



**HAL**  
open science

# On the Wasserstein geodesic principal component analysis of probability measures

Nina Vesseron, Elsa Cazelles, Alice Le Brigant, Thierry Klein

## ► To cite this version:

Nina Vesseron, Elsa Cazelles, Alice Le Brigant, Thierry Klein. On the Wasserstein geodesic principal component analysis of probability measures. International Conference on Learning Representations (ICLR), Apr 2026, Rio de Janeiro, Brazil. <hal-05133250>

**HAL Id: hal-05133250**

**<https://hal.science/hal-05133250v1>**

Submitted on 27 Jun 2025

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

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



HAL Authorization

---

# ON THE WASSERSTEIN GEODESIC PRINCIPAL COMPONENT ANALYSIS OF PROBABILITY MEASURES

---

**Nina Vesperon**  
CREST-ENSAE, IP Paris  
nina.vesperon@ensae.fr

**Elsa Cazelles\***  
CNRS, IRIT, Université de Toulouse  
elsa.cazelles@irit.fr

**Alice Le Brigant\***  
Université Paris 1 Panthéon Sorbonne  
alice.le-brigant@univ-paris1.fr

**Thierry Klein**  
ENAC, IMT, Université de Toulouse  
thierry.klein@math.univ-toulouse.fr

## ABSTRACT

This paper focuses on Geodesic Principal Component Analysis (GPCA) on a collection of probability distributions using the Otto-Wasserstein geometry. The goal is to identify geodesic curves in the space of probability measures that best capture the modes of variation of the underlying dataset. We first address the case of a collection of Gaussian distributions, and show how to lift the computations in the space of invertible linear maps. For the more general setting of absolutely continuous probability measures, we leverage a novel approach to parameterizing geodesics in Wasserstein space with neural networks. Finally, we compare to classical tangent PCA through various examples and provide illustrations on real-world datasets.

## 1 Introduction

In this paper, we are interested in computing the main modes of variation of a dataset of absolutely continuous (a.c.) probability measures supported in  $\mathbb{R}^d$ . For data points living in an arbitrary Hilbert space, the classical approach defined by Principal Component Analysis (PCA) consists in finding a sequence of nested affine subspaces on which the projected data retain a maximal part of the variance of the original dataset, or equivalently, yield best lower-dimensional approximations. When dealing with a set of a.c. probability distributions, a natural choice is to identify the probability measures with their probability density functions and to perform PCA on these using the  $L^2$  Hilbert metric. Unfortunately, as highlighted in [10], the components computed in this manner fail to capture the intrinsic structure of the distributions of the dataset. Using the Wasserstein metric  $W_2$  instead has proven to be a proper way to overcome these limitations, taking into account the geometry of the data.

The Wasserstein metric endows the space of probability distributions with a *Riemannian-like* structure, framing the problem as PCA on a (positively) curved Riemannian manifold. A first approach to solve this task, known as *Tangent PCA* (TPCA), consists in embedding the data into the tangent space at a reference point, and applying classical PCA in this flat space [12]. In the Wasserstein space, this approach is equivalent to using the linearized Wasserstein distance [39, 6]. TPCA is computationally advantageous but can generically induce distortion in the embedded data, depending on the curvature of the manifold at the reference point and the dispersion of the data. A more geometrically coherent approach is *Geodesic PCA* (GPCA) [15, 16], where principal modes of variations are geodesics that minimize the variance of the projection residuals. For a set of probability measures  $\nu_1, \dots, \nu_n$ , the first geodesic component solves

$$\inf_{t \rightarrow \mu(t) \text{ geodesic}} \sum_{i=1}^n \inf_t W_2^2(\mu(t), \nu_i). \quad (1)$$

Interestingly, unlike in the Hilbert setting, this criterion is *not* equivalent to maximizing the variance of the projections, which leads to a different notion of PCA on Riemannian manifolds (see [34, 35]).

---

\*These authors contributed equally to this work.

For one-dimensional probability measures, geodesic PCA (1) and its linearized approximation coincide, as the embedding into a tangent space is then an isometry when constrained to a convex set [5]. An algorithm in this case has been proposed in [10], with an approximate extension in dimension 2. For higher-dimensional measures, performing exact GPCA in the Wasserstein space remains challenging, starting with geodesic parametrization. In [33], the authors point out the difficulty of parametrizing tangent vectors at a given point –a necessary step to parametrize geodesic components– leading them to replace geodesics by generalized geodesics [3]. In this paper, we show that the delicate task of performing exact GPCA using the Wasserstein Riemannian structure can be bypassed, without changing the geometry, by lifting the probability distributions to the space of (non necessarily optimal) maps that pushforward a given reference measure, as described by Otto [28]. This approach is independent of the chosen reference measure and yields a convenient way to parametrize geodesics and define orthogonality with respect to the Wasserstein metric.

**Main contributions** We perform GPCA of a set of measures in the Wasserstein space in two cases: centered Gaussian distributions and a.c. probability measures. The method is exact in the sense that it does not rely on a linearization of the Wasserstein space, and the components are true geodesics that minimize the sum of squared norms of the projection residuals (1). In the Gaussian case, we leverage the Otto-Wasserstein geometry to lift the computations in the flat space of invertible matrices, in the spirit of [15]. We show that GPCA generically yields results very similar to those of tangent PCA, and give an illustration of when exception to this rule occurs. In the general case of a.c. probability distributions, we propose a novel parameterization of Wasserstein geodesics based on neural networks that, to the best of our knowledge, is the first to leverage Otto’s geometry for modeling geodesic paths in the Wasserstein space. We use multilayer perceptrons (MLPs), trained to minimize the cost function (1). We illustrate our methods on images and 3D point clouds.

**Organization of the paper** In Section 2, we present the Wasserstein metric and its restriction to Gaussian distributions, as well as the related Otto-Wasserstein geometries. We present GPCA for centered Gaussian distributions in Section 3, and the general case of a.c. probability measures is displayed in Section 4. Experiments are presented in Section 5, and then the paper ends with a discussion in Section 6. All the proofs and additional experiments are deferred in the appendices.

## 2 Background

**The Wasserstein distance** Optimal transport is about finding the optimal way to transport mass from one distribution  $\mu$  on  $\mathbb{R}^d$  to another  $\nu$  with respect to a ground cost, say the Euclidean squared distance. The total transport cost defines the *Wasserstein distance*  $W_2$  between a.c. measures  $\mu, \nu$  with moment of order 2, whose Monge formulation [27] is given by

$$W_2^2(\mu, \nu) = \int_{\mathbb{R}^d} \|x - T_\mu^\nu(x)\|^2 d\mu(x), \quad (2)$$

and where the map  $T_\mu^\nu$  is the  $\mu$ -a.s. unique gradient of a convex function verifying  $T_\mu^\nu \# \mu = \nu$  [8]. When the distributions  $\mu$  and  $\nu$  are centered (non-degenerate) Gaussian distributions, they can be identified with their covariance matrices  $\Sigma_\mu, \Sigma_\nu$  and (2) is referred to as the *Bures-Wasserstein distance*  $BW_2$  on the manifold  $S_d^{++}$  of symmetric positive definite (SPD) matrices (see e.g. [26, 4]):

$$BW_2^2(\Sigma_\mu, \Sigma_\nu) = \text{tr} \left[ \Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{1/2} \Sigma_\nu \Sigma_\mu^{1/2})^{1/2} \right]. \quad (3)$$

Both (2) and (3) can be induced by a Riemannian metric on their respective manifolds, i.e. the space of a.c. distributions and  $S_d^{++}$ , as we will see in the following. For more details, see Appendix B.

**Otto-Wasserstein geometry of centered Gaussian distributions** The set of centered non-degenerate Gaussian distributions on  $\mathbb{R}^d$  is identified with the manifold  $S_d^{++}$  of SPD matrices. The Riemannian geometry of the Bures-Wasserstein metric (3) can be described by considering  $S_d^{++}$  as the quotient of the manifold  $GL_d$  of invertible matrices by the right action of the orthogonal group  $O_d$ . In this geometry,  $GL_d$  is decomposed into equivalence classes called *fibers*. The fiber over  $\Sigma \in S_d^{++}$  is defined to be the pre-image of  $\Sigma$  under the projection  $\pi : A \in GL_d \mapsto AA^\top \in S_d^{++}$ , and can be obtained as the result of the action of  $O_d$  on a representative, e.g.  $\Sigma^{1/2}$  the only SPD square root of  $\Sigma$ :

$$\pi^{-1}(\Sigma) = \{A \in GL_d, AA^\top = \Sigma\} = \Sigma^{1/2} O_d. \quad (4)$$

Tangent vectors to  $GL_d$  are said to be *horizontal* if they are orthogonal to the fibers with respect to the Frobenius metric, i.e. if they belong to the space

$$\text{Hor}_A := \{X \in \mathbb{R}^{d \times d}, X^\top A - A^\top X = 0\}, \quad (5)$$

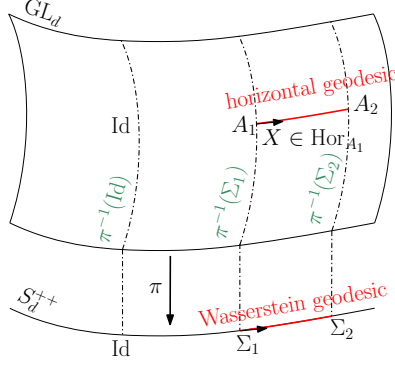


Figure 1: The Otto-Wasserstein geometry of centered non-degenerate Gaussian distributions. Figure inspired by [17].

for a given point  $A \in GL_d$ . Then the projection  $\pi$  defines an isometry between the horizontal subspace  $\text{Hor}_A$  equipped with the Frobenius inner product  $\langle X, Y \rangle := \text{tr}(XY^\top)$ , and  $S_d^{++}$  equipped with a Riemannian metric that induces the Bures-Wasserstein distance (3) as the geodesic distance. In particular, this means that moving horizontally along straight lines in the top space  $GL_d$  is equivalent to moving along geodesics in the bottom space  $S_d^{++}$  (see Figure 1), as recalled in the following proposition.

**Proposition 1** ([36, 23, 4]). *Any geodesic  $t \mapsto \Sigma(t)$  in  $S_d^{++}$  for the Bures-Wasserstein metric (3) is the  $\pi$ -projection of a horizontal line segment in  $GL_d$ , that is*

$$\Sigma(t) = \pi(A + tX) = (A + tX)(A + tX)^\top, \quad A \in GL_d, X \in \text{Hor}_A, \quad (6)$$

where  $t$  is defined in a certain time interval  $(t_{\min}, t_{\max})$ . Also, the Bures-Wasserstein distance between two covariance matrices  $\Sigma_1, \Sigma_2 \in S_d^{++}$  is given by the minimal distance between their fibers

$$BW_2(\Sigma_1, \Sigma_2) = \inf_{Q_1, Q_2 \in O_d} \|\Sigma_1^{1/2}Q_1 - \Sigma_2^{1/2}Q_2\| = \inf_{Q \in SO_d} \|\Sigma_1^{1/2} - \Sigma_2^{1/2}Q\|, \quad (7)$$

where  $\|\cdot\|$  is the Frobenius norm and  $SO_d$  is the special orthogonal group.

It is essential to note that the geodesic equation (6) cannot be extended at all time  $t \in \mathbb{R}$ , as the only geodesic lines are those obtained by translation [18, Proposition 3.6]. Therefore, (6) is only defined on a time interval  $(t_{\min}, t_{\max})$  that depends on the eigenvalues of  $XA^{-1}$  (see Appendix B.3). We refer the interested reader to Appendix B.2 for a more detailed presentation of this geometry.

**Otto-Wasserstein geometry of a.c. probability measures** The Riemannian structure described for Gaussian distributions is a special case of Otto's [28] more general construction : the bottom space becomes the space  $\text{Prob}(\Omega)$  of a.c. distributions supported in a compact set  $\Omega \subset \mathbb{R}^d$  while the top space is the space of diffeomorphisms  $\text{Diff}(\Omega)$  endowed with the  $L^2$  metric with respect to a fixed reference measure  $\rho$  (see Figure 13 in Appendix B; note that the compactness assumption on  $\Omega$  can be replaced by integrality conditions on the densities.) The *fibers* of  $\text{Diff}(\Omega)$  are then defined to be the pre-images under the projection  $\pi : \varphi \in \text{Diff}(\Omega) \mapsto \pi(\varphi) = \varphi\#\rho \in \text{Prob}(\Omega)$ . In this setting, *horizontal* displacements in  $\text{Diff}(\Omega)$  are along vector fields that are gradients of functions, and once again horizontal line segments project to Wasserstein geodesics :

**Proposition 2** ([28]). *Any geodesic  $t \mapsto \mu(t)$  for the Wasserstein metric (2) is the  $\pi$ -projection of a line segment in  $\text{Diff}(\Omega)$  going through a diffeomorphism  $\varphi$  at horizontal speed  $\nabla f \circ \varphi$  for some smooth function  $f \in \mathcal{C}(\mathbb{R}^d)$ . That is, for  $t$  defined in a certain interval  $(t_{\min}, t_{\max})$ ,*

$$\mu(t) = \pi(\varphi + t\nabla f \circ \varphi) = (\text{id} + t\nabla f)\#(\varphi\#\rho). \quad (8)$$

We emphasize that  $f$  need not be convex in (8), contrary to the more classical parametrization of geodesics due to McCann [24] between two distributions  $\mu_0$  and  $\mu_1 = \nabla u\#\mu_0$  :

$$\mu(t) = (\text{id} + t(\nabla u - \text{id}))\#\mu_0, \text{ with } t \in [0, 1] \text{ and } u \text{ a convex function.} \quad (9)$$

Equation (8) parametrizes geodesics provided that  $\text{id} + t\nabla f$  is a diffeomorphism, and thus it is defined on a time interval that depends on the eigenvalues of the hessian of  $f$ . On the other hand, the convexity condition on the function  $u$  in parametrization (9) ensures that time  $t$  is defined on  $[0, 1]$ . Both are completely equivalent (see Appendix B.3 for details).

