CREST

CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-85

Log-PCA versus Geodesic PCA of histograms in the Wasserstein space

E. CAZELLES[1]
V. SEGUY[2]
J. BIGOT[3]
M. CUTURI[4]
N. PAPADAKIS[5]

[1] Institut de Mathématiques de Bordeaux et CNRS; IMB-UMR5251; Université de Bordeaux, E-mail: elsa.cazelles@u-bordeaux.fr

[2] Graduate School of Informatics; Kyoto University, E-mail: vivien.seguy@iip.ist.i.kyoto-u.ac.jp

[3] Institut de Mathématiques de Bordeaux et CNRS; IMB-UMR5251; Université de Bordeaux, E-mail: jeremie.bigot@math.u-bordeaux.fr

[4] CREST; ENSAE; Université Paris-Saclay, E-mail: marco.cuturi@ensae.fr

[5] Institut de Mathématiques de Bordeaux et CNRS; IMB-UMR5251; Université de Bordeaux, E-mail: nicolas.papadakis@math.u-bordeaux.fr

Log-PCA versus Geodesic PCA of histograms in the Wasserstein space.

Elsa Cazelles[¶*] Vivien Seguy[¶†] Jérémie Bigot[*] Marco Cuturi[‡]
Nicolas Papadakis[*]

August 29, 2017

Abstract

This paper is concerned by the statistical analysis of data sets whose elements are random histograms. For the purpose of learning principal modes of variation from such data, we consider the issue of computing the PCA of histograms with respect to the 2-Wasserstein distance between probability measures. To this end, we propose to compare the methods of log-PCA and geodesic PCA in the Wasserstein space as introduced in [BGKL15, SC15]. Geodesic PCA involves solving a non-convex optimization problem. To solve it approximately, we propose a novel forward-backward algorithm. This allows a detailed comparison between log-PCA and geodesic PCA of one-dimensional histograms, which we carry out using various datasets, and stress the benefits and drawbacks of each method. We extend these results for two-dimensional data and compare both methods in that setting.

Keywords: Geodesic Principal Componant Analysis, Wasserstein Space, Non-convex optimization
AMS classifications: 62-07, 68R10, 62H25

1 Introduction

Most datasets describe multivariate data, namely vectors of relevant features that can be modeled as random elements sampled from an unknown distribution. In that setting, Principal Component Analysis (PCA) is certainly the simplest and most widely used approach to reduce the dimension of such datasets. We consider in this work the statistical analysis of data sets whose elements are histograms supported on the real line. Just as with PCA, our main goal in that setting is to compute the principal modes of variation of histograms around their mean element and therefore facilitate the visualization of such datasets. However, since the number, size or locations of significant bins in the histograms of interest may vary from one histogram

[*]Institut de Mathématiques de Bordeaux et CNRS, IMB-UMR5251, Université de Bordeaux
[†]Graduate School of Informatics, Kyoto University.
[‡]CREST, ENSAE, Université Paris-Saclay.
[¶]These authors contributed equally.

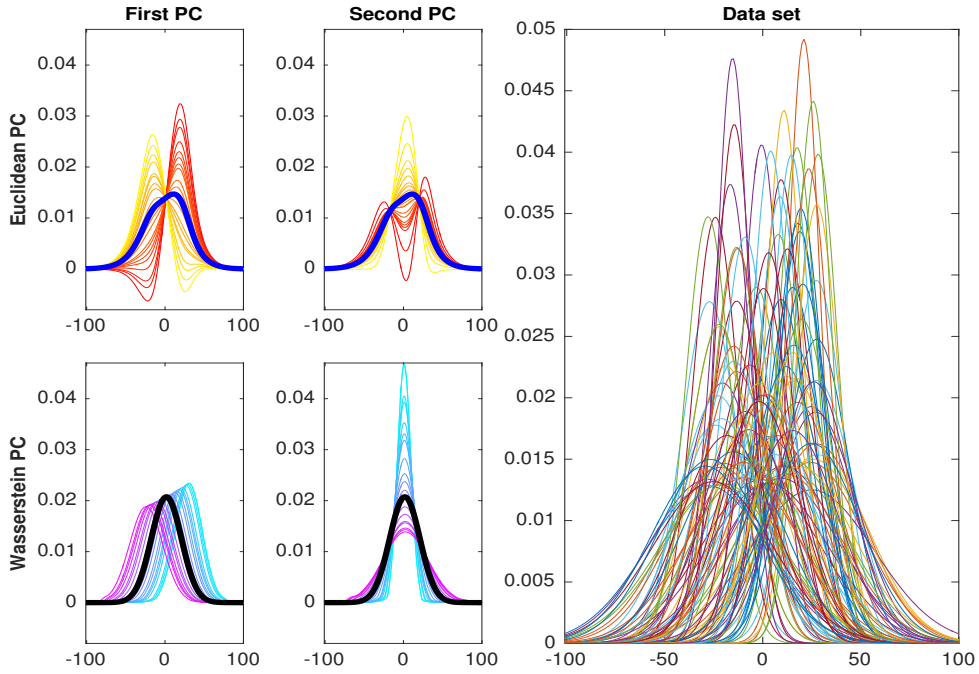

Figure 1: Synthetic example. (Right) A data set of $n = 100$ Gaussian histograms randomly translated and scaled. (Top-left) Standard PCA of this data set with respect to the Euclidean metric. The Euclidean barycenter of the data set is depicted in blue. (bottom-left) Geodesic PCA with respect to the Wasserstein metric using the iterative geodesic algorithm (4.1). The black curve represents the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $W_2(\Omega)$.

to another, using standard PCA on histograms (with respect to the Euclidean metric) is bound to fail (see for instance Figure 1).

