ell^{\mathcal{E}})) + n\beta_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}_q^{\mathcal{E}}) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}_\ell^{\mathcal{E}})),$$

with α_g and β_g defined in (9). *2.*

Proof The result is a direct application of Theorem 7 and (38) using the basis $\{\boldsymbol{\xi}_j^{\mathcal{E}}\}_{j=1}^{p^2}$. □

Remark that this corresponds to the Euclidean Cramér-Rao bounds obtained for several distributions in [42, 75, 16, 64]. Also notice that we retrieve the same result as [85, Theorem 5] for the Gaussian distribution, i.e., $\alpha_g = 1$ and $\beta_g = 0$.

4.2.2 Natural Riemannian distance

The natural Riemannian distance refers to the distance induced by the affine invariant metric (18) with the standard choice of coefficients $\alpha = 1$ and $\beta = 0$. The elementary tools for this metric for $\mathcal{H}_p^{++}$ are

$$\begin{aligned} &\text{Metric:} && \langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_\Sigma^{(1,0)} = \Re\mathfrak{e}(\operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta})) \\ &\text{Logarithm:} && \log_\Sigma(\hat{\Sigma}) = \Sigma \log(\Sigma^{-1}\hat{\Sigma}) \\ &\text{Distance:} && \delta_{(1,0)}^2(\Sigma, \hat{\Sigma}) = \| \log(\Sigma^{-1}\hat{\Sigma})\|_2^2. \end{aligned} \tag{40}$$

Recall that the full description of this geometry is provided in Section 2. A basis of the tangent space $T_\Sigma \mathcal{H}_p^{++}$ that is orthonormal with respect to the metric in (40) can

be obtained by coloring the canonical basis of previous section as

$$\boldsymbol{\xi}_q^{(1,0)} = \Sigma^{1/2}\boldsymbol{\xi}_q^{\mathcal{E}}\Sigma^{1/2}. \tag{41}$$

The whole basis is denoted $\{\boldsymbol{\xi}_q^{(1,0)}\}_{q=1}^{p^2}$. We then have the following result:

Theorem 9 (Cramér-Rao bound on natural Riemannian distance) *Let $\hat{\Sigma}$ an unbiased estimator of $\Sigma \in \mathcal{H}_p^{++}$ built from iid data $\{\mathbf{x}_i\}_{i=1}^n$ drawn from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$. The Riemannian distance between $\hat{\Sigma}$ and Σ is bounded in expectation as*

$$E\left[\delta_{(1,0)}^2(\hat{\Sigma}, \Sigma)\right] \geq \frac{1}{n}\left(\frac{p^2-1}{\alpha_g} + \frac{1}{\alpha_g + p\beta_g}\right), \tag{42}$$

with α_g and β_g defined in (9).

Proof Plugging the basis $\{\boldsymbol{\xi}_q^{(1,0)}\}_{q=1}^{p^2}$ of $T_\Sigma\mathcal{H}_p^{++}$ defined in (41) in (37) yields

$$\left[\mathbf{F}_{(1,0)}\right]_{q\ell} = \langle \boldsymbol{\xi}_q^{(1,0)}, \boldsymbol{\xi}_\ell^{(1,0)}\rangle_\Sigma^{\text{FIM}} = \Re\mathfrak{e}(n\alpha_g \operatorname{Tr}(\boldsymbol{\xi}_q^{\mathcal{E}}\boldsymbol{\xi}_\ell^{\mathcal{E}}) + n\beta_g \operatorname{Tr}(\boldsymbol{\xi}_q^{\mathcal{E}})\operatorname{Tr}(\boldsymbol{\xi}_\ell^{\mathcal{E}})).$$

Hence, from the relations

$$\operatorname{Tr}(\boldsymbol{\xi}_q^{\mathcal{E}}\boldsymbol{\xi}_\ell^{\mathcal{E}}) = \delta_{q\ell} \qquad \text{and} \qquad \operatorname{Tr}(\boldsymbol{\xi}_q^{\mathcal{E}})\operatorname{Tr}(\boldsymbol{\xi}_\ell^{\mathcal{E}}) = \begin{cases} 1 \text{ if } (q,\ell) \in [\![1,p]\!]^2 \\ 0 \text{ otherwise,} \end{cases}$$

we obtain the Fisher information matrix

$$\mathbf{F}_{(1,0)} = n\alpha_g \mathbf{I}_{p^2} + n\beta_g \begin{bmatrix} \mathbf{1}_{p\times p} & \mathbf{0}_{p\times p(p-1)} \\ \mathbf{0}_{p(p-1)\times p} & \mathbf{0}_{p(p-1)\times p(p-1)} \end{bmatrix},$$

which is expressed as $\mathbf{F}_{(1,0)} = n\alpha_g\mathbf{I}_{p^2} + np\beta_g\mathbf{v}\mathbf{v}^T$ with unitary vector $\mathbf{v} = \frac{1}{\sqrt{p}}\left[\ \mathbf{1}_p \mid \mathbf{0}_{p(p-1)}\ \right]$, i.e. $\mathbf{v}^T\mathbf{v} = 1$. Hence, the inverse of the Fisher information matrix can be obtained by the Sherman-Morrison formula. In particular, its vector of eigenvalues can be directly identified as $\frac{1}{n}\left[(\alpha_g + p\beta_g)^{-1}, \alpha_g^{-1}, \ldots, \alpha_g^{-1}\right]$ and summed to obtain its trace. Theorem 7 and (38) are then applied to conclude. □

Remark Contrarily to the Euclidean case of Theorem 8, the bound on the natural Riemannian distance in Theorem 9 does not depend on the parameter Σ. This is generally a desirable property, as it offers an interpretation grounded solely on intrinsic dimensions of the problem. Additionally, simulation examples in Section 4.3 show that assessing the error with such criterion (that is more in accordance with the nature of the parameter) can also reveal unexpected properties of the estimates.

4.2.3 Fisher-Rao distance

The Fisher-Rao distance refers to the geodesic distance associated with the Fisher information metric (cf. Section 2.3). A subtlety is that we voluntarily omit the

dependency on n of the Fisher information metric of Theorem 2, i.e., the bound will be obtained for using a generic metric in (18) with $\alpha = \alpha_g$ and $\beta = \beta_g$. This distinction has two main reasons: i) it appears more logical to evaluate performance with a distance whose expression does not vary with the sample support of the scenario n; ii) this allows us to also stress that, though identical, two metrics play a separate role in the derivations: one is inherent to the statistical model, the other is a choice made to measure estimation accuracy. Hence, the elementary tools on $\mathcal{H}_p^{++}$ are

$$\begin{aligned}
&\text{Metric:} && \langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\Sigma}^{(\alpha_g,\beta_g)} = \Re\mathfrak{e}(\alpha_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}\Sigma^{-1}\boldsymbol{\eta}) + \beta_g \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\xi}) \operatorname{Tr}(\Sigma^{-1}\boldsymbol{\eta})) \\
&\text{Logarithm:} && \log_{\Sigma}(\hat{\Sigma}) = \Sigma \log(\Sigma^{-1}\hat{\Sigma}) \\
&\text{Distance:} && \delta^2_{(\alpha_g,\beta_g)}(\Sigma, \hat{\Sigma}) = \alpha_g \| \log(\Sigma^{-1}\hat{\Sigma})\|_2^2 + \beta_g (\log\det(\Sigma^{-1}\hat{\Sigma}))^2.
\end{aligned} \tag{43}$$

