Information Geometry (2018) 1:137–179
https://doi.org/10.1007/s41884-018-0014-4

RESEARCH PAPER

Wasserstein Riemannian geometry of Gaussian densities

Luigi Malagò[1] · **Luigi Montrucchio**[2] · **Giovanni Pistone**[3]

Received: 29 May 2018 / Revised: 9 October 2018 / Published online: 20 November 2018
© Springer Nature Singapore Pte Ltd. 2018

Abstract
The Wasserstein distance on multivariate non-degenerate Gaussian densities is a Riemannian distance. After reviewing the properties of the distance and the metric geodesic, we present an explicit form of the Riemannian metrics on positive-definite matrices and compute its tensor form with respect to the trace inner product. The tensor is a matrix which is the solution to a Lyapunov equation. We compute the explicit formula for the Riemannian exponential, the normal coordinates charts and the Riemannian gradient. Finally, the Levi-Civita covariant derivative is computed in matrix form together with the differential equation for the parallel transport. While all computations are given in matrix form, nonetheless we discuss also the use of a special moving frame.

Keywords Information geometry · Gaussian distribution · Wasserstein distance · Riemannian metrics · Natural gradient · Riemannian exponential · Normal coordinates · Levi-Civita covariant derivative · Optimization on positive-definite symmetric matrices

Mathematics Subject Classification 15B48 · 53C23 · 53C25 · 60D05

✉ Giovanni Pistone
giovanni.pistone@carloalberto.org

Luigi Malagò
malago@rist.ro

Luigi Montrucchio
luigi.montrucchio@unito.it

[1] Romanian Institute of Science and Technology, RIST, Str. Virgil Fulicea nr. 17, 400022 Cluj-Napoca, Romania

[2] Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Turin, Italy

[3] de Castro Statistics, Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Turin, Italy

1 Introduction

Given two probability measures ν_1 and ν_2 on $\mathbb{R}^n$, with finite second moments, consider the set $\mathscr{P}(\nu_1, \nu_2)$ of probability measures on the product sample space $\mathbb{R}^{2n}$, such that the two n-dimensional margins have the prescribed distributions, $X_1 \sim \nu_1$ and $X_2 \sim \nu_2$. The index

$$W^2 = \inf \left\{ \mathbb{E}_\mu \left[\|X_1 - X_2\|^2 \right] \middle| \mu \in \mathscr{P}(\nu_1, \nu_2) \right\}$$

as a measure of dissimilarity between distributions has been considered by many classical authors e.g., C. Gini, P. Levy, and M.R. Fréchet. There is considerable contemporary literature discussing the index W, which is usually called Wasserstein distance. E.g., the monograph by C. Villani [37]. We want also to mention Y. Brenier [9] and R.J. McCann [27].

There is an important particular case, where the above problem reduces to the Monge transport problem. Borrowing the argument from M. Knott and C.S. Smith [18], assume $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function and $\nabla\Phi(X_1) \sim \nu_2$. Clearly, the condition

$$\mathbb{E}_\mu \left[\|X_1 - \nabla\Phi(X_1)\|^2 \right] \leq \mathbb{E}_\mu \left[\|X_1 - X_2\|^2 \right], \quad \mu \in \mathscr{P}(\nu_1, \nu_2),$$

turns out to be equivalent to $\mathbb{E}_\mu [X_1 \cdot \nabla\Phi(X_1)] \geq \mathbb{E}_\mu [X_1 \cdot X_2]$. Latter inequality shows that the minimum quadratic distance is attained. In view of this new formulation, let Φ^* be the convex conjugate of Φ. By the Young inequality we have

$$X_1 \cdot X_2 \leq \Phi(X_1) + \Phi^*(X_2)$$

as well as the Young equality

$$X_1 \cdot \nabla\Phi(X_1) = \Phi(X_1) + \Phi^*(\nabla\Phi X_1).$$

By assumption $X_2 \sim \nabla\Phi(X_1)$, so that

$$\begin{aligned} \mathbb{E}_\mu [X_1 \cdot \nabla\Phi(X_1)] = \mathbb{E}_\mu \left[\Phi(X_1) + \Phi^*(\nabla\Phi(X_1)) \right] = \\ \mathbb{E}_\mu \left[\Phi(X_1) + \Phi^*(X_2) \right] \geq \mathbb{E}_\mu [X_1 \cdot X_2], \end{aligned}$$

which proves that $\nabla\Phi(X_1)$ solves the Monge problem.

This argument, including an existence proof, is in Y. Brenier [9]. In the present paper we shall study the same problem where all the involved distributions are Gaussian. It would be feasible to reduce the Gaussian case to the general one. However, we resort to methods specially suited for this case.

1.1 The Gaussian case

Given two Gaussian distributions $\nu_i = \mathrm{N}_n(\mu_i, \Sigma_i)$, $i = 1, 2$, consider the set $\mathscr{G}(\nu_1, \nu_2)$ of Gaussian distributions on $\mathbb{R}^{2n}$ such that the two n-dimensional margins have the prescribed distributions, $X_i \sim \nu_i$. The corresponding index is

$$W^2 = \inf\left\{\mathbb{E}_\mu\left[\|X_1 - X_2\|^2\right] \middle| \mu \in \mathscr{G}(\nu_1, \nu_2)\right\}. \tag{1}$$

Observe that if $\mu_1 = \mu_2 = 0$ and U is a symmetric matrix such that $U\Sigma_1 U = \Sigma_2$, then the previous argument applies by means of the convex function $\Phi(x) = \frac{1}{2}x^t U x$.

The value of W^2 in Eq. (1) as a function of the mean and the dispersion matrix has been computed by some authors, in particular: I. Olkin and F. Pukelsheim [28], D. C. Dowson and B. V. Landau [12], C. R. Givens and R. M. Shortt [14], M. Gelbrich [13]. They found the (equivalent) forms

$$\begin{aligned} W^2 &= \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right) \\ &= \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1\Sigma_2\right)^{1/2}\right). \end{aligned} \tag{2}$$

Further interpretations of W are available. R. Bhatia et al. [8] showed that W is also the solution of constrained minimization problems for the Frobenius matrix norm $\|M\| = \sqrt{\mathrm{Tr}\,(M^* M)}$, when $\mu_1 = \mu_2 = 0$. Especially,

$$W = \min\left\{\left\|\Sigma_1^{1/2}U - \Sigma_2^{1/2}V\right\| \middle| U \text{and} V \text{orthogonal}\right\}.$$

Notice that $\Sigma^{1/2}U$ is the generic transformation of the standard Gaussian to the Gaussian with dispersion matrix Σ.

Because of the exponent 2 in Eq. (1), the W distance is more precisely called L^2-Wasserstein distance. Other exponents or other distances could be used in the definition. The quadratic case is particularly relevant as W is a Riemannian distance. More references will be given later.

In an Information Geometry perspective, we can mimic the argument of the seminal paper by Amari [4], who derived the notion of both Fisher metric and natural gradient, from the second order approximation of the Kullback-Leibler divergence.

It will be shown (see Sect. 2) that the value W^2 of Eq. (2) has the differential second-order expansion for small H:

$$\mathrm{Tr}\left(\Sigma + (\Sigma + H) - 2\left(\Sigma^{1/2}(\Sigma + H)\Sigma^{1/2}\right)^{1/2}\right) \simeq \mathrm{Tr}\left(\mathscr{L}_\Sigma[H]\,\Sigma\,\mathscr{L}_\Sigma[H]\right), \tag{3}$$

where $\mathscr{L}_\Sigma[H] = X$ is the solution to the Lyapunov equation $X\Sigma + \Sigma X = H$.

The quadratic form in the RHS of Eq. (3) provides a candidate to be the Riemannian inner product associated with the distance W. In addition, if f is a smooth real function

defined on a small W-sphere. i.e., $W(\Sigma, \Sigma + H) = \epsilon$ for small ϵ, then the increment $f(\Sigma + H) - f(\Sigma)$ is maximized along the direction

$$\text{grad } f(\Sigma) = \nabla f(\Sigma)\Sigma + \Sigma \nabla f(\Sigma),$$

where here ∇ denotes the Euclidean gradient. The operator grad is Amari's natural gradient, i.e., the Riemannian gradient.

It is remarkable that all geometric objects shown in the previous equations above may be expressed as matrix operations. In this paper, we proceed in developing systematically the Wasserstein geometry of Gaussian models according to such a formalism.

1.2 Relations with the literature on the general transport theory

The Wasserstein distance and its relevant geometry can be studied non-parametrically also for general distributions. We do not pursue in this direction and refer to the monograph by C. Villani [37]. The L^2-Wasserstein metric geometry has been shown to be Riemannian by F. Otto [29, §4] and J. Lott [21]. Cf. the earlier account by J.D. Lafferty [19].

Let us briefly discuss Otto's approach in the language of Information Geometry, i.e., with reference to S. Amari and H. Nagaoka [3]. In view of the non-parametric approach first introduced in [33], and denoted by $\mathscr{M}$ the set of n-dimensional Gaussian densities with zero mean, the vector bundle

$$H\mathscr{M} = \left\{ (\rho, \phi) \middle| \rho \in \mathscr{M}, \phi \in L^2(\rho), \int \phi \, \rho = 0 \right\}$$

is the Amari Hilbert bundle on $\mathscr{M}$. The Hilbert bundle contains the statistical bundle whose fibers consist of the scores $\frac{d}{dt} \log \rho(t)\big|_{t=0}$ for all smooth curves $t \mapsto \rho(t) \in \mathscr{M}$ with $\rho(0) = \rho$. In turn, the statistical bundle is the tangent space of $\mathscr{M}$ considered as an exponential manifold, see [32,33].

In our present case, since the model $\mathscr{M}$ is an exponential family, the natural parameter is the concentration matrix $C = \Sigma^{-1}$. The log-likelihood is

$$\log \rho(y; C) = -\frac{1}{2} \log 2\pi + \frac{1}{2} \log \det C - \frac{1}{2} y^* C y.$$

If V is a symmetric matrix, the derivative of $C \mapsto \log \rho(y; C)$ in the direction V is

$$d_V \log \rho(y; C) = \frac{1}{2} \text{Tr} \left(C^{-1} V \right) - \frac{1}{2} y^* V y = \text{Tr} \left(\phi(y; C) V \right)$$

where $\phi(y; C) = \frac{1}{2}(C^{-1} - yy^*)$ is a symmetric matrix identified with a linear operator on symmetric matrices Sym (n), equipped with the Frobenius inner product. The fiber at $\rho(\cdot; C)$ consists of the vector space of functions Tr $(\phi(\cdot; C)V)$, $V \in$ Sym (n). The inner product in the Hilbert bundle, restricted to the parameterized statistical bundle, is the Fisher metric

$$F_C(U,V) = \int d_U \log \rho(y;C) d_V \log \rho(y;C)\ \rho(y;C)\ dy =$$
$$-\int d_U \operatorname{Tr}(\phi(y;C)V)\ \rho(y;C)\ dx = \frac{1}{2}\operatorname{Tr}\left(UC^{-1}VC^{-1}\right). \tag{4}$$

The study of the Fisher metric in the Gaussian case has been done first by L.T. Skovgaard [35].

F. Otto [29, §1.3], who was motivated by the study of a class of partial differential equation, considered a inner product defined on smooth functions of the ρ-fiber of the Hilbert bundle, as

$$(\phi_1,\phi_2) \mapsto \int \nabla\phi_1(x)\cdot\nabla\phi_2(x)\ \rho(x)\ dx. \tag{5}$$

In the non-parametric case, Otto's metric of Eq. (5) is related to the Wasserstein distance, for a detailed study of such a metric see J. Lott [21].

If we apply this definition to our score $\operatorname{Tr}(\phi(y;C)V) = \operatorname{Tr}\left(\frac{1}{2}(C^{-1}-yy^*)V\right)$ and $V \in \operatorname{Sym}(n)$, the gradient is $\nabla \operatorname{Tr}(\phi(y;C)V) = -Vy$ and the metric becomes

$$G_C(U,V) = \int \nabla \operatorname{Tr}(\phi(y;C)U)\cdot\nabla \operatorname{Tr}(\phi(y;C)V)\ \rho(y;C)\ dy$$
$$= \int y^*VUy\ \rho(y;\Sigma)\ dy = \operatorname{Tr}\left(UC^{-1}V\right). \tag{6}$$

The equivalence between the metric in Eq. (6) and the one in Eq. (4) can be seen by a change of parameterization both in $\mathscr{M}$ and in each fiber. First, one must define the inner product at Σ to be the inner product computed in the bijection $\Sigma \leftrightarrow C$, to get $\operatorname{Tr}(U\Sigma V)$, which is the form of the metric provided by A. Takatsu [36, Prop. A]. Second, one has to change the parameterization on each fiber of the statistical bundle by $U \mapsto U\Sigma + \Sigma U$. The involved change of parameterization in the statistical bundle $(C,U) \mapsto (C^{-1}, UC^{-1} + C^{-1}U)$ whose inverse is $(\Sigma, X) \mapsto (\Sigma^{-1}, \mathscr{L}_\Sigma[X])$ produces the desired inner product.

We mention also that the Machine Learning literature discusses a divergence introduced by A. Hyvärinen [16], which is related to Otto's metric. Precisely, in the concentration parameterization the Hyvärinen divergence is

$$\operatorname{DH}(D|C) = \frac{1}{2}\int |\nabla \log \rho(y;D) - \nabla \log \rho(y;C)|^2\ \rho(y;C)\ dy$$
$$= \frac{1}{2}\int |Dy - Cy|^2\ \rho(y;C)\ dy = \operatorname{Tr}\left(C^{-1}(D-C)^2\right),$$

and the second derivative of $D \mapsto \operatorname{DH}(D|C)$ at C is

$$d^2 \operatorname{DH}(C|C)[X,Y] = \operatorname{Tr}\left(XC^{-1}Y\right).$$

In Statistics, Hyvärinen divergence is related to local proper scoring rules, see M. Parry et al. [31].

1.3 Overview

The first two sections of the paper are mostly review of known material. In Sec. 2 we recall some properties of the space of symmetric matrices. In particular, we study the Riccati equation, the Lyapunov equation, and we calculate derivatives for the two mappings $\mathrm{sq}\colon A \mapsto A^2$ and $\mathrm{sqrt}\colon A \mapsto A^{1/2}$. The mapping $\sigma\colon A \mapsto AA^*$, where A is a non-singular square matrix is shown to be a submersion and the horizontal vectors at each point is computed. Despite of our manifold being finite dimensional, there is no need of choosing a basis, as all operations of interest are matrix operations. For that reason, we rely on the language of non-parametric differential geometry of W. Klingenberg [17] and S. Lang [20].

In Sec. 3 we discuss known results about the metric geometry induced by the Wasserstein distance. These results are re-stated in Prop. 3 and, for sake of completeness, we provide a further proof inspired by [12]. Prop. 4 provides an explicit metric geodesic, as done by R.J. McCann [27, Example 1.7].

The space of non-degenerate Gaussian measures (or, equivalently, the space of positive definite matrices) can be endowed with a Riemann structure that induces the Wasserstein distance. This is elaborated in Sec. 4, where we use the presentation given by [36], cf. also [8], which in turn adapts to the Gaussian case the original work [29, §4].

The remaining part of the paper is offered as a new contribution to this topic. The Wasserstein Riemannian metric turns out to be

$$W_\Sigma(U, V) = \mathrm{Tr}\left(\mathscr{L}_\Sigma[U]\,\Sigma\,\mathscr{L}_\Sigma[V]\right) = \frac{1}{2}\,\mathrm{Tr}\left(\mathscr{L}_\Sigma[U]\,V\right), \tag{7}$$

at each matrix Σ, and where U, V are symmetric matrices. By submersion methods we study the more general problem of the horizontal surfaces in $\mathrm{GL}(n)$, characterized in Prop. 8. As a specialized case we get the Riemannian geodesic which agrees with the metric geodesic of Sect. 3.

The explicit form of Riemannian exponential is obtained in Sect. 5. The natural (Riemannian) gradient is discussed in Sect. 6 and some applications to optimization are provided in Sect. 6.1. The analysis of the second-order geometry is treated in Sect. 7, where we compute the Levi-Civita covariant derivative, the Riemannian Hessian, and discuss other related topics. However, the curvature tensor will not be taken into consideration in the present paper.

In the final Sect. 8, we discuss the results in view of applications and in Information Geometry of statistical sub-models of the Gaussian manifold.

2 Symmetric matrices

The set $\mathscr{G}^n$ of Gaussian distributions on $\mathbb{R}^n$ is in 1-to-1 correspondence with the space of its parameters $\mathscr{G}^n \ni \mathrm{N}_n(\mu, \Sigma) \leftrightarrow (\mu, \Sigma) \in \mathbb{R}^n \times \mathrm{Sym}^+(n)$. Moreover, $\mathscr{G}^n$ is closed for the weak convergence and the identification is continuous in both directions. A reference for Gaussian distributions is the monograph T.W. Anderson [6].

For ease of later reference, we recall a few results on spaces of matrices. General references are the monographs by P. R. Halmos [15], J. R. Magnus and H. Neudecker [22], and R. Bhatia [7].

The vector space of $n \times m$ real matrices is denoted by $\mathrm{M}(n \times m)$, while square matrices are denoted $\mathrm{M}(n) = \mathrm{M}(n \times n)$. It is an Euclidean space of dimension nm and the vectorization mapping $\mathrm{M}(n \times m) \ni A \mapsto \mathbf{vec}\,(A) \in \mathbb{R}^{nm}$ is an isometry for the Frobenius inner product $\langle A, B\rangle = (\mathbf{vec}\,(A))^*(\mathbf{vec}\,(B)) = \mathrm{Tr}\,(AB^*)$.