In this paper, we propose to use the 2-Wasserstein metric [Vil03, §7.1] to measure the distance between histograms, and to compute their modes of variation accordingly. In our approach, histograms are seen as piecewise constant probability density functions (pdf) supported in a given interval Ω of the real line. In this setting, the variability in a set of histograms can be analyzed via the notion of Geodesic PCA (GPCA) of probability measures in the Wasserstein space $W_2(\Omega)$ admitting these histograms as pdf. That approach has been recently proposed in the statistics and machine learning literature in [BGKL15] for probability measures supported on the real line, and in [SC15, WSB$^+$13] for discrete probability measures on $\mathbb{R}^d$. However, implementing GPCA remains a challenging computational task even in the simplest case of pdf's supported on $\mathbb{R}$. The purpose of this paper is to provide a fast algorithm to perform GPCA of probability measures supported on the real line, and to compare its performances with log-PCA, namely standard PCA in the tangent space at the Wasserstein barycenter of the data [FLPJ04, PM16].

1.1 Related results

Foundations of Geodesic PCA in the Wasserstein space. The space of probability measures (with finite second moment) endowed with the 2-Wasserstein distance is not a Hilbert space. Therefore, standard PCA, which involves computing a covariance matrix, cannot be applied directly to compute principal mode of variations in a Wasserstein sense. Nevertheless, a meaningful notion of PCA can still be defined by relying on the pseudo-Riemannian structure of the Wasserstein space, which was extensively studied in [AGS04] and [AGS06]. Following this principle, a framework for GPCA of probability measures supported on a interval $\Omega \subset \mathbb{R}$ was introduced in [BGKL15]. GPCA is defined as the problem of estimating a principal geodesic subspace (of a given dimension) which maximizes the variance of the projection of the data to that subspace. In that approach the base point of that subspace is the Wasserstein barycenter of the data as introduced in [AC11], which is also known as a Fréchet mean. Existence, consistency and a detailed characterization of GPCA in $W_2(\Omega)$ were studied in [BGKL15]. In particular, the authors have shown that this approach is equivalent to map the data in the tangent space of $W_2(\Omega)$ at the Fréchet mean, and then to perform a PCA in this Hilbert space that is constrained to lie in a convex and closed subset of functions. Mapping the data to this tangent space is not difficult in the one-dimensional case as it amounts to computing a set of optimal maps between the data and their Wasserstein barycenter, for which a closed form is available using their quantile functions (see for example [Vil03, §2.2]). To perform PCA on the mapped data, [BGKL15] fell short of proposing an algorithm to minimize that problem, which has a non-convex and non-differentiable objective function as well as involved constraints. Only a numerical approximation to the computation of GPCA was proposed in [BGKL15], which amounts to applying log-PCA, namely a standard PCA of the dataset mapped beforehand to the tangent space of $W_2(\Omega)$ at its Fréchet mean.

Previous work in the one-dimensional case. PCA of histograms with respect to the Wasserstein metric has also been proposed in [VIB15] in the context of symbolic data analysis. Their approach consists in computing a standard PCA in the Hilbert space $L^2([0,1])$ of the quantile functions associated to the histograms. Therefore, the algorithm in [VIB15] corresponds to log-PCA of probability measures as suggested in [BGKL15], but it does not solve the problem of convex-constrained PCA in a Hilbert space associated to an exact GPCA in $W_2(\Omega)$.

PGA and log-PCA in Riemannian manifolds The method of GPCA proposed in [BGKL15] clearly shares similarities with analogs of PCA for data belonging to a Riemannian manifold $\mathcal{M}$ of finite dimension. These methods, generally referred to as Principal Geodesic Analysis (PGA) extend the notion of classical PCA in Euclidean spaces for the purpose of analyzing data belonging to curved Riemannian manifolds (see e.g. [FLPJ04, SLHN10]). This generalization of PCA proceeds by replacing Euclidean concepts of vector means, lines and orthogonality by the more general notions in Riemannian manifolds of Fréchet mean, geodesics, and orthogonality in tangent spaces.