Recall that full details on this geometry are provided in Section 2. Contrary to previous geometries, since the considered metric is the Fisher information one, we do not actually need to compute a basis of the tangent space $T_{\Sigma}\mathcal{H}_p^{++}$ to obtain the bound. However, notice that if needed, such a basis can be obtained using the Gram-Schmidt orthogonalization process. In this case, the CramÃľr-Rao bound is:

Theorem 10 (CramÃľr-Rao bound on Fisher-Rao distance) *Let $\hat{\Sigma}$ be an unbiased estimator of $\Sigma \in \mathcal{H}_p^{++}$ built from iid data $\{\mathbf{x}_i\}_{i=1}^n$ drawn from $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma, g)$. The Fisher-Rao distance between $\hat{\Sigma}$ and Σ is bounded in expectation as*

$$E\left[\delta^2_{(\alpha_g,\beta_g)}(\hat{\Sigma}, \Sigma)\right] \geq \frac{p^2}{n}.$$

Proof By definition, we have $\langle \cdot, \cdot \rangle_{\Sigma}^{\mathrm{FIM}} = n \langle \cdot, \cdot \rangle_{\Sigma}^{(\alpha_g,\beta_g)}$. Hence, since the basis of interest, denoted $\{\boldsymbol{\xi}_q^{(\alpha_g,\beta_g)}\}_{q=1}^{p^2}$, is orthonormal according to $\langle \cdot, \cdot \rangle_{\Sigma}^{(\alpha_g,\beta_g)}$, it follows that $\mathbf{F}_{(\alpha_g,\beta_g)} = n\mathbf{I}_{p^2}$. The trace of its inverse is therefore p^2/n and the proof is concluded by applying Theorem 7 and (38). □

We notice that Theorems 9 and 10 coincide in the Gaussian case ($\alpha_g = 1$ and $\beta_g = 0$).

Remark Theorem 10 actually exemplifies a more of universal result, which illustrates that the Fisher-Rao distance is the most in accordance with the underlying statistical model. Indeed, the proof strategy of Theorem 10 holds for any geometry induced by a statistical model (parameter manifold and probability density function). Thus, the Fisher-Rao distance will always be bounded by a ratio between the intrinsic problem dimension and the number of samples.

4.3 Simulation examples

This section illustrates the results of Theorems 8-10 for the multivariate t-distribution (cf. example of Section 1.1), and various covariance matrix estimators. In the following, the scatter matrix is built as a $p \times p$ (with $p = 10$) Toeplitz matrix $[\Sigma_T]_{ij} = \rho^{|i-j|}$ with $\rho = {}^{0.9(1+\sqrt{-1})}/\sqrt{2}$. For samples distributed as $\mathbf{x} \sim$ C-CES$(\mathbf{0}, \Sigma_T, g_d)$, where g_d is the density generator of the t-distribution with d degrees of freedom, we study the performance of the following estimators of Σ_T:

- SCM: the usual sample covariance matrix, defined as $\hat{\Sigma}_{\text{SCM}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^H$.
- MLE: The estimator $\hat{\Sigma}_{\text{MLE}}$ defined in (26) using the appropriate function $\psi(t) = -\phi(t)$, with ϕ defined in (12).
- Mismatched MLE: the M-estimator $\hat{\Sigma}_{\text{m-MLE}}$ constructed as the MLE, except that the parameter d is different from the true parameter. Here, $d = 10$ is set regardless of the underlying distribution.

These performances are evaluated with respect to n (n ranging from 11 to 10^3) through the mean squared distances $\delta^2_{\mathcal{E}}$, $\delta^2_{(1,0)}$ and $\delta^2_{(\alpha_{g_d},\beta_{g_d})}$ (evaluated on 10^4 Monte-Carlo simulations) and are compared to the corresponding CramÃľr-Rao lower bounds from Theorems 8-10.

The left column of Figure 8 displays the results for a t-distribution with $d = 100$ degrees of freedom. Notice that, in this case, data almost follow a Gaussian distribution (it is usually admitted that $d > 30$ allows to assume Gaussianity of the data). In this setting, $\hat{\Sigma}_{\text{MLE}} \simeq \hat{\Sigma}_{\text{SCM}}$ so these estimators reach similar performances. For all performance measurements (different distances), the mismatched MLE appears not efficient at high sample support, which is due to a bias induced on the scale through the wrong choice of parameter d. Also, $\alpha \simeq 1$ and $\beta \simeq 0$, so $\langle \cdot, \cdot \rangle_{\cdot}^{(1,0)}$ and $\langle \cdot, \cdot \rangle_{\cdot}^{(\alpha_{g_d},\beta_{g_d})}$ generate almost identical distances and corresponding bounds, as observed in Figure 8. Interestingly, as noted in [85], these performance criteria show that the studied estimators are not efficient at low sample support. The natural metric is able to reflect some empirical results in terms of application – the SCM is known to provide an inaccurate estimation at low sample support –, while the Euclidean metric is apparently not, i.e., the CramÃľr-Rao bound and MSE on the Euclidean metric appear non-informative here.

The right column of Figure 8 displays the same results for a t-distribution with $d = 3$ degrees of freedom. Here, the distribution is heavy tailed and the SCM, as well as the mismatched MLE, fail to provide an accurate estimator of the scatter matrix. In this case, the study of the Euclidean metric reveals that the MLE is not efficient at low sample support, however it converges to the bound as n grows. We notice that the convergence towards this regime appears to be slower through the study of the natural and C-CES Fisher-Rao metric, which may be an interesting point in order to quantify the number of samples needed to achieve good performance in terms of application purpose.