Symmetric matrices $\mathrm{Sym}\,(n)$ form a vector subspace of $M(n)$ whose orthogonal complement is the space of anti-symmetric matrices $\mathrm{Sym}^{\perp}\,(n)$. We will find it convenient the use, with regard to symmetric matrices, of the equivalent inner product $\langle A, B\rangle_2 = \frac{1}{2}\,\mathrm{Tr}\,(AB)$, see e.g. Eq. (18) below. The closed pointed cone of non-negative-definite symmetric matrices is denoted by $\mathrm{Sym}^{+}\,(n)$ and its interior, the open cone of the positive-definite symmetric matrices, by $\mathrm{Sym}^{++}\,(n)$.

Given $A, B \in \mathrm{Sym}\,(n)$, the equation $TAT = B$ is called Riccati equation. If $A \in \mathrm{Sym}^{++}\,(n)$ and $B \in \mathrm{Sym}^{+}\,(n)$, then the equation $TAT = B$ has unique solution $T \in \mathrm{Sym}^{+}\,(n)$. In fact, from $TAT = B$ it follows $A^{1/2}TA^{1/2}A^{1/2}TA^{1/2} = A^{1/2}BA^{1/2}$ and, in turn, $A^{1/2}TA^{1/2} = \left(A^{1/2}BA^{1/2}\right)^{1/2}$ because $T \in \mathrm{Sym}^{+}\,(n)$. Hence, the solution to Riccati equation is

$$T = A^{-1/2}\left(A^{1/2}BA^{1/2}\right)^{1/2}A^{-1/2}. \tag{8}$$

Notice that $\det\,(T) = \det\,(A)^{-1/2}\det\,(B)^{1/2}$, consequently $\det\,(T) > 0$ if $\det\,(B) > 0$. In terms of random variables, if $X \in \mathrm{N}_n\,(0, A)$ and $Y = \mathrm{N}_n\,(0, B)$, then T is the unique matrix of $\mathrm{Sym}^{+}\,(n)$ such that $Y \sim TX$.

A more compact closed-form solution of the Riccati equation is available. Given $A \in \mathrm{Sym}^{++}\,(n)$ and $B \in \mathrm{Sym}^{+}\,(n)$, observe that $AB = A^{1/2}(A^{1/2}BA^{1/2})A^{-1/2}$. By similarity, the eigenvalues of AB are non-negative, hence the square root

$$(AB)^{1/2} = A^{1/2}(A^{1/2}BA^{1/2})^{1/2}A^{-1/2} \tag{9}$$

is well defined, see [7, Ex. 4.5.2]. Therefore, an equivalent formulation of Eq. (8) is

$$T = A^{-1}A^{1/2}\left(A^{1/2}BA^{1/2}\right)^{1/2}A^{-1/2} = A^{-1}(AB)^{1/2}. \tag{10}$$

Since $AB = A(BA)A^{-1}$, the eigenvalues of AB and BA are identical, so that the same argument used before yields too

$$T = (BA)^{1/2}A^{-1}. \tag{11}$$

The square mapping $\mathrm{sq}\colon A \mapsto A^2$ is an injection of $\mathrm{Sym}^{++}\,(n)$ onto itself with derivative $d_X\,\mathrm{sq}(A) = XA + AX$. Hence, the derivative operator $d\,\mathrm{sq}(A)$ is invertible. An alternative notation for the derivative we find convenient to use now and then is $d_X\,\mathrm{sq}(A) = d\,\mathrm{sq}(A)[X]$.

For each assigned matrix $V \in \mathrm{Sym}\,(n)$, the matrix $X = (d\,\mathrm{sq}(A))^{-1} V$ is the unique solution X in the space $\mathrm{Sym}\,(n)$ to the Lyapunov equation

$$V = XA + AX. \tag{12}$$

Its solution will be written $X = \mathscr{L}_A[V]$. Clearly we have also

$$V = \mathscr{L}_A[V]\,A + A\mathscr{L}_A[V] \quad \text{and} \quad X = \mathscr{L}_A[XA + AX]. \tag{13}$$

The Lyapunov operator itself can be seen as a derivative. In fact, the inverse of the square mapping sq is the square root mapping $\mathrm{sqrt}\colon \Sigma \to \Sigma^{1/2}$. By the derivative-of-the-inverse rule,

$$d_V\,\mathrm{sqrt}(\Sigma) = (d\,\mathrm{sq}(\mathrm{sqrt}(\Sigma)))^{-1}[V] = \mathscr{L}_{\Sigma^{1/2}}[V]. \tag{14}$$

If Σ is the dispersion of a non-singular Gaussian distribution, then $C = \Sigma^{-1} \in \mathrm{Sym}^{++}\,(n)$ is the concentration matrix and represents an alternative and useful parameterization. From the Lyapunov equation $V = X\Sigma + \Sigma X$ we obtain $\Sigma^{-1} V \Sigma^{-1} = \Sigma^{-1} X + X \Sigma^{-1}$, hence

$$\mathscr{L}_\Sigma[V] = \mathscr{L}_{\Sigma^{-1}}\left[\Sigma^{-1} V \Sigma^{-1}\right] \quad \text{and} \quad \mathscr{L}_{\Sigma^{-1}}[U] = \mathscr{L}_\Sigma[\Sigma U \Sigma].$$

Likewise, another useful formula is

$$\mathscr{L}_\Sigma[V] = \Sigma^{-1/2}\mathscr{L}_\Sigma\left[\Sigma^{-1/2} V \Sigma^{-1/2}\right]\Sigma^{-1/2}. \tag{15}$$

There is also a relation between the Lyapunov equation and the trace. From $X\Sigma + \Sigma X = V$, it follows $\Sigma^{-1} X \Sigma + X = \Sigma^{-1} V$. Then

$$\mathrm{Tr}\,(\mathscr{L}_\Sigma[V]) = \frac{1}{2}\,\mathrm{Tr}\left(\Sigma^{-1} V\right). \tag{16}$$

We will later need the derivative of the mapping $A \mapsto \mathscr{L}_A[V]$, for a fixed V. Differentiating the first identity in Eq. (13) in the direction U, we have

$$0 = d_U\mathscr{L}_A[V]\,A + \mathscr{L}_A[V]\,U + U\mathscr{L}_A[V] + A\,d_U\mathscr{L}_A[V].$$

Hence $d_U\mathscr{L}_A[V]$ is the solution to the Lyapunov equation

$$d_U\mathscr{L}_A[V]\,A + A\,d_U\mathscr{L}_A[V] = -(\mathscr{L}_A[V]\,U + U\mathscr{L}_A[V]),$$

so that we get

$$d_U\mathscr{L}_A[V] = -\mathscr{L}_A[\mathscr{L}_A[V]\,U + U\mathscr{L}_A[V]]. \tag{17}$$

It will be useful in the following to evaluate the second derivative of the mapping $\mathrm{sqrt}\colon \Sigma \mapsto \Sigma^{1/2}$. From Eqs. (14) and (17) it follows

$$d^2\,\mathrm{sqrt}(\Sigma)[U,V] = \mathscr{L}_{\Sigma^{1/2}}\left[\mathscr{L}_{\Sigma^{1/2}}[V]\,\mathscr{L}_{\Sigma^{1/2}}[U] + \mathscr{L}_{\Sigma^{1/2}}[U]\,\mathscr{L}_{\Sigma^{1/2}}[V]\right].$$

Lyapunov equation plays a crucial role, as the linear operator $\mathscr{L}_A$ enters the expression of the Riemannian metric with respect to the standard inner product, see Eq. (7). As a consequence, the numerical implementation of the inner product $W_\Sigma(U,V)$ will require the computation of the matrix $\mathscr{L}_\Sigma[U]$. There are many ways to write down the closed-form solution to Eq. (12). They are discussed in [7]. However, efficient numerical solutions are not based on the closed forms, but rely on specialized numerical algorithms, as discussed by E. L. Wachspress [38] and by V. Simoncini [34].

We now turn to the computation of the second-order approximation of W^2 in Eq. (2).

Fix $\Sigma \in \mathrm{Sym}^{++}(n)$ and let $H \in \mathrm{Sym}(n)$ so that $(\Sigma \pm H) \in \mathrm{Sym}^{++}(n)$. Hence, $\Sigma + \theta H \in \mathrm{Sym}^{++}(n)$ for all $\theta \in [-1,+1]$. Consider the expression of W^2 with $\mu_1 = \mu_2 = 0$, $\Sigma_1 = \Sigma$, $\Sigma_2 = \Sigma + \theta H$, namely

$$\theta \mapsto W^2(\Sigma, \Sigma+\theta H) = 2\,\mathrm{Tr}\,(\Sigma) + \theta\,\mathrm{Tr}\,(H) - 2\,\mathrm{Tr}\left(\left(\Sigma^2 + \theta\Sigma^{1/2}H\Sigma^{1/2}\right)^{1/2}\right).$$

By Eq. (14) and Eq. (16), the first-order derivative is

$$\begin{aligned}\frac{d}{d\theta}W^2(\Sigma,\Sigma+\theta H) &= \mathrm{Tr}\,(H) - 2\,\mathrm{Tr}\left(\mathscr{L}_{(\Sigma^2+\theta\Sigma^{1/2}H\Sigma^{1/2})^{1/2}}\left[\Sigma^{1/2}H\Sigma^{1/2}\right]\right)\\ &= \mathrm{Tr}\,(H) - \mathrm{Tr}\left(\left(\Sigma^2+\theta\Sigma^{1/2}H\Sigma^{1/2}\right)^{-1/2}\left(\Sigma^{1/2}H\Sigma^{1/2}\right)\right).\end{aligned}$$

Observe that $\frac{d}{d\theta}W^2(\Sigma,\Sigma+\theta H)\big|_{\theta=0} = 0$.

The second derivative is

$$\frac{d^2}{d\theta^2}W^2(\Sigma,\Sigma+\theta H) = \mathrm{Tr}\left(\frac{d}{d\theta}\left(\Sigma^2+\theta\Sigma^{1/2}H\Sigma^{1/2}\right)^{-1/2}\left(\Sigma^{1/2}H\Sigma^{1/2}\right)\right)$$

with

$$\begin{aligned}\frac{d}{d\theta}\left(\Sigma^2+\theta\Sigma^{1/2}H\Sigma^{1/2}\right)^{-1/2} &= \left(\Sigma^2+\theta\Sigma^{1/2}H\Sigma^{1/2}\right)^{-1/2}\\ &\times\mathscr{L}_{(\Sigma^2+\theta\Sigma^{1/2}H\Sigma^{1/2})^{1/2}}\left[\Sigma^{1/2}H\Sigma^{1/2}\right]\left(\Sigma^2+\theta\Sigma^{1/2}H\Sigma^{1/2}\right)^{-1/2},\end{aligned}$$

so that

$$\begin{aligned}\frac{d^2}{d\theta^2}W^2(\Sigma,\Sigma+\theta H)\Big|_{\theta=0} &= \mathrm{Tr}\left(\Sigma^{-1}\mathscr{L}_\Sigma\left[\Sigma^{1/2}H\Sigma^{1/2}\right]\Sigma^{-1}\Sigma^{1/2}H\Sigma^{1/2}\right)\\ &= \mathrm{Tr}\left(\Sigma^{-1/2}\mathscr{L}_\Sigma\left[\Sigma^{1/2}H\Sigma^{1/2}\right]\Sigma^{-1/2}H\right) = \mathrm{Tr}\,(\mathscr{L}_\Sigma[H]\,H),\end{aligned}$$

where Eq. (15) has been used. Finally, observe that

$$\begin{aligned}\mathrm{Tr}\left(\mathscr{L}_\Sigma[H]\,\Sigma\,\mathscr{L}_\Sigma[H]\right) &= \mathrm{Tr}\left(\mathscr{L}_\Sigma[H]\,\mathscr{L}_\Sigma[H]\,\Sigma\right)\\ &= \frac{1}{2}\,\mathrm{Tr}\left(\mathscr{L}_\Sigma[H]\left(\mathscr{L}_\Sigma[H]\,\Sigma+\Sigma\,\mathscr{L}_\Sigma[H]\right)\right) = \frac{1}{2}\,\mathrm{Tr}\left(\mathscr{L}_\Sigma[H]\,H\right) \qquad (18)\end{aligned}$$

We can conclude that

$$\begin{aligned}W^2(\Sigma,\,\Sigma+\theta H) &= \frac{\theta^2}{2}\,\mathrm{Tr}\left(\mathscr{L}_\Sigma[H]\,H\right)+\mathrm{o}(\theta^2)\\ &= \theta^2\,\mathrm{Tr}\left(\mathscr{L}_\Sigma[H]\,\Sigma\,\mathscr{L}_\Sigma[H]\right)+\mathrm{o}(\theta^2).\end{aligned}$$

Therefore, the bi-linear form in the RHS suggests the form of the Riemannian metric to be derived.

2.1 The mapping $A \mapsto AA^*$

We study now the extension of the square operation to general invertible matrices, namely the mapping $\text{œ} : \mathrm{GL}(n) \to \mathrm{Sym}^{++}(n)$, defined by $\text{œ}(A) = AA^*$. Next proposition shows that this operation is a submersion. We recall first its definition, see [10, Ch. 8, Ex. 8–10] or [20, §II.2].

Let $\mathscr{O}$ be an open set of the Hilbert space H, and $f : \mathscr{O} \to \mathscr{N}$ a smooth surjection from the Hilbert space H onto a manifold $\mathscr{N}$, i.e., assume that for each $A \in \mathscr{O}$ the derivative at A, $df(A) : H \to T_{f(A)}\mathscr{N}$ is surjective. In such a case, for each $C \in \mathscr{N}$, the fiber $f^{-1}(C)$ is a sub-manifold. Assigned a point $A \in f^{-1}(C)$, a vector $U \in H$ is called vertical if it is tangent to the manifold $f^{-1}(C)$. Each such a tangent vector U is the velocity at $t = 0$ of some smooth curve $t \mapsto \gamma(t)$ with $\gamma(0) = A$ and $\dot\gamma(0) = U$. Precisely, from $f(\gamma(t)) = C$ for all t we derive the characterization of vertical vectors. We have $df(A)[\dot\gamma(0)] = 0$ i.e., the tangent space at A is $T_A f^{-1}(f(A)) = \mathrm{Ker}(df(A))$. The orthogonal space to the tangent space $T_A f^{-1}(f(A))$ is called the space of horizontal vectors at A,

$$\mathscr{H}_A = \mathrm{Ker}(df(A))^\perp = \mathrm{Im}\left(df(A)^*\right).$$

Let us apply this argument to our specific case. Let $\mathrm{GL}(n) \subset \mathrm{M}(n)$ be the open set of invertible matrices; $\mathrm{O}(n)$ the subgroup of $\mathrm{GL}(n)$ of orthogonal matrices; $\mathrm{Sym}^\perp(n)$ the subspace of $\mathrm{M}(n)$ of anti-symmetric matrices.

Proposition 1 *1. For each given $A \in \mathrm{GL}(n)$ we have the orthogonal splitting*

$$\mathrm{M}(n) = \mathrm{Sym}(n)\,A \oplus \mathrm{Sym}^\perp(n)\,(A^*)^{-1}.$$

2. The mapping

$$\sigma :\ \mathrm{GL}(n) \ni A \mapsto AA^* \in \mathrm{Sym}^{++}(n)$$

has derivative at A given by $d_X\sigma(A) = XA^* + AX^*$. It is a submersion with fibers

$$\sigma^{-1}(C) = \left\{ C^{1/2}R \middle| R \in \mathrm{O}(n) \right\}.$$

3. *The kernel of the differential is*

$$\mathrm{Ker}(d\sigma(A)) = \mathrm{Sym}^{\perp}(n)\,(A^*)^{-1}$$

and its orthogonal complement, $\mathscr{H}_A = \mathrm{Ker}(d\sigma(A))^{\perp}$, is

$$\mathscr{H}_A = \mathrm{Sym}(n)\,A.$$

4. *The orthogonal projection of $X \in M(n)$ onto $\mathscr{H}_A$ is $\mathscr{L}_{AA^*}\left[XA^* + AX^*\right]A$.*

Proof We provide here the proof for sake of completeness. See also [36] and [8].

1. If $\langle B, CA\rangle = 0$, for all $C \in \mathrm{Sym}(n)$ i.e., $CA \in \mathrm{Sym}^{+}(n)\,A$, then $\mathrm{Tr}(BA^*C) = 0$, so that $BA^* \in \mathrm{Sym}^{\perp}(n)$ that is, $B \in \mathrm{Sym}^{\perp}(n)\,(A^*)^{-1}$.
2. Let the matrix A be an element in the fiber manifold $\sigma^{-1}(AA^*)$. The derivative of σ at A, $X \mapsto XA^* + AX^*$, is surjective, because for each $W \in \mathrm{Sym}(n)$ we have $d\sigma(A)\left[\frac{1}{2}W(A^*)^{-1}\right] = W$. Hence σ is a submersion and the fiber $\sigma^{-1}(AA^*) = \left\{(AA^*)^{1/2}R \middle| R \in \mathrm{O}(n)\right\}$ is a sub-manifold of $\mathrm{GL}(n)$.
3. Let us compute the splitting of $\mathrm{M}(n)$ into the kernel of $d\sigma(A)$ and its orthogonal: $\mathrm{M}(n) = \mathrm{Ker}(d\sigma(A)) \oplus \mathscr{H}_A$. The vector space tangent to $\sigma^{-1}(AA^*)$ at A is the kernel of the derivative at A:

$$\begin{aligned}\mathrm{Ker}\,(d\sigma(A)) &= \left\{X \in \mathrm{M}(n) |\; XA^* + AX^* = 0\right\}\\ &= \left\{X \in \mathrm{M}(n) |\; (AX^*)^* = -AX^*\right\}.\end{aligned}$$

   Therefore, $X \in \mathrm{Ker}(d\sigma(A))$ if, and only if, $AX^* \in \mathrm{Sym}^{\perp}(n)$, i.e., $\mathrm{Ker}(d\sigma(A)) = \mathrm{Sym}^{\perp}(n)\,(A^*)^{-1}$. We have just proved that this implies $\mathscr{H}_A = \mathrm{Sym}(n)\,A$.
4. Consider the decomposition of X into the horizontal and the vertical part: $X = CA + D(A^*)^{-1}$ with $C \in \mathrm{Sym}(n)$ and $D \in \mathrm{Sym}^{\perp}(n)$. By transposition, we get $X^* = A^*C - A^{-1}D$. From the previous two equations, we obtain the two equations $XA^* = C(AA^*) + D$ and $AX^* = (AA^*)C - D$. The sum of the two previous equations is $XA^* + AX^* = C(AA^*) + (AA^*)C$, which is a Lyapunov equation having solution $C = \mathscr{L}_{AA^*}\left[XA^* + AX^*\right]$. It follows that the projection is $CA = \mathscr{L}_{AA^*}\left[XA^* + AX^*\right]A$ □

3 Wasserstein distance

The aim of this section is to discuss the Wasserstein distance for the Gaussian case as well as the equation for the associated metric geodesic. Most of its content is an exposition of known results.