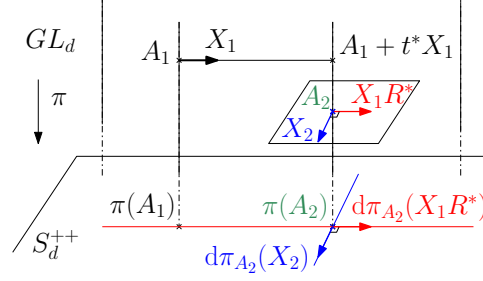


Figure 2: First (red) and second (blue) geodesic components of Gaussian GPCA, where  $d\pi_A$  denotes the differential of the projection  $\pi : A \mapsto AA^\top$  at  $A \in GL_d$ .

### 3 Geodesic PCA on centered Gaussian distributions

In this section, we perform exact GPCA on a set of centered (non-degenerate) Gaussian distributions with covariance matrices  $\Sigma_1, \dots, \Sigma_n \in S_d^{++}$  using the Bures-Wasserstein metric (3). Following [15], we define the first component as the geodesic  $t \mapsto \Sigma(t) \in S_d^{++}$  that minimizes the sum of squared residuals of the  $BW_2$ -projections of the data:

$$\inf_{t \mapsto \Sigma(t) \text{ geodesic}} \sum_{i=1}^n \inf_{t_i} BW_2^2(\Sigma(t_i), \Sigma_i). \quad (10)$$

The second principal component is defined to be the geodesic that minimizes the same cost function, with the constraint of intersecting the previous component orthogonally. The subsequent principal components have the additional constraint of going through the intersection of the first two principal geodesics. This definition does not impose that the geodesic components go through the Wasserstein barycenter (or Fréchet mean [1]), and in Section 5 we show an example where this is indeed not verified. This gives an observation of the phenomenon already described in [16] for spherical geometry. The proofs of this section are deferred to Appendix D.

**Learning the geodesic components** Following Proposition 1, we lift the GPCA problem (10) to the total space  $GL_d$  of Otto's fiber bundle. This has several advantages: the Bures-Wasserstein distance in the cost function (10) is replaced by the Frobenius norm  $\|\cdot\|$ , the geodesic is replaced by a horizontal line segment, and the projection times  $t_i$  become explicit. The price to pay is an optimization over variables  $(Q_i)_{i=1}^n \in SO_d$ , needed to represent the covariance matrices  $\Sigma_i$  by invertible matrices  $\Sigma_i^{1/2} Q_i$  in their respective fibers.

**Proposition 3.** *Let  $\pi : GL_d \rightarrow S_d^{++}$ ,  $A \mapsto AA^\top$  and  $(A_1, X_1, (Q_i)_{i=1}^n)$  be a solution of*

$$\inf F(A_1, X_1, (Q_i)_{i=1}^n) := \sum_{i=1}^n \|A_1 + p_{A_1, X_1}(t_i)X_1 - \Sigma_i^{1/2} Q_i\|^2, \quad (11)$$

$$\text{subject to } A_1 \in GL_d, X_1 \in \text{Hor}_{A_1}, \|X_1\|^2 = 1, Q_1, \dots, Q_n \in SO_d.$$

*Then there exist  $t_{\min}, t_{\max} \in \mathbb{R}$  such that the geodesic  $\Sigma : t \in [t_{\min}, t_{\max}] \mapsto \pi(A_1 + tX_1)$  in  $S_d^{++}$  minimizes (10).*

Here the  $t_i$  are the projection times given by  $t_i = \langle \Sigma_i^{1/2} Q_i - A_1, X_1 \rangle$ , and  $p_{A, X}$  is a projection operator that clips any  $t \in \mathbb{R}$  onto a closed interval  $[t_{\min}, t_{\max}]$  depending on  $A$  and  $X$ , such that  $A + p_{A, X}(t)X$  is invertible for any  $t$  in this interval. Clipping the time parameter of the line segment is necessary to ensure it remains within  $GL_d$  and projects onto a geodesic in the bottom space. The second component is a geodesic of  $S_d^{++}$  that orthogonally intersects the first component. Lifting again the problem in  $GL_d$ , this boils down to searching for a horizontal line  $t \mapsto A_2 + tX_2$  where  $A_2 = (A_1 + t^* X_1)R^*$  for a rotation matrix  $R^*$ , a time  $t^* \in [t_{\min}, t_{\max}]$  and a horizontal vector  $X_2 \in \text{Hor}_{A_2}$  such that  $\langle X_2, X_1 R^* \rangle = 0$ . The equation for  $A_2$  ensures that the  $\pi$ -projections of the first two horizontal lines intersect, while the condition on  $X_2$  ensures that they intersect orthogonally (since  $X_1 R^*$  is horizontal at  $A_2$  as can easily be checked). See Figure 2. The second component is thus defined by  $\Sigma_2(t) = \pi(A_2 + tX_2)$ , found by solving:

$$\begin{aligned} & \inf F(A_2, X_2, (Q_i)_{i=1}^n) \\ \text{subject to } & A_2 = (A_1 + t^* X_1)R^*, R^* \in SO_d, t^* \in [t_{\min}, t_{\max}] \\ & X_2 \in \text{Hor}_{A_2}, \|X_2\|^2 = 1, \langle X_2, X_1 R^* \rangle = 0, Q_1, \dots, Q_n \in SO_d. \end{aligned} \quad (12)$$

Note that this step requires to find new rotation matrices  $(Q_i)_{i=1}^n$ . The first two components fix the intersection point  $\pi(A_2)$  through which all other geodesic components will pass, see Figure 2. For every higher order component, we

search for a velocity vector  $X_k$  that is horizontal at some point in the fiber over  $\pi(A_2)$  and orthogonal to the lifts of the velocity vectors of the previous components. Details on the implementation of these components are given in Appendix D.2.

**On the restriction to the space of Gaussian distributions** Geodesic PCA can also be defined in the more general space of a.c. probability distributions, as presented in Section 4. A natural question that arises is whether performing GPCA in the whole space of probability distributions gives the same result as restricting to the space of Gaussian distributions, which is totally geodesic. To our knowledge, the answer to this question is not known in general, although it is true in one dimension.

**Proposition 4.** *Let  $\nu_i = \mathcal{N}(m_i, \sigma_i^2)$  for  $i = 1, \dots, n$ , be  $n$  univariate Gaussian distributions. The first principal geodesic component  $t \in [0, 1] \mapsto \mu(t)$  solving (1) remains in the space of Gaussian distributions for all  $t \in [0, 1]$ .*

## 4 Geodesic PCA on a.c. probability measures: GPCAGEN

We now tackle the task of performing GPCA on a set of a.c. probability measures  $\nu_1, \dots, \nu_n$  using the Otto-Wasserstein geometry. Similarly to the Gaussian case (10), the first geodesic principal component is a geodesic  $t \mapsto \mu(t)$  that solves

$$\inf_{t \mapsto \mu(t) \text{ geodesic}} \sum_{i=1}^n \inf_{t_i} W_2^2(\mu(t_i), \nu_i), \quad (13)$$

and the following components are defined as in the previous section. Here we assume that the probability measures  $\nu_1, \dots, \nu_n$  are known through samples i.e., for each  $\nu_i$  we are given a batch of samples  $\{x_1^{(i)}, \dots, x_{m_i}^{(i)}\}$  of size  $m_i$ . We propose a parameterization of the geodesic principal components based on Otto’s formulation, leveraging neural networks. Additionally, we introduce a dedicated cost function to optimize the different geodesic components.

**Parameterizing geodesics** Following Proposition 2 and equation (8), any geodesic  $t \mapsto \mu(t)$  in the Wasserstein space  $(\text{Prob}(\Omega), W_2)$  can be expressed as  $\mu(t) = (\varphi + t\nabla f \circ \varphi) \# \rho$ , for  $t$  in some interval  $[t_{\min}, t_{\max}]$ ,  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$  a diffeomorphism,  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  a smooth function, and  $\rho$  a fixed reference measure, taken to be the standard Gaussian distribution in this work. Using multilayer perceptrons (MLPs) to parametrize the functions  $\varphi$  and  $f$ , denoted  $\varphi_\theta$  and  $f_\psi$ , respectively, the curve

$$t \mapsto \mu_{\theta, \psi}(t) = (\text{id} + t\nabla f_\psi) \# (\varphi_\theta \# \rho)$$

is a geodesic for  $t \in [t_{\min}, t_{\max}]$ , provided that  $\text{id} + t\nabla f_\psi \in \text{Diff}(\Omega)$  for all  $t$  in this interval. Equivalently, this condition holds if the Hessian matrix  $I_d + tH_{f_\psi}(x)$  is positive definite for all  $x \in \mathbb{R}^d$  and  $t \in [t_{\min}, t_{\max}]$ , where  $H_{f_\psi}(x)$  denotes the Hessian of  $f_\psi$  at  $x$ . In practice, we enforce this constraint by monitoring the eigenvalues of  $I_d + tH_{f_\psi}(x)$  (see Appendix B.3) and either clipping  $t$  or adjusting the interval  $[t_{\min}, t_{\max}]$  to ensure that all eigenvalues remain positive. This representation enables to sample from the distributions along the geodesic. Specifically, given the learned vector field  $\varphi_\theta$  and function  $f_\psi$ , one can sample from  $\mu_{\theta, \psi}(t)$  by first drawing  $x \sim \rho$  and then applying the transformations  $\varphi_\theta$  and  $\text{id} + t\nabla f_\psi$  sequentially as  $\varphi_\theta(x) + t\nabla f_\psi(\varphi_\theta(x)) \sim \mu_{\theta, \psi}(t)$ .

**Learning the geodesic components** The first principal component in GPCA minimizes the objective in equation (13). The scalar variables  $t_i$  specify the projection time of each distribution  $\nu_i$  onto the geodesic  $t \mapsto \mu(t)$ . Leveraging the explicit form of Otto’s geodesic, (13) can be reformulated as:

$$\inf_{\substack{f \in \mathcal{C}(\mathbb{R}^d), \varphi \in \text{Diff}(\Omega) \\ t_1, \dots, t_n \in [t_{\min}, t_{\max}]}} \mathcal{L}(f, \varphi, t_1, \dots, t_n) := \sum_{i=1}^n W_2^2((\text{id} + t_i \nabla f) \# (\varphi \# \rho), \nu_i). \quad (14)$$

We jointly learn the parameters  $t_i$  together with the neural networks  $\varphi_\theta$  and  $f_\psi$  to minimize the objective (14). In practice, we approximate the squared Wasserstein distance  $W_2^2$  with the Sinkhorn divergence  $S_\varepsilon$ , and represent the distributions  $\rho$  and  $\nu_i$  using batches of  $m$  samples  $x_k \sim \rho$  and  $y_j \sim \nu_i$ . The optimization proceeds by updating the parameters based on a single distribution  $\nu_i$  sampled at each iteration, as detailed in Algorithm 1. To compute  $t_{\min}$  and  $t_{\max}$  on line 5 of Algorithm 1, we approximate the extremal eigenvalues of  $H_{f_\psi}$  by evaluating the largest and smallest eigenvalues over the finite set  $\{H_{f_\psi}(x_k)\}_{k=1}^m$ , and substitute these estimates into the theoretical bounds from Appendix B.3.

---

**Algorithm 1** Geodesic PCA algorithm for a.c. measures: GPCAGEN
 

---

- 1: Initialize  $\varphi_\theta$ ,  $f_\psi$  and the  $t_i$  for  $1 \leq i \leq n$
  - 2: **while** not converged **do**
  - 3:   **for**  $i = 1$  to  $n$  **do**
  - 4:     Draw  $m$  i.i.d samples  $y_j^{(i)} \sim \nu_i$  and draw  $m$  i.i.d samples  $x_k \sim \rho \quad 1 \leq j, k \leq m$
  - 5:     Estimate  $t_{\min}, t_{\max}$  with  $\{H_{f_\psi}(x_k)\}_{k=1}^m$  and set  $t'_i = \min(\max(t_i, t_{\min}), t_{\max})$
  - 6:      $z_k^{(i)} \leftarrow (\text{id} + t'_i \nabla f_\psi) \circ (\varphi_\theta)(x_k)$  for  $1 \leq k \leq m$
  - 7:      $\mathcal{L}_{\theta, \psi, t_i} \leftarrow S_\varepsilon \left( \frac{1}{m} \sum_{k=1}^m \delta_{z_k^{(i)}}, \frac{1}{m} \sum_{j=1}^m \delta_{y_j^{(i)}} \right)$
  - 8:     Update  $\varphi_\theta$ ,  $f_\psi$  and the  $t_i$  with  $\nabla \mathcal{L}_{\theta, \psi, t_i}$
  - 9:   **end for**
  - 10: **end while**
- 

The second principal component minimizes the objective in (13) subject to the constraint that it intersects the first component orthogonally. Similar to the first component, we use two MLP,  $f_{\psi_2}$  and  $\varphi_{\theta_2}$ , to parameterize the geodesic  $t \mapsto \mu_{\theta_2, \psi_2}(t)$ , along with  $n$  scalar variables  $t_i^2$ , to optimize the objective (14). We also introduce two additional scalar variables,  $t_{\text{inter}}^1$  and  $t_{\text{inter}}^2$ , which define the intersection times of the two geodesics, along with the regularization terms:

$$\mathcal{I}(\mu_1, \mu_2, t_{\text{inter}}^1, t_{\text{inter}}^2) = W_2^2(\mu_1(t_{\text{inter}}^1), \mu_2(t_{\text{inter}}^2)) \quad \text{and} \quad \mathcal{O}(g, h) = \frac{\langle g, h \rangle_{L^2(\rho)}^2}{\|g\|_{L^2(\rho)}^2 \|h\|_{L^2(\rho)}^2},$$

where  $\mathcal{I}$  enforces the geodesics  $\mu_1 = \mu_{\theta, \psi}$  and  $\mu_2 = \mu_{\theta_2, \psi_2}$  to intersect at the respective times  $t_{\text{inter}}^1$  and  $t_{\text{inter}}^2$ , and  $\mathcal{O}(g, h)$  ensures orthogonality between the corresponding horizontal vector fields  $g = \nabla f_\psi(\varphi_\theta)$  and  $h = \nabla f_{\psi_2}(\varphi_{\theta_2})$  in  $L^2(\rho)$ . The total objective used to optimize the second principal component incorporates these regularization terms and is given by:

$$\mathcal{L}(f_{\psi_2}, \varphi_{\theta_2}, t_1^2, \dots, t_n^2) + \lambda_I \mathcal{I}(\mu_{\theta, \psi}, \mu_{\theta_2, \psi_2}, t_{\text{inter}}^1, t_{\text{inter}}^2) + \lambda_O \mathcal{O}(\nabla f_\psi(\varphi_\theta), \nabla f_{\psi_2}(\varphi_{\theta_2}))$$

where  $\lambda_I$  and  $\lambda_O$  are the regularization parameters controlling the trade-off between the intersection and orthogonality regularization terms, respectively. The training algorithm used to optimize the second principal component follows the same structure as Algorithm 1, except for the seventh line, where the regularization terms, estimated using the minibatch  $x_k \sim \rho$ , are added to the loss function.