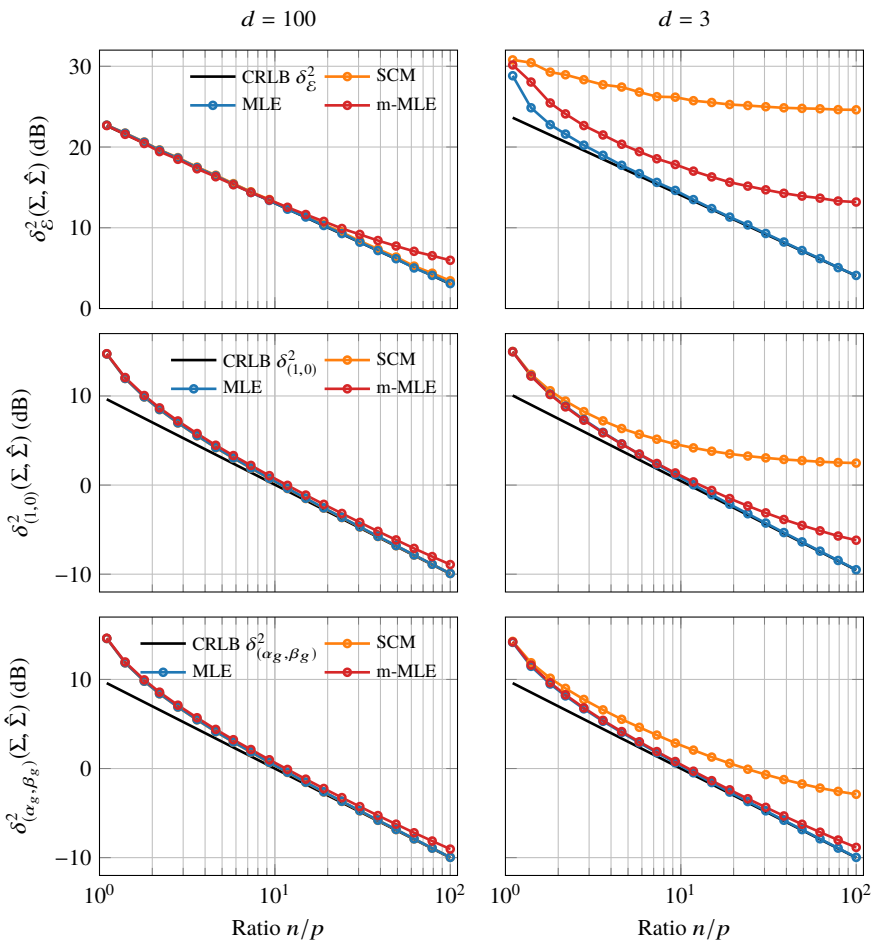

Fig. 8: From top to bottom: Euclidean, Natural, CES Fisher-Rao CRLB and mean squared distance scatter matrix for t-distribution with d degrees of freedom versus n/p for $p = 10$. On the left, $d = 100$ (close to Gaussian case) and, on the right, $d = 3$.

5 Riemannian classification with the Fisher-Rao distance

Classification is a ubiquitous task in machine learning. From a statistical point of view, the problem generally consists of attributing a class to each sample (or batch of samples) from an unlabelled mixture of different distributions. The Fisher-Rao geometry provides a tool that can be efficiently leveraged in this context: as most classification methods are based on the Euclidean distance between samples, these can be transposed to the Riemannian setting by using the Fisher-Rao distance on the statistical feature space (i.e., the parameters of the assumed model). Such

transposition is often beneficial as it leverages a metric that is in accordance with the model (e.g., it can account for its natural geometric invariance). In this regard, Section 5.1 presents a generic framework driven by the Fisher-Rao geometry. An example based on CES models and the nearest centroÃŕd classifier is derived in Section 5.2 and applied to EEG recordings in Section 5.3.

5.1 A Fisher-Rao Riemannian classification framework

The use of statistical features (or descriptors) is common in batch sample classification, as these tend to be more discriminative than raw data. Interestingly, when assuming a statistical model for the batches, the model parameters appear as a natural choice for such statistical features, and the Fisher-Rao distance as a natural tool to compare them. For example, assuming two C-CES models with the same probability density function f, but different parameters Σ_1 and Σ_2, the Fisher-Rao distance (cf. Theorem 5 and (43)) acts distance between statistical models through the following relation:

$$\underbrace{\delta_{\mathrm{FR}}(f(\mathbf{x}|\Sigma_1), f(\mathbf{x}|\Sigma_2))}_{\text{dist. between models}} \stackrel{\text{def}}{=} \underbrace{\delta_{\mathrm{FR}}(\Sigma_1, \Sigma_2)}_{\text{FR-dist. between parameters}}. \tag{44}$$

In practice, we handle empirical distributions (i.e., batches of samples), so this distance can be evaluated as:

$$\underbrace{\hat{\delta}_{\mathrm{FR}}(\{\mathbf{x}_{i,1}\}_{i=1}^n, \{\mathbf{x}_{i,2}\}_{i=1}^n)}_{\text{dist. between batches}} \stackrel{\text{def}}{=} \underbrace{\delta_{\mathrm{FR}}(\hat{\Sigma}_1, \hat{\Sigma}_2)}_{\text{FR-dist. between estimated covariances}}, \tag{45}$$

where $\{\mathbf{x}_{i,1}\}_{i=1}^n$ (resp. $\{\mathbf{x}_{i,2}\}_{i=1}^n$) denotes a sample batch, and $\hat{\Sigma}_1$ (resp. $\hat{\Sigma}_2$) denotes an estimate of its covariance matrix, such as the maximum likelihood estimator presented in Section 3. From this perspective, a batch classification problem then turns into a problem of classifying covariance matrices on $\mathcal{H}_p^{++}$. Such a task can be achieved by using a standard classification algorithm in which criteria and objects are carefully transposed according to the Fisher-Rao distance (rather than the Euclidean one). For examples related to this setup: the Riemannian nearest centroÃŕd (or minimum distance to mean) classifier [11, 94]; the Riemannian K-means on $\mathcal{H}_p^{++}$ was, e.g., used in [32, 46], Kernel methods based on Riemannian distances were studied in [12, 49, 50], and Riemannian Gaussian mixture models on $\mathcal{H}_p^{++}$ were proposed in [81, 82]. The following section presents the Riemannian counterpart of the nearest centroid classifier for $\mathcal{H}_p^{++}$.

Remark Beyond C-CES models, the presented framework generalizes to a generic (model-driven) Riemannian classification methodology, which can be summarized as follows: *i*) Model selection: we assume an underlying statistical model, whose parameters should differ between classes; *ii*) Statistical Feature extraction: we es-

timate the corresponding parameters for each batch; *iii*) Riemannian classification: the extracted features are classified by leveraging the Fisher-Rao distance.

5.2 Nearest centroÃŕd classifier on $\mathcal{H}_p^{++}$ for C-CES models

The Riemannian center of mass corresponding to the framework discussed in Section 5.1 when assuming a Gaussian model has been the reference method to classify electroencephalography (EEG) recordings for the past decade [11]. This section extends this methodology to the C-CES distributions, and presents the necessary tools to compute the Riemannian center of masses on $\mathcal{H}_p^{++}$.

Formally, we focus here on the supervised classification of batches of data. Formally, given an unknown batch of a n-sample $\{\mathbf{x}_i\}_{i=1}^n$ and z fixed classes, a classifier $C : (\mathbb{C}^p)^n \to [\![1, \ldots, z]\!]$ infers the class label $y \in [\![1, \ldots, z]\!]$, i.e.,

$$y = C\left(\{\mathbf{x}_i\}_{i=1}^n\right). \tag{46}$$

To provide accurate results, the classifier C is trained on m batches of n samples $\{\{\mathbf{x}_{i,j}\}_{i=1}^n, y_j\}_{j=1}^m$ associated to known class labels $y_j \in [\![1, \ldots, z]\!]$. In practice, one usually aims to evaluate the accuracy of a classifier on some dataset $\mathcal{T}$. To do so, the dataset is split into training and test sets, denoted $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$, respectively. The classifier C is trained on $\mathcal{T}_{\text{train}}$ and prediction is performed on the testing set $\mathcal{T}_{\text{test}}$. Predicted labels are then compared to actual labels, which yields the accuracy of C on the considered dataset. Notice that there are different ways to build $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$ from $\mathcal{T}$, see e.g., the documentation of scikit-learn [76] for more details.