3.1 Block-Gaussian

Let us suppose that the dispersion matrix $\Sigma \in \mathrm{Sym}^+(2n)$ is partitioned into $n \times n$ blocks, and consider random variables X and Y such that

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}_{2n}(\mu, \Sigma), \quad \Sigma = \begin{bmatrix} \Sigma_1 & K \\ K^* & \Sigma_2 \end{bmatrix},$$

so that $K_{ij} = \mathrm{Cov}\left(X_i, Y_j\right)$ if $i = 1, \ldots, n$ and $j = (n+1), \ldots, 2n$. It follows that $K_{ij}^2 \le (\Sigma_1)_{ii}(\Sigma_2)_{jj} \le \frac{1}{2}\left((\Sigma_1)_{ii} + (\Sigma_2)_{jj}\right)$, which in turn imply the bounds

$$\|K\|_2^2 \le \mathrm{Tr}(\Sigma_1)\,\mathrm{Tr}(\Sigma_2) \quad \text{and} \quad \sup_{ij}\left|K_{ij}\right| \le \frac{1}{2}(\mathrm{Tr}(\Sigma_1) + \mathrm{Tr}(\Sigma_2)). \tag{19}$$

For mean vectors $\mu_1, \mu_2 \in \mathbb{R}^2$ and dispersion matrices $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^+(n)$, define the set of jointly Gaussian distributions with given marginals to be

$$\mathscr{G}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \left\{\mathrm{N}_{2n}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & K \\ K^* & \Sigma_2 \end{bmatrix}\right)\right\},$$

and the Gini dissimilarity index

$$\begin{aligned} &W^2((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) \\ &\quad = \inf\left\{\mathbb{E}\left[\|X - Y\|^2\right] \middle| \begin{bmatrix} X \\ Y \end{bmatrix} \sim \gamma, \gamma \in \mathscr{G}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))\right\} \\ &\quad = \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}(\Sigma_1) + \mathrm{Tr}(\Sigma_2) - 2\sup\left\{\mathrm{Tr}(K) \middle| \begin{bmatrix} \Sigma_1 & K \\ K^* & \Sigma_2 \end{bmatrix} \in \mathrm{Sym}^+(2n)\right\} \end{aligned} \tag{20}$$

Actually, in view of either of the bounds in Eq. (19), the set $\mathscr{G}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))$ is compact and the inf is attained.

It is easy to verify that

$$\begin{aligned} &W((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) \\ &\quad = \sqrt{\min\left\{\mathbb{E}\left[\|X - Y\|^2\right] \middle| \begin{bmatrix} X \\ Y \end{bmatrix} \sim \gamma, \gamma \in \mathscr{G}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))\right\}} \end{aligned}$$

defines a distance on the space $\mathscr{G}_n \simeq \mathbb{R}^n \times \mathrm{Sym}^+(n)$. The symmetry of W is clear as well as the triangle inequality, by considering Gaussian distributions on $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$ with given marginals. To conclude, assume that the min is reached at some $\overline{\gamma}$. Then

$$0 = W((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \mathbb{E}_{\overline{\gamma}}\left[|X - Y|^2\right] \Leftrightarrow \mu_1 = \mu_2 \quad \text{and} \quad \Sigma_1 = \Sigma_2.$$

A further observation is that distance W is homogeneous i.e.,

$$W((\lambda\mu_1, \lambda^2\Sigma_1), (\lambda\mu_2, \lambda^2\Sigma_2)) = \lambda W((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)), \quad \lambda \geq 0.$$

3.2 Computing the quadratic dissimilarity index

We will present a proof as given by Dowson and Landau [12], but with some corrections.

Given $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^+(n)$, each admissible K's in (20) belongs to a compact set of $\mathrm{M}(n)$ thanks to bound (19), so the maximum of the function $2\,\mathrm{Tr}(K)$ is reached. Therefore, we are led to study the problem

$$\begin{cases} \alpha(\Sigma_1, \Sigma_2) = \max\limits_{K \in \mathrm{M}(n)} 2\,\mathrm{Tr}(K) \\ \text{subject to} \\ \Sigma = \begin{bmatrix} \Sigma_1 & K \\ K^* & \Sigma_2 \end{bmatrix} \in \mathrm{Sym}^+(2n) \end{cases} \tag{21}$$

The value of the similar problem with max replaced by min will be denoted by $\beta(\Sigma_1, \Sigma_2)$.

Proposition 2 *1. Let* $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^+(n)$. *Then*

$$\alpha(\Sigma_1, \Sigma_2) = 2\,\mathrm{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right) \text{ and } \beta(\Sigma_1, \Sigma_2) = -\alpha(\Sigma_1, \Sigma_2).$$

2. If moreover $\det(\Sigma_1) > 0$, *then*

$$\alpha(\Sigma_1, \Sigma_2) = 2\,\mathrm{Tr}\left((\Sigma_1\Sigma_2)^{1/2}\right).$$

Proof (point (1)) A symmetric matrix $\Sigma \in \mathrm{Sym}(2n)$ is non-negative defined if, and only if, it is of the form $\Sigma = SS^*$, with $S \in \mathrm{M}(2n)$. Given the block structure of Σ in (21), we can write

$$\begin{bmatrix} \Sigma_1 & K \\ K^* & \Sigma_2 \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} \begin{bmatrix} A^* & B^* \end{bmatrix} = \begin{bmatrix} AA^* & AB^* \\ BA^* & BB^* \end{bmatrix},$$

where A and B are two matrices in $\mathrm{M}(n \times 2n)$.

Therefore, problem (21) becomes

$$\begin{cases} \alpha(\Sigma_1, \Sigma_2) = \max\limits_{A, B \in \mathrm{M}(n \times 2n)} 2\,\mathrm{Tr}\left(AB^*\right) \\ \text{subject to} \\ \Sigma_1 = AA^*, \quad \Sigma_2 = BB^* \end{cases}$$

We have already observed that the optimum exists, so the necessary conditions of Lagrange theorem allows us to characterize this optimum. However, the two constraints $\Sigma_1 = AA^*$ and $\Sigma_2 = BB^*$ are not necessarily regular at every point (i.e., the Jacobian of the transformation may fail to be of full rank at some point), so we must take into account that the optimum could be an irregular point. To this purpose, as a customary, we shall adopt Fritz John first-order formulation for the Lagrangian (see [25]).

We shall initially assume that both Σ_1 and Σ_2 are non-singular.

Let then $(\nu_0, \Lambda, \Gamma) \in \{0, 1\} \times \mathrm{Sym}\,(n) \times \mathrm{Sym}\,(n)$, $(\nu_0, \Lambda, \Gamma) \neq (0, 0, 0)$, where the symmetric matrices Λ and Γ are the Lagrange multipliers. The Lagrangian function will be

$$\begin{aligned} L &= 2\nu_0 \operatorname{Tr}\left(AB^*\right) - \operatorname{Tr}\left(\Lambda AA^*\right) - \operatorname{Tr}\left(\Gamma BB^*\right) \\ &= 2\nu_0 \operatorname{Tr}\left(AB^*\right) - \operatorname{Tr}\left(A^*\Lambda A\right) - \operatorname{Tr}\left(B^*\Gamma B\right) \end{aligned}$$

The first-order conditions of L lead to

$$\begin{cases} \nu_0 B = \Lambda A, & \nu_0 A = \Gamma B \\ \Sigma_1 = AA^*, & \Sigma_2 = BB^* \end{cases}. \tag{22}$$

In the case $\nu_0 = 1$, i.e., the case of stationary regular points, Eq. (22) becomes

$$\begin{cases} B = \Lambda A, & A = \Gamma B \\ \Sigma_1 = AA^*, & \Sigma_2 = BB^* \end{cases}, \tag{23}$$

which in turn implies

$$\begin{cases} \Lambda\Sigma_1\Lambda = \Sigma_2 \\ \Gamma\Sigma_2\Gamma = \Sigma_1 \end{cases}, \quad \Lambda, \Gamma \in \mathrm{Sym}\,(n) \tag{24}$$

and further

$$K = \Sigma_1\Lambda = \Gamma\Sigma_2.$$

Of course, Eqs. (24) could be more general than Eqs. (23) and thus possibly contain undesirable solutions. In this light, we establish the following facts, in which both matrices Σ_1 and Σ_2 must be nonsingular. Notice that in this case Eqs. (24) imply that both Λ and Γ are nonsingular as well.

Claim 1: If (Γ, Λ) *is a solution to* (24) *and* $\Lambda^{-1} = \Gamma$, *then the couple* (Γ, Λ) *are Lagrange multipliers of Problem* (21).

Actually, let $\Sigma_1 = AA^*$, $A \in \mathrm{M}(n \times 2n)$ be any representation of the matrix Σ_1. Define $B = \Lambda A$ so that $A = \Lambda^{-1}B = \Gamma B$. Moreover

$$BB^* = \Lambda AA^*\Lambda = \Lambda\Sigma_1\Lambda = \Sigma_2,$$

and so (Λ, Γ) are multipliers associated with the feasible point (A, B).

Claim 2: The set of solutions to (24)*, such that* $\Gamma^{-1} = \Lambda$*, is not empty. In particular, there is a unique pair* $\left(\widetilde{\Lambda}, \widetilde{\Gamma}\right)$ *where both* $\widetilde{\Lambda}$ and $\widetilde{\Gamma}$ *are positive definite.*

We have already observed that Eqs. (24) imply that Λ and Γ are nonsingular. Moreover, we have $\Gamma^{-1}\Sigma_1\Gamma^{-1} = \Sigma_2$. Recalling that Riccati's equation has one and only one solution in the class of positive definite matrices, then $X = \Lambda = \Gamma^{-1}$.

Now we proceed to study the solutions to $\Lambda\Sigma_1\Lambda = \Sigma_2$ and we shall show that Eq (24) has infinitely many solutions. In correspondence to each one Λ, the value of the objective function will be given by $2\,\mathrm{Tr}\,(K) = 2\,\mathrm{Tr}\,(\Sigma_1\Lambda)$. Therefore, we must select the matrix Λ such that $\mathrm{Tr}\,(\Sigma_1\Lambda)$ be maximized.

Following [12], we define

$$R = \Sigma_1^{1/2}\Lambda\Sigma_1^{1/2} \in \mathrm{Sym}\,(n)\,,$$

so that, in view of (24), we have

$$R^2 = \Sigma_1^{1/2}\Lambda\Sigma_1^{1/2}\Sigma_1^{1/2}\Lambda\Sigma_1^{1/2} = \Sigma_1^{1/2}\Lambda\Sigma_1\Lambda\Sigma_1^{1/2} = \Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} \in \mathrm{Sym}^+\,(n)\,. \tag{25}$$

Moreover,

$$\mathrm{Tr}\,(R) = \mathrm{Tr}\left(\Sigma_1^{1/2}\Lambda\Sigma_1^{1/2}\right) = \mathrm{Tr}\left(\Sigma_1^{1/2}\Sigma_1^{1/2}\Lambda\right) = \mathrm{Tr}\,(\Sigma_1\Lambda) = \mathrm{Tr}\,(K)\,.$$

Eq. (25) shows that, though the Lagrangian can have many rest points (i.e., many solutions Λ) the matrix $R^2 = \Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} \in \mathrm{Sym}^+\,(n)$ remains constant. Not so the value of the objective function $\mathrm{Tr}\,(K) = \mathrm{Tr}\,(R)$ which depends on R (i.e., on Λ).

Let

$$R^2 = \sum_k \lambda_k E_k$$

denote the spectral decomposition of R^2, then the solutions to R will be

$$R = \sum_k \varepsilon_k \lambda_k^{1/2} E_k$$

with $\varepsilon_k = \pm 1$. Hence $\mathrm{Tr}\,(K) = \mathrm{Tr}\,(R)$ will be maximized whenever $\varepsilon_k \equiv 1$ and so $R \in \mathrm{Sym}^+\,(n)$. Clearly the objective function will be minimized if $\varepsilon_k \equiv -1$. From now on the proof of the min statement follows similarly.

Hence the maximum of the trace occurs at

$$R = \left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2},$$

namely $\Lambda = \Sigma_1^{-1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\Sigma_1^{-1/2}$. Thanks to Claims 1-2 this matrix is a multiplier of the Lagrangian and so we would have

$$\alpha\left(\Sigma_1, \Sigma_2\right) = 2\mathrm{Tr}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}, \tag{26}$$

as long as the optimum is attained at a regular point. In fact, to complete the proof, we must still examine the case $\nu_0 = 0$, for which Eq. (22) becomes

$$\Lambda A = 0, \quad \Gamma B = 0.$$

It follows

$$\begin{aligned}\Lambda\Sigma_1 &= \Lambda A A^* = 0\\ \Gamma\Sigma_2 &= \Gamma B B^* = 0,\end{aligned}$$

and consequently $\Lambda = \Gamma = 0$. Therefore there is no irregular point, provided Σ_1 and Σ_2 are not singular matrices. So we have proved the relation (26) under the above assumptions.

Last step will be that of extending our result to possibly singular matrices Σ_1 and Σ_2.

Given the two matrices $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^+(n)$, set

$$\Sigma_1(\varepsilon) = \Sigma_1 + \varepsilon I_n \text{ and } \Sigma_2(\varepsilon) = \Sigma_2 + \varepsilon I_n, \text{ with } \varepsilon \in [0, 1].$$

If $\varepsilon > 0$, then

$$\det\left(\Sigma_i + \varepsilon I\right) = \prod_{j=1}^{n}(\lambda_{i,j} + \varepsilon) > 0, \quad i = 1, 2.$$

where $\lambda_{i,j}$, $j = 1, \ldots, n$ is a set of eigenvalues of Σ_i , $i = 1, 2$. Let us consider the parametric programming problem

$$\begin{cases} \alpha(\Sigma_1(\varepsilon), \Sigma_2(\varepsilon)) = \max\limits_{K \in \mathrm{M}(n)} 2\,\mathrm{Tr}\,(K) \\ \quad \text{subject to} \\ \quad \begin{bmatrix} \Sigma_1(\varepsilon) & K \\ K^* & \Sigma_2(\varepsilon) \end{bmatrix} \in \mathrm{Sym}^+(2n) \end{cases}$$

Observe that the feasible region is contained in a compact set independent of $\varepsilon \in [0, 1]$ because of the bound (19).

Now the continuity of the optimal value $\varepsilon \mapsto \alpha(\Sigma_1(\varepsilon), \Sigma_2(\varepsilon))$ follows easily from Berge maximum theorem, see for instance [2, Th. 17.31]. Hence

$$\alpha(\Sigma_1, \Sigma_2) = \lim_{\varepsilon \to 0} \alpha(\Sigma_1(\varepsilon), \Sigma_2(\varepsilon)) = 2\,\mathrm{Tr}\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)$$

and the assertion is proved for any $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^+(n)$. □

Proof (point (2)) From Eq. (9) we have

$$\begin{aligned}\operatorname{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right) &= \operatorname{Tr}\left(\Sigma_1^{1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\Sigma_1^{-1/2}\right)\\ &= \operatorname{Tr}\left((\Sigma_1\Sigma_2)^{1/2}\right).\end{aligned}$$

□

The following result provides exact both lower and upper bounds of $\mathbb{E}\left[\|X-Y\|^2\right]$.

Proposition 3 *Let X, Y be multivariate Gaussian random variables taking values in $\mathbb{R}^n$ and having means μ_1 and μ_2 and dispersion matrices Σ_1 and Σ_2 respectively. Then*

$$\begin{aligned}&\|\mu_1-\mu_2\|^2+\operatorname{Tr}\left(\Sigma_1+\Sigma_2-2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)\le\mathbb{E}\left[\|X-Y\|^2\right]\le\\ &\|\mu_1-\mu_2\|^2+\operatorname{Tr}\left(\Sigma_1+\Sigma_2+2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right).\end{aligned}$$

If $\det\Sigma_1\neq 0$*, then the extremal values are attained at the joint distribution of*

$$\begin{bmatrix}X\\ \mu_2\pm T(X-\mu_1)\end{bmatrix}\sim$$
$$\mathrm{N}_{2n}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\begin{bmatrix}\Sigma_1 & \pm T\Sigma_1\\ \pm\Sigma_1 T & \Sigma_2\end{bmatrix}\right)=\mathrm{N}_{2n}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\begin{bmatrix}\Sigma_1 & \pm(\Sigma_2\Sigma_1)^{1/2}\\ \pm(\Sigma_1\Sigma_2)^{1/2} & \Sigma_2\end{bmatrix}\right),$$

respectively, where $T\in\mathrm{Sym}^{+}(n)$ is the solution to the Riccati equation $T\Sigma_1 T=\Sigma_2$.

Proof From Proposition 2 and Eq. (20), it follows

$$\begin{aligned}\min\left[\|X-Y\|^2\right]&=\|\mu_1-\mu_2\|^2+\operatorname{Tr}(\Sigma_1)+\operatorname{Tr}(\Sigma_2)-2\operatorname{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right),\\ \max\left[\|X-Y\|^2\right]&=\|\mu_1-\mu_2\|^2+\operatorname{Tr}(\Sigma_1)+\operatorname{Tr}(\Sigma_2)+2\operatorname{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right).\end{aligned}$$

To check the extremal points it suffices to observe that, in view of relation (8):

$$\operatorname{Tr}(T\Sigma_1)=\operatorname{Tr}\left(\Sigma_1^{-1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\Sigma_1^{1/2}\right)=\operatorname{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right).$$

Hence it is verified that the extremal values are attained at $Y=\mu_2\pm T(X-\mu_1)$. In the second form of the distribution we are using Eq. (10) and Eq. (11). □

The W-distance defines on $\mathbb{R}\times\mathrm{Sym}^{++}(n)$ a metric geometry with geodesics. This result is due to [27].