**Remark 1.** Note that without Otto's geometry, giving sense to a second GPCA component in the space  $\text{Prob}(\Omega)$  solving (13) is delicate as the notion of orthogonality does not exist in  $(\text{Prob}(\Omega), W_2)$ .

Higher-order components are estimated by solving (13), while enforcing orthogonality with previously computed components at the fixed intersection point  $\mu^* = \mu_{\theta, \psi}(t_{\text{inter}}^1) \approx \mu_{\theta_2, \psi_2}(t_{\text{inter}}^2)$ . With  $\mu^*$  fixed, the  $k$ -th component is optimized similarly to the second, introducing a new intersection time  $t_{\text{inter}}^k$ , but without updating  $t_{\text{inter}}^1$ . The optimization incorporates the same regularization terms to enforce intersection at  $\mu^*$  and orthogonality with all previously estimated components.

## 5 Experiments

### 5.1 Experiments on centered Gaussian distributions

In this section, we consider toy examples in  $S_2^{++}$  and compare GPCA to its widely used linearized approximation, TPCA (see Appendix C). We use two equivalent coordinate systems for covariance matrices in  $S_2^{++}$ : the first comes from the spectral decomposition

$$(a, b, \theta) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} \mapsto \Sigma(a, b, \theta) = P_\theta \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix} P_\theta^\top, \quad (15)$$

where  $P_\theta$  is the rotation matrix of angle  $\theta$ , and the second maps any SPD matrix to a point in the interior of the cone  $\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3, z > 0, z^2 < x^2 + y^2\}$  through the bijective parametrization

$$(x, y, z) \in \mathcal{C} \mapsto \begin{pmatrix} x+y & z \\ z & x-y \end{pmatrix} \in S_2^{++}. \quad (16)$$

Generically, GPCA and TPCA yield very similar results: for sets of  $n = 50$  covariance matrices randomly generated using a uniform distribution on the spectral parameter space as in (15), GPCA reduces the objective (10) of less than 1% w.r.t. TPCA, on average for 100 trials. This suggests that TPCA is generally a very good approximation of GPCA. Two extreme cases are described below : (i) GPCA is equivalent to TPCA and (ii) GPCA differs drastically from TPCA.

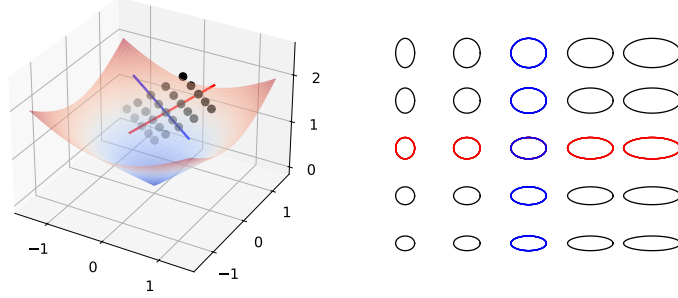


Figure 3: GPCA on a set of diagonal covariance matrices  $\Sigma_{ij}$  with varying eigenvalues  $1 \leq a_i^2 \leq 3$ ,  $1 \leq b_j^2 \leq 2$ . In cone coordinates (16), the matrices form a planar grid inside the cone of SPD matrices (**left**), and correspond to ellipses of varying width and height (**right**). The first component (red) captures the variation in  $a$ , while the second component (blue) captures the variation in  $b$ .

**Matrices with same orientation** If we consider a set of covariance matrices that live in the subspace  $\theta = \text{constant}$  in the parametrization (15), then both GPCA and TPCA yield exactly the same results, namely that of linear PCA in the  $(a, b)$ -coordinates. This is because any such subspace has zero curvature for the Wasserstein metric, and geodesics are straight lines in the  $(a, b)$ -coordinates (Appendix D.1). Figure 3 shows the geodesic components obtained for a set of matrices in the subspace  $\theta = 0$  that form a regular rectangular grid in the  $(a, b)$  coordinates, i.e.  $\Sigma_{ij} = \text{diag}(a_i^2, b_j^2)$  where the  $a_i$ 's and  $b_j$ 's are equally spaced. They are indeed straight lines that capture the variations in  $a$  and  $b$  respectively.

**Matrices with same eigenvalues** Now we consider covariance matrices that all have the same eigenvalues but different orientations. Specifically, we choose  $\Sigma_i = \Sigma(a, b, \theta_i)$  as defined in (15), for positive reals  $a > b$  and where  $\theta_i = i\pi/n$  for  $i = 0, \dots, n-1$  and an even number  $n$ . In the cone coordinates (16), the covariance matrices are displayed on a circle of equation  $x = \text{constant}$  (constant trace) and  $y^2 + z^2 = \text{constant}$  (constant determinant), as shown in Figure 4 (in practice, we choose a slightly open circle to break the symmetry). Then the Bures-Wasserstein barycenter [1] of the covariance matrices  $\Sigma_1, \dots, \Sigma_n$  is given by  $\bar{\Sigma} = (a+b)^2/4 I$  (see Proposition 15 in Appendix D.1 for a proof). When performing tangent PCA on  $\Sigma_1, \dots, \Sigma_n$  at the barycenter  $\bar{\Sigma}$ , the radial distances between  $\bar{\Sigma}$  and  $\Sigma_i$  are preserved, but not the pairwise distances between the  $\Sigma_i$ 's. The following result evaluates the level of this distortion.

**Proposition 5.** *Let  $\Sigma \in S_2^{++}$  with eigenvalues  $a^2, b^2$  and  $\Sigma' = P_\theta \Sigma P_\theta^\top$  where  $P_\theta$  is the rotation matrix of angle  $\theta$ . Then, denoting  $\bar{\Sigma} = ((a+b)/2)^2 I$ , we have*

$$\frac{BW_2^2(\Sigma, \Sigma')}{BW_{2, \bar{\Sigma}}^2(\Sigma, \Sigma')} = 1 - \left(\frac{a-b}{a+b}\right)^2 \cos^2 \theta + O((a-b)^4), \quad (17)$$

where  $BW_{2, \bar{\Sigma}}$  is the linearized Bures-Wasserstein distance at  $\bar{\Sigma}$  recalled in equation (30).

For a given  $\theta$ , Equation (17) shows that the distortion induced by linearization is most important for  $|a-b|/|a+b|$  close to 1, which corresponds to covariance matrices that are close to the border of the cone (since  $(a-b)^2/(a+b)^2 = (x^2 + y^2)/z^2$ ), see Figure 4 (**left**). Indeed, in that case, the results of GPCA can be very different from those of TPCA and the first component may not even go through the Wasserstein barycenter  $\bar{\Sigma}$ , see Figure 4 (**middle**) and Figure 8 in Appendix A. In that case GPCA may be seen as worse-behaved as TPCA, as some of the Gaussian distributions will project onto the first geodesic component boundaries, yielding a poor separation. Figure 4 (**right**) shows the percentage of improvement of the cost (10) (in terms of minimization) of GPCA with respect to TPCA, in the setting previously described for different values of the ratio  $|a-b|/|a+b|$ , thus backing up formula (17). GPCA is run 5 times for each value of the ratio, and the best result is kept.

## 5.2 Experiments on absolutely continuous distributions

We conduct a preliminary experiment on a synthetic dataset with known geodesics to verify that our algorithm, GPCAGEN (Section 4), accurately recovers the first two principal components. We then apply GPCAGEN to 3D point clouds from the ModelNet40 dataset [40] and to color distributions of images from the Landscape Pictures dataset [31]. For these experiments,  $f_\psi$  and  $\varphi_\theta$  are MLPs with four hidden layers of size 128 and an output layer of size 1 and  $d$  respectively. We found that setting the regularization coefficients  $\lambda_I$  and  $\lambda_O$  to 1.0 ensures the algorithm works as

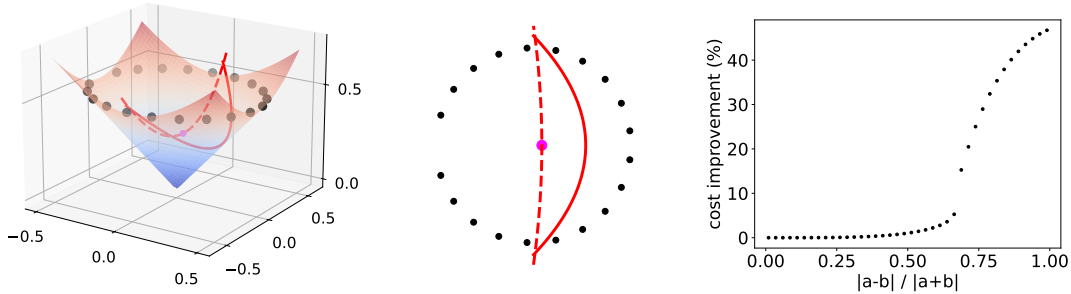


Figure 4: Comparison between tangent and geodesic PCA on a set of  $n = 20$  covariance matrices with same eigenvalues  $a^2, b^2$  and different orientations  $\theta$ . **(left)** They are equally spaced on an (open) circle in a horizontal plane inside the cone of SPD matrices. The first component of TPCA (dashed red line) goes through the Fréchet mean  $\bar{\Sigma}$  (magenta dot), a multiple of the identity, while the component of GPCA (solid red line) does not. Here  $|a - b|/|a + b| \approx 0.8$ . **(middle)** Representation of the left figure in the  $(x, y)$  coordinates. **(right)** Evolution of the first component cost improvement (in the sense of minimization) of GPCA with respect to TPCA, as a function of the ratio  $|a - b|/|a + b|$ .

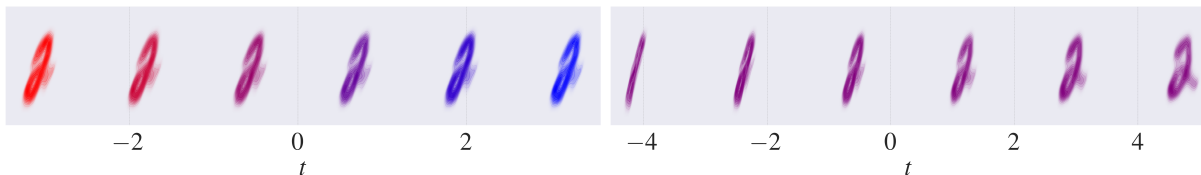


Figure 5: Densities of probability distributions uniformly sampled along the first and second principal geodesics components. GPCAGEN successfully recovers the two orthogonally intersecting geodesics constructed from MNIST data. The first component **(left)** captures variation in color space, while the second component **(right)** recovers the interpolation from the digit "1" to the digit "2".

expected in all experiments. A discussion of the regularization coefficients, along with details on the architecture and hyperparameters, is provided in Appendix E.

**MNIST geodesics.** We represent each image from the MNIST dataset [19] as a probability measure over  $\mathbb{R}^4$ . The grayscale pixel intensities define a normalized density over spatial coordinates  $(x, y) \in \mathbb{R}^2$ , and we enrich this representation by assigning each pixel two additional values corresponding to red and blue color channels. We construct two orthogonal geodesics: the first one interpolates between a digit "1" and a digit "2", both assigned a fixed purple by setting the color channels to 0.5. The second one is defined from the midpoint of the first, by linearly interpolating the color from red to blue. As shown in Figures 5 and 10, GPCAGEN successfully recovers the two geodesics intersecting orthogonally. A second experiment on the MNIST dataset is displayed in Appendix A.

**3D point cloud.** We use the ModelNet40 3D point cloud dataset [40] and apply GPCA to a subset of 100 randomly selected lamp point clouds. Figure 6 **(middle row)** and Figure 7 **(left)** demonstrate that the first principal component captures the distinction between hanging lamps (chandeliers) and standing lamps (floor lamps), while the second component reflects variations in the thickness of the lamp structure. We conduct a similar experiment on 100 point clouds from ModelNet40 representing different chairs. As shown in Figure 6 **(top row)** and Figure 11, the first principal component captures the height of the seat, while the second component distinguishes between chairs and armchairs.

**Landscape images.** We use 39 images from the Landscape Pictures dataset [31] and compute GPCAGEN to the corresponding point clouds, where each point cloud represents color distribution in the image. Figure 6 **(bottom row)** and Figure 7 **(right)** show that the first component captures variations in overall brightness, ranging from bright to dark images, while the second component distinguishes between images that are predominantly green and those that are predominantly blue.