For the model selection step, we consider that each batch $\{\mathbf{x}_{i,j}\}_{i=1}^n$ is distributed according to $\mathbf{x} \sim \text{C-CES}(\mathbf{0}, \Sigma_j, g)$. The statistical parameter extraction step is then performed by maximum likelihood estimation on each batch (cf. Section 3). From there, the feature classification problem is set as $\mathcal{T} = \{\hat{\Sigma}_j, y_j\}_{j=1}^m$ on $\mathcal{H}_p^{++}$. We then exploit the Fisher-Rao distance δ of C-CES distribution defined in Theorem 5 to generalize the nearest centroÃŕd classifier, also referred to as minimum distance to mean (MDM) classifier, to $\mathcal{H}_p^{++}$, . This classification algorithm consists of two steps:

- First, it computes the center of mass of each class, also called class center, from covariance matrices in the training set $\mathcal{T}_{\text{train}}$.
- Then, it assigns the label of the closest class center to each covariance in $\mathcal{T}_{\text{test}}$. □

Since the covariance matrices lie on the Riemannian manifold $\mathcal{H}_p^{++}$, the geodesic distance δ from Theorem 5 is leveraged in both steps.

We now detail the first step. For every class $y \in [\![1, \ldots, z]\!]$, one must compute the class center $\bar{\Sigma}^{(y)}$ from the training set $\mathcal{T}_{\text{train}}$. It is the center of mass of the set $\{\hat{\Sigma}_j \in \mathcal{T}_{\text{train}} : y_j = y\}$. We thus need to be able to compute the center of mass $\bar{\Sigma}$ of a set $\{\Sigma_j\}_{j=1}^m$ of matrices in $\mathcal{H}_p^{++}$ according to the Fisher-Rao distance in Theorem 5. Following [54], the Riemannian center of mass is defined in the following Definition 1.

Definition 1 (Riemannian center of mass on $\mathcal{H}_p^{++}$) The center of mass $\bar{\Sigma}^\star$ of $\{\Sigma_i\}_{i=1}^m$ on $\mathcal{H}_p^{++}$ is defined as the minimizer of the variance computed with the geodesic distance

$$\bar{\Sigma}^\star = \underset{\bar{\Sigma}\in\mathcal{H}_p^{++}}{\arg\min} \quad V(\bar{\Sigma}) \tag{47}$$

with $\quad V(\bar{\Sigma}) \overset{\text{def}}{=} \frac{1}{2m}\sum_{j=1}^m \delta^2(\bar{\Sigma},\Sigma_j)$.

Remark that if the Riemannian distance δ is replaced by its Euclidean counterpart, $\delta_{\mathcal{E}}(\Sigma,\hat{\Sigma}) = \|\Sigma-\hat{\Sigma}\|_2$, then the minimizer of V becomes the arithmetic mean $\bar{\Sigma} = \frac{1}{m}\sum_{j=1}^m \Sigma_j$. Unfortunately, for the Riemannian case, a closed-form solution of (47) remains unknown [65] except in very specific cases ($m = 2$, commuting matrices, ...). Hence, one must turn to an iterative optimization procedure. As in [78], we focus here on a Riemannian gradient descent on $\mathcal{H}_p^{++}$. Recall from Section 3 that, to employ this algorithm, we need to compute the Riemannian gradient of (47), choose a retraction and a step size rule. The Riemannian gradient of V was derived in [54, 65], and is provided in Proposition 3.

Proposition 3 (Riemannian gradient of V) *the Riemannian gradient* $\operatorname{grad} V(\bar{\Sigma})$ *of the variance V defined in* (47) *at $\bar{\Sigma}\in\mathcal{H}_p^{++}$ is*

$$\operatorname{grad} V(\bar{\Sigma}) = -\frac{1}{m}\sum_{j=1}^m \log_{\bar{\Sigma}}(\Sigma_j) = -\frac{1}{m}\sum_{j=1}^m \bar{\Sigma}^{1/2}\log(\bar{\Sigma}^{-1/2}\Sigma_j\bar{\Sigma}^{-1/2})\bar{\Sigma}^{1/2}. \tag{48}$$

Proof In [65], a technical proof directly deriving the distance of Theorem 5 for $\alpha = 1$ and $\beta = 0$ is provided. Here, we propose a more general Riemannian geometry proof, which do not depend on the distance or the manifold. The proved result is well-known and can for instance be found in [77] without proof. Given $\Sigma\in\mathcal{H}_p^{++}$, we aim to show that the gradient of the function $v(\bar{\Sigma}) = \frac{1}{2}\delta^2(\bar{\Sigma},\Sigma)$ is $\operatorname{grad} v(\bar{\Sigma}) = -\log_{\bar{\Sigma}}(\Sigma)$, where $\log_{.}(\cdot)$ is the Riemannian logarithm mapping corresponding to the Riemannian distance $\delta(\cdot,\cdot)$. Let $\bar{\Sigma}(t)$ the geodesic such that $\bar{\Sigma}(0) = \bar{\Sigma}$ and $\dot{\bar{\Sigma}}(0) = \boldsymbol{\xi}\in\mathcal{H}_p$. It follows that $\mathrm{D}\,v(\bar{\Sigma})[\boldsymbol{\xi}] = \frac{\mathrm{d}}{\mathrm{d}t}v(\bar{\Sigma}(t))\big|_{t=0}$. Let γ_t the geodesic joining $\bar{\Sigma}(t)$ to Σ. By construction, $H(s,t) = \gamma_t(s)$ is a variation of the geodesic γ_0 [41, Definition 3.24]. Furthermore, we have $\frac{\mathrm{d}}{\mathrm{d}t}v(\bar{\Sigma}(t))\big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}E(\gamma_t)\big|_{t=0}$, where $E(\gamma_t) = \frac{1}{2}\int_0^1\langle\dot{\gamma}_t(s),\dot{\gamma}_t(s)\rangle_{\gamma_t(s)}\mathrm{d}s$ is the energy of the geodesic γ_t. Let $Y(s)$ such that $H(t,s) = \exp_{\gamma_0(s)}(tY(s))$. From [41, Theorem 3.31], we get the first variation formula of energy

$$\frac{\mathrm{d}}{\mathrm{d}t}E(\gamma_t) = [\langle Y(s),\dot{\gamma}_0(s)\rangle_{\gamma_0(s)}]_0^1 - \int_0^1\langle Y(s),\nabla_{\dot{\gamma}_0(s)}\dot{\gamma}_0(s)\rangle_{\gamma_0(s)}\mathrm{d}s.$$

Since γ_0 is a geodesic, $\nabla_{\dot{\gamma}_0(s)}\dot{\gamma}_0(s) = 0$. Hence, the second term vanishes. Moreover, $Y(1) = \frac{1}{t}\log_{\gamma_0(1)}(\gamma_t(1))$. Since $\gamma_0(1) = \gamma_t(1) = \Sigma$, $Y(1) = 0$. We also have $Y(0) = \frac{1}{t}\log_{\gamma_0(0)}(\gamma_t(0)) = \frac{1}{t}\log_{\bar{\Sigma}}(\bar{\Sigma}(t)) = \frac{1}{t}t\boldsymbol{\xi} = \boldsymbol{\xi}$. It follows that

$$\frac{\mathrm{d}}{\mathrm{d}t} E(\gamma_t) = -\langle \boldsymbol{\xi}, \dot{\gamma}_0(0) \rangle_{\gamma_0(0)} = \langle \boldsymbol{\xi}, -\log_{\bar{\Sigma}}(\Sigma) \rangle_{\bar{\Sigma}}.$$