Proposition 4 *The relation*

$$W((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \sqrt{\|\mu_1 - \mu_2\|^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)} \tag{27}$$

defines a distance on $\mathbb{R}^n \times \mathrm{Sym}^+(n)$. *The geodesic from* (μ_1, Σ_1) *to* (μ_2, Σ_2), *with* $(\mu_1, \Sigma_1), (\mu_2, \Sigma_2) \in \mathbb{R}^n \times \mathrm{Sym}^{++}(n)$, *is the curve*

$$\Gamma : [0, 1] \ni t \mapsto (\mu(t), \Sigma(t)),$$

where $\mu(t) = (1-t)\mu_1 + t\mu_2$ *and*

$$\begin{aligned}\Sigma(t) &= ((1-t)I + tT)\Sigma_1((1-t)I + tT)\\ &= (1-t)^2\Sigma_1 + t^2\Sigma_2 + t(1-t)\left((\Sigma_1\Sigma_2)^{1/2} + (\Sigma_2\Sigma_1)^{1/2}\right),\end{aligned}$$

and T *is the (unique) non-negative definite solution to the Riccati equation* $T\Sigma_1 T = \Sigma_2$.

Proof Clearly, $\Gamma(0) = (\mu_1, \Sigma_1)$ and $\Gamma(1) = (\mu_2, \Sigma_2)$. Let us compute the distance between $\Gamma(0)$ and the point

$$\Gamma(t) = (\mu(t), \Sigma(t)) = (\mu_1 + t(\mu_2 - \mu_1), ((1-t)I + tT)\Sigma_1((1-t)I + tT)).$$

We have

$$\begin{aligned}\Sigma_1^{1/2}\Sigma(t)\Sigma_1^{1/2} &= \Sigma_1^{1/2}((1-t)I + tT)\Sigma_1((1-t)I + tT)\Sigma_1^{1/2}\\ &= \left(\Sigma_1^{1/2}((1-t)I + tT)\Sigma_1^{1/2}\right)\left(\Sigma_1^{1/2}((1-t)I + tT)\Sigma_1^{1/2}\right),\end{aligned}$$

so that

$$\left(\Sigma_1^{1/2}\Sigma(t)\Sigma_1^{1/2}\right)^{1/2} = \Sigma_1^{1/2}((1-t)I + tT)\Sigma_1^{1/2},$$

and hence

$$\begin{aligned}&\mathrm{Tr}\left(\left(\Sigma_1^{1/2}\Sigma(t)\Sigma_1^{1/2}\right)^{1/2}\right)\\ &= \mathrm{Tr}\left(\Sigma_1^{1/2}((1-t)I + tT)\Sigma_1^{1/2}\right) = (1-t)\,\mathrm{Tr}(\Sigma_1) + t\,\mathrm{Tr}(T\Sigma_1).\end{aligned}$$

We have

$$\begin{aligned}\mathrm{Tr}(\Sigma(t)) &= \mathrm{Tr}(((1-t)I + tT)\Sigma_1((1-t)I + tT))\\ &= (1-t)^2\,\mathrm{Tr}(\Sigma_1) + 2t(1-t)\,\mathrm{Tr}(T\Sigma_1) + t^2\,\mathrm{Tr}(\Sigma_2)\end{aligned}$$

Collecting all the above results,

$$\begin{aligned}
&\operatorname{Tr}\left(\Sigma_1+\Sigma(t)-2\left(\Sigma_1^{1/2}\Sigma(t)\Sigma_1^{1/2}\right)^{1/2}\right)\\
&\quad=\operatorname{Tr}(\Sigma_1)+(1-t)^2\operatorname{Tr}(\Sigma_1)+2t(1-t)\operatorname{Tr}(T\Sigma_1)\\
&\qquad+t^2\operatorname{Tr}(\Sigma_2)-2(1-t)\operatorname{Tr}(\Sigma_1)-2t\operatorname{Tr}(T\Sigma_1)\\
&\quad=t^2\operatorname{Tr}(\Sigma_1)+t^2\operatorname{Tr}(\Sigma_2)-2t^2\operatorname{Tr}(T\Sigma_1)\\
&\quad=t^2\operatorname{Tr}\left(\Sigma_1+\Sigma_2-2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right).
\end{aligned}$$

In conclusion,

$$\begin{aligned}
&W(\Gamma(0),\Gamma(t))\\
&=\sqrt{\|\mu(0)-\mu(t)\|^2+\operatorname{Tr}\left(\Sigma(0)+\Sigma(t)-2\left(\Sigma(0)^{1/2}\Sigma(t)\Sigma(0)^{1/2}\right)^{1/2}\right)}\\
&=t\,W(\Gamma(0),\Gamma(1)).
\end{aligned}$$

□

We end this section by adding a few remarks.

In metric spaces, the definition of geodesic we use here is related to Merger convexity property, see [30, p. 78]. A stronger definition requires the proportionality of the distance between couple of points on the curve, i.e.,

$$W(\Gamma(s),\Gamma(t))=|t-s|\,W(\Gamma(0),\Gamma(1)),$$

for $s,t\in[0,1]$. It will be proved later that in fact our geodesics enjoy such a stronger property.

Clearly Proposition 4 still holds under the only assumption that Σ_1 is not singular, but the case in which both the distributions are degenerate remains excluded.

The simplest example occurs when the two subspaces, Range Σ_1 and Range Σ_2, are orthogonal. In this case, for all joint distribution of the random vector (X,Y), with marginals $X\sim \mathrm{N}_2(0,\Sigma_1)$ and $Y\sim \mathrm{N}_2(0,\Sigma_2)$, the values of X and Y will lie into orthogonal subspaces, so that $XY^*=0$. Hence $\|X-Y\|^2=\|X\|^2+\|Y\|^2$, and

$$\mathbb{E}\,\|X-Y\|^2=\mathbb{E}\,\|X\|^2+\mathbb{E}\,\|Y\|^2=\operatorname{Tr}(\Sigma_1)+\operatorname{Tr}(\Sigma_2).$$

So any joint distribution (X,Y) attains the optimal value $\sqrt{\operatorname{Tr}(\Sigma_1)+\operatorname{Tr}(\Sigma_2)}$.

If we now define $X(t)=(1-t)X+tY$, then

$$\mathbb{E}\left[\|X-X(t)\|^2\right]=\mathbb{E}\left[t^2\|X-Y\|^2\right]=t^2\left[\operatorname{Tr}(\Sigma_1)+\operatorname{Tr}(\Sigma_2)\right],$$

consequently $X(t)$ is the geodesic joining the two random vectors X and Y.

The previous example can be extended by taking two singular matrices

$$\Sigma_1 = \sigma_1^2 v v^* \text{ and } \Sigma_2 = \sigma_2^2 w w^*$$

where $v \neq w \in \mathbb{R}^n$ and $\|v\| = \|w\| = 1$. Clearly, Range $\Sigma_1 \cap$ Range $\Sigma_2 = \{0\}$ and they are one-dimensional spaces spanned by vectors v and w, respectively (it is not restrictive to assume $v^* w \geq 0$, too). By Eq. (27),

$$G(\Sigma_1, \Sigma_2) = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1 \sigma_2 v^* w}.$$

Despite singularity of these matrices, it can be directly found the point realizing the minimum in (20), which is the singular matrix in $\mathrm{Sym}^+(2n)$:

$$\begin{bmatrix} \sigma_1^2 v v^* & \sigma_1 \sigma_2 v w^* \\ \sigma_1 \sigma_2 w v^* & \sigma_2^2 w w^* \end{bmatrix} = \begin{bmatrix} \sigma_1 v \\ \sigma_2 w \end{bmatrix} \begin{bmatrix} \sigma_1 v^* & \sigma_2 w^* \end{bmatrix}.$$

4 Wasserstein Riemannian geometry

We have seen how to compute the geodesic for the distance W. Since the component $\mathbb{R}^n$ carries the standard Euclidean geometry, we focus on the geometry of the matrix part, i.e., we shall restrict our analysis to 0-mean distributions $\mathrm{N}_n(0, \Sigma)$. Moreover, Σ will be assumed to be positive definite. Our purpose is to endow the open set $\mathrm{Sym}^{++}(n)$ with a structure of Riemannian manifold whose metric tensor generates the Wasserstein distance. The Riemannian metric is obtained by pushing forward the Euclidean geometry of square matrices to the space of dispersion matrices via the mapping $\sigma : A \mapsto AA^* = \Sigma$. This approach has been introduced by F. Otto [29] in the general non-parametric case and developed in the Gaussian case by A. Takatsu [36] and R. Bhatia [8].

In view of Prop. 1, $\sigma : \mathrm{GL}(n) \to \mathrm{Sym}^{++}(n) \subset \mathrm{M}(n)$ is a submersion and $\mathscr{H}_A = \mathrm{Sym}(n) A$ is the space of horizontal vectors at A.

We recall that a submersion $f : \mathrm{GL}(n) \to \mathrm{Sym}^{++}(n)$ is called Riemannian if for all A the differential restricted to horizontal vectors

$$df(A)|_{\mathscr{H}_A} : \mathscr{H}_A \to T_{f(A)} \mathrm{Sym}^{++}(n) = \mathrm{Sym}(n)$$

is an isometry i.e.,

$$U, V \in \mathscr{H}_A \Rightarrow \langle df(A)[U], df(A)[V] \rangle_{f(A)} = \langle U, V \rangle. \tag{28}$$

A linear isometry is always 1-to-1 and, if it is onto, we can write backward that

$$X, Y \in T_{f(A)} \mathrm{Sym}^{++}(n) \Rightarrow \langle X, Y \rangle_{f(A)} = \left\langle \left(df(A)|_{\mathscr{H}_A} \right)^{-1} X, \left(df(A)|_{\mathscr{H}_A} \right)^{-1} Y \right\rangle.$$

Conversely, the previous equation provides the definition of a metric on $\mathrm{Sym}^{++}(n)$ for which the submersion f is Riemannian.

If U_A is the projection of U on $\mathscr{H}_A$, then $df(A)[U] = df(A)[U_A]$ and Eq. (28) becomes

$$U, V \in \mathrm{Sym}\,(n) \Rightarrow \langle df(A)[U], df(A)[V]\rangle_{f(A)} = \langle df(A)[U_A], df(A)[V_A]\rangle_{f(A)} = \langle U_A, V_A\rangle\,.$$

In general, a submersion induces a local diffeomorphisms from horizontal spaces to the image manifold. In our case, the submersion σ provides a global parameterization of the manifold of symmetric matrices. Fix a matrix $A \in \mathrm{GL}(n)$ such that $\sigma(A) = AA^* = \Sigma$, and consider the open convex cone

$$\mathscr{H}_A^{++} = \mathrm{Sym}^{++}\,(n)\,A \subset \mathscr{H}_A.$$

We denote by σ_A the restriction to $\mathscr{H}_A^{++}$ of σ.

Proposition 5 *For all $A \in \mathrm{GL}\,(n)$, the mapping*

$$\sigma_A\colon \mathscr{H}_A^{++} \ni B \mapsto BB^* = C \in \mathrm{Sym}^{++}\,(n)$$

is a surjective bijection, with inverse

$$\sigma_A^{-1}(C) = C^{-1/2}(C^{1/2}\Sigma C^{1/2})^{1/2}C^{-1/2}A.$$

Proof For each $C \in \mathrm{Sym}^{++}\,(n)$, the equation

$$C = BB^* = (BA^{-1}A)(BA^{-1}A)^* = (BA^{-1})\Sigma(BA^{-1})^*$$

is a Riccati equation for BA^{-1}. As $B \in \mathrm{Sym}^{++}\,(n)\,A$, we have $BA^{-1} \in \mathrm{Sym}^{++}\,(n)$ and

$$BA^{-1} = C^{-1/2}(C^{1/2}\Sigma C^{1/2})^{1/2}C^{-1/2}$$

is the unique solution. □

We come now to the point, i.e., the construction of a metric based on horizontal vectors at a given matrix Σ. We are here using Prop. 1.

Proposition 6 *The inner product*

$$\langle U, V\rangle_\Sigma \equiv W_\Sigma(U, V) = \mathrm{Tr}\,(\mathscr{L}_\Sigma\,[U]\,\Sigma\,\mathscr{L}_\Sigma\,[V])\,, \quad U, V \in \mathrm{Sym}\,(n)\,,$$

defines a metric on $\mathrm{Sym}^{++}\,(n)$ *such that $\sigma\colon A \mapsto AA^*$ is a Riemannian submersion.*

Proof Let $X \in \mathrm{M}(n)$ and consider the decomposition of $X = X_V + X_H$ with X_V vertical at A and X_H horizontal at A. Then $d\sigma(A)[X] = d\sigma(A)[X_H]$ and the restriction of the derivative $d\sigma(A)$ to the vector space $\mathscr{H}_A$ of horizontal vectors at A is 1-to-1

onto the tangent space of $\mathrm{Sym}^{++}(n)$ at AA^*, that is, $\mathrm{Sym}(n)$. For such a restriction, for each $H \in \mathscr{H}_A$,

$$U = d\sigma(A)[H] = HA^* + AH^* = HA^{-1}AA^* + A(HA^{-1}A)^* \\ = (HA^{-1})AA^* + AA^*(HA^{-1})^* = (HA^{-1})AA^* + AA^*(HA^{-1}),$$

so that the inverse mapping of the restriction is given by

$$H = \left(d\sigma(A)|_{\mathscr{H}_A}\right)^{-1}(U) = \mathscr{L}_{AA^*}[U]A, \tag{29}$$

Let us push-forward the inner product from $\mathscr{H}_A$ to $T_{AA^*}\,\mathrm{Sym}^{++}(n)$.

From Eq. (29), we have

$$W_{AA^*}(U, V) = \left\langle \left(d\sigma(A)|_{\mathscr{H}_A}\right)^{-1}(U), \left(d\sigma(A)|_{\mathscr{H}_A}\right)^{-1}(V) \right\rangle = \\ \langle \mathscr{L}_{AA^*}[U]A, \mathscr{L}_{AA^*}[V]A \rangle = \mathrm{Tr}\left(\mathscr{L}_{AA^*}[U]AA^*\mathscr{L}_{AA^*}[V]\right).$$

which depends on $AA^* = \Sigma$ only. □

Next proposition provides a useful tensorial form of Wasserstein Riemannian metric.

Proposition 7 *It holds*

$$W_\Sigma(U, V) = \frac{1}{2}\langle \mathscr{L}_\Sigma[U], V \rangle \equiv \langle \mathscr{L}_\Sigma[U], V \rangle_2 .$$

Proof We have

$$\mathrm{Tr}\left(\mathscr{L}_\Sigma[U]\Sigma\mathscr{L}_\Sigma[V]\right) = \mathrm{Tr}\left(\mathscr{L}_\Sigma[V]\Sigma\mathscr{L}_\Sigma[U]\right) = \mathrm{Tr}\left(\mathscr{L}_\Sigma[U]\mathscr{L}_\Sigma[V]\Sigma\right),$$

and, taking the semi-sum of the first and the last term of the previous equation,

$$W_\Sigma(U, V) = \frac{1}{2}\mathrm{Tr}\left\{\mathscr{L}_\Sigma[U]\left[\mathscr{L}_\Sigma[V]\Sigma + \Sigma\mathscr{L}_\Sigma[V]\right]\right\} = \frac{1}{2}\mathrm{Tr}\left\{\mathscr{L}_\Sigma[U]V\right\}.$$

□

After having shown in Prop. 4 the existence of a metric geodesic for the Wasserstein distance, connecting a pair of matrices $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^{++}(n)$, we prove that the same curve is a Riemannian geodesic, see R.J. McCann [26] and also [8,36].

More generally, let us discuss the existence of affine horizontal surfaces in $\mathrm{GL}(n)$ and the existence of geodesically convex surfaces in $\mathrm{Sym}^{++}(n)$. As a particular case, the results give rise to the desired Riemannian geodesics.

A surface $\theta \mapsto A(\theta) \in \mathrm{GL}(n)$, with $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^n$, is called horizontal for the submersion $\sigma : A \mapsto AA^*$, if $\partial/\partial\theta_j A(\theta) \in \mathscr{H}_{A(\theta)}$ for each j and θ, i.e.,

$$\left(\frac{\partial}{\partial\theta_j}A(\theta)\right)A(\theta)^{-1} \in \mathrm{Sym}\,(n)\,. \tag{30}$$

A surface is horizontal if, and only if, every smooth curve which lies in it is horizontal.

Proposition 8 *1. The surface $\Theta \ni \theta \mapsto A(\theta) \in \mathrm{GL}\,(n)$ is horizontal for σ if, and only if,*

$$\frac{\partial}{\partial\theta_j}A^*(\theta)A(\theta) = A^*(\theta)\frac{\partial}{\partial\theta_j}A(\theta), \quad j = 1,\ldots,k, \quad \theta \in \Theta\,. \tag{31}$$

2. Let

$$A(\theta) = A_0 + \sum_{i=1}^{k}\theta_i(A_i - A_0), \quad \theta \in \Theta, \tag{32}$$

be a surface in $\mathrm{GL}\,(n)$ with the k-simplex of $\mathbb{R}^k$ contained in Θ. The surface is horizontal if, and only if,

$$A_j^*A_i = A_i^*A_j, \quad i,j = 0,\ldots,k.$$

3. Let be given $\Sigma_0, \Sigma_1 \in \mathrm{Sym}^{++}\,(n)$ and choose A_0, A_1 such that $\Sigma_0 = A_0A_0^*$ and $\Sigma_1 = A_1A_1^*$. The line

$$A(\theta) = (1-\theta)A_0 + \theta A_1 \tag{33}$$

is horizontal for θ in an open interval containing 0 and 1 if, and only if, $A_1 = TA_0$ with $T \in \mathrm{Sym}^{++}\,(n)$. This implies T is the solution of the Riccati equation $T\Sigma_0T = \Sigma_1$.