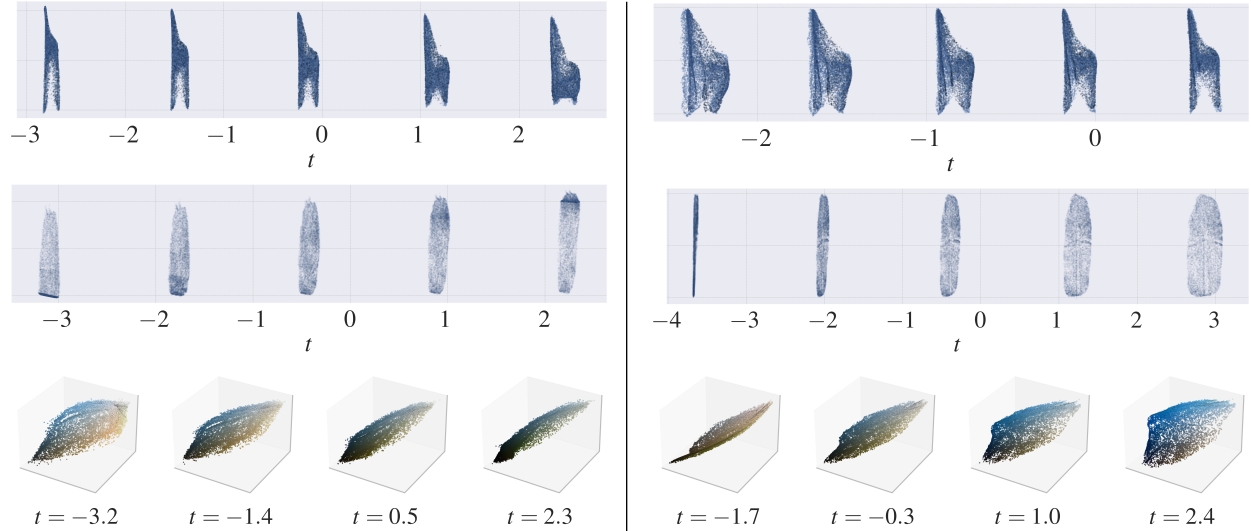


Figure 6: Empirical distributions uniformly sampled along the geodesics corresponding to the first (**left**) and second (**right**) principal components, as computed by GPCAGEN in the 3D point cloud of chairs experiment (**top row**), the 3D point cloud of lamps experiment (**middle row**) and the Landscape images experiment (**bottom row**).

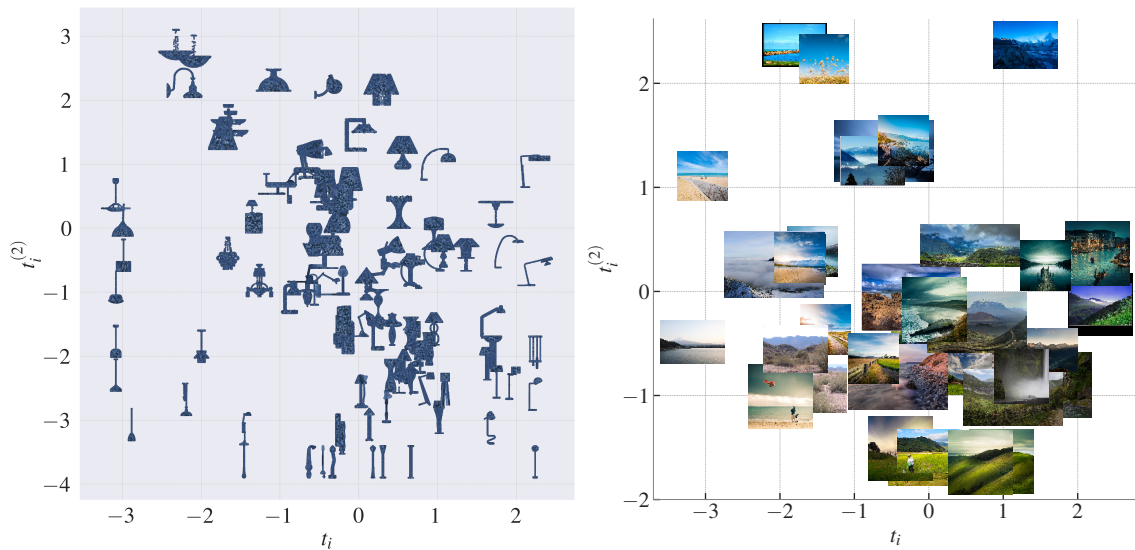


Figure 7: Each lamp point cloud (**left**) and each image (**right**) is embedded in the plane according to its projection times onto the first and second principal components computed by GPCAGEN.

## 6 Discussion

We have proposed two methods for computing exact GPCA : one tailored for Gaussian distributions and the other for the more general case of a.c. probability distributions. In the Gaussian case, our experiments suggest that GPCA and TPCA generically yield very similar results, except for distributions with covariance matrices that are close to the boundary of the SPD cone, for which GPCA can yield undesirable effects as suggested by the pathological example of Figure 4. In the general case of a.c. probability measures, a key advantage of our approach is that it operates directly on continuous distributions, avoiding the need for empirical approximations of the  $\nu_i$ , which would require equal sample sizes and can introduce discretization artifacts in the recovered components. Additionally, our method enables sampling from any point along the geodesic components—something not possible with discrete approximations commonly used in TPCA. Otto’s parametrization also allowed us to avoid relying on input convex neural networks (ICNNs) by not

requiring convex functions, with the trade-off being the need to estimate the eigenvalues of the Hessian of  $f$ . This perspective opens new directions for parametrizing convex functions without imposing hard architectural constraints.

## Acknowledgments

This work benefited from financial support from the French government managed by the National Agency for Research under the France 2030 program, with the reference ANR-23-PEIA-0004. This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-103245). This work was partially supported by Hi! Paris through the PhD funding of Nina Vesseron.

## References

- [1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- [2] Luigi Ambrosio, Alberto Bressan, Dirk Helbing, Axel Klar, Enrique Zuazua, Luigi Ambrosio, and Nicola Gigli. A user’s guide to optimal transport. *Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009*, Editors: Benedetto Piccoli, Michel Rasche, pages 1–155, 2013.
- [3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- [4] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.
- [5] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l’Institut Henri Poincaré - Probabilités et Statistiques*, 53(1):1–26, 2017.
- [6] Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution’s template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.
- [7] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.
- [8] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.
- [9] Anna Calissano, Aasa Feragen, and Simone Vantini. Populations of unlabelled networks: Graph space geometry and generalized geodesic principal components. *Biometrika*, pages 147–170, 2024.
- [10] Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, and Nicolas Papadakis. Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456, 2018.
- [11] David G Ebin and Jerrold Marsden. Groups of diffeomorphisms and the motion of an incompressible fluid. *Annals of Mathematics*, 92(1):102–163, 1970.
- [12] P. Thomas Fletcher, Conglin Lu, and Sarang Joshi. Statistics of shape via principal geodesic analysis on Lie groups. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.
- [13] Philipp Harms, Elodie Maignant, and Stefan Schlager. Approximation of Riemannian distances and applications to distance-based learning on manifolds. *arXiv preprint :1904.11860*, 2019.
- [14] Wen Huang and Ke Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194(1):371–413, 2022.
- [15] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, pages 1–58, 2010.
- [16] Stephan Huckemann and Herbert Ziezold. Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 38(2):299–319, 2006.
- [17] Boris Khesin, Gerard Misiołek, and Klas Modin. Geometric hydrodynamics and infinite-dimensional Newton’s equations. *Bulletin of the American Mathematical Society*, 58(3):377–442, 2021.
- [18] Benoit Kloeckner. A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9(2):297–323, 2010.
- [19] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- [20] Cyril Letrouit and Quentin Mérigot. Gluing methods for quantitative stability of optimal transport maps. *arXiv preprint arXiv:2411.04908*, 2024.
- [21] Chi-Kwong Li and Fuzhen Zhang. Eigenvalue continuity and Gersgorin’s theorem. *Electronic Journal of Linear Algebra*, 2019.
- [22] Yingjie Ma, Xi Gao, Chao Liu, and Jie Li. Improved SQP and SLSQP algorithms for feasible path-based process optimisation. *Computers & Chemical Engineering*, page 108751, 2024.

- [23] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry*, 1:137–179, 2018.
- [24] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.
- [25] Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. Geomstats: A Python package for Riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020.
- [26] Klas Modin. Geometry of matrix decompositions seen through optimal transport and information geometry. *Journal of Geometric Mechanics*, 9(3):335–390, 2017.
- [27] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.
- [28] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 2001.
- [29] Michael J.D. Powell. On search directions for minimization algorithms. *Mathematical programming*, 4:193–201, 1973.
- [30] James O. Ramsay and Bernard W. Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2002.
- [31] Arnaud Rougetet. Landscape pictures dataset. <https://www.kaggle.com/datasets/arnaud58/landscape-pictures>, 2020.
- [32] Klaus D Schmidt. On inequalities for moments and the covariance of monotone functions. *Insurance: Mathematics and Economics*, 55:91–95, 2014.
- [33] Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. *Advances in Neural Information Processing Systems*, 28, 2015.
- [34] Stefan Sommer, François Lauze, and Mads Nielsen. The differential of the exponential map, Jacobi fields and exact principal geodesic analysis. *CoRR, abs/1008.1902*, 2010.
- [35] Stefan Sommer, François Lauze, and Mads Nielsen. Optimization over geodesics for exact principal geodesic analysis. *Advances in Computational Mathematics*, 40:283–313, 2014.
- [36] Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka J. Math.*, 48:1005–1026, 2011.
- [37] Yann Thanwerdas. *Riemannian and stratified geometries on covariance and correlation matrices*. PhD thesis, Université Côte d’Azur, 2022.
- [38] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [39] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101:254–269, 2013.
- [40] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes, 2015.
- [41] Juan Zhou, Kangkang Deng, Hongxia Wang, and Zheng Peng. Inexact Riemannian gradient descent method for nonconvex optimization. *arXiv preprint arXiv:2409.11181*, 2024.

## A Additional experiments and figures

In this section, we present additional figures to further explain the experiments described in the paper as well as an additional experiment on a.c. distributions.

Figure 8 concerns the experiment on Gaussian distributions with diagonal covariances described in Section 5.1 corresponding to Figure 4. It shows all three principal components found by tangent PCA (**left**) and geodesic PCA, in two equally optimal solutions (**middle, right**).

Figure 10 displays on the plane the two first geodesic components of the MNIST experiment of Section 5.2, while Figure 11 shows the planar representation of the 3D point cloud of chairs experiment given by the projection onto the first two geodesic components found by GPCAGEN algorithm and depicted in Figure 6 (**top row**).

Finally, we present an additional experiment on the MNIST dataset. We use the same color construction as in the experiment presented in Section 5.2, we then apply GPCAGEN to a dataset of 20 red digits "1", 20 blue digits "1", 20 red digits "2", and 20 blue digits "2" (see Figure 12). As shown in Figures 9 and 12, GPCAGEN again identifies two orthogonal geodesics: the first primarily captures variation in color, while the second captures variation in shape—from digit "2" to digit "1".

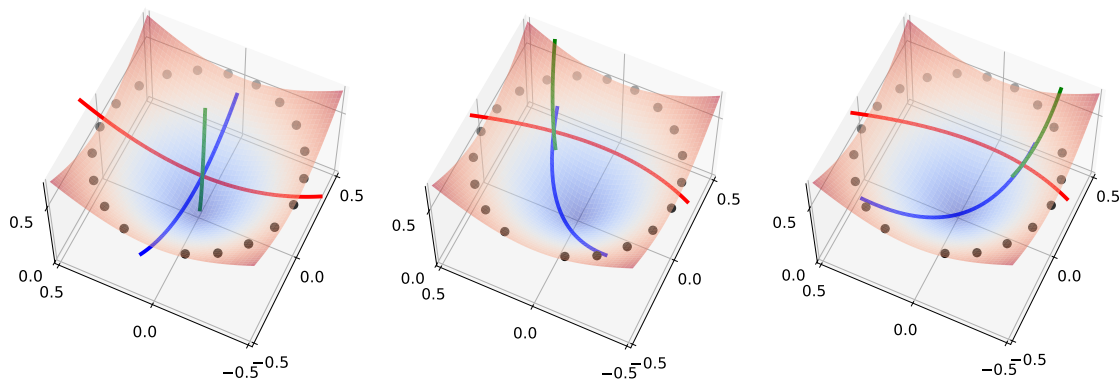


Figure 8: Principal geodesic components of a set of Gaussian distributions whose covariance matrices have same eigenvalues and different orientations, as described in Section 5.1. Tangent PCA yields a unique solution (**left**) where geodesic components cross at the barycenter, while geodesic PCA yields two equally optimal solutions (**middle, right**) where the geodesic components cross at another point. The first geodesic component is shown in red, the second in blue, the third in green.

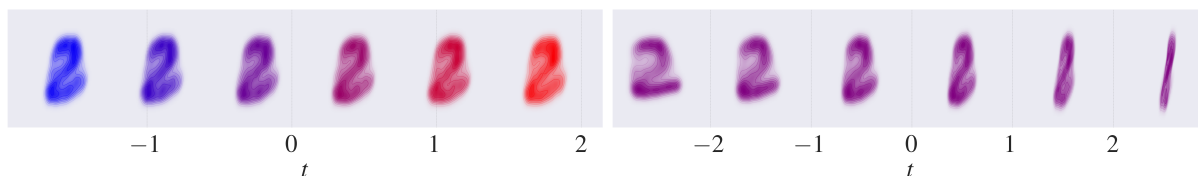


Figure 9: Densities of probability distributions uniformly sampled along the geodesics corresponding to the first and second principal components. The first component (**left**) returned by GPCAGEN captures variation in color space, while the second component (**right**) recovers the interpolation between digit "2" and digit "1".

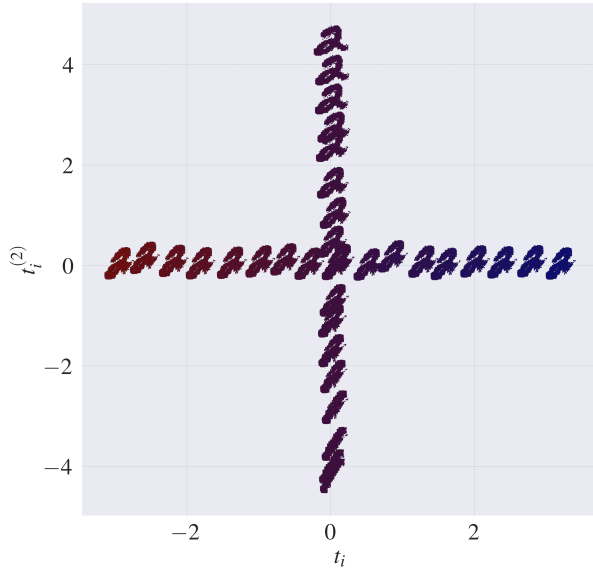


Figure 10: Each point cloud, corresponding to a distribution along one of the artificially constructed geodesics, is embedded in the plane according to its projection times onto the first and second geodesics returned by the GPCAGEN algorithm. We observe that GPCAGEN successfully recovers the two orthogonally intersecting geodesics designed from MNIST-based interpolations of digit shape and color.