We thus get $\mathrm{D}\, v(\bar{\Sigma})[\boldsymbol{\xi}] = \langle \boldsymbol{\xi}, -\log_{\bar{\Sigma}}(\Sigma) \rangle_{\bar{\Sigma}}$. The result follows by identification. One can then conclude the proof of the proposition by using the sum property of the gradient operator. □

Then, the most common choice for the retraction is to take the Riemannian exponential mapping (21). Furthermore, the stepsize in this case is often simply set to 1. It follows that, given some initialization $\bar{\Sigma}_{(0)}$, the sequence of iterates is

$$\begin{aligned} \bar{\Sigma}_{(k+1)} &= \exp_{\bar{\Sigma}_{(k)}} \left(\tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log_{\bar{\Sigma}_{(k)}}(\Sigma_j) \right) \\ &= \bar{\Sigma}_{(k)}^{1/2} \exp \left(\tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log(\bar{\Sigma}_{(k)}^{-1/2} \Sigma_j \bar{\Sigma}_{(k)}^{-1/2}) \right) \bar{\Sigma}_{(k)}^{1/2}. \end{aligned} \tag{49}$$

The variance V (47) is a strictly geodesically convex function over $\mathcal{H}_p^{++}$ [89]. Hence, its minimizer is unique.

Remark Notice that there is no dependence on α and β in (49). This means that the Riemannian center of mass according to the Fisher-Rao distance in Theorem 5 is the same for every C-CES distribution.

The computation of the class centers being solved, we now turn to the second step of the nearest centroÃŕd classifier: the assignment to a class y_j of each estimated covariance matrix $\hat{\Sigma}_j$ belonging to the test set $\mathcal{T}_{\text{test}}$. This is achieved by taking the class that corresponds to the minimal geodesic distance with respect to all class centers, i.e.,

$$y_j = \underset{y \in [\![1,\ldots,z]\!]}{\arg\min} \left\{ \delta^2(\bar{\Sigma}^{(y)}, \hat{\Sigma}_j) \right\}_{y \in [\![1,\ldots,z]\!]}. \tag{50}$$

The resulting nearest centroÃŕd classifier on $\mathcal{H}_p^{++}$ is summarized in Algorithm 1.

Algorithm 1: Nearest centroÃŕd classifier on $\mathcal{H}_p^{++}$

Input: A training set $\mathcal{T}_{\text{train}} = \{(\hat{\Sigma}_j, y_j)\}_{j=1}^{m_{\text{train}}}$ and a test set $\mathcal{T}_{\text{test}} = \{\hat{\Sigma}_j\}_{j=1}^{m_{\text{test}}}$.
Output: Predictions of the test set $\{y_j\}_{j=1}^{m_{\text{test}}}$.
# Training
for $y = 1$ *to* z **do**
  Compute the center of mass $\bar{\Sigma}^{(y)}$ of $\{\hat{\Sigma}_j \in \mathcal{T}_{\text{train}} : y_j = y\}$ with (49).
# Testing
for $j = 1$ *to* m_{test} **do**
  Assign $\hat{\Sigma}_j$ to the class with the nearest class center $\bar{\Sigma}^{(y)}$ with (50).

Remark The Gaussian assumption allows recover the classification algorithm from [11], as in this case: *i*) the maximum likelihood estimator is the sample covariance matrix $\hat{\Sigma}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{i,j} \mathbf{x}_{i,j}^H$; *ii*) $\alpha = 1$ and $\beta = 0$ in the Fisher-Rao distance δ of Theorem 5.

5.3 Application to EEG classification

One usually needs to classify EEG recordings in the context of brain-computer interfaces (BCI), where a subject interacts with a computer through brain activity. There are several paradigms for BCI based on EEG. The three main ones are: steady-states visually evoked potentials (SSVEP) [53], motor imagery (MI) [91], and event-related potentials (ERP) [5]. This example focuses on ERP data, where subjects are exposed to some stimuli (most often a visual one). These induce a signal response in the brain: the so-called P300, which is a positive wave occurring 300 ms after the stimulus. An ERP dataset consists in a set of trials separated into two classes: a target class (TA), for which the subject is exposed to a stimulus; and a non-target class (NT), for which there is no stimulus. More specifically, we consider the BNCI2014_009 dataset [5], which is available on the MOABB platform[5]. This dataset contains data from 10 subjects, with 3 sessions each. Data were acquired on 16 electrodes at 256 Hz and bandpass filtered between 0.1 Hz and 20 Hz. Recordings were then downsampled to 128 Hz. Each session of each subject contains 1728 trials of 0.8s: 288 target and 1440 non-target ones. Hence, each dataset (one session of one subject) yields $\mathcal{T} = \{\mathbf{X}_j, y_j\}_{j=1}^{m}$ in $\mathbb{R}^{p\times n} \times \{\mathrm{TA}, \mathrm{NT}\}$, where $p = 16$, $n = 102$, $m = 1728$ and $z = 2$.

To perform classification of ERPs, raw data are not directly used. Instead, following [13], augmented data are leveraged. Given the training set $\mathcal{T}_{\text{train}} = \{\mathbf{X}_j\}_{j=1}^{m_{\text{train}}}$, we compute the average target ERP with

$$\mathbf{P}_{\mathrm{TA}} = \frac{1}{m_{\mathrm{TA}}} \sum_{\substack{j=1 \\ y_j=\mathrm{TA}}}^{m_{\text{train}}} \mathbf{X}_j, \tag{51}$$

where m_{TA} is the number of target trials in the training set $\mathcal{T}_{\text{train}}$. From there, augmented trials are defined as

$$\tilde{\mathbf{X}}_j = \begin{bmatrix} \mathbf{P}_{\mathrm{TA}} \\ \mathbf{X}_j \end{bmatrix}. \tag{52}$$

Covariance matrices $\hat{\Sigma}_j$ are then estimated from these augmented trials both in the training and testing sets. Finally, the nearest centroÃŕd classifier in Algorithm 1 is applied on these augmented covariance. We compare two different versions here:

1. **Gaussian version**: covariance matrices estimated through the sample covariance matrix (SCM) and nearest centroÃŕd classifier employed with $\alpha = 1$ and $\beta = 0$.
2. ***t*-distribution version**: covariance matrices estimated with the MLE of the t-distribution with $\nu = 2.1$ degrees of freedom and nearest centroÃŕd classifier used with $\alpha = \frac{\nu+p}{\nu+p+2}$ and $\beta = \alpha - 1$.