4. Let be given $\Sigma_j = A_jA_j^* \in \mathrm{Sym}^{++}\,(n)$, $j = 0, 1\ldots,k$. The surface

$$\theta \mapsto A_0 + \sum_{j=0}^{k}\theta_k(A_j - A_0)$$

is horizontal in an open set of parameters containing the k-simplex if, and only if, $A_i = T_{ij}A_j$ with $T_{ij} \in \mathrm{Sym}^{++}\,(n)$, $i,j = 0,\ldots,k$.

Proof 1. Eq. (30) is equivalent to $A^*(\theta)^{-1}\partial/\partial\theta_jA^*(\theta) = \partial/\partial\theta_jA(\theta)A(\theta)^{-1}$ hence to $\partial/\partial\theta_jA^*(\theta)A(\theta) = A^*(\theta)\partial/\partial\theta_jA(\theta)$.

2. For the surface in Eq. (32) we have $\partial/\partial\theta_jA(\theta) = A_j$ so that Eq. (31) becomes

$$A_j^*(\theta)A(\theta) = A^*(\theta)A_j(\theta), \quad j = 1,\ldots,k\,, \quad \theta \in \Theta.$$

If $\theta = 0$, it holds $A_j^*A_0 = A_0^*A_j$, $j = 1,\ldots,k$. If $\theta = e_i$ then it holds $A_j^*A_i = A_iA_j^*$ for $i,j = 1,\ldots,k$. The converse holds by linearity.

3. Assume $\theta \mapsto A(\theta)$ of Eq. (33) is horizontal on Θ. Then, from the previous item we know $A_1^* A_0 = A_0^* A_1$. In turn, this implies $A_0^{*-1} A_1^* = A_1 A_0^{-1}$, hence $T = A_1 A_0^{-1} \in \mathrm{Sym}\,(n)$. It follows $T\Sigma_0 T = A_1 A_0^{-1} \Sigma_0 (A_0^*)^{-1} A_1^* = \Sigma_1$. It remains to show that T is positive definite. Actually, it holds

$$(1-\theta)A_0 + \theta A_1 = ((1-\theta)I + \theta T)\, A_0 \in \mathrm{GL}\,(n)\,, \quad \theta \in \Theta.$$

If λ_i are eigenvalues of the matrix T, then the eigenvalues of the matrix $(1-\theta)I + \theta T$ are $(1-\theta) + \theta\lambda_i$. As they are never zero for any $\theta \in [0,1]$, it follows that no λ_i can be negative. The λ_i are not zero by assumption and the conclusion $T \in \mathrm{Sym}^{++}\,(n)$ follows.
Conversely, if $T \in \mathrm{Sym}^{++}\,(n)$ and $TA_0 = A_1$, then $A_1^* A_0 = A_0^* T A_0$ is symmetric. Consequently, for all θ such that $(1-\theta)A_0 + \theta A_1 \in \mathrm{GL}\,(n)$ the curve is horizontal. On the other hand, $(1-\theta)I + \theta T$ is the convex combination of positive definite matrices then it is positive definite on an open interval containing $[0,1]$.
4. Conversely, The proof follows exactly the same arguments as in the 2-points case of the previous item. □

The previous proposition shows that there is equality between the metric geodesic derived from the Wasserstein distance and the the geodesic we obtain from the submersion argument. Moreover, in the next Corollary we also characterize the existence of geodesically convex surfaces with given vertices.

Corollary 1 *1. Given $\Sigma_0, \Sigma_1 \in \mathrm{Sym}^{++}\,(n)$, there exists an open interval $\Theta \supset [0,1]$ such that the curve*

$$\Sigma(\theta) = ((1-\theta)I + \theta T)\, \Sigma_0\, ((1-\theta)I + \theta T)\,, \quad \theta \in \Theta, \tag{34}$$

is the Wasserstein Riemannian geodesic through Σ_0 and Σ_1, with $T\Sigma_0 T = \Sigma_1$.
2. Let $\Sigma_0, \ldots, \Sigma_k \in \mathrm{Sym}^{++}\,(n)$, there exists an open set Θ containing the k-simplex such that the surface

$$\Sigma(\theta) = \left(I + \sum_{j=1}^{k} \theta(T_j - I)\right) \Sigma_0 \left(I + \sum_{j=1}^{k} \theta(T_j - I)\right), \quad \theta \in \Theta,$$

is the Wasserstein Riemannian geodesic surface through $\Sigma_0, \ldots, \Sigma_k$ if, and only if, the matrices T_j, which are the positive definite solution of the Riccati equations $T_j \Sigma_0 T_j$, $j = 1, \ldots, k$, pairwise commute.

Proof 1. Pick $A_0 = \Sigma_0^{1/2} U$, with $U \in \mathrm{O}(n)$, and $A_1 = TA_0$, where T is the positive definite solution of the Riccati equation $T\Sigma_0 T = \Sigma_1$ and so $A_1 A_1^* = \Sigma_1$. By Prop. 8, Item 3, $\theta \mapsto A(\theta)$ is horizontal in $\mathrm{GL}\,(n)$. Consequently, $\Sigma(\theta) = A(\theta)A^*(\theta)$ is a geodesic.
2. In view of Prop. 8, Item 4, $T_{ij} = T_i T_j^{-1}$. The surface is horizontal if, and only if, each T_{ij} is symmetric, that is, $T_i T_j^{-1} = T_j T_i^{-1}$, which, in turn, is equivalent to $T_i T_j = T_j T_i$. □

Unlike the two-points case, the commutativity condition puts severe restrictions on the set of matrices $\Sigma_0,...,\Sigma_k$ generating a geodesic surface, when $k > 1$. For instance, if $\Sigma_0 = I$, then we have $T_i = \Sigma_i^{1/2}$. Hence, Corollary 9 entails that the matrices $I,\Sigma_1,...,\Sigma_k$ generate a geodesic surface if, and only if, they pairwise commute.

5 Wasserstein Riemannian exponential

We aim now at reformulating a Riemannian geodesic in terms of the exponential map. In other words, the purpose is that of writing the geodesic arc passing through a given point and having a given velocity at the point itself.

The velocity of the geodesic of Eq. (34) is

$$\dot{\Sigma}(\theta) = (T - I)\Sigma_0 + \Sigma_0(T - I) + 2\theta(T - I)\Sigma_0(T - I).$$

Using the horizontal lift $\Sigma(\theta) = A(\theta)A^*(\theta)$, the velocity turns out to be

$$\dot{\Sigma}(\theta) = \dot{A}(\theta)A^*(\theta) + A(\theta)\dot{A}^*(\theta) = \dot{A}(\theta)A^{-1}(\theta)\Sigma(\theta) + \Sigma(\theta)A^*(\theta)^{-1}\dot{A}^*(\theta),$$

where $\dot{A}(\theta)A^{-1}(\theta) \in \mathrm{Sym}\,(n)$ by Eq. (30). Therefore,

$$\dot{A}(\theta)A^{-1}(\theta) = A^*(\theta)^{-1}\dot{A}^*(\theta) = \mathscr{L}_{\Sigma(\theta)}\left[\dot{\Sigma}(\theta)\right].$$

In particular, the initial velocity is

$$\dot{\Sigma}(0) = (T - I)\Sigma(0) + \Sigma(0)(T - I). \tag{35}$$

and $T - I = \mathscr{L}_{\Sigma(0)}\left[\dot{\Sigma}(0)\right]$.

Let us compute the norm of the velocity in the Riemannian metric. The value of $W^2(\dot{\Sigma}, \dot{\Sigma})$ at $\Sigma(\theta)$ is

$$\begin{aligned}
&\mathrm{Tr}\left(\mathscr{L}_{\Sigma(\theta)}\left[\dot{\Sigma}(\theta)\right]\Sigma(\theta)\mathscr{L}_{\Sigma(\theta)}\left[\dot{\Sigma}(\theta)\right]\right)\\
&\quad= \mathrm{Tr}\left(\dot{A}(\theta)A^{-1}(\theta)A(\theta)A^*(\theta)A^*(\theta)^{-1}\dot{A}^*(\theta)\right)\\
&\quad= \mathrm{Tr}\left(\dot{A}(\theta)\dot{A}^*(\theta)\right) = \mathrm{Tr}\left((T - I)\Sigma(0)(T - I)\right).
\end{aligned}$$

It is constant, as we expect from the definition by isometric submersion. Also, we can confirm that the length of the geodesic is

$$\begin{aligned}
&\sqrt{\mathrm{Tr}\left((T - I)\Sigma(0)(T - I)\right)} = \sqrt{\mathrm{Tr}\left(\Sigma_0 + \Sigma_1 + T\Sigma_0 + \Sigma_0 T\right)}\\
&= \sqrt{\mathrm{Tr}\left(\Sigma_0 + \Sigma_1 + 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\right)}.
\end{aligned}$$

The last equality follows from the relation $\Sigma_0^{1/2}T\Sigma_0^{1/2} = (\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}$.

By substituting Eq. (35) into the equation of the geodesic (34), we get

$$\begin{aligned}\Sigma(\theta) &= \Sigma(0) + \theta\left[(T-I)\Sigma(0) + \Sigma(0)(T-I)\right] + \theta^2(T-I)\Sigma(0)(T-I)\\ &= \Sigma(0) + \theta\dot{\Sigma}(0) + \theta^2\mathscr{L}_{\Sigma(0)}[\dot{\Sigma}(0)]\Sigma(0)\mathscr{L}_{\Sigma(0)}[\dot{\Sigma}(0)].\end{aligned}$$

We are so led to the following definition, see [1, p. 101–102]) for example.

Definition 1 For any $C \in \mathrm{Sym}^{++}(n)$ and $V \in \mathrm{Sym}(n) \simeq T_C\,\mathrm{Sym}^{++}(n)$, the Wasserstein Riemannian exponential is

$$\mathrm{Exp}_C(V) = C + V + \mathscr{L}_C[V]C\mathscr{L}_C[V] = (\mathscr{L}_C[V] + I)C(\mathscr{L}_C[V] + I), \quad (36)$$

Next proposition collects some properties of the Riemannian exponential.

Proposition 9

1. *All geodesics emanating from a point $C \in \mathrm{Sym}^{++}(n)$ are of the form $\Sigma(\theta) = \mathrm{Exp}_C(\theta V)$, with $\theta \in J_V$, where J_V is the open interval about the origin:*

$$J_V = \left\{\theta \in \mathbb{R} \middle| I + \theta\mathscr{L}_C[V] \in \mathrm{Sym}^{++}(n)\right\}.$$

2. *The map $V \mapsto \mathrm{Exp}_C(V)$, restricted to the open set*

$$\Theta = \left\{V \in \mathrm{Sym}(n) : I + \mathscr{L}_C[V] \in \mathrm{Sym}^{++}(n)\right\},$$

is a diffeomorphism of Θ into $\mathrm{Sym}^{++}(n)$ with inverse

$$\mathrm{Log}_C(B) = (BC)^{1/2} + (CB)^{1/2} - 2C\,;$$

3. *The derivative of the Riemannian exponential is*

$$d_X\left(V \longmapsto \mathrm{Exp}_C(V)\right) = X + \mathscr{L}_C[X]C\mathscr{L}_C[V] + \mathscr{L}_C[V]C\mathscr{L}_C[X].$$

Remark 1 Notice that $I + \theta\mathscr{L}_C[V] = \mathscr{L}_C\left[\frac{1}{2}C^{-1} + \theta V\right]$ hence, $\theta \in J_V$ if $\frac{1}{2}C^{-1} + \theta V \in \mathrm{Sym}^{++}(n)$.

Clearly, $0 \in J_V$ and $\mathrm{Exp}_C(0) = C$ and the maximal open interval containing 0 in which $\mathrm{Exp}_C(\theta V) \in \mathrm{Sym}^{++}(n)$ is precisely J_V. Moreover, the interval J_V is unbounded from the right, i.e., it is of the kind $J_V = \left(\bar{\theta}, +\infty\right)$, provided $V \in \mathrm{Sym}^{+}(n)$. Likewise, $J_V = \left(-\infty, \bar{\theta}\right)$, if $-V \in \mathrm{Sym}^{+}(n)$. Similarly, Θ is an open set containing the origin and so $V \mapsto \mathrm{Exp}_C(V)$ is a local diffeomorphism around the origin.

Since the geodesics are not defined for all the values of the parameter $t \in \mathbb{R}$, we infer that the Riemannian manifold $\mathrm{Sym}^{++}(n)$ is geodesically incomplete. Of course this is not a surprising fact: $\mathrm{Sym}^{++}(n)$ is not a complete metric space, and hence Hopf-Rinow theorem implies that it cannot be geodesically complete, see M.P. do Carmo [10].

Proof 1. Let

$$\Sigma(\theta) = \mathrm{Exp}_C(\theta V) = C + \theta V + \theta^2 \mathscr{L}_C[V] C \mathscr{L}_C[V], \quad \theta \in J_V.$$

Clearly, $\Sigma(0) = C$ and $\dot{\Sigma}(0) = V$. Pick a scalar $\bar{\theta} \in J_V$ and consider the two matrices $\Sigma(0)$ and $\Sigma\left(\bar{\theta}\right)$ belonging to the curve Σ. Introduce the new parameterization $\tilde{\Sigma}(\tau) = \Sigma\left(\tau\bar{\theta}\right)$, so that $\tilde{\Sigma}(0) = \Sigma(0)$ and $\tilde{\Sigma}(1) = \Sigma\left(\bar{\theta}\right)$. We have,

$$\tilde{\Sigma}(\tau) = C + \tau(\bar{\theta}V) + \tau^2 \mathscr{L}_C\left[\bar{\theta}V\right] C \mathscr{L}_C\left[\bar{\theta}V\right]. \tag{37}$$

Setting $\tilde{T} - I = \mathscr{L}_C\left[\bar{\theta}V\right]$, we have $\tilde{T} \in \mathrm{Sym}^{++}(n)$ and

$$\tilde{T}C\tilde{T} = (I + \mathscr{L}_C\left[\bar{\theta}\right])C(I + \mathscr{L}_C\left[\bar{\theta}\right]) = \tilde{\Sigma}(1),$$

and the Eq. (37) above becomes

$$\begin{aligned}\tilde{\Sigma}(\tau) &= C + \tau(\tilde{T} - I)C + \tau C(\tilde{T} - I) + \tau^2(\tilde{T} - I)C(\tilde{T} - I) = \\ &= \left[(1-\tau)I + \tau\tilde{T}\right] C \left[(1-\tau)I + \tau\tilde{T}\right],\end{aligned}$$

which is the geodesic connecting $\Sigma(0) = \tilde{\Sigma}(0) = C$ to $\tilde{\Sigma}(1) = \Sigma(\bar{\theta})$.

2. By Eq. (36) the solution to Riccati equation

$$\mathrm{Exp}_C(V) = (I + \mathscr{L}_C[V])C(I + \mathscr{L}_C[V]) = B$$

is

$$I + \mathscr{L}_C[V] = C^{-1/2}(C^{1/2}BC^{1/2})^{1/2}C^{-1/2}$$

provided $I + \mathscr{L}_C[V] \in \mathrm{Sym}^{++}(n)$. This is true in a sufficiently small neighborhood $\|V\| < r$ of the origin. The inversion of the operator $\mathscr{L}_C[\cdot]$ and Eq. (9) provide the desired formula for $\mathrm{Log}_C(B)$.

3. The derivative follows from a simple bilinear computation.

□

The second order properties of the geodesic and the Riemannian exponential will be established in Sect. 7.6.

6 Natural gradient

We have found the form of the Riemannian metric associated to Wasserstein distance. In turn, the inner product equals the second order approximation of W^2. This is a

general fact, whose interpretation is based on the discussion of the natural gradient of the metric as solution to the problem

$$\begin{cases} \max f(X+H) - f(X) \\ \text{subject to} \\ W^2(X, X+H) = \varepsilon \text{ (small and fixed)} \end{cases}$$

which allows the identification of the direction of the maximal increase of the function f with the natural gradient, according to the name introduced by Amari [4], i.e., the Riemannian gradient as defined below.

The Riemannian gradient is the gradient with respect to the inner product of the metric. We denote by ∇ the gradient with respect to the inner product $\langle \cdot, \cdot \rangle_2$ and by grad the gradient with respect to the Riemannian metric. By Prop. 7, $W_\Sigma(X, Y) = \langle \mathscr{L}_\Sigma [X], Y \rangle_2$, hence for each smooth scalar field ϕ we have

$$\operatorname{grad} \phi(\Sigma) = \mathscr{L}_\Sigma^{-1}[\nabla \phi(\Sigma)] = \nabla \phi(\Sigma)\Sigma + \Sigma \nabla \phi(\Sigma),$$

where the second equality follows from the definition of $\mathscr{L}_\Sigma$. Conversely,

$$\mathscr{L}_\Sigma \left[\operatorname{grad} \phi(\Sigma)\right] = \nabla \phi(\Sigma).$$

The gradient flow of a smooth scalar field ϕ is the flow generated by the vector field

$$\gamma \mapsto (\gamma, -\operatorname{grad} \phi(\gamma)),$$

that is, the flow of the differential equation

$$\dot{\gamma}(\theta) = -\operatorname{grad} \phi(\gamma(\theta)) = -\left(\nabla \phi(\gamma(\theta))\gamma(\theta) + \gamma(\theta)\nabla \phi(\gamma(\theta))\right).$$

The gradient flow equation is the model for many optimization problems which are based on various discrete time approximations of the gradient flow. It should be noted that the expression of the natural gradient in the Wasserstein Riemannian metric is simple and does not require any time-consuming operation as it is the case in optimization methods using the Fisher Riemannian metric. We do not discuss this issue here and refer to [1,4,24].