Figure 11: Each chair point cloud is embedded in the plane according to its projection times onto the first and second geodesics returned by the GPCAGEN algorithm.

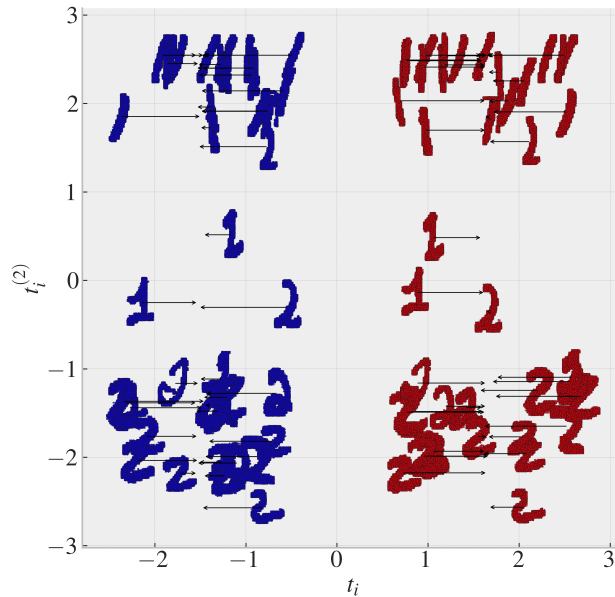


Figure 12: Each MNIST digit is embedded in the plane (the arrows indicate the exact position of each digit) according to its projection times onto the first and second geodesics returned by the GPCAGEN algorithm. We observe that the first principal component recovered by GPCAGEN captures variation in color, while the second component reflects the transformation from digit "2" to digit "1".

## B The Otto-Wasserstein geometry

In this section, we briefly describe the fiber bundle structure over the Wasserstein space due to Otto [28], that is behind the Riemannian interpretation of the Wasserstein distance. We then present its restriction to the space of centered non-degenerate Gaussian distributions, which coincides with the Bures-Wasserstein Riemannian geometry on SPD matrices. Finally, we relate Otto's parametrization of geodesics to McCann's interpolation.

We present these well-known results without proofs and refer the interested reader to [28, 17] and [2, Section 6.1] for more details in the general setting and to [36, 23, 4] for details and proofs in the Gaussian setting.

### B.1 The Otto-Wasserstein geometry of a.c. distributions

Consider the space  $\text{Prob}(\Omega)$  of absolutely continuous probability measures with smooth densities with respect to the Lebesgue measure, and support included in a compact set  $\Omega \subset \mathbb{R}^d$ , as well as the space  $\text{Diff}(\Omega)$  of diffeomorphisms on  $\Omega$ . These spaces can be equipped with an infinite-dimensional manifold structure, see e.g. [11], that we will not describe here. The tangent space of  $\text{Diff}(\Omega)$  at  $\varphi \in \text{Diff}(\Omega)$  is given by

$$T_\varphi \text{Diff}(\Omega) = \{v \circ \varphi, v : \Omega \rightarrow \mathbb{R}^d \text{ vector field}\}.$$

We fix a reference measure  $\rho \in \text{Prob}(\Omega)$  and equip  $\text{Diff}(\Omega)$  with the  $L^2$ -metric with respect to  $\rho$ , defined for any tangent vectors  $u \circ \varphi, v \circ \varphi \in T_\varphi \text{Diff}(\Omega)$  as

$$\langle u \circ \varphi, v \circ \varphi \rangle_{L^2(\rho)} := \int (u \circ \varphi) \cdot (v \circ \varphi) d\rho = \int u \cdot v d\mu,$$

where  $\mu = \varphi_\# \rho$ . Then the space of diffeomorphisms can be decomposed into *fibers*, defined to be equivalence classes under the projection

$$\pi : \text{Diff}(\Omega) \rightarrow \text{Prob}(\Omega), \quad \varphi \mapsto \varphi_\# \rho.$$

Specifically, the fiber over  $\mu \in \text{Prob}(\Omega)$  is given by  $\pi^{-1}(\mu) = \{\varphi \in \text{Diff}(\Omega), \varphi_\# \rho = \mu\}$ , see Figure 13 (**right**). The tangent space to the fiber  $\pi^{-1}(\mu)$  at  $\varphi \in \text{Diff}(\Omega)$  and its orthogonal with respect to the  $L^2(\rho)$ -metric are referred to as the *vertical* and *horizontal* spaces respectively :

$$\text{Ver}_\varphi := \ker d\pi_\varphi, \quad \text{Hor}_\varphi := (\text{Ver}_\varphi)^\perp,$$

where  $d\pi_\varphi : T_\varphi \text{Diff}(\Omega) \rightarrow T_{\pi(\varphi)} \text{Prob}(\Omega)$  denotes the differential of  $\pi$  at  $\varphi$ . Moving along vertical vectors in  $\text{Diff}(\Omega)$  means staying in the same fiber, i.e. projecting always to the same measure  $\mu$  in the bottom space. On the contrary, moving along horizontal vectors means moving orthogonally to the fibers, i.e., in the direction that gets fastest away from the fiber. The following proposition gives the form of vertical and horizontal vectors.

**Proposition 6.** *Let  $\varphi \in \text{Diff}(\Omega)$ . Then*

$$\begin{aligned} \text{Ver}_\varphi &= \{w \circ \varphi, \nabla \cdot (w\mu) = 0\}, \\ \text{Hor}_\varphi &= \{\nabla f \circ \varphi, f \in C^\infty(\Omega)\}. \end{aligned}$$

The following results state that line segments and  $L^2(\rho)$ -distances in  $\text{Diff}(\Omega)$  can be used to compute Wasserstein geodesics and distances in the space of probability measures  $\text{Prob}(\Omega)$ , provided we restrict to horizontal displacements.

**Proposition 7.** *The projection  $\pi : \text{Diff}(\Omega) \rightarrow \text{Prob}(\Omega)$  is a Riemannian submersion, i.e.  $d\pi_\varphi : \text{Hor}_\varphi \rightarrow T_{\pi(\varphi)} \text{Prob}(\Omega)$  is an isometry for any  $\varphi \in \text{Diff}(\Omega)$ .*

This implies the following.

**Proposition 8** (Proposition 2 in main). *Any geodesic  $t \mapsto \mu(t)$  for the Wasserstein metric (2) is the  $\pi$ -projection of a line segment in  $\text{Diff}(\Omega)$  going through a diffeomorphism  $\varphi$  at horizontal speed  $\nabla f \circ \varphi$  for some smooth function  $f \in C(\mathbb{R}^d)$ . That is, for  $t$  defined in a certain interval  $(t_{\min}, t_{\max})$ ,*

$$\mu(t) = \pi(\varphi + t\nabla f \circ \varphi) = (\text{id} + t\nabla f)_\#(\varphi_\# \rho). \quad (18)$$

We comment on the link between this parametrization and McCann's interpolation in Section B.3.

## B.2 The Otto-Wasserstein geometry of Gaussian distributions

The Bures-Wasserstein distance (3) on the space  $S_d^{++}$  of symmetric positive definite (SPD) matrices is the geodesic distance induced by a Riemannian metric  $g^{BW}$ , which can be written in different ways. Here we use the expression from [37, Table 4.7], defined for  $\Sigma = PDP^\top \in S_d^{++}$  and  $U = PU'P^\top \in S_d$ , by

$$g_\Sigma^{BW}(U, U) = \frac{1}{2} \sum_{1 \leq i, j \leq d} \frac{1}{d_i + d_j} U'_{ij}{}^2, \quad (19)$$

where the  $d_i$ 's are the diagonal elements of  $D$ . The associated Riemannian geometry can be described by Otto's fiber bundle restricted to the space of centered Gaussian distributions, in the following way.

In this setting, diffeomorphisms are restricted to invertible linear maps  $\varphi : u \mapsto Au$  for some invertible matrix  $A$ , i.e. the space of diffeomorphisms is replaced by the Lie group of invertible matrices  $GL_d$ . Tangent vectors are then given by linear maps  $u \mapsto Xu$  for any matrix  $X \in \mathbb{R}^{d \times d}$ . Fixing the standard normal distribution  $\rho = \mathcal{N}(0, \text{Id})$  as reference measure, the  $L^2$ -metric with respect to  $\rho$  between  $u \mapsto Xu$  and  $u \mapsto Yu$  is then written, for any  $X, Y \in \mathbb{R}^{d \times d}$ :

$$\int_{\mathbb{R}^d} \varphi(u)^\top \psi(u) d\rho(u) = \int_{\mathbb{R}^d} \text{tr}(\varphi(u)\psi(u)^\top) d\rho(u) = \text{tr} \left( \int_{\mathbb{R}^d} Xu u^\top Y^\top d\rho(u) \right) = \text{tr}(XY^\top),$$

yielding the standard Frobenius inner product on (the tangent space of)  $GL_d$ . We obtain a fibration of the top space  $GL_d$  over the bottom space  $S_d^{++}$  by considering the following projection

$$\pi : GL_d \rightarrow S_d^{++}, \quad A \mapsto AA^\top, \quad (20)$$

see Figure 13 (left). The fiber over  $\Sigma \in S_d^{++}$  is

$$\pi^{-1}(\Sigma) = \{A \in GL_d, AA^\top = \Sigma\} = \Sigma^{1/2} O_d, \quad (21)$$

where  $O_d$  denotes the space of orthogonal matrices and  $\Sigma^{1/2}$  denotes the only SPD square root of the SPD matrix  $\Sigma$ . The differential of the projection  $\pi(A) = AA^\top$  is given by

$$d\pi_A(X) = XA^\top + AX^\top. \quad (22)$$

Therefore, vertical vectors, which are those tangent to the fibers, or equivalently, those belonging to the kernel of  $d\pi_A(X)$ , are given by

$$\begin{aligned} \text{Ver}_A &:= \{X \in \mathbb{R}^{d \times d}, XA^\top + AX^\top = 0\} \\ &= \{X \in \mathbb{R}^{d \times d}, XA^\top \text{ is antisymmetric}\} \\ &= \{X = K(A^\top)^{-1}, K \in S_d^\perp\} = S_d^\perp (A^\top)^{-1}. \end{aligned}$$

where  $S_d^\perp$  denotes the space of antisymmetric matrices of size  $d$ . Once again, moving along vertical vectors in  $GL_d$  means staying in the same fiber, i.e. projecting always to the same SPD matrix in the bottom space  $S_d^{++}$ . Horizontal vectors are those that are orthogonal to all vertical vectors (for the Frobenius metric), i.e. matrices  $X$  such that for any antisymmetric matrix  $K$ :

$$0 = \langle X, K(A^\top)^{-1} \rangle = \text{tr}(XA^{-1}K^\top)$$

which is equivalent to  $XA^{-1}$  symmetric (this can be seen by taking for  $K$  the basis elements of  $S_d^\perp$  in the above equation), yielding

$$\begin{aligned} \text{Hor}_A &:= \{X \in \mathbb{R}^{d \times d}, (A^\top)^{-1}X^\top = XA^{-1}\} \\ &= \{X \in \mathbb{R}^{d \times d}, X^\top A - A^\top X = 0\} \\ &= \{X = KA, K \in S_d\} = S_d A \end{aligned}$$

where  $S_d$  denotes the space of symmetric matrices.

**Proposition 9.** *The projection  $\pi : GL_d \rightarrow S_d^{++}$ ,  $A \mapsto AA^\top$  is a Riemannian submersion, i.e.  $d\pi_A$  is an isometry from  $\text{Hor}_A$  equipped with the Frobenius inner product to  $T_{\pi(A)}S_d^{++}$  equipped with the inner product  $g_{\pi(A)}^{BW}$ , for any  $A \in GL_d$ .*

Just like in the general case, this yields a way to lift the computation of geodesics and distances.

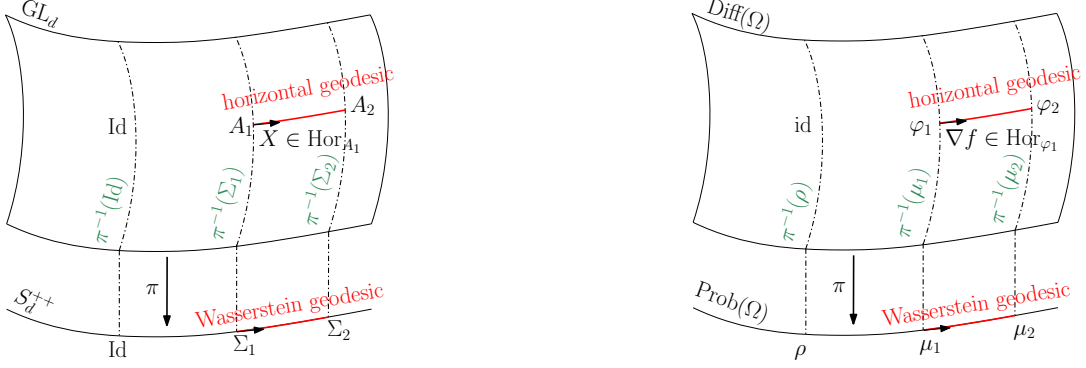


Figure 13: The Otto-Wasserstein geometry of **(left)** centered Gaussian distributions and **(right)** a.c. probability distributions. Figures inspired by [17].