Achieved accuracies are presented in Figure 9. One can observe that both classifiers feature very good performance on this dataset. One can further notice that they have very similar performance. Indeed, on average, the nearest centroÃŕd classifier

[5] https://github.com/NeuroTechX/moabb – A standard benchmark platform for BCI.

with the t-distribution is better by 0.12%. Considering that the SCM is much simpler to compute than the MLE of the t-distribution, one can argue that the nearest centroÃŕd classifier associated with the Gaussian distribution is more advantageous on this dataset. Due to the biological nature of the data, which can be expected to be noisy and contain a non-negligible amount of outliers, one could have expected that a heavy-tail distribution such as the t with $\nu = 2.1$ perform significantly better. However, the dataset at hand has been curated and the preprocessing has been designed for the Gaussian distribution to work well. Leveraging the t-distribution might be advanategous on real world non-curated data.

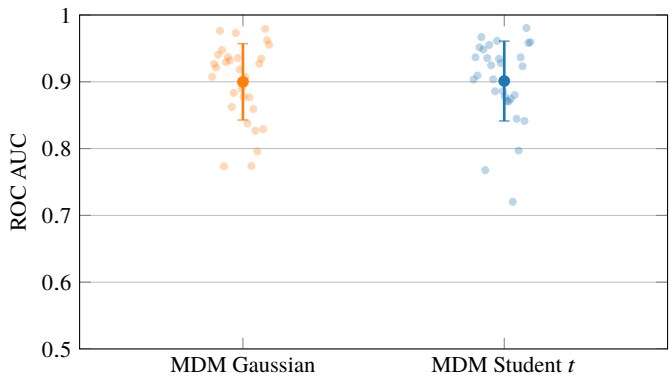

Fig. 9: AUC of ROC plots for the nearest centroÃŕd classifiers exploiting the Gaussian distribution (left) and Student t-distribution with $\nu = 2.1$ degrees of freedom (right) applied on the BNCI2014_009 dataset [5] (10 subjects, 3 sessions each).

6 Conclusion

This chapter presented the Fisher-Rao geometry of C-CES distributions, and its practical uses in statistical signal processing and machine learning. Remark that the methodology that consists in obtaining a Riemannian geometry from the Fisher information metric generalizes to any statistical model (assuming that the parameter space is a smooth manifold). Hence, the approaches presented in this introduction can extend to many other models and applications. Among other examples, such intrinsic analysis has been conducted for the estimation of rotations matrices [27] and for other Lie groups related to tracking problems [56, 57]. In other scopes more directly related to elliptical distributions, we can also mention that geometric tools were used for:

- **Structured covariance matrices**: In many applications, the covariance matrix is known to satisfy some form of structural constraint, that can be exploited to reduce

the dimension of the estimation problem (see, e.g., [98, 87]). Geometric tools can then be leveraged by expressing the constrained space as a sub-manifold of $\mathcal{H}_p^{++}$. For example: the Fisher information metric was used to obtain structured estimators in [62, 61]; A geometry of Toeplitz matrices was studied in [6]; A framework for in probabilistic component analysis (low-rank structured covariance matrices) in C-CES was proposed in [21]; Kronecker products preserve geodesic convexity [96], ans such structure was considered in online covariance matrix estimation in [22]; geometry and structured covariance have also been considered for blind source separation [20].

- **Non-centered models**: The geodesics and Fisher-Rao distance of the model $\mathbf{x} \sim \mathcal{CES}(\boldsymbol{\mu}, \Sigma)$ for the mean-and-covariance product manifold $\mathbb{C}^p \times \mathcal{H}_p^{++}$ remains intractable in the general case. Even for the Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, only special cases and approximations from geodesic triangles can be obtained [30, 90, 34]. Numerical methods to evaluate these geodesics and corresponding distances were proposed in [69, 68]. Concerning estimation problems, Riemannian optimization was leveraged for non-centered mixture of scaled Gaussian distributions (a sub-family of C-CES distributions) in [33, 36].
- **Mixture models**: Mixtures of C-CES can occur within the samples (the observation is the sum of multiple independent contributions) or within batches (the sample set aggregating multiple classes of C-CES). The within-sample mixture is typically used to cast robust models for probabilistic principal component analysis [31, 88, 47]. In this context, geometric tools were developed for low-rank scaled Gaussian signal corrupted by white Gaussian noise in [32]. The within-batch mixture corresponds to a typical sample-wise classification problem. For this purpose, g-convex relaxations for Gaussian mixture models were studied in [48].

As a final note, we also point out that *information geometry* also refers to a much broader field than the scope covered by this chapter [3, 4]. For comprehensive overviews of the many geometric structures behind families of probability distributions, we refer the readers to [66, 67].

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press (2009)
2. Amari, S.I.: Natural gradient works efficiently in learning. Neural computation **10**(2), 251–276 (1998)
3. Amari, S.i.: Information geometry and its applications, vol. 194. Springer (2016)
4. Amari, S.i.: Information geometry. Japanese Journal of Mathematics **16**, 1–48 (2021)
5. Aricò, P., Aloise, F., Schettini, F., Salinari, S., Mattia, D., Cincotti, F.: Influence of P300 latency jitter on event related potential-based brain–computer interface performance. Journal of neural engineering **11**(3), 035008 (2014)
6. Arnaudon, M., Barbaresco, F., Yang, L.: Riemannian medians and means with applications to radar signal processing. IEEE Journal of Selected Topics in Signal Processing **7**(4), 595–604 (2013)

7. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-Euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **56**(2), 411–421 (2006)
8. Atkinson, C., Mitchell, A.F.: Rao's distance measure. Sankhyā: The Indian Journal of Statistics, Series A pp. 345–365 (1981)
9. Auderset, C., Mazza, C., Ruh, E.A.: Angular gaussian and cauchy estimation. Journal of multivariate analysis **93**(1), 180–197 (2005)
10. BANGS II, W.J.: Array processing with generalized beam-formers. Yale University (1971)
11. Barachant, A., Bonnet, S., Congedo, M., Jutten, C.: Multiclass brain–computer interface classification by Riemannian geometry. IEEE Transactions on Biomedical Engineering **59**(4), 920–928 (2011)
12. Barachant, A., Bonnet, S., Congedo, M., Jutten, C.: Classification of covariance matrices using a Riemannian-based kernel for BCI applications. Neurocomputing **112**, 172–178 (2013)
13. Barachant, A., Congedo, M.: A plug&play P300 BCI using information geometry. arXiv preprint arXiv:1409.0107 (2014)
14. Barrau, A., Bonnabel, S.: A note on the intrinsic Cramér-Rao bound. In: Geometric Science of Information, pp. 377–386. Springer (2013)
15. Berkane, M., Oden, K., Bentler, P.M.: Geodesic estimation in elliptical distributions. Journal of Multivariate Analysis **63**(1), 35–46 (1997)
16. Besson, O., Abramovich, Y.I.: On the fisher Information Matrix for multivariate elliptically contoured distributions. IEEE Signal Processing Letters **20**(11), 1130–1133 (2013)
17. Bhatia, R.: Positive definite matrices. Princeton university press (2009)
18. Bhatia, R., Jain, T., Lim, Y.: On the Bures–Wasserstein distance between positive definite matrices. Expositiones Mathematicae **37**(2), 165–191 (2019)
19. Bonnabel, S.: Stochastic gradient descent on Riemannian manifolds. IEEE Transactions on Automatic Control **58**(9), 2217–2229 (2013)
20. Bouchard, F., Breloy, A., Ginolhac, G., Renaux, A.: A Riemannian approach to blind separation of t-distributed sources. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp. 965–969. IEEE (2021)
21. Bouchard, F., Breloy, A., Ginolhac, G., Renaux, A., Pascal, F.: A riemannian framework for low-rank structured elliptical models. IEEE Transactions on Signal Processing **69**, 1185–1199 (2021)
22. Bouchard, F., Breloy, A., Mian, A., Ginolhac, G.: On-line Kronecker product structured covariance estimation with Riemannian geometry for t-distributed data. In: 2021 29th European Signal Processing Conference (EUSIPCO), pp. 856–859. IEEE (2021)
23. Bouchard, F., Mian, A., Zhou, J., Said, S., Ginolhac, G., Berthoumieu, Y.: Riemannian geometry for compound gaussian distributions: Application to recursive change detection. Signal Processing **176**, 107716 (2020)
24. Boumal, N.: On intrinsic Cramér-Rao bounds for riemannian submanifolds and quotient manifolds. IEEE transactions on signal processing **61**(7), 1809–1821 (2013)
25. Boumal, N.: Optimization and estimation on manifolds. Ph.D. thesis, Université catholique de Louvain (2014)
26. Boumal, N.: An introduction to optimization on smooth manifolds. Cambridge University Press (2023)
27. Boumal, N., Singer, A., Absil, P.A., Blondel, V.D.: Cramér–Rao bounds for synchronization of rotations. Information and Inference: A Journal of the IMA **3**(1), 1–39 (2014)
28. Breloy, A., Ginolhac, G., Renaux, A., Bouchard, F.: Intrinsic cramér–rao bounds for scatter and shape matrices estimation in ces distributions. IEEE Signal Processing Letters **26**(2), 262–266 (2018)
29. Breloy, A., Ollila, E., Pascal, F.: Spectral shrinkage of Tyler's m-estimator of covariance matrix. In: 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 535–538. IEEE (2019)
30. Calvo, M., Oller, J.M.: An explicit solution of information geodesic equations for the multivariate normal model. Statistics & Risk Modeling **9**(1-2), 119–138 (1991)