6.1 Gradient flow and optimization

With reference to the full Gaussian distribution, one can consider smooth functions defined on $\mathbb{R}^n \times \mathrm{Sym}^{++}(n)$. The first component of the gradient does not require a special gradient as the Riemannian structure is the Euclidean one. The full gradient will thus have two components:

$$\begin{aligned} \operatorname{grad} \phi(\mu, \Sigma) &= \left(\nabla_1 \phi(\mu, \Sigma), \operatorname{grad}_2 \phi(\mu, \Sigma)\right) \\ &= \left(\nabla_1 \phi(\mu, \Sigma), \nabla_2 \phi(\mu, \Sigma)\Sigma + \Sigma \nabla_2 \phi(\mu, \Sigma)\right). \end{aligned} \tag{38}$$

An important example is based on the gradient flow of the mean value of an objective function $f\colon \mathbb{R}^n \to \mathbb{R}$. Its Euler scheme is used in optimization, see [1, Ch. 4] and [23]. In the second example in Sect. 6.2 we discuss the gradient flow of the entropy function of a centered Gaussian.

We call relaxation to the full Gaussian model of the objective function $f : \mathbb{R}^n \to \mathbb{R}$ the function

$$\phi(\mu, \Sigma) = \mathbb{E}\left[f(X)\right], \quad X \sim \mathrm{N}_n(\mu, \Sigma).$$

If we would include the Dirac measures in the Gaussian model, then $f(x) = \phi(x, 0)$ and the function ϕ would actually be an extension of the given function. However, we consider only $\Sigma \in \mathrm{Sym}^{++}(n)$ in order to work with a function defined on our manifold.

There are two ways to calculate the expected value as a function of μ and Σ. Each of them leads to a peculiar expression of the natural gradient.

The first one arises from the relation

$$\phi(\mu, \Sigma) = \mathbb{E}\Big[f(\Sigma^{1/2}Z + \mu)\Big], \quad Z \sim \mathrm{N}_n(0, I).$$

which will lead to an equation for the gradient involving the derivatives of f. The second one uses

$$\phi(\mu, \Sigma) = \int f(x)(2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^* \Sigma^{-1}(x-\mu)\right) dx.$$

In the second case, the natural gradient will be achieved by an equation not involving the gradient of the function f. Both forms have their own field of application.

Condider the first approach, under standard conditions regarding the derivation under the expectation sign. We have

$$\nabla_1 \phi(\mu, \Sigma) = \mathbb{E}\Big[\nabla f(\Sigma^{1/2}Z + \mu)\Big] = \mathbb{E}\left[\nabla f(X)\right].$$

By means of Eq. (14), it is straightforward to compute $d_U(\Sigma \mapsto \phi(\mu, \Sigma))$.

Note that ∇f is the column vector and so $\nabla^* f$ will be a row vector. We have

$$\begin{aligned} d_U\phi(\mu, \Sigma) &= \mathbb{E}\Big[df(\Sigma^{1/2}Z + \mu)[\mathscr{L}_{\Sigma^{1/2}}(U)\, Z]\Big] \\ &= \mathbb{E}\Big[\nabla^* f(\Sigma^{1/2}Z + \mu)\mathscr{L}_{\Sigma^{1/2}}(U)\, Z\Big] \\ &= \mathbb{E}\Big[\mathrm{Tr}\, \nabla^* f(\Sigma^{1/2}Z + \mu)\mathscr{L}_{\Sigma^{1/2}}(U)\, Z\Big]. \end{aligned}$$

Under symmetrization (and setting $X = \Sigma^{1/2}Z + \mu$):

$$\begin{aligned} d_U\phi(\mu, \Sigma) &= \frac{1}{2}\mathbb{E}\left[\mathrm{Tr}\, \mathscr{L}_{\Sigma^{1/2}}(U)\left(Z\nabla^* f(X) + \nabla f(X) Z\right)\right] \\ &= \left\langle U, \mathbb{E}\left(\left(Z\nabla^* f(X) + \nabla f(X) Z\right)\right)\right\rangle_{\Sigma^{1/2}} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\mathbb{E}\,\mathrm{Tr}\,\mathscr{L}_{\Sigma^{1/2}}\left(Z\nabla^* f(X) + \nabla f(X)Z\right)U \\
&= \left\langle \mathbb{E}\mathscr{L}_{\Sigma^{1/2}}\left(Z\nabla^* f(X) + \nabla f(X)Z\right), U\right\rangle_2 .
\end{aligned}
$$

It follows that

$$
\nabla_2\phi(\mu, \Sigma) = \mathbb{E}\left[\mathscr{L}_{\Sigma^{1/2}}\left(Z\nabla^* f(X) + \nabla f(X)Z\right)\right].
$$

Calculating the natural gradient:

$$
\begin{aligned}
&\mathrm{grad}_2\,\phi(\mu, \Sigma) \\
&\quad = \Sigma\mathbb{E}\left[\mathscr{L}_{\Sigma^{1/2}}\left(Z\nabla^* f(X) + \nabla f(X)Z\right)\right] + \mathbb{E}\left[\mathscr{L}_{\Sigma^{1/2}}\left(Z\nabla^* f(X) + \nabla f(X)Z\right)\right]\Sigma.
\end{aligned}
$$

If we set $\Xi = \mathbb{E}\left[Z\nabla^* f(X) + \nabla f(X)Z\right]$, the natural gradient admits the representation

$$
\mathrm{grad}_2\,\phi(\mu, \Sigma) = \Sigma\mathscr{L}_{\Sigma^{1/2}}(\Xi) + \mathscr{L}_{\Sigma^{1/2}}(\Xi)\,\Sigma.
$$

We move on to consider the second procedure. Following the standard computation of the Fisher score and starting from the log-density $p(x; \mu, \Sigma)$ of $\mathrm{N}_n(\mu, \Sigma)$, we have

$$
\begin{aligned}
\log p(x; \mu, \Sigma) &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(x-\mu)^*\Sigma^{-1}(x-\mu) \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma - \frac{1}{2} - \mathrm{Tr}\left(\Sigma^{-1}(x-\mu)(x-\mu)^*\right).
\end{aligned}
\tag{39}
$$

Denoting the partial derivative $d_u\,(\mu \longmapsto \log p(x; \mu, \Sigma))$ as $d_u \log p(x; \mu, \Sigma)$, and the other derivative $d_U\,(\Sigma \longmapsto \log p(x; \mu, \Sigma))$ as $d_U \log p(x; \mu, \Sigma)$, we get:

$$
\begin{aligned}
d_u \log p(x; \mu, \Sigma) &= (x-\mu)^*\Sigma^{-1}u = \left\langle \Sigma^{-1}(x-\mu), u\right\rangle \\
d_U \log p(x; \mu, \Sigma) &= -\frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}U\right) + \frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}U\,\Sigma^{-1}(x-\mu)(x-\mu)^*\right) \\
&= \frac{1}{2}\left\langle \Sigma^{-1}(x-\mu)(x-\mu)^*\Sigma^{-1} - \Sigma^{-1}, U\right\rangle \\
&= \left\langle \Sigma^{-1}\left((x-\mu)(x-\mu)^* - \Sigma\right)\Sigma^{-1}, U\right\rangle_2
\end{aligned}
$$

So that

$$
\begin{aligned}
d_u\phi(\mu, \Sigma) &= \int f(x)\, d_u \log p(x; \mu, \Sigma)\; p(x; \mu; \Sigma)\, dx \\
&= \left\langle \Sigma^{-1}\int f(x)(x-\mu)p(x; \mu; \Sigma)\, dx, u\right\rangle
\end{aligned}
$$

and

$$d_U\phi(\mu,\Sigma) = \int f(x)\, d_U \log p(x;\mu,\Sigma)\; p(x;\mu,\Sigma)\, dx$$
$$= \left\langle \Sigma^{-1} \int f(x)\left((x-\mu)(x-\mu)^* - \Sigma\right) p(x;\mu,\Sigma)\, dx\; \Sigma^{-1}, U \right\rangle_2 .$$

At last, thanks to Eq. (38), the natural gradient of $\phi(\mu,\Sigma)$ will be

$$\nabla_1\phi(\mu,\Sigma) = \Sigma^{-1}\int f(x)(x-\mu)p(x;\mu;\Sigma)\, dx$$
$$\operatorname{grad}_2\phi(\mu,\Sigma) = \int f(x)\left((x-\mu)(x-\mu)^* - \Sigma\right) p(x;\mu,\Sigma)\, dx\; \Sigma^{-1}$$
$$+ \Sigma^{-1}\int f(x)\left((x-\mu)(x-\mu)^* - \Sigma\right) p(x;\mu,\Sigma)\, dx.$$

6.2 Entropy gradient flow

The flow of entropy can be easily calculated by Eq. (39). We have

$$\mathscr{E}(\mu,\Sigma) = -\int \log p(x;\mu,\Sigma) p(x;\mu,\Sigma)\, dx$$
$$= \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det\Sigma - \frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}\Sigma\right)$$
$$= \frac{n}{2}(\log 2\pi - 1) + \frac{1}{2}\log\det\Sigma.$$

The entropy does not depend on μ so that $\nabla_1\mathscr{E}(\mu,\Sigma) = 0$. Moreover (see [22, §8.3]) we know that $\nabla\mathscr{E}(\Sigma) = \Sigma^{-1}$, so that

$$\operatorname{grad}\mathscr{E}(\Sigma) = (\Sigma^{-1}\Sigma + \Sigma\Sigma^{-1}) = 2I.$$

The entropic flow will be solution to the equations

$$\dot\mu(t) = 0, \quad \dot\Sigma(t) + 2I = 0,$$

that is

$$\mu(t) = \mu(0), \quad \Sigma(t) = \Sigma(0) - 2tI.$$

The integral curve is defined for all t such that $2t < \lambda_*$, λ_* being the minimum of the spectrum of $\Sigma(0)$.

7 Second order geometry

Recall that $\mathrm{Sym}^{++}(n)$ as an open set of the Hilbert space $\mathrm{Sym}(n)$, endowed with the inner product $\langle X, Y\rangle_2 = \frac{1}{2}\mathrm{Tr}(XY)$. Prop. 7 states that the Wasserstein Riemannian metric W can be expressed through the inner product of $\mathrm{Sym}(n)$, as

$$W_\Sigma(X, Y) = \langle X, Y\rangle_\Sigma = \langle \mathscr{L}_\Sigma[X], Y\rangle_2,$$

for each (Σ, X) and (Σ, Y) in the trivial tangent bundle $T\,\mathrm{Sym}^{++}(n) \simeq \mathrm{Sym}^{++}(n) \times \mathrm{Sym}(n)$. In the equation above, $\mathscr{L}\colon\ \mathrm{Sym}^{++}(n) \mapsto L(\mathrm{Sym}(n), \mathrm{Sym}(n))$ is the field of linear operators defining the Wasserstein metric with respect to the standard inner product.

In the trivial chart, a smooth vector field X is a smooth mapping $X\colon\ \mathrm{Sym}^{++}(n) \to \mathrm{Sym}(n)$. The action of the vector field X on the scalar field f that is, Xf, is expressed in the trivial chart by $d_X f$, i.e., the scalar field whose value at point Σ is the derivative of f in the direction $X(\Sigma)$. Similarly, $d_Y X$ denotes the vector field whose value at point Σ is the derivative at Σ of X in the direction $Y(\Sigma)$. The Lie bracket $[X, Y]$ of two smooth vector fields X, Y is given by $d_X Y - d_Y X$.

7.1 The moving frame

While we prefer to express our computation by matrix algebra, in some cases it may be useful to employ a vector basis. Let us now introduce a field of vector bases of particular interest.

The set of symmetric matrices

$$E^{p,q} = e_p e_q^* + e_q e_p^*, \quad p, q = 1, \ldots, n, \tag{40}$$

e_p being the p-th element of the standard basis of $\mathbb{R}^n$, spans the vector space $\mathrm{Sym}(n)$. Notice that $\mathrm{Tr}(E^{p,q}) = 2\delta_{p,q}$, where δ is the Kronecker symbol. To avoid repeated elements, a unique enumeration is obtained by taking indexes in the set A of the parts of $\{1, \ldots, n\}$ having 1 or 2 elements.

The generating set of Eq. (40) is related to the symmetric product of matrices by the equation

$$E^{p,q}E^{r,s} + E^{r,s}E^{p,q} = \delta_{q,r}E^{p,s} + \delta_{q,s}E^{p,r} + \delta_{p,r}E^{q,s} + \delta_{p,s}E^{q,r},$$

where δ is the Kronecker symbol.

In particular, if we take the trace of the equation above, we get

$$\left\langle E^{p,q}, E^{r,s}\right\rangle_2 = \delta_{p,r}\delta_{q,s} + \delta_{p,s}\delta_{q,r},$$

which in turn implies

$$\left\langle E^{p,q}, E^{r,s}\right\rangle_2 = \begin{cases} 0 & \text{if } \{p, q\} \neq \{r, s\}, \\ 1 & \text{if } \{p, q\} = \{r, s\} \text{ and } p \neq q, \\ 2 & \text{if } \{p, q\} = \{r, s\} \text{ and } p = q \end{cases}$$

In the sequel, we denote by $(E^\alpha)_{\alpha\in A}$ the vector basis above, properly normalized to obtain an orthonormal basis. We do not write down the normalizing constants in order to simplify the notation.

For each $\Sigma \in \mathrm{Sym}^{++}(n)$ the sequence

$$\mathscr{E}^\alpha(\Sigma) = E^\alpha \Sigma + \Sigma E^\alpha, \quad \alpha \in A, \tag{41}$$

is a vector basis of $\mathrm{Sym}(n) \simeq T_\Sigma \mathrm{Sym}^{++}(n)$, because it is the image of a vector basis under a linear mapping which is onto. We will call such a sequence of vector fields the (principal) moving frame.

Notice the following properties:

$$\mathscr{E}^\alpha = d_{E^\alpha}\Sigma^2\,; \quad \mathscr{L}_\Sigma\left[\mathscr{E}^\alpha(\Sigma)\right] = E^\alpha\,; \quad \mathscr{E}^\alpha(I) = 2E^\alpha\,.$$

At a generic point Σ, we can express each $\mathscr{E}^\alpha$ in the $(E^\beta)_\beta$'s orthonormal basis as

$$\mathscr{E}^\alpha(\Sigma) = \sum_\beta g_{\alpha,\beta}(\Sigma)E^\beta\,, \quad g_{\alpha,\beta}(\Sigma) = \mathrm{Tr}\left(E^\alpha \Sigma E^\beta\right). \tag{42}$$

Since

$$W_\Sigma(\mathscr{E}^\alpha, \mathscr{E}^\beta) = \mathrm{Tr}\left(\mathscr{L}_\Sigma\left[\mathscr{E}^\alpha(\Sigma)\right]\Sigma\mathscr{L}_\Sigma\left[\mathscr{E}^\beta(\Sigma)\right]\right) = \mathrm{Tr}\left(E^\alpha\Sigma E^\beta\right),$$

the matrix $[g_{\alpha,\beta}]_{\alpha,\beta}$ is the expression of the Riemannian metric in such a moving frame. Namely, if X, Y are vector fields expressed in the moving frame as $X = \sum_\alpha x_\alpha \mathscr{E}^\alpha$ and $Y = \sum_\beta y_\beta \mathscr{E}^\beta$, then

$$\begin{aligned} W_\Sigma(X,Y) &= \mathrm{Tr}\left(\mathscr{L}_\Sigma\left[\sum_\alpha x_\alpha(\Sigma)\mathscr{E}^\alpha\right]\Sigma(\Sigma)\,\mathscr{L}_\Sigma\left[\sum_\beta y_\beta(\Sigma)\mathscr{E}^\beta\right](\Sigma)\right) \\ &= \mathrm{Tr}\left(\left(\sum_\alpha x_\alpha(\Sigma)E^\alpha\right)\Sigma\left(\sum_\beta y_\beta(\Sigma)E^\beta\right)\right) = \sum_{\alpha,\beta} x_\alpha(\Sigma)y_\beta(\Sigma)g_{\alpha,\beta}(\Sigma). \end{aligned}$$

This expression of the inner product is to be compared to that used in [36].

In this way, any vector field X has two representations: one with respect to the moving frame $(\mathscr{E}^\alpha)_\alpha$ and another one with respect to the basis $(E^\alpha)_\alpha$. These two representations are related to each other as follows. We have

$$X = \sum_\alpha x_\alpha \mathscr{E}^\alpha = \sum_\alpha x_\alpha \sum_\beta g_{\alpha,\beta}E^\beta = \sum_\beta\left(\sum_\alpha x_\alpha g_{\alpha,\beta}\right)E^\beta,$$

so that

$$\langle X, E^\gamma \rangle_2 = \frac{1}{2} \operatorname{Tr}\left(X E^\gamma\right) = \sum_\beta \left(\sum_\alpha x_\alpha g_{\alpha,\beta} \right) \operatorname{Tr}\left(E^\beta E^\gamma\right) = \sum_\alpha x_\alpha g_{\alpha,\gamma},$$

hence, by applying the inverse matrix $[g^{\alpha,\beta}(\Sigma)] = [g_{\alpha,\beta}(\Sigma)]^{-1}$, we have

$$x_\alpha = \sum_\gamma g^{\alpha,\gamma} \langle X, E^\gamma \rangle_2 \ . \tag{43}$$

For example, $\mathscr{L}_\Sigma[V] = \sum_\alpha \ell^\alpha_\Sigma(V) \mathscr{E}^\alpha(\Sigma)$, with

$$\ell^\alpha_\Sigma(V) = \sum_\gamma g^{\alpha,\gamma}(\Sigma) \langle \mathscr{L}_\Sigma[V], E^\gamma \rangle_2 = W_\Sigma(V, \sum_\gamma g^{\alpha,\gamma} E^\gamma).$$

7.2 Covariant derivative in the moving frame

If X and Y are vector fields, denote by $D_Y X$ the action of a covariant derivative, namely, a bilinear operator satisfying, for each scalar field f, the following two conditions:

(CD1) $D_{fY} X = f D_Y X$,
(CD2) $D_Y(fX) = (d_Y f)X + f D_Y X$.

see e.g [10, Sect. 3] or [20, Ch. 8.4].