**Proposition 10** (Proposition 1 in main). *Any geodesic  $t \mapsto \Sigma(t)$  in  $S_d^{++}$  for the Bures-Wasserstein metric (3) is the  $\pi$ -projection of a horizontal line segment in  $GL_d$ , that is*

$$\Sigma(t) = \pi(A + tX) = (A + tX)(A + tX)^\top, \quad A \in GL_d, X \in \text{Hor}_A, \quad (23)$$

where  $t$  is defined in a certain time interval  $(t_{\min}, t_{\max})$ . Also, the Bures-Wasserstein distance between two covariance matrices  $\Sigma_1, \Sigma_2 \in S_d^{++}$  is given by the minimal distance between their fibers

$$BW_2(\Sigma_1, \Sigma_2) = \inf_{Q_1, Q_2 \in O_d} \|\Sigma_1^{1/2} Q_1 - \Sigma_2^{1/2} Q_2\| = \inf_{Q \in SO_d} \|\Sigma_1^{1/2} - \Sigma_2^{1/2} Q\|, \quad (24)$$

where  $\|\cdot\|$  is the Frobenius norm and  $SO_d$  is the special orthogonal group.

Formula (23) and the first equality of (24) are direct consequences of the fact that  $\pi$  is a Riemannian submersion. To obtain the second equality of (24), we first notice that optimizing on  $Q_1, Q_2 \in O_d$  is equivalent to optimizing on a single  $Q \in O_d$  thanks to the invariance of the Frobenius metric w.r.t. the right action of  $O_d$ . And second, that the infimum is attained at (see [4, Equations 3 and 35])

$$Q^* = \Sigma_2^{-1/2} T \Sigma_1^{1/2}, \quad \text{where } T = \Sigma_1^{-1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \Sigma_1^{-1/2}$$

is the Monge map from  $\Sigma_1$  to  $\Sigma_2$  (see [23, equation 8]), and so  $Q^*$  has positive determinant and belongs to  $SO_d$ .

Thus the closest element of the fiber  $\pi^{-1}(\Sigma_2)$  to  $\Sigma_1^{1/2}$  is given by  $\Sigma_2^{1/2} Q^* = T \Sigma_1^{1/2}$ , i.e. by left multiplying  $\Sigma_1^{1/2}$  by the Monge map  $T$ . This is more generally true for any representative of  $\Sigma_1$ :

**Proposition 11.** *Let  $\Sigma_1, \Sigma_2 \in S_d^{++}$ ,  $T$  the Monge map from  $\Sigma_1$  to  $\Sigma_2$ ,  $A_1 \in \pi^{-1}(\Sigma_1)$ . Then  $A_2 := T A_1$  is said to be aligned with respect to  $A_1$ , that is, it is the closest point in  $\pi^{-1}(\Sigma_2)$  to  $A_1$ . More precisely, we have*

1.  $A_2 - A_1 = (T - I)A_1 \in \text{Hor}_{A_1}$
2.  $\text{Log}_{\Sigma_1}(\Sigma_2) := d\pi_{A_1}((T - I)A_1) = (T - I)\Sigma_1 + \Sigma_1(T - I)$
3.  $BW_2(\Sigma_1, \Sigma_2) = \|\text{Log}_{\Sigma_1} \Sigma_2\|_{\Sigma_1}^{BW} = \|(T - I)A_1\|$

where  $\text{Log}$  is the Riemannian logarithm map,  $\|\cdot\|_{\Sigma}^{BW} = \sqrt{g_{\Sigma}^{BW}(\cdot, \cdot)}$  and  $\|\cdot\|$  is the Frobenius norm.

This means that to compute the Bures-Wasserstein distance between two covariance matrices  $\Sigma_1$  and  $\Sigma_2$ , one can consider any representative  $A_1$  in the fiber over  $\Sigma_1$ , compute the representative  $A_2$  of  $\Sigma_2$  aligned to  $A_1$  (using the Monge map) and finally compute the Frobenius norm of  $A_2 - A_1$ .

### B.3 Geodesic parametrization

There are two classical parameterizations for Wasserstein geodesics in the space of a.c. probability measures.

**McCann's interpolation** The first one, due to McCann [24], is given between two probability distributions  $\mu_0$  and  $\mu_1$ , and depends on the optimal transport map in (2), obtained as the gradient of a convex function  $u$ , that is  $T_{\mu_0}^{\mu_1} = \nabla u$  and

$$\mu_t = ((1 - t) \text{id} + t \nabla u)_{\#} \mu_0 = (\text{id} + t(\nabla u - \text{id}))_{\#} \mu_0, \quad t \in [0, 1]. \quad (25)$$

**Otto's geodesic** The second one, exploiting Otto's fiber bundle geometry [28], consists in writing a geodesic in the Wasserstein space as the projection of a horizontal geodesic in the total space of diffeomorphisms. Such a horizontal geodesic is a line segment going through a diffeomorphism  $\varphi$  with a horizontal speed  $\nabla f \circ \varphi$ , where  $f$  is any smooth function (not necessarily convex). Therefore we get

$$\mu_s = (\varphi + s\nabla f \circ \varphi)_{\#}\rho = (\text{id} + s\nabla f)_{\#}(\varphi_{\#}\rho), \quad s \in (s_0, s_1). \quad (26)$$

In this second expression, the bounds on the time  $s$  depends on the function  $f$ . Indeed, for  $\mu_s$  to be a geodesic,  $\text{id} + s\nabla f$  needs to remain in the space of diffeomorphisms for a given  $s$ , which means that  $\text{id} + s\text{Hess } f$  needs to be positive definite. Therefore, we get the following conditions depending on the minimum  $\lambda_{\min}$  and maximum  $\lambda_{\max}$  eigenvalues of  $\text{Hess } f$ :

$$\begin{cases} s \in ]-\infty, -1/\lambda_{\min}[ & \text{if } \lambda_{\max} < 0, \\ s \in ]-1/\lambda_{\max}, +\infty[ & \text{if } \lambda_{\min} > 0, \\ s \in ]-1/\lambda_{\max}, -1/\lambda_{\min}[ & \text{if } \lambda_{\min} < 0 < \lambda_{\max}. \end{cases} \quad (27)$$

It is clear that equation (25) is a particular case of equation (26), where we choose  $\varphi_{\#}\rho = \mu_0$  and  $\nabla f = \nabla u - \text{id}$ . Conversely, one can write equation (26) under the form of equation (25). For a given diffeomorphism  $\varphi$  and function  $f$ , consider the geodesic given by (26), and set  $\mu_0 = \varphi_{\#}\rho$ . Assume that we are in the case where all eigenvalues of  $\text{Hess } f$  are negative, then  $s$  must be in  $]-\infty, -1/\lambda_{\min}[$ . Consider  $s^* \in ]0, -1/\lambda_{\min}[$ , and define  $\mu_1 := \mu_{s^*} = (\text{id} + s^*\nabla f)_{\#}\mu_0$ . Setting  $t = s/s^*$  we have that the geodesic between  $\mu_0$  and  $\mu_1$  is written

$$\mu_t = (\text{id} + ts^*\nabla f)_{\#}\mu_0 = (\text{id} + t(\nabla u - \text{id}))_{\#}\mu_0, \quad t \in [0, 1].$$

for  $u(x) = s^*f + \|x\|^2/2$ . Now for any eigenvalue  $\lambda_i$  of  $H_f$  the hessian of  $f$ , we have

$$\lambda_i > \lambda_{\min} > -1/s^* \quad \text{i.e.} \quad s^*\lambda_i + 1 > 0.$$

by the interval of definition of  $s^*$ . This means that the hessian  $H_u = s^*H_f + \text{id}$  is positive definite, which means that  $u$  is necessarily convex. The other cases work similarly.

**The Gaussian case** Transposing Otto's formulation (26) to the case of a geodesic between Gaussian distributions means that for  $A \in GL_d$  and  $X \in \text{Hor}_A$  such that  $\|X\| = 1$ , the interval of definition of a geodesic depends on the invertibility of  $A + sX$ . In turn, the maximal interval of definition of  $s \in (s_0, s_1)$  is defined from the eigenvalues of  $XA^{-1}$ , through the same formula (27).

## C Linearized optimal transport and tangent PCA

In this section, we provide the definition of linearized Wasserstein distance and details on how to perform tangent PCA for both Gaussian distributions and general a.c. distributions. Tangent PCA is a widely used approach to compute PCA on the Wasserstein space, that consists in embedding probability distributions into the tangent space at some reference measure  $\rho$ , and performing PCA in the tangent space with respect to the linearized Wasserstein distance.

### C.1 The case of centered Gaussian distributions

We consider  $n$  covariance matrices  $\Sigma_1, \dots, \Sigma_n$  and their Bures-Wasserstein barycenter (or Fréchet mean)  $\bar{\Sigma}$ , that is, the SPD matrix verifying [1]:

$$\bar{\Sigma} = \underset{\Sigma \in S_d^{++}}{\text{argmin}} \sum_{i=1}^n BW_2^2(\Sigma, \Sigma_i). \quad (28)$$

The idea behind tangent PCA is to represent each data point by the corresponding tangent vector, given by the Riemannian logarithm map, in the tangent space at the reference point  $\bar{\Sigma}$ , i.e.

$$\{\text{Log}_{\bar{\Sigma}}\Sigma_i\}_{i=1}^n \subset T_{\bar{\Sigma}}S_d^{++}. \quad (29)$$

Now, one can lift the computations from the tangent space at  $\bar{\Sigma}$  to the horizontal space at a point in the fiber over  $\bar{\Sigma}$ , say  $A := \bar{\Sigma}^{1/2}$ , by aligning all representatives to  $A$ , see Proposition 11. The key point is that the tangent space at  $\bar{\Sigma}$  equipped with the Bures-Wasserstein Riemannian metric is isometric to  $\text{Hor}_A := S_d A$  equipped with the Frobenius inner product – where we recall that  $S_d$  is the space of symmetric matrices. This means that instead of performing PCA for the Bures-Wasserstein inner product on the tangent vectors (29), we can instead perform linear PCA on their pre-images by  $d\pi_A$ , see Proposition 11:

$$\{(T_i - I)A\}_{i=1}^n \subset \text{Hor}_{A_1}, \quad \text{where } T_i = \Sigma_i^{-1/2}(\Sigma_i^{1/2}\bar{\Sigma}\Sigma_i^{1/2})^{1/2}\Sigma_i^{-1/2}.$$

$T_i$  is the optimal transport map from  $\bar{\Sigma}$  to  $\Sigma_i$ , see Section B.2. Now, noticing that

$$\langle K_1 A, K_2 A \rangle = \text{Tr}(K_1 A A^\top K_2^\top) = \text{Tr}(K_1 \bar{\Sigma} K_2^\top), \quad \forall K_1, K_2 \in S_d,$$

we see that the space  $\text{Hor}_A$  equipped with the Frobenius inner product is itself isometric to  $S_d$  equipped with the Frobenius inner product weighted by  $\bar{\Sigma}$ . Therefore, tangent PCA is performed through Euclidean PCA on the (centered) vectors  $\{T_i - I\}_{i=1}^n$ , in the vector space  $S_d$ , with respect to the Frobenius metric weighted by  $\bar{\Sigma}$ . Another way to see this is by noticing that the linearized Bures-Wasserstein distance  $BW_{2,\bar{\Sigma}}$  with respect to  $\bar{\Sigma}$  is given by

$$\begin{aligned} BW_{2,\bar{\Sigma}}(\Sigma_1, \Sigma_2) &:= \|\text{Log}_{\bar{\Sigma}} \Sigma_1 - \text{Log}_{\bar{\Sigma}} \Sigma_2\|_{\bar{\Sigma}}^{BW} \\ &= \|d\pi_{\bar{\Sigma}^{1/2}}((T_1 - I)\bar{\Sigma}^{1/2}) - d\pi_{\bar{\Sigma}^{1/2}}((T_2 - I)\bar{\Sigma}^{1/2})\|_{\bar{\Sigma}}^{BW} \\ &= \|(T_1 - I)\bar{\Sigma}^{1/2} - (T_2 - I)\bar{\Sigma}^{1/2}\| \\ &= \|(T_1 - T_2)\bar{\Sigma}^{1/2}\| \end{aligned}$$

where  $\|\cdot\|^{BW}$  denotes the norm associated to the Bures Wasserstein Riemannian metric (19),  $\pi$  is Otto's projection (20), and we have used Propositions 9 and 11. Finally,

$$BW_{2,\bar{\Sigma}}(\Sigma_1, \Sigma_2) := \|\text{Log}_{\bar{\Sigma}} \Sigma_1 - \text{Log}_{\bar{\Sigma}} \Sigma_2\|_{\bar{\Sigma}}^{BW} = \|T_1 - T_2\|_{\bar{\Sigma}}, \quad (30)$$

where  $\|\cdot\|_{\bar{\Sigma}}$  denotes the Frobenius norm weighted by  $\bar{\Sigma}$ .

## C.2 The case of a.c. distributions

Similarly, one can embed a.c. probability distributions  $\nu_1, \dots, \nu_n$  into the  $L^2(\rho)$  space at some a.c. reference measure  $\rho$  through the optimal maps  $\nu_i \mapsto T_\rho^{\nu_i}$  in the Monge problem (2). Then, the Wasserstein distance can be approximated by the linearized Wasserstein distance [39] given by

$$W_{2,\rho}(\nu_1, \nu_2) = \|T_\rho^{\nu_1} - T_\rho^{\nu_2}\|_{L^2(\rho)}. \quad (31)$$

Note that as previously mentioned, this metric induces distortions : while the radial distances from  $\rho$  to any  $\mu_i$  are preserved, that is  $\|\text{id} - T_\rho^{\nu_i}\|_{L^2(\rho)} = W_2(\rho, \nu_i)$ , other distances are not  $\|T_\rho^{\nu_1} - T_\rho^{\nu_2}\|_{L^2(\rho)} \neq W_2(\nu_1, \nu_2)$ . A recent paper [20] proved however, that under some assumptions,  $W_{2,\rho}$  is bi-Hölder equivalent to  $W_2$ , which indicates that the distortion effect can be controlled.