31. Chen, T., Martin, E., Montague, G.: Robust probabilistic PCA with missing data and contribution analysis for outlier detection. Computational Statistics & Data Analysis **53**(10), 3706–3716 (2009)
32. Collas, A., Bouchard, F., Breloy, A., Ginolhac, G., Ren, C., Ovarlez, J.P.: Probabilistic PCA from heteroscedastic signals: geometric framework and application to clustering. IEEE Transactions on Signal Processing **69**, 6546–6560 (2021)
33. Collas, A., Bouchard, F., Breloy, A., Ren, C., Ginolhac, G., Ovarlez, J.P.: A Tyler-type estimator of location and scatter leveraging Riemannian optimization. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5160–5164. IEEE (2021)
34. 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: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5697–5701. IEEE (2022)
35. Collas, A., Breloy, A., Ginolhac, G., Ren, C., Ovarlez, J.P.: Robust geometric metric learning. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 1447–1451. IEEE (2022)
36. Collas, A., Breloy, A., Ren, C., Ginolhac, G., Ovarlez, J.P.: Riemannian optimization for non-centered mixture of scaled gaussian distributions. IEEE Transactions on Signal Processing (2023)
37. Couillet, R., Pascal, F., Silverstein, J.W.: The random matrix regime of MaronnaâĂŹs m-estimator with elliptically distributed samples. Journal of Multivariate Analysis **139**, 56–78 (2015)
38. Drašković, G., Breloy, A., Pascal, F.: On the asymptotics of maronna's robust PCA. IEEE Transactions on Signal Processing **67**(19), 4964–4975 (2019)
39. Drašković, G., Pascal, F.: New insights into the statistical properties of m-estimators. IEEE Transactions on Signal Processing **66**(16), 4253–4263 (2018)
40. Duembgen, L., Tyler, D.E.: Geodesic convexity and regularized scatter estimators. arXiv preprint arXiv:1607.05455 (2016)
41. Gallot, S., Hulin, D., Lafontaine, J.: Riemannian geometry. Springer (1990)
42. Greco, M., Gini, F.: Cramér-Rao lower bounds on covariance matrix estimation for complex elliptically symmetric distributions. IEEE Transactions on Signal Processing **61**(24), 6401–6409 (2013)
43. Han, A., Mishra, B., Jawanpuria, P.K., Gao, J.: On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. Advances in Neural Information Processing Systems **34**, 8940–8953 (2021)
44. Higham, N.J.: Functions of matrices: theory and computation. SIAM (2008)
45. Hippert-Ferrer, A., Bouchard, F., Mian, A., Vayer, T., Breloy, A.: Learning Graphical Factor Models with Riemannian optimization. arXiv preprint arXiv:2210.11950 (2022)
46. Hippert-Ferrer, A., El Korso, M.N., Breloy, A., Ginolhac, G.: Robust low-rank covariance matrix estimation with a general pattern of missing values. Signal Processing **195**, 108460 (2022)
47. Hong, D., Gilman, K., Balzano, L., Fessler, J.A.: HePPCAT: Probabilistic PCA for data with heteroscedastic noise. IEEE Transactions on Signal Processing **69**, 4819–4834 (2021)
48. Hosseini, R., Sra, S.: Matrix manifold optimization for gaussian mixtures. Advances in neural information processing systems **28** (2015)
49. Jayasumana, S., Hartley, R., Salzmann, M.: Kernels on Riemannian manifolds. In: Riemannian computing in computer vision, pp. 45–67. Springer (2016)
50. Jayasumana, S., Hartley, R., Salzmann, M., Li, H., Harandi, M.: Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 73–80 (2013)
51. Jeuris, B., Vandebril, R., Vandereycken, B.: A survey and comparison of contemporary algorithms for computing the matrix geometric mean. Electronic Transactions on Numerical Analysis **39**, 379–402 (2012)