A convenient way to express a covariant derivative in the moving frame (41) is to define Christoffel symbols in the moving frame as

$$\sum_\gamma \Gamma^\gamma_{\alpha,\beta} \mathscr{E}^\gamma = D_{\mathscr{E}^\alpha} \mathscr{E}^\beta = E^\beta E^\alpha + E^\alpha E^\beta.$$

Each $\Gamma^\gamma_{\alpha,\beta}$ is to be computed by means of Eq. (43).

If $X = \sum_\alpha x_\alpha \mathscr{E}^\alpha$ and $Y = \sum_\beta y_\beta \mathscr{E}^\beta$, by using *(CD1)*, *(CD2)*, and Eq. (42), we obtain

$$\begin{aligned} D_X Y &= \sum_{\alpha,\beta} x_\alpha D_{\mathscr{E}^\alpha}(y_\beta \mathscr{E}^\beta) = \sum_{\alpha,\beta} x_\alpha \left(\left(d_{\mathscr{E}^\alpha} y_\beta\right) \mathscr{E}^\beta + y_\beta \left(D_{\mathscr{E}^\alpha} \mathscr{E}^\beta\right) \right) \\ &= \sum_{\alpha,\gamma} x_\alpha d_{\mathscr{E}^\alpha} y_\gamma \mathscr{E}^\gamma + \sum_{\alpha,\beta,\gamma} y_\beta \Gamma^\gamma_{\alpha,\beta} \mathscr{E}^\gamma = \sum_\gamma \sum_{\alpha,\beta} x_\alpha \left(d_{\mathscr{E}^\alpha} y_\gamma + y_\beta \Gamma^\gamma_{\alpha,\beta} \right) \mathscr{E}^\gamma. \end{aligned}$$

The inner product of $D_X Y$ and $Z = \sum_\delta z_\delta \mathscr{E}^\delta$ is

$$\langle D_X Y, Z \rangle_\Sigma = \sum_{\alpha,\beta,\gamma,\delta} x_\alpha \left(d_{\mathscr{E}^\alpha} y_\gamma + y_\beta \Gamma^\gamma_{\alpha,\beta} \right) g_{\delta,\gamma} z_\delta.$$

7.3 Levi-Civita derivative

The Levi-Civita (covariant) derivative of a vector field, is the unique covariant derivative D that, for all vector fields X, Y, Z, is:

(LC1) compatible with the metric, $d_X W(Y, Z) = W(D_X Y, Z) + W(Y, D_X Z)$,
(LC2) torsion-free, $D_Y X - D_X Y = [X, Y] = d_Y X - d_X Y$.

In order to keep a compact notation, it will be convenient to make use of the symmetrized of a matrix $A \in \mathrm{M}(n)$, defined by $\{A\}_S = \frac{1}{2}(A + A^*)$. If either A or B is symmetric, then $\mathrm{Tr}(\{A\}_S B) = \mathrm{Tr}(AB)$.

We denote by X, Y, Z smooth vector fields on $\mathrm{Sym}^{++}(n)$ and we shall use frequently the derivative of the vector field $\Sigma \mapsto \mathscr{L}_\Sigma[X]$. In view of Eq. (17) and under our notation for the symmetrization, we have

$$d_Y \mathscr{L}_\Sigma[X] = -2\mathscr{L}_\Sigma\left[\{\mathscr{L}_\Sigma[X]\, Y\}_S\right].$$

Proposition 10 *The Levi-Civita derivative $D_X Y$ is implicitly defined by*

$$\begin{aligned}\langle D_X Y, Z\rangle_\Sigma &= \langle d_X Y, Z\rangle_\Sigma + \left\langle X, \{\mathscr{L}_\Sigma[Y]\, Z\}_S\right\rangle_\Sigma \\ &\quad - \left\langle X, \{\mathscr{L}_\Sigma[Z]\, Y\}_S\right\rangle_\Sigma - \left\langle Y, \{\mathscr{L}_\Sigma[Z]\, X\}_S\right\rangle_\Sigma \\ &= \langle d_X Y, Z\rangle_\Sigma + \frac{1}{2}\mathrm{Tr}\left(\mathscr{L}_\Sigma[X]\, Z \mathscr{L}_\Sigma[Y]\right) \\ &\quad - \frac{1}{2}\mathrm{Tr}\left(\mathscr{L}_\Sigma[X]\, Y \mathscr{L}_\Sigma[Z]\right) - \frac{1}{2}\mathrm{Tr}\left(\mathscr{L}_\Sigma[Y]\, X \mathscr{L}_\Sigma[Z]\right), \end{aligned} \tag{44}$$

while the Levi-Civita derivative itself is given by

$$\begin{aligned} D_X Y = d_X Y &- \{\mathscr{L}_\Sigma[X]\, Y + \mathscr{L}_\Sigma[Y]\, X\}_S \\ &+ \{\Sigma \mathscr{L}_\Sigma[X]\, \mathscr{L}_\Sigma[Y] + \Sigma \mathscr{L}_\Sigma[Y]\, \mathscr{L}_\Sigma[X]\}_S . \end{aligned}$$

Proof In our case, Eq. *MD3* of [20, p. 205] becomes

$$\begin{aligned} 2\langle D_X Y, \mathscr{L}_\Sigma[Z]\rangle_2 &= 2\langle d_X Y, \mathscr{L}_\Sigma[Z]\rangle_2 + \langle Y, d_X \mathscr{L}_\Sigma[Z]\rangle_2 \\ &\quad + \langle X, d_Y \mathscr{L}_\Sigma[Z]\rangle_2 - \langle X, d_Z \mathscr{L}_\Sigma[Y]\rangle_2 . \end{aligned} \tag{45}$$

By Eq. (17) we have

$$\langle Y, d_X \mathscr{L}_\Sigma[Z]\rangle_2 = -2\left\langle Y, \mathscr{L}_\Sigma\left[\{\mathscr{L}_\Sigma[Z]\, X\}_S\right]\right\rangle_2 = -2\left\langle Y, \{\mathscr{L}_\Sigma[Z]\, X\}_S\right\rangle_\Sigma ,$$

and, analogously,

$$\begin{aligned} \langle X, d_Y \mathscr{L}_\Sigma[Z]\rangle_2 &= -2\left\langle X, \{\mathscr{L}_\Sigma[Z]\, Y\}_S\right\rangle_\Sigma , \\ \langle X, d_Z \mathscr{L}_\Sigma[Y]\rangle_2 &= -2\left\langle X, \{\mathscr{L}_\Sigma[Y]\, Z\}_S\right\rangle_\Sigma . \end{aligned}$$

This way, Eq. (45) becomes the first part of Eq. (44).

The second part of Eq. (44) is then easily obtained. For instance,

$$\left\langle X, \{\mathscr{L}_\Sigma[Z]\}_S \right\rangle_\Sigma = \frac{1}{2}\operatorname{Tr}\left(\mathscr{L}_\Sigma[X]\{Z\mathscr{L}_\Sigma[Y]\}_S\right) = \frac{1}{2}\operatorname{Tr}\left(\mathscr{L}_\Sigma[X]\, Z\mathscr{L}_\Sigma[Y]\right).$$

Regarding the explicit formula of the Levi-Civita derivative (10), observe that

$$\begin{aligned}
&\frac{1}{2}\operatorname{Tr}\left(\mathscr{L}_\Sigma[X]\, Z\mathscr{L}_\Sigma[Y]\right) = \frac{1}{2}\operatorname{Tr}\left(\mathscr{L}_\Sigma[Y]\,\mathscr{L}_\Sigma[X]\, Z\right)\\
&\quad= \frac{1}{2}\operatorname{Tr}\left(\{\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y]\}_S\, Z\right)\\
&\quad= \frac{1}{2}\operatorname{Tr}\left(\mathscr{L}_\Sigma\left[\{\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y]\}_S\,\Sigma + \Sigma\,\{\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y]\}_S\right] Z\right)\\
&\quad= \left\langle \{\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y]\}_S\,\Sigma + \Sigma\,\{\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y]\}_S, Z\right\rangle_\Sigma\\
&\quad= \left\langle \{\Sigma\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y]\}_S + \{\Sigma\mathscr{L}_\Sigma[Y]\,\mathscr{L}_\Sigma[X]\}_S, Z\right\rangle_\Sigma\\
&\quad= \left\langle \{\Sigma\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y] + \Sigma\mathscr{L}_\Sigma[Y]\,\mathscr{L}_\Sigma[X]\}_S, Z\right\rangle_\Sigma.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\frac{1}{2}\operatorname{Tr}\left(\mathscr{L}_\Sigma[X]\, Y\mathscr{L}_\Sigma[Z]\right) + \frac{1}{2}\operatorname{Tr}\left(\mathscr{L}_\Sigma[Y]\, X\mathscr{L}_\Sigma[Z]\right)\\
&\quad= \frac{1}{2}\operatorname{Tr}\left(\{\mathscr{L}_\Sigma[X]\, Y + \mathscr{L}_\Sigma[Y]\, X\}_S\,\mathscr{L}_\Sigma[Z]\right)\\
&\quad= \left\langle \{\mathscr{L}_\Sigma[X]\, Y + \mathscr{L}_\Sigma[Y]\, X\}_S, Z\right\rangle_\Sigma.
\end{aligned}$$

Therefore, Eq. (44) can be written as

$$\begin{aligned}
\langle D_X Y, Z\rangle_\Sigma = \langle &\Sigma d_X Y - \{\mathscr{L}_\Sigma[X]\, Y + \mathscr{L}_\Sigma[Y]\, X\}_S\\
&+ \{\Sigma\mathscr{L}_\Sigma[X]\,\mathscr{L}_\Sigma[Y] + \Sigma\mathscr{L}_\Sigma[Y]\,\mathscr{L}_\Sigma[X]\}_S\, Z\,\rangle,
\end{aligned}$$

and the desired result obtains. □

Observe that we have computed the Levi-Civita covariant derivative using its explicit expression in term of derivatives of the metric. However is easy to check the result directly using the properties of the Lyapunov operator.

7.4 Levi-Civita derivative in a moving frame

Let us explicit the Levi-Civita derivative in the moving frame (41). Note that $X(\Sigma) = \mathscr{E}^\alpha(\Sigma) = E^\alpha\Sigma + \Sigma E^\alpha$ and $Y(\Sigma) = \mathscr{E}^\beta(\Sigma) = E^\beta\Sigma + \Sigma E^\beta$ are vector fields.

Proposition 11 *For the Levi-Civita covariant derivative D, it holds*

$$D_{\mathscr{E}^\alpha}\mathscr{E}^\beta = E^\beta E^\alpha\Sigma + \Sigma E^\alpha E^\beta.$$

Proof Eq. (10) yields

$$D_{\mathscr{E}^\alpha}\mathscr{E}^\beta = d_{\mathscr{E}^\alpha}\mathscr{E}^\beta - \left\{\mathscr{L}_\Sigma\left[\mathscr{E}^\alpha\right]\mathscr{E}^\beta + \mathscr{L}_\Sigma\left[\mathscr{E}^\beta\right]\mathscr{E}^\alpha\right\}_S + \left\{\Sigma\mathscr{L}_\Sigma\left[\mathscr{E}^\alpha\right]\mathscr{L}_\Sigma\left[\mathscr{E}^\beta\right] + \Sigma\mathscr{L}_\Sigma\left[\mathscr{E}^\beta\right]\mathscr{L}_\Sigma\left[\mathscr{E}^\alpha\right]\right\}_S . \tag{46}$$

We are going to compute one by one the three terms in this equation.

The first term of Eq. (46) is

$$\begin{aligned} d_{\mathscr{E}^\alpha}\mathscr{E}^\beta &= d_{(E^\alpha\Sigma+\Sigma E^\alpha)}(E^\beta\Sigma + \Sigma E^\beta) \\ &= E^\beta(E^\alpha\Sigma + \Sigma E^\alpha) + (E^\alpha\Sigma + \Sigma E^\alpha)E^\beta \\ &= E^\beta E^\alpha\Sigma + E^\beta\Sigma E^\alpha + E^\alpha\Sigma E^\beta + \Sigma E^\alpha E^\beta . \end{aligned}$$

The second one is

$$\begin{aligned} &-\left\{\mathscr{L}_\Sigma\left[\mathscr{E}^\alpha\right]\mathscr{E}^\beta + \mathscr{L}_\Sigma\left[\mathscr{E}^\beta\right]\mathscr{E}^\alpha\right\}_S \\ &= -\left\{E^\alpha(E^\beta\Sigma + \Sigma E^\beta) + E^\beta(E^\alpha\Sigma + \Sigma E^\alpha)\right\}_S \\ &= -\left\{E^\alpha E^\beta\Sigma + E^\alpha\Sigma E^\beta + E^\beta E^\alpha\Sigma + E^\beta\Sigma E^\alpha\right\}_S \\ &= -\frac{1}{2}\left(E^\alpha E^\beta\Sigma + E^\beta E^\alpha\Sigma + \Sigma E^\beta E^\alpha + \Sigma E^\alpha E^\beta\right) - \left(E^\alpha\Sigma E^\beta + E^\beta\Sigma E^\alpha\right). \end{aligned}$$

Their sum is

$$\frac{1}{2}\left(E^\beta E^\alpha\Sigma + \Sigma E^\alpha E^\beta\right) - \frac{1}{2}\left(E^\alpha E^\beta\Sigma + \Sigma E^\beta E^\alpha\right).$$

The third term is

$$\begin{aligned} &\left\{\Sigma\mathscr{L}_\Sigma\left[\mathscr{E}^\alpha\right]\mathscr{L}_\Sigma\left[\mathscr{E}^\beta\right] + \Sigma\mathscr{L}_\Sigma\left[\mathscr{E}^\beta\right]\mathscr{L}_\Sigma\left[\mathscr{E}^\alpha\right]\right\}_S = \left\{\Sigma E^\alpha E^\beta + \Sigma E^\beta E^\alpha\right\}_S \\ &= \frac{1}{2}\left(\Sigma E^\alpha E^\beta + \Sigma E^\beta E^\alpha + E^\beta E^\alpha\Sigma + E^\alpha E^\beta\Sigma\right). \end{aligned}$$

□

The computation of the Christoffel symbols $\sum_\gamma \Gamma^\sigma_{\alpha,\beta}\mathscr{E}^\gamma = D_{\mathscr{E}^\alpha}\mathscr{E}^\beta$ would require the solution of the equations

$$E^\beta E^\alpha\Sigma + \Sigma E^\alpha E^\beta = \sum_\gamma \Gamma^\gamma_{\alpha,\beta}(\Sigma)\left(E^\gamma\Sigma + \Sigma E^\gamma\right).$$

We do not discuss that here.

Instead, let us take now $X = x_\alpha\mathscr{E}^\alpha$ and $Y = y_\beta\mathscr{E}^\beta$. Properties *(CD1)* and *(CD2)* lead to

$$\begin{aligned} D_{(x_\alpha\mathscr{E}^\alpha)}(y_\beta\mathscr{E}^\beta) &= x_\alpha D_{E^\alpha}(y_\beta E^\beta) = x_\alpha\left(d_{E^\alpha}y_\beta E^\beta + y_\beta D_{E^\alpha}E^\beta\right) \\ &= x_\alpha d_{E^\alpha}y_\beta E^\beta + x_\alpha y_\beta\left(E^\beta E^\alpha\Sigma + \Sigma E^\alpha E^\beta\right). \end{aligned}$$

Finally, for general X and Y,

$$D_X Y = \sum_{\alpha,\beta} x_\alpha d_{E^\alpha} y_\beta E^\beta + \sum_{\alpha,\beta} x_\alpha y_\beta \left(E^\beta E^\alpha \Sigma + \Sigma E^\alpha E^\beta \right)$$

which is the desired result.

7.5 Parallel transport

The expression of the Levi-Civita derivative in Eq. (44) can be re-written as

$$\langle D_X Y, Z \rangle_\Sigma = \langle d_X Y, Z \rangle_\Sigma + \langle \Gamma(\Sigma; X, Y), Z \rangle_\Sigma ,$$

where $\Gamma(\Sigma; \cdot, \cdot)$ is the symmetric tensor field defined by

$$\begin{aligned}
&\langle \Gamma(\Sigma; X, Y), Z \rangle_\Sigma = \\
&\quad \frac{1}{2} \operatorname{Tr} \left(\mathscr{L}_\Sigma [X] Z \mathscr{L}_\Sigma [Y] \right) - \frac{1}{2} \operatorname{Tr} \left(\mathscr{L}_\Sigma [X] Y \mathscr{L}_\Sigma [Z] \right) - \frac{1}{2} \operatorname{Tr} \left(\mathscr{L}_\Sigma [Y] X \mathscr{L}_\Sigma [Z] \right) \\
&\quad = \frac{1}{2} \operatorname{Tr} \left(\mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] Z \right) - \frac{1}{2} \operatorname{Tr} \left(\left(\mathscr{L}_\Sigma [X] Y + \mathscr{L}_\Sigma [Y] X \right) \mathscr{L}_\Sigma [Z] \right) \\
&\quad = \frac{1}{2} \operatorname{Tr} \left(\mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] \left(\mathscr{L}_\Sigma [Z] \Sigma + \Sigma \mathscr{L}_\Sigma [Z] \right) \right) \\
&\qquad - \frac{1}{2} \operatorname{Tr} \left(\left(\mathscr{L}_\Sigma [X] Y + \mathscr{L}_\Sigma [Y] X \right) \mathscr{L}_\Sigma [Z] \right) \\
&\quad = \frac{1}{2} \operatorname{Tr} \left(\left(\Sigma \mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] + \mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] \Sigma - \mathscr{L}_\Sigma [X] Y - \mathscr{L}_\Sigma [Y] X \right) \mathscr{L}_\Sigma [Z] \right) \\
&\quad = \left\langle \left\{ \Sigma \mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] + \mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] \Sigma - \mathscr{L}_\Sigma [X] Y - \mathscr{L}_\Sigma [Y] X \right\}_S , Z \right\rangle_\Sigma .
\end{aligned}$$

We have

$$\Gamma(\Sigma; X, Y) = \left\{ \Sigma \mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] + \mathscr{L}_\Sigma [Y] \mathscr{L}_\Sigma [X] \Sigma - \mathscr{L}_\Sigma [X] Y - \mathscr{L}_\Sigma [Y] X \right\}_S ,$$

and, on the diagonal,

$$\Gamma(\Sigma; X, X) = \Sigma \mathscr{L}_\Sigma [X] \mathscr{L}_\Sigma [X] + \mathscr{L}_\Sigma [X] \mathscr{L}_\Sigma [X] \Sigma - \mathscr{L}_\Sigma [X] X - X \mathscr{L}_\Sigma [X] .$$

$\Gamma(\Sigma; X, Y)$ is the expression in the trivial chart of the Christoffel symbol of the Levi-Civita derivative as in [17]. In [20], $-\Gamma$ is called the spray of the Levi-Civita derivative.