Then, denoting  $\bar{\nu}_n$  the Wasserstein barycenter [1] of  $\nu_1, \dots, \nu_n$ , that is the solution of

$$\bar{\nu}_n \in \underset{\nu}{\text{argmin}} \sum_{i=1}^n W_2^2(\nu, \nu_i), \quad (32)$$

tangent PCA consists in performing classical PCA [30] of  $(T_{\bar{\nu}_n}^{\nu_i} - \text{id})_{i=1}^n$  in the Hilbert space  $L^2(\bar{\nu}_n)$ .

## D Geodesic PCA for Gaussian distributions

In this section, we present the proofs related to geodesic PCA for Gaussian distributions and the implementation of our algorithm in this case.

### D.1 Proofs related to GPCA for Gaussian distributions

We first proof the existence of mimimizers for the GPCA problems lifted in Otto's fiber bundle.

**Lemma 1.** *The GPCA problem (11) for the first component admits a global minimum.*

*Proof.* First, let us define the set of normalized matrices  $\mathbb{B} := \{X \in \mathbb{R}^{d \times d}, \|X\| = 1\}$ . By denoting  $\lambda_{\min}$  (resp.  $\lambda_{\max}$ ) the smallest (resp. largest) eigenvalue of  $XA^{-1}$ , extending the geodesic  $t \mapsto A + tX$  as far as possible (see Section B.3) means that the closed interval  $[t_{\min}, t_{\max}]$  is defined for some fixed  $\varepsilon > 0$  by

$$\begin{cases} (-\infty, -1/\lambda_{\min} - \varepsilon] & \text{if } \lambda_{\max} < 0, \\ [-1/\lambda_{\max} + \varepsilon, +\infty) & \text{if } \lambda_{\min} > 0, \\ [-1/\lambda_{\max} + \varepsilon, -1/\lambda_{\min} - \varepsilon] & \text{if } \lambda_{\min} < 0 < \lambda_{\max}. \end{cases} \quad (33)$$

Let us now consider the function

$$F : GL_d \times \mathbb{B} \times (\mathbb{R}^{d \times d})^n \longrightarrow \mathbb{R}$$

$$(A, X, (Q_i)_{i=1}^n) \longmapsto \sum_{i=1}^n \|A + p_{(A,X)}(t_i)X - \Sigma_i^{1/2}Q_i\|^2 =: \sum_{i=1}^n g_i(A, X, Q_i),$$

where  $t_i = \langle \Sigma_i^{1/2}Q_i - A, X \rangle$  and  $p_{(A,X)} : \mathbb{R} \rightarrow \mathbb{R}$  is the projection operator that clips a point  $t$  into  $[t_{\min}, t_{\max}]$ , which depends on  $A$  and  $X$ . Then the function  $F$  is continuous on  $GL_d \times \mathbb{B} \times (\mathbb{R}^{d \times d})^n$  as composition of linear and continuous functions. Note that the function  $(A, X) \mapsto p_{(A,X)}(t_i)$  is continuous by eigenvalue continuity [21]. Additionally, the function  $F$  is coercive on  $GL_d \times \mathbb{B} \times (\mathbb{R}^{d \times d})^n$ . Indeed, on a diagonal  $\{A = \Sigma_i^{1/2}Q_i, \text{ for } (A, Q_i) \in GL_d \times \mathbb{R}^{d \times d}\}$  for some  $i \in \{1, \dots, n\}$ , we have  $t_i = 0$ , and therefore we have either  $g_i(A, X, Q_i) = 0$  if  $p_{(A,X)}(0) = 0$ , or  $g_i(A, X, Q_i) = \varepsilon \|X\|^2 = \varepsilon$  otherwise. This would imply that  $g_i(A, X, Q_i)$  doesn't go to infinity when the norm  $\|(A, X, Q_i)\| \rightarrow \infty$ . However, in this case, we have  $g_j(A, X, Q_j) \rightarrow \infty$  when  $\|(A, X, Q_j)\| \rightarrow \infty$  for any  $j \neq i$ . Moreover, as  $p_{(A,X)}(t_i)$  is a clipping, it won't play a role in the coercivity. We conclude by the fact that the function  $(A, X) \mapsto X^\top A - A^\top X$  is continuous, implying that the set of constraint  $\{(A, X) \in GL_d \times \mathbb{R}^{d \times d} : X^\top A - A^\top X = 0\}$  is closed and  $\mathbb{B}$  and  $SO_d$  are compact. The optimization problem (11) thus admits a global minimum.  $\square$

Note that this result also applies for the second component (12) and the higher order components.

**Proposition 12** (Proposition 3 in main). *Let  $\pi : GL_d \rightarrow S_d^{++}$ ,  $A \mapsto AA^\top$  and  $(A_1, X_1, (Q_i)_{i=1}^n)$  be a solution of*

$$\inf F(A_1, X_1, (Q_i)_{i=1}^n) := \sum_{i=1}^n \|A_1 + p_{A_1, X_1}(t_i)X_1 - \Sigma_i^{1/2}Q_i\|^2,$$

*subject to*  $A_1 \in GL_d, X_1 \in \text{Hor}_{A_1}, \|X_1\|^2 = 1, Q_1, \dots, Q_n \in SO_d.$

*Then there exist  $t_{\min}, t_{\max} \in \mathbb{R}$  such that the geodesic  $\Sigma : t \in [t_{\min}, t_{\max}] \mapsto \pi(A_1 + tX_1)$  in  $S_d^{++}$  minimizes (10).*

*Proof.* A horizontal geodesic in  $GL_d$  is a straight line going through a base point  $A \in GL_d$  in the direction of a horizontal vector  $X \in \text{Hor}_A$  (that we consider normalized, ie.  $\|X\|^2 = 1$ ), i.e.  $t \mapsto A + tX \in GL_d$ . Denoting  $[t_{\min}, t_{\max}]$  the interval constructed in (33) which depends on the eigenvalues of  $XA^{-1}$ , we have that  $(\pi(A + tX))_{t \in [t_{\min}, t_{\max}]}$  is a geodesic in the Bures-Wasserstein sense, see Proposition 1, and

$$\begin{aligned} \min_{t \in [t_{\min}, t_{\max}]} BW_2^2(\pi(A + tX), \Sigma_i) &= \min_{t \in [t_{\min}, t_{\max}]} \inf_{Q_i \in SO_d} \|A + tX - \Sigma_i^{1/2}Q_i\|^2 \\ &= \inf_{Q_i \in SO_d} \|A + p_{(A,X)}(t_i)X - \Sigma_i^{1/2}Q_i\|^2, \end{aligned}$$

where  $t_i = \langle \Sigma_i^{1/2}Q_i - A, X \rangle$  is the (orthogonal) projection time of  $\Sigma_i^{1/2}Q_i$  onto the line  $t \mapsto A + tX$ .

We therefore deduce that a set of solution  $(A, X, (Q_i)_{i=1}^n)$  of (11) defines a proper geodesic  $(\pi(A + tX))_{t \in [t_{\min}, t_{\max}]}$ , solution of problem (10).  $\square$

**Proposition 13** (Proposition 4 in main). *Let  $\nu_i = \mathcal{N}(m_i, \sigma_i^2)$  for  $i = 1, \dots, n$  be  $n$  univariate Gaussian distributions. The first principal geodesic component  $t \in [0, 1] \mapsto \mu(t)$  solving (1) remains in the geodesic space of Gaussian distributions for all  $t \in [0, 1]$ .*

*Proof.* Let  $\text{Prob}_2(\mathbb{R})$  be the set of a.c. probability measures on  $\mathbb{R}$  that have finite second moment, and  $\mathcal{Q}$  the set of corresponding quantile functions :

$$\mathcal{Q} = \{F_\nu^{-1}; \nu \in \text{Prob}_2(\mathbb{R})\}$$

$\mathcal{Q}$  is the set of increasing, left-continuous functions  $q : (0, 1) \rightarrow \mathbb{R}$ , and a convex cone in  $L^2([0, 1])$ , the set of square-integrable functions on  $[0, 1]$ . The mapping

$$\Phi : \nu \mapsto F_\nu^{-1} \tag{34}$$

defines an isometry between  $\text{Prob}_2(\mathbb{R})$  equipped with the Wasserstein metric, and  $\mathcal{Q}$  equipped with the  $L^2$  metric (see e.g. [5]), that is, for any  $\mu, \nu \in \text{Prob}_2(\mathbb{R})$ ,

$$W_2(\mu, \nu) = \|F_\mu^{-1} - F_\nu^{-1}\|_{L^2([0,1])}.$$

The map  $\Phi$  in (34) also defines an isometry from the set of (univariate) Gaussian distributions to the set of all Gaussian quantile functions  $\mathcal{G}$ . This space  $\mathcal{G}$  is the upper-half of the plane  $\mathcal{F}$  spanned by the constant function  $\mathbf{1}$  and the quantile function  $F_0^{-1}$  of the standard normal distribution:

$$\mathcal{G} = \mathbb{R} \cdot \mathbf{1} + \mathbb{R}_+^* \cdot F_0^{-1} \subset \mathcal{F} := \text{span}(\mathbf{1}, F_0^{-1}).$$

Now, consider  $n$  normal distributions  $\nu_1, \dots, \nu_n$ , and  $(\mu(t))_{t \in [0,1]}$  the first principal geodesic component found by minimizing equation (13), the sum of squared residuals in  $\text{Prob}_2(\mathbb{R})$ . Since  $\mu$  is a Wasserstein geodesic in  $\text{Prob}_2(\mathbb{R})$  and  $\Phi$  is an isometry, the curve  $t \mapsto \Phi(\mu)(t) = F_{\mu(t)}^{-1}$  is an  $L^2([0, 1])$ -geodesic in  $\mathcal{Q}$ , i.e. a line segment

$$t \in [0, 1] \mapsto F_{\mu(t)}^{-1} = (1-t)F_{\mu(0)}^{-1} + tF_{\mu(1)}^{-1}.$$

Since  $\{\mathbf{1}, F_0^{-1}\}$  forms an orthonormal basis of  $\mathcal{F}$ , the orthogonal projection of this line segment on  $\mathcal{F}$  is given by

$$t \in [0, 1] \mapsto \langle F_{\mu(t)}^{-1}, \mathbf{1} \rangle \mathbf{1} + \langle F_{\mu(t)}^{-1}, F_0^{-1} \rangle F_0^{-1},$$

which lies in  $\mathcal{G}$ . To see this, we need to show that the following value is positive:

$$\langle F_{\mu(t)}^{-1}, F_0^{-1} \rangle = \int_0^1 F_{\mu(t)}^{-1}(y) F_0^{-1}(y) dy = \int_{\mathbb{R}} x F_0^{-1} \circ F_{\mu(t)}(x) d\mu(t)(x) = \mathbb{E}(XT(X)),$$

where  $X \sim \mu(t)$  and  $T = F_0^{-1} \circ F_{\mu(t)}$  is the Monge map from  $\mu(t)$  to the standard normal distribution. Since  $T$  is increasing, we indeed have  $\mathbb{E}(XT(X)) > 0$  (see e.g. the proof of Theorem 2.2 in [32]).

Finally, since  $\Phi(\mu)$  orthogonally projects from  $\mathcal{Q}$  to  $\mathcal{G}$  w.r.t the  $L^2$  metric and  $\Phi$  defines an isometry, we get that the geodesic  $\mu$  orthogonally projects to a geodesic  $\pi(\mu)$  in the space of Gaussian distributions, w.r.t. the Wasserstein metric. By the distance minimizing property of orthogonal projections, we know that the cost function (13) evaluated at  $\pi(\mu)$  is no larger than its value at  $\mu$ . Since  $\mu$  is optimal, we get that  $\mu = \pi(\mu)$  and  $\mu$  belongs to the space of Gaussian distributions.  $\square$

**Proposition 14.** *Let  $\Sigma_1, \Sigma_2$  two SPD matrices that are diagonalizable in the same orthonormal basis, i.e.*

$$\Sigma_1 = P \begin{pmatrix} a_1^2 & 0 \\ 0 & b_1^2 \end{pmatrix} P^\top \quad \text{and} \quad \Sigma_2 = P \begin{pmatrix} a_2^2 & 0 \\ 0 & b_2^2 \end{pmatrix} P^\top,$$

where  $P$  is orthogonal. Then  $BW_2^2(\Sigma_1, \Sigma_2) = (a_1 - a_2)^2 + (b_1 - b_2)^2$ , and thus the Bures-Wasserstein geodesic between  $\Sigma_1$  and  $\Sigma_2$  is given by

$$\Sigma(t) = P \begin{pmatrix} ((1-t)a_1 + tb_1)^2 & 0 \\ 0 & ((1-t)a_2 + tb_2)^2 \end{pmatrix} P^\top, \quad 0 \leq t \leq 1.$$

*Proof.* This is a straightforward computation using equation (3).  $\square$

**Proposition 15.** *Let us consider  $n = 2p$  covariance matrices  $\Sigma_i = \Sigma(a, b, \theta_i)$  as defined in (15), where  $\theta_i = i\pi/n$  for  $i = 0, \dots, n-1$ . Then, the Bures-Wasserstein barycenter (28) of these covariance matrices is given by  $\bar{\Sigma} = (a+b)^2/4 I$ .*

*Proof.* Each pair of covariance matrices

$$\Sigma_i = P_{\theta_i} \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix} P_{\theta_i}^\top, \quad \text{and} \quad \Sigma_{i+p} = P_{\theta_i+\pi/2} D P_{\theta_i+\pi/2}^\top = P_{\theta_i} \begin{pmatrix} b^2 & 0 \\ 0 & a^2 \end{pmatrix} P_{\theta_i}^\top$$

are diagonalizable in the same basis, and so by Proposition 14, the geodesic from  $\Sigma_i$  to  $\Sigma_{i+p}$  is

$$\Sigma(t) = P_{\theta_i} \begin{pmatrix} ((1-t)a + tb)^2 & 0 \\ 0 & ((1-t)b + ta)^2 \end{pmatrix} P_{\theta_i}^\top, \quad 0 \leq t \leq 1.$$

In particular, the Fréchet mean is given by  $\bar{\Sigma} = \Sigma(1/2) = ((a+b)/2)^2 I$ . Since each pair of covariance matrices has the same Fréchet mean, the Fréchet mean of the whole set  $\Sigma_1, \dots, \Sigma_n$  is also given by  $\bar{\Sigma}$ .  $\square$

**Proposition 16** (Proposition 5 in main). *Let  $\Sigma \in S_2^{++}$  with eigenvalues  $a^2, b^2$  and  $\Sigma' = P_\theta \Sigma P_\theta^\top$  where  $P_\theta$  is the rotation matrix of angle  $\theta$ . Then, denoting  $\bar{\Sigma} = ((a+b)/2)^2 I$  we have*

$$\frac{BW_2^2(\Sigma, \Sigma')}{BW_{2, \bar{\Sigma}}^2(\Sigma, \Sigma')} = 1 - \left( \frac{a-b}{a+b} \right)^2 \cos^2 \theta + O((a-b)^4). \quad (35)$$

*Proof.* Recall that the linearized Bures-Wasserstein distance at  $\bar{\Sigma}$  between  $\Sigma$  and  $\Sigma'$  is given by the distance between their images by the Riemannian logarithm map  $U := \text{Log}_{\bar{\Sigma}}\Sigma$  and  $U' := \text{Log}_{\bar{\Sigma}}\Sigma'$  in the tangent space at  $\bar{\Sigma}$ , i.e.

$$BW_{2,\bar{\Sigma}}(\Sigma, \Sigma') = \|U - U'\|_{\bar{\Sigma}}^{BW},$$

where  $\|\cdot\|^{BW}$  denotes the norm associated to the Bures-Wasserstein Riemannian metric (19). As in any Riemannian manifold, the true geodesic distance can be approximated by this linearized distance in the tangent space, corrected by the curvature (see e.g. Lemma 1 in [13]):

$$BW_2^2(\Sigma, \Sigma') = (\|U - U'\|_{\bar{\Sigma}}^{BW})^2 - \frac{1}{3}R_{\bar{\Sigma}}(U, U', U, U') + O(\|U\|_{\bar{\Sigma}}^{BW} + \|U'\|_{\bar{\Sigma}}^{BW})^6, \quad (36)$$

where  $R_{\bar{\Sigma}}$  is the curvature tensor.