52. Kai-Tai, F., Yao-Ting, Z.: Generalized multivariate analysis. Science Press Beijing and Springer-Verlag, Berlin (1990)
53. Kalunga, E.K., Chevallier, S., Barthélemy, Q., Djouani, K., Monacelli, E., Hamam, Y.: Online SSVEP-based BCI using Riemannian geometry. Neurocomputing **191**, 55–68 (2016)
54. Karcher, H.: Riemannian center of mass and mollifier smoothing. Communications on pure and applied mathematics **30**(5), 509–541 (1977)
55. Kay, S.M.: Fundamentals of statistical signal processing. Prentice Hall PTR (1993)
56. Labsir, S., Giremus, A., Yver, B., Benoudiba-Campanini, T.: Joint shape and centroid position tracking of a cluster of space debris by filtering on Lie groups. Signal Processing **183**, 108027 (2021)
57. Labsir, S., Renaux, A., Vilà-Valls, J., Chaumette, E.: Barankin, McAulay–Seidman and Cramér–Rao bounds on matrix lie groups. Automatica **156**, 111199 (2023)
58. Lang, S.: Differential and Riemannian manifolds. Springer (2012)
59. Lee, J.M.: Riemannian manifolds: an introduction to curvature. Springer (2006)
60. Maronna, R.A., Yohai, V.J.: Robust estimation of multivariate location and scatter. Wiley StatsRef: Statistics Reference Online (1976)
61. Mériaux, B., Ren, C., Breloy, A., El Korso, M.N., Forster, P.: Matched and mismatched estimation of Kronecker product of linearly structured scatter matrices under elliptical distributions. IEEE Transactions on Signal Processing **69**, 603–616 (2020)
62. Meriaux, B., Ren, C., El Korso, M.N., Breloy, A., Forster, P.: Robust estimation of structured scatter matrices in (mis) matched models. Signal Processing **165**, 163–174 (2019)
63. Micchelli, C.A., Noakes, L.: Rao distances. Journal of Multivariate Analysis **92**(1), 97–115 (2005)
64. Mitchell, A.E.: The information matrix, skewness tensor and a-connections for the general multivariate elliptic distribution. Annals of the Institute of Statistical Mathematics **41**, 289–304 (1989)
65. Moakher, M.: A differential geometric approach to the geometric mean of symmetric positive-definite matrices. SIAM journal on matrix analysis and applications **26**(3), 735–747 (2005)
66. Nielsen, F.: An elementary introduction to information geometry. Entropy **22**(10), 1100 (2020)
67. Nielsen, F.: The many faces of information geometry. Not. Am. Math. Soc **69**(1), 36–45 (2022)
68. Nielsen, F.: Fisher-Rao distance and pullback SPD cone distances between multivariate normal distributions. arXiv preprint arXiv:2307.10644 (2023)
69. Nielsen, F.: A simple approximation method for the Fisher–Rao distance between multivariate normal distributions. Entropy **25**(4), 654 (2023)
70. Ollila, E., Eriksson, J., Koivunen, V.: Complex elliptically symmetric random variables - generation, characterization, and circularity tests. IEEE Transactions on Signal Processing **59**(1), 58–69 (2011)
71. Ollila, E., Soloveychik, I., Tyler, D.E., Wiesel, A.: Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization. arXiv preprint arXiv:1608.08126 (2016)
72. Ollila, E., Tyler, D.E.: Regularized m-estimators of scatter matrix. IEEE Transactions on Signal Processing **62**(22), 6059–6070 (2014)
73. Ollila, E., Tyler, D.E., Koivunen, V., Poor, H.V.: Complex elliptically symmetric distributions: Survey, new results and applications. IEEE Transactions on signal processing **60**(11), 5597–5625 (2012)
74. Pascal, F., Chitour, Y., Quek, Y.: Generalized robust shrinkage estimator and its application to STAP detection problem. IEEE Transactions on Signal Processing **62**(21), 5640–5651 (2014)
75. Pascal, F., Renaux, A.: Statistical analysis of the covariance matrix MLE in K-distributed clutter. Signal Processing **90**(4), 1165–1175 (2010)
76. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. the Journal of machine Learning research **12**, 2825–2830 (2011)

77. Pennec, X.: Hessian of the Riemannian squared distance. Preprint. https://www-sop.inria.fr/members/Xavier.Pennec/AOS-DiffRiemannianLog.pdf (2017)
78. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. International Journal of computer vision **66**, 41–66 (2006)
79. Rao, C.R.: Information and accuracy attainable in the estimation of statistical parameters. Kotz S & Johnson NL (eds.), Breakthroughs in Statistics Volume i: Foundations and Basic Theory, 235–248 (1945)
80. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. In: Breakthroughs in Statistics: Foundations and basic theory, pp. 235–247. Springer (1992)
81. Said, S., Bombrun, L., Berthoumieu, Y., Manton, J.H.: Riemannian gaussian distributions on the space of symmetric positive definite matrices. IEEE Transactions on Information Theory **63**(4), 2153–2170 (2017)
82. Said, S., Hajri, H., Bombrun, L., Vemuri, B.C.: Gaussian distributions on Riemannian symmetric spaces: statistical learning with structured covariance matrices. IEEE Transactions on Information Theory **64**(2), 752–772 (2017)
83. Skovgaard, L.T.: A Riemannian geometry of the multivariate normal model. Scandinavian journal of statistics pp. 211–223 (1984)
84. Slepian, D.: Estimation of signal parameters in the presence of noise. Transactions of the IRE Professional Group on Information Theory **3**(3), 68–89 (1954)
85. Smith, S.T.: Covariance, subspace, and intrinsic Cramér-Rao bounds. IEEE Transactions on Signal Processing **53**(5), 1610–1630 (2005)
86. Sun, Y., Babu, P., Palomar, D.P.: Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms. IEEE Transactions on Signal Processing **62**(19), 5143–5156 (2014)
87. Sun, Y., Babu, P., Palomar, D.P.: Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. IEEE Transactions on Signal Processing **64**(14), 3576–3590 (2016)
88. Sun, Y., Breloy, A., Babu, P., Palomar, D.P., Pascal, F., Ginolhac, G.: Low-complexity algorithms for low rank clutter parameters estimation in radar systems. IEEE Transactions on Signal Processing **64**(8), 1986–1998 (2015)
89. Tang, M., Rong, Y., Chen, C.: Riemannian Lp center of mass for scatter matrix estimation in complex elliptically symmetric distributions. In: 2021 IEEE 24th International Conference on Information Fusion (FUSION), pp. 1–8 (2021). DOI 10.23919/FUSION49465.2021.9626967
90. Tang, M., Rong, Y., Zhou, J., Li, X.R.: Information geometric approach to multisensor estimation fusion. IEEE Transactions on Signal Processing **67**(2), 279–292 (2018)
91. Tangermann, M., Müller, K.R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K.J., Mueller-Putz, G., et al.: Review of the BCI competition IV. Frontiers in neuroscience p. 55 (2012)
92. Thanwerdas, Y.: Riemannian and stratified geometries on covariance and correlation matrices. Theses, Université Côte d'Azur (2022). URL https://hal.science/tel-03698752
93. Thanwerdas, Y., Pennec, X.: O (n)-invariant Riemannian metrics on SPD matrices. Linear Algebra and its Applications **661**, 163–201 (2023)
94. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on Riemannian manifolds. IEEE transactions on pattern analysis and machine intelligence **30**(10), 1713–1727 (2008)
95. Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. The Annals of Statistics pp. 234–251 (1987)
96. Wiesel, A.: Geodesic convexity and covariance estimation. IEEE Transactions on Signal Processing **60**(12), 6182–6189 (2012). DOI 10.1109/TSP.2012.2218241
97. Wiesel, A.: Unified framework to regularized covariance estimation in scaled gaussian models. IEEE Transactions on Signal Processing **60**(1), 29–38 (2012)
98. Wiesel, A., Zhang, T., et al.: Structured robust covariance estimation. Foundations and Trends® in Signal Processing **8**(3), 127–216 (2015)
99. Zhang, H., J Reddi, S., Sra, S.: Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. Advances in Neural Information Processing Systems **29** (2016)

100. Zhang, T., Cheng, X., Singer, A.: Marchenko-Pastur law for TylerâĂŹs and MaronnaâĂŹs m-estimators. arXiv preprint arXiv:1401.3424 (2014)
101. Zhang, T., Wiesel, A., Greco, M.S.: Multivariate generalized gaussian distribution: Convexity and graphical models. IEEE Transactions on Signal Processing **61**(16), 4141–4148 (2013)