Given the Christoffel symbol, the linear differential equation of the parallel transport along a curve $t \mapsto \Sigma(t)$ is

$$\begin{cases} \dot{U}_V(t) + \Gamma(\Sigma(t); \dot{\Sigma}(t), U_V(t)) = 0, \\ U_V(0) = V, \end{cases}$$

see [20, VIII, §3 and §4]. Recall that the parallel transport for the Levi-Civita derivative is isometric.

We do not discuss here the representation in the moving frame of Eq. (7.5). We limit ourselves to mention that the action of the Christoffel symbol on vector fields expressed in the moving frame can be computed from

$$\begin{aligned}
&\Gamma(\Sigma; \mathscr{E}^\alpha, \mathscr{E}^\beta) \\
&\quad = \left\{ \Sigma \mathscr{L}_\Sigma \left[\mathscr{E}^\beta\right] \mathscr{L}_\Sigma \left[\mathscr{E}^\alpha\right] + \mathscr{L}_\Sigma \left[\mathscr{E}^\beta\right] \mathscr{L}_\Sigma \left[\mathscr{E}^\alpha\right] \Sigma \right. \\
&\qquad \left. - \mathscr{L}_\Sigma \left[\mathscr{E}^\alpha\right] \mathscr{E}^\beta - \mathscr{L}_\Sigma \left[\mathscr{E}^\beta\right] \mathscr{E}^\alpha \right\} \\
&\quad = \left\{ \Sigma E^\beta E^\alpha + E^\beta E^\alpha \Sigma - E^\alpha (E^\beta \Sigma + \Sigma E^\beta) - E^\beta (E^\alpha \Sigma + \Sigma E^\alpha) \right\}_S \\
&\quad = \left\{ \Sigma E^\beta E^\alpha + E^\beta E^\alpha \Sigma - E^\alpha E^\beta \Sigma - E^\alpha \Sigma E^\beta - E^\beta E^\alpha \Sigma - E^\beta \Sigma E^\alpha \right\}_S \\
&\quad = -(E^\alpha \Sigma E^\beta + E^\beta \Sigma E^\alpha).
\end{aligned}$$

7.6 Riemannian Hessian

According to [1, Def. 5.5.1] and [10, p. 141], the Riemannian Hessian of a smooth scalar field $\phi \colon \mathrm{Sym}^{++}(n) \to \mathbb{R}$, is the Levi-Civita covariant derivative of the natural gradient $\operatorname{grad}\phi$. Namely, for each vector field X, it is the vector field $\mathrm{Hess}_X \phi$ whose value at Σ is

$$\mathrm{Hess}_X \phi(\Sigma) = D_X(\operatorname{grad}\phi)(\Sigma) = D_X(\nabla\phi(\Sigma)\Sigma + \Sigma\nabla\phi(\Sigma)).$$

The associated symmetric bilinear form is (see [1, Prop. 5.5.3])

$$\mathrm{Hess}\,\phi(\Sigma)(X, Y) = \langle D_X(\operatorname{grad}\phi)(\Sigma), Y\rangle_\Sigma .$$

To our purpose it will be enough to compute the diagonal of the symmetric form. Therefore, letting $X = Z = V$ in the second part of Eq. (44), we obtain

$$\begin{aligned}
&\mathrm{Hess}\,\phi(\Sigma)(V, V) \\
&\quad = \langle d_V Y, V\rangle_\Sigma + \frac{1}{2}\operatorname{Tr}\left[\mathscr{L}_\Sigma[V]\, V \mathscr{L}_\Sigma[Y]\right] \\
&\qquad - \frac{1}{2}\operatorname{Tr}\left[\mathscr{L}_\Sigma[V]\, Y \mathscr{L}_\Sigma[V]\right] - \frac{1}{2}\operatorname{Tr}\left[\mathscr{L}_\Sigma[V]\, V \mathscr{L}_\Sigma[Y]\right] \\
&\quad = \langle d_V Y, V\rangle_\Sigma - \frac{1}{2}\operatorname{Tr}\left[\mathscr{L}_\Sigma[V]\, Y \mathscr{L}_\Sigma[V]\right],
\end{aligned}$$

where $Y = \operatorname{grad}\phi(\Sigma)$. After plugging $Y = \operatorname{grad}\phi(\Sigma) = \Sigma\nabla\phi(\Sigma) + \nabla\phi(\Sigma)\Sigma$ into it, we get easily

$$\begin{aligned}
\mathrm{Hess}\,\phi(\Sigma)(V, V) &= \left\langle \nabla_V^2\phi(\Sigma)\Sigma + \Sigma\nabla_V^2\phi(\Sigma), V\right\rangle_\Sigma \\
&\quad + \operatorname{Tr}\left[\nabla\phi(\Sigma)\, V \mathscr{L}_\Sigma[V]\right] - \operatorname{Tr}\left[\mathscr{L}_\Sigma[V]\nabla\phi(\Sigma)\Sigma\mathscr{L}_\Sigma[V]\right].
\end{aligned}$$

Plugging $V = \mathscr{L}_\Sigma[V]\,\Sigma + \Sigma\mathscr{L}_\Sigma[V]$ into the second term of the RHS, we have at last

$$\begin{aligned}&\operatorname{Hess}\phi(\Sigma)(V,V)\\&= \left\langle \nabla_V^2\phi(\Sigma)\,\Sigma + \Sigma\nabla_V^2\phi(\Sigma), V\right\rangle_\Sigma + \operatorname{Tr}\left[\nabla\phi(\Sigma)\,\mathscr{L}_\Sigma[V]\,\Sigma\mathscr{L}_\Sigma[V]\right]. \quad (47)\end{aligned}$$

Relation (47) substantiates the following important property that links the Hessian to the derivative along a geodesic (see the proof of Prop. 5.5.4 of [1]).

Proposition 12 *Let $\phi : \mathrm{Sym}^{++}(n) \to \mathbb{R}$ be a smooth scalar field and define*

$$\varphi(t) = \phi\left(\exp_\Sigma(tV)\right).$$

It holds

$$\ddot{\varphi}(0) = \operatorname{Hess}\phi(\Sigma)(V,V).$$

Proof By Prop. 9

$$\Sigma(t) = \operatorname{Exp}_\Sigma(tV) = \Sigma + tV + t^2\mathscr{L}_\Sigma[V]\Sigma\mathscr{L}_\Sigma[V]$$

where $\Sigma(0) = \Sigma$ and $\dot{\Sigma}(0) = V$. Hence $\dot{\varphi}(t) = \left\langle \nabla\phi(\Sigma(t)), \dot{\Sigma}(t)\right\rangle_2$, and

$$\ddot{\varphi}(t) = \left\langle \nabla^2\phi(\Sigma(t))[\dot{\Sigma}(t)], \dot{\Sigma}(t)\right\rangle_2 + \left\langle \nabla\phi(\Sigma(t)), \ddot{\Sigma}(t)\right\rangle_2$$

that evaluated at $t = 0$, provides

$$\ddot{\varphi}(0) = \left\langle \nabla^2\phi(\Sigma)[V], V\right\rangle_2 + 2\left\langle \nabla\phi(\Sigma), \mathscr{L}_\Sigma(V)\Sigma\mathscr{L}_\Sigma(V)\right\rangle_2.$$

In view of Eq. (47),

$$\operatorname{Hess}\phi(\Sigma)(V,V) = \left\langle \nabla_V^2\phi(\Sigma), V\right\rangle_2 + 2\left\langle \nabla\phi(\Sigma), \mathscr{L}_\Sigma[V]\,\Sigma\mathscr{L}_\Sigma[V]\right\rangle = \ddot{\varphi}(0).$$

□

8 Conclusion

In the present paper we have discussed in some detail the Wasserstein geometric properties of the Gaussian densities manifold. We have followed a known argument based on the geometric notion of submersion and we have improved upon what is known in the literature by offering a number of further results. In particular, we have studied the geodesic surfaces and provided an explicit form for the Riemannian exponential. More important, a new formulation of the metric based on the field of operators

$\Sigma \mapsto \mathscr{L}_\Sigma [\cdot]$ is introduced. This field of operator gives the Riemannian metric by the Frobenius inner product: $W_\Sigma(X, Y) = \langle \mathscr{L}_\Sigma [X], Y \rangle_2$. This gives rise to an explicit identification of the Riemannian gradient as well as to the calculation of the Levi-Civita covariant derivative, through the partial derivatives of the metric. The equations of the parallel transport and of the Riemannian Hessian have been also derived.

While the form of the natural gradient is simple and may be a source of applications such as those of interest in Machine Learning, the Levi-Civita covariant derivative turns out to be more involved and it is not clear how to use it in applications. However, we have produced a simpler form by the introduction of a special moving frame. In view of this issue, we have not proceeded in this paper to compute other geometrical quantities of interest, like the curvature tensor.

Numerical as well as simulation methods for the relevant equations of the geometry, like geodesics, parallel transport, Hessians, should be also considered. Applications of special interest are in the area of the linear optimization, by means of the natural gradient as direction of increase and by using the Riemannian exponential as a retraction, cf. [1] and in Amari monograph [5]. Also, second order optimization methods (Newton method), via the Riemannian Hessian and the Riemannian exponential, cf. [1] and [5], are source of promising researches.

The issue of a comparison between Fisher and Wasserstein metric is not taken into account here as it is, for example, in Chevallier et al. [11].

From the point of view of applications in Statistics and Machine Learning, the use of the full Gaussian model is in many cases not realistic. We expect our results to be used to compute the Wasserstein geometry induced on parsimonious sub-manifolds such as those listed below.

1. Sub-manifold of the correlation matrices i.e, with unitary diagonal elements. In this case, the tangent space at each point is the space of symmetric matrices with zero diagonal.
2. Sub-manifold of trace 1 matrices. This case is of particular interest in Physics and prompts for a generalization of the theory to complex Gaussians i.e., Gaussians densities on $\mathbb{C}^n$. Such distributions have Hermitian covariant matrices, a case that is discussed in [8].
3. Sub-manifold of the concentration matrices with a given sparsity pattern. Notice that concentration matrices and dispersion matrices are both elements of the same space $\mathrm{Sym}^{++}(n)$. In this case the statistical interpretation of the Wasserstein distance is not available but nevertheless other interpretations of the distance are mentioned in the Introduction.

Acknowledgements The authors wish to thank two anonymous referees for helpful comments. G. Pistone acknowledges the support of de Castro Statistics and Collegio Carlo Alberto. He is a member of GNAMPA-INdAM.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008). (with a foreword by Paul Van Dooren)

2. Aliprantis, C.D., Border, K.C.: Infinite Dimensional Analysis. A Hitchhiker's Guide, 3rd edn. Springer, Berlin (2006)
3. Amari, S., Nagaoka, H.: Methods of information geometry. American Mathematical Society, Providence (2000). **(translated from the 1993 Japanese original by Daishi Harada)**
4. Amari, S.I.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998). https://doi.org/10.1162/089976698300017746
5. Amari, S.I.: Information geometry and its applications. Appl. Math. Sci. 194 (2016). https://doi.org/10.1007/978-4-431-55978-8
6. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. Wiley Series in Probability and Statistics, 3rd edn. Wiley, Hoboken (2003)
7. Bhatia, R.: Positive Definite Matrices. Princeton Series in Applied Mathematics. Princeton University Press, Princeton (2007). ([2015] paperback edition of the 2007 original [MR2284176])
8. Bhatia, R., Jain, T., Lim, Y.: On the Bures-Wasserstein distance between positive definite matrices. Expositiones Mathematicae (2018). https://doi.org/10.1016/j.exmath.2018.01.002arXiv:1712.01504**(in press)**
9. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math. **44**(4), 375–417 (1991). https://doi.org/10.1002/cpa.3160440402
10. do Carmo, M.P.: Riemannian geometry. Mathematics: Theory and Applications. Birkhuser Boston Inc., Cambridge (1992). (translated from the second Portuguese edition by Francis Flaherty)
11. Chevallier, E., Kalunga, E., Angulo, J.: Kernel density estimation on spaces of Gaussian distributions and symmetric positive definite matrices. SIAM J. Imaging Sci. **10**(1), 191–215 (2017). https://doi.org/10.1137/15M1053566
12. Dowson, D.C., Landau, B.V.: The Fréchet distance between multivariate normal distributions. J. Multivar. Anal. **12**(3), 450–455 (1982). https://doi.org/10.1016/0047-259X(82)90077-X
13. Gelbrich, M.: On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. Math. Nachr. **147**, 185–203 (1990). https://doi.org/10.1002/mana.19901470121
14. Givens, C.R., Shortt, R.M.: A class of Wasserstein metrics for probability distributions. Michigan Math. J. **31**(2), 231–240 (1984). https://doi.org/10.1307/mmj/1029003026
15. Halmos, P.R.: Finite-dimensional vector spaces. The University Series in Undergraduate Mathematics, 2nd edn. D. Van Nostrand Co., Inc., Princeton-Toronto-New York-London (1958)
16. Hyvrinen, A.: Estimation of non-normalized statistical models by score matching. J. Mach. Learn. Res. **6**, 695–709 (2005)
17. Klingenberg, W.P.A.: Riemannian Geometry, *De Gruyter Studies in Mathematics*, vol. 1, 2nd edn. Walter de Gruyter & Co., Berlin (1995). https://doi.org/10.1515/9783110905120
18. Knott, M., Smith, C.S.: On the optimal mapping of distributions. J. Optim. Theory Appl. **43**(1), 39–49 (1984). https://doi.org/10.1007/BF00934745
19. Lafferty, J.D.: The density manifold and configuration space quantization. Trans. Am. Math. Soc. **305**(2), 699–741 (1988). https://doi.org/10.2307/2000885
20. Lang, S.: Differential and Riemannian manifolds, Graduate Texts in Mathematics, vol. 160, 3rd edn. Springer, Berlin Heidelberg (1995)
21. Lott, J.: Some geometric calculations on Wasserstein space. Comm. Math. Phys. **277**(2), 423–437 (2008). https://doi.org/10.1007/s00220-007-0367-3
22. Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley Series in Probability and Statistics. Wiley, Chichester (1999). (Revised reprint of the 1988 original)
23. Malagò, L., Pistone, G.: Combinatorial optimization with information geometry: Newton method. Entropy **16**, 4260–4289 (2014)
24. Malagò, L., Pistone, G.: Information geometry of the Gaussiandistributionin view of stochastic optimization. In: Proceedings of FOGA'15, held on January 17-20, 2015, Aberystwyth,Wales, 2015 (2015)
25. Mangasarian, O.L., Fromovitz, S.: The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. J. Math. Anal. Appl. **17**, 37–47 (1967). https://doi.org/10.1016/0022-247X(67)90163-1
26. McCann, R.J.: A convexity principle for interacting gases. Adv. Math. **128**(1), 153–179 (1997). https://doi.org/10.1006/aima.1997.1634
27. McCann, R.J.: Polar factorization of maps on Riemannian manifolds. Geom. Funct. Anal. **11**(3), 589–608 (2001). https://doi.org/10.1007/PL00001679

28. Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra Appl. **48**, 257–263 (1982). https://doi.org/10.1016/0024-3795(82)90112-4
29. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differential Equations **26**(1-2), 101–174 (2001)
30. Papadopoulos, A.: Metric spaces, convexity and non-positive curvature, *IRMA Lectures in Mathematics and Theoretical Physics*, vol. 6, 2nd edn. European Mathematical Society (EMS), Zürich (2014). https://doi.org/10.4171/132
31. Parry, M., Dawid, A.P., Lauritzen, S.: Proper local scoring rules. Ann. Stat. **40**(1), 561–592 (2012). https://doi.org/10.1214/12-AOS971
32. Pistone, G.: Nonparametric information geometry. In: F. Nielsen, F. Barbaresco (eds.) Geometric Science of Information, *Lecture Notes in Comput. Sci.*, vol. 8085, pp. 5–36. Springer, Heidelberg (2013). First International Conference, GSI 2013 Paris, France, August 28-30 (2013) **(proceedings)**
33. Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. **23**(5), 1543–1561 (1995)
34. Simoncini, V.: Computational methods for linear matrix equations. SIAM Rev. **58**(3), 377–441 (2016). https://doi.org/10.1137/130912839
35. Skovgaard, L.T.: A Riemannian geometry of the multivariate normal model. Scand. J. Stat. **11**(4), 211–223 (1984)
36. Takatsu, A.: Wasserstein geometry of Gaussian measures. Osaka J. Math. **48**(4), 1005–1026 (2011)
37. Villani, C.: Optimal Transport: Old and New. Grundlehren der mathematischen Wissenschaften. Springer, Berlin Heidelberg (2008)
38. Wachspress, E.L.: Trail to a Lyapunov equation solver. Comput. Math. Appl. **55**(8), 1653–1659 (2008). https://doi.org/10.1016/j.camwa.2007.04.048