Recall from Equation (19) that the Bures-Wasserstein norm of a vector  $U$  is expressed in an eigenvector basis of the base point, here  $\bar{\Sigma}$ . Since any basis is an eigenvector basis of  $\bar{\Sigma}$ , it is convenient to choose that of  $\Sigma$ , which we can assume without loss of generality to be the canonical basis. Thus we write  $\Sigma = D$  where  $D = \text{diag}(a^2, b^2)$  and  $\Sigma' = P_\theta D P_\theta^\top$ , and the norm associated to the Bures-Wasserstein Riemannian metric is given by

$$\|U\|_{\bar{\Sigma}}^{BW} = \frac{1}{2} \sum_{1 \leq i, j \leq 2} \frac{1}{d_i + d_j} U_{ij}^2$$

where the  $d_i$ 's are the eigenvalues of  $\bar{\Sigma}$ , given here by  $d_1 = d_2 = ((a+b)/2)^2$ . From Proposition 11 we have

$$\begin{aligned} U &:= \text{Log}_{\bar{\Sigma}}\Sigma = (T - I)\bar{\Sigma} + \bar{\Sigma}(T - I), \\ U' &:= \text{Log}_{\bar{\Sigma}}\Sigma' = (T' - I)\bar{\Sigma} + \bar{\Sigma}(T' - I), \end{aligned}$$

where

$$\begin{aligned} T &:= \bar{\Sigma}^{-1/2}(\bar{\Sigma}^{1/2}\Sigma\bar{\Sigma}^{1/2})^{1/2}\bar{\Sigma}^{-1/2} = \frac{2}{a+b}D^{1/2}, \\ T' &:= \bar{\Sigma}^{-1/2}(\bar{\Sigma}^{1/2}\Sigma'\bar{\Sigma}^{1/2})^{1/2}\bar{\Sigma}^{-1/2} = \frac{2}{a+b}P_\theta D^{1/2}P_\theta^\top, \end{aligned}$$

and easily get

$$U = \frac{a^2 - b^2}{2}J, \quad U' = \frac{a^2 - b^2}{2}P_\theta J P_\theta^\top, \quad \text{where } P_\theta J P_\theta^\top = \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}$$

and  $J = \text{diag}(1, -1)$ . Thus after some computations we obtain

$$\begin{aligned} \|U\|_{\bar{\Sigma}}^{BW} &= \|U'\|_{\bar{\Sigma}}^{BW} = |a - b|\sqrt{2}, \\ BW_{2,\bar{\Sigma}}(\Sigma, \Sigma') &= \|U - U'\|_{\bar{\Sigma}}^{BW} = \sqrt{2}|(a - b) \sin \theta|. \end{aligned} \quad (37)$$

To compute the curvature tensor, we use the following formula from [37, Table 4.7]

$$R_{\bar{\Sigma}}(U, U', U, U') = \frac{3}{2} \sum_{i,j} \frac{d_i d_j}{d_i + d_j} [U_0, U'_0]_{ij}^2$$

where  $[A, B] = AB - BA$  is the Lie bracket of matrices,  $U_0$  and  $U'_0$  are the only symmetric matrices verifying the Sylvester equations  $U = U_0\bar{\Sigma} + \bar{\Sigma}U_0$  and  $U' = U'_0\bar{\Sigma} + \bar{\Sigma}U'_0$  respectively. Since  $\bar{\Sigma}$  is a multiple of the identity, we easily get

$$U_0 = \frac{a-b}{a+b}J, \quad U'_0 = \frac{a-b}{a+b}P_\theta J P_\theta^\top$$

and straightforward computations yield

$$R_{\bar{\Sigma}}(U, U', U, U') = \frac{3}{2} \frac{(a-b)^4}{(a+b)^2} \sin^2 2\theta. \quad (38)$$

Finally, putting together (36), (37) and (38) and we obtain

$$BW_2^2(\Sigma, \Sigma') = BW_{2,\bar{\Sigma}}^2(\Sigma, \Sigma') - 2 \frac{(a-b)^4}{(a+b)^2} \sin^2 \theta \cos^2 \theta + O((a-b)^6),$$

and dividing by the squared linearized optimal transport distance yields the desired result.  $\square$

## D.2 Implementation of GPCA for Gaussian distributions

As described in Section 3, the first and second components of geodesic PCA are respectively found by solving the minimization problems (11) and (12). The geodesic components are given by

$$\Sigma_i(t) = (A_i + tX_i)(A_i + tX_i)^\top, \quad \text{for } i = 1, 2,$$

where  $A_1 \in GL_d$  and  $X_1 \in \text{Hor}_{A_1}$  are minimizers of (11), and  $A_2 \in GL_d$  and  $X_2 \in \text{Hor}_{A_2}$  minimizers of (12). The matrix  $\pi(A_2)$  is the crossing point through which all geodesic components intersect, see Figure 2. The higher order components are found in a analogous way: for the  $k$ -th component, we search for a horizontal segment  $t \mapsto A_k + tX_k$  where  $A_k$  belongs to the fiber over the intersection point (we parametrize it w.r.t. the previous position in the fiber, i.e.  $A_k = A_{k-1}R_{k-1}$  for a certain  $R_{k-1} \in SO_d$ ) and the horizontal velocity vector  $X_k$  is orthogonal to the lifts of the velocity vectors of the previous component. Thus, the  $k$ -th component,  $k \geq 3$ , solves:

$$\begin{aligned} & \inf F(A_k, X_k, (Q_i)_{i=1}^n) \\ \text{subject to } & A_k = A_{k-1}R_{k-1}, R_{k-1} \in SO_d, X_k \in \text{Hor}_{A_k}, \|X_k\|^2 = 1, \\ & \langle X_k, X_{k-\ell}R_{k-\ell} \dots R_{k-1} \rangle = 0, 1 \leq \ell \leq k-1, Q_1, \dots, Q_n \in SO_d. \end{aligned} \quad (39)$$

Following [15] and [9], we propose an iterative algorithm to implement these components, that, for each component, alternates two steps:

(Step 1) minimization of the objective function  $F$  (see (11)) with respect to  $(Q_i)_{i=1}^n$  for fixed  $(A, X)$ ,

(Step 2) minimization of the objective function  $F$  with respect to  $(A, X)$  for fixed  $(Q_i)_{i=1}^n$ .

In dimension  $d = 2$ , any rotation matrix  $Q$  can be parametrized by a scalar angle  $\theta$  and both steps are solved using the Sequential Least Squares Programming (SLSQP) algorithm [22] available on the *scipy python library* [38]. In higher dimension, each minimization with respect to a rotation matrix is performed using Riemannian gradient descent on  $SO_d$ , relying on the Riemannian geometry of  $SO_d$  induced by the standard Frobenius metric of the ambient space  $\mathbb{R}^{d \times d}$ . In particular we use the exponential map implemented in the Python library *geomstats* [25]. More details on the Riemannian geometry of  $SO_d$  and the Riemannian gradient descent procedure can be found e.g. in [7, Sections 7.4 and 4.3].

Unfortunately, we cannot ensure the convergence of the iterates of the proposed block alternating algorithm, as classical arguments require uniqueness of the minimizer at each iterations [29]. This is unachievable in our problem: the line with base point  $A$  and direction  $X \in \text{Hor}_A$  and the line with base point  $AQ$  and direction  $XQ \in \text{Hor}_{AQ}$  for  $Q \in O_d$  project onto the same geodesic in the bottom space. However, regarding (Step 1), and thanks to Theorem 3.7 in [14], we have for fixed  $(A, X)$  that the cost function  $f : (Q_1, \dots, Q_n) \mapsto F(A, X, (Q_i)_{i=1}^n)$  has the Riemannian Kurdyka-Lojasiewicz property at any point of  $(O_d)^n$ . Finally, we have the convergence of the iterates towards an accumulation point thanks to Theorem 3.14 in [41]. The three assumptions in this theorem are verified in our case : Assumption (3.5) ( $L$ -Retraction Smoothness) is obtained because  $\text{grad} f$  is Lipschitz, and Corollary 10.54 in [7]; Assumption (3.7) (bounded from below) directly holds because  $f \geq 0$ ; Assumption (3.8) (individual Retraction Lipschitzness) is verified thanks to Corollary 10.47 in [7].

## E Hyperparameters

All experiments were conducted on a single V100 GPU with 32GB of memory, using a shared set of hyperparameters detailed in Table 1. The same hyperparameters are used for computing both the first and second geodesic components, except for the number of gradient steps (see Table 1), which is increased for the second component. This is likely due to the additional complexity introduced by the intersection and orthogonality constraints enforced through regularization. Both  $f_\psi$  and  $\varphi_\theta$  are implemented as standard multilayer perceptrons (MLPs) with four hidden layers of width 128. We use ELU activation functions in  $f_\psi$  because its gradient is used to parameterize a transport map in our formulation, and ELUs are commonly employed in such settings. The Sinkhorn divergence  $S_\varepsilon$  is used in the loss function as a surrogate for the squared Wasserstein distance to compute the geodesic components. The regularization parameter  $\varepsilon$  must be adapted to the scale of the data; we set it as  $\varepsilon = 0.01 \mathbb{E}_{x, x' \sim \nu_i} \|x - x'\|^2$ , where the expectation is approximated via Monte Carlo using the current minibatch samples. Note that setting  $\varepsilon$  this way is the default configuration in the OTT-JAX library. For computing the second geodesic component, we fix the regularization coefficients  $\lambda_O$  and  $\lambda_I$  to 1.0, which we found to be robust across all experiments. While increasing them (e.g., to 10.0) typically yields similar results, excessively large values may degrade performance. Conversely, if these regularization terms are too small, the algorithm tends to recover the first component as the second, due to its lower cost. In practice, we monitor the regularization terms during optimization to ensure they decrease sufficiently relative to their initial values,

confirming that the optimization effectively optimize the intersection and orthogonality constraints. To determine the hyperparameters in Table 1, we performed a grid search over the optimizer learning rate for the  $t_i$  in  $10^{-4}, 10^{-3}, 10^{-2}$ , and over the regularization coefficients  $\lambda_{\mathcal{O}}$  and  $\lambda_{\mathcal{I}}$  in 0.1, 1.0, 10.0, 100.0. We found that setting both regularization terms to 1.0 consistently yielded good performance across all experiments.

Hyperparameter	Value
$f_{\psi}$ architecture	dense MLP $d \rightarrow 128 \rightarrow 128 \rightarrow 128 \rightarrow 128 \rightarrow 1$ ELU activation functions
$f_{\psi}$ optimizer	Adam step size = 0.0005 $\beta_1 = 0.9$ $\beta_2 = 0.999$
$\varphi_{\theta}$ architecture	dense MLP $d \rightarrow 128 \rightarrow 128 \rightarrow 128 \rightarrow 128 \rightarrow d$ RELU activation functions
$\varphi_{\theta}$ optimizer	Adam step size = 0.0005 $\beta_1 = 0.9$ $\beta_2 = 0.999$
$t_i$ optimizer	Adam step size = 0.001 $\beta_1 = 0.9$ $\beta_2 = 0.999$
batch size	1024
number of gradient steps first component	120,000
number of gradient steps second component	200,000
$\lambda_{\mathcal{O}}$	1.0
$\lambda_{\mathcal{I}}$	1.0

Table 1: Hyperparameters used across all experiments.

**Note on  $\varphi$  parameterization.** Note that although  $\varphi$  is theoretically required to be a diffeomorphism in Otto’s parameterization of geodesics (equation 8), we parameterize it using a simple MLP. Initially, we experimented with normalizing flows to ensure invertibility, but observed that a standard MLP yielded similar results. In Otto’s geodesic framework,  $\varphi$  serves to modify the reference measure  $\rho$  and define the measure at  $t = 0$  along the geodesic. If  $\varphi$  is not a diffeomorphism and the pushforward  $\varphi_{\#}\rho$  is not absolutely continuous, the resulting geodesic becomes degenerate, which may hinder optimization of the loss equation (13). In practice, however, we found that the MLP  $\varphi_{\theta}$  reliably produces absolutely continuous measures, which is sufficient for our method.