In [FLPJ04], linearized PGA, which we refer to as log-PCA, is defined as follows. In a first step, data are mapped to the tangent space $T_{\bar{x}}\mathcal{M}$ at their Fréchet mean $\bar{x}$ by applying the logarithmic map $\log_{\bar{x}}$ to each data point. Then, in a second step, standard PCA in the Euclidean

space $T_{\bar{x}}\mathcal{M}$ can be applied. This provides a family of orthonormal tangent vectors. Principal components of variation in $\mathcal{M}$ can then defined by back-projection of these tangent vectors on $\mathcal{M}$ by using the exponential map at $\bar{x}$, that is known to parameterize geodesics at least locally. Log-PCA (e.g. linearized PGA) has low computational cost, but this comes at the expense of two simplifications and drawbacks:

(1) First, log-PCA amounts to substituting geodesic distances between data points by the linearized distance in $T_{\bar{x}}\mathcal{M}$, which may not always be a good approximation because of the curvature of $\mathcal{M}$, see e.g. [SLHN10].

(2) Secondly, the exponential map at the Fréchet mean parameterizes geodesics only locally, which implies that principal components in $\mathcal{M}$ obtained with log-PCA may not be geodesic along the typical range of the dataset.

Numerical approaches to GPCA and log-PCA in the Wasserstein space. Computational methods have been introduced in [SC15, WSB$^+$13] to extend the concepts of PGA in a Riemannian manifold to that of the space $W_2(\mathbb{R}^d)$ of probability measures supported on $\mathbb{R}^d$ endowed with the Wasserstein metric. [WSB$^+$13] propose to compute a notion of template measure (using k-means clustering) of a set of (possibly discrete) probability measures, and to consider then the optimal transport plans from each measure in the data set to that template measure. Computation of the barycentric projection of each optimal transport plan leads to a set of Monge maps over which a standard PCA can be applied, resulting in an orthonormal family of tangent vectors defined on the support of the template measure. Principal components of variation in $\mathbb{R}^d$ can then be obtained through the push-forward operator, namely by moving the mass along these tangent vectors. This approach, analog to log-PCA on Riemannian manifolds, suffers from the main drawbacks mentioned above: For $d > 1$, the linearized Wasserstein distance may be a crude approximation of the Wasserstein distance, and there is no guarantee that the computed tangent vectors parameterize geodesics of sufficient length to summarize most of the variability in the dataset. Losing geodesicity means that the principal components are curves in $W_2(\mathbb{R}^d)$ along which the mass may not be transported optimally. Therefore, this may significantly reduce the interpretability of these principal components. A different approach was proposed in [SC15], in which the notion of generalized geodesics in $W_2(\mathbb{R}^d)$ (see e.g. Chapter 9 in [AGS06]) is used to define a notion of PGA of discrete probability measures. In [SC15], generalized geodesics are parameterized using two velocity fields defined on the support of the Wasserstein barycenter. The authors proposed to minimize directly the distances from the measures in the dataset to these generalized geodesics, by updating these velocity fields which are constrained to be in opposite directions. This approach is more involved computationally than log-PCA, but it avoids some of the drawbacks highlighted above. Indeed, the resulting principal components yield curves in $W_2(\mathbb{R}^d)$ that are guaranteed to be approximately geodesics. Nevertheless, the computational method in [SC15] requires several numerical approximations (both on the optimal transport metric and on geodesics) for the algorithm to work at large scales. Therefore, it does not solve exactly the problem of computing geodesic PCA in $W_2(\mathbb{R}^d)$.

1.2 Main contributions

In this paper, we focus on computing an exact GPCA on probability measures supported on $\Omega \subset \mathbb{R}$. The case $d = 1$ has the advantage that the linearized Wasserstein distance in the tangent space is equal to the Wasserstein distance in the space $W_2(\Omega)$. The main challenge is thus to obtain principal curves which are geodesics along the range of the dataset.

The first contribution of this paper is to propose two fast algorithms for GPCA in $W_2(\Omega)$. The first algorithm finds iteratively geodesics such that the Wasserstein distance between the dataset and the parameterized geodesic is minimized with respect to $W_2(\Omega)$. This approach is thus somewhat similar to the one in [SC15]. However, a heuristic barycentric projection is used in [SC15] to remain in the feasible set of constraints during the optimization process. In our approach, we rely on proximal operators of both the objective function and the constraints to obtain an algorithm which is guaranteed to converge to a critical point of the objective function. Moreover, we show that the global minimum of our objective function for the first principal geodesic curve corresponds indeed to the solution of the exact GPCA problem defined in [BGKL15]. While this algorithm is able to find iteratively orthogonal principal geodesics, there is not guarantee that several principal geodesics parameterize a surface which is also geodesic. This is why we propose a second algorithm which computes all the principal geodesics at once by parameterizing a geodesic surface as a convex combination of optimal velocity fields, by relaxing the orthogonality constraint between principal geodesics. Both algorithms are variant of the proximal Forward-Backward algorithm. They converge to a stationary point of the objective function, as shown by recent results in non-convex optimization based on proximal methods [ABS13, OCBP14].

Our second contribution is to numerically compare log-PCA in $W_2(\Omega)$ as done in [BGKL15] (for $d = 1$) or [WSB$^+$13], with our approach to solve the exact Wasserstein GPCA problem.

Finally, we discuss some extensions of these results to the comparison of log-PCA and geodesic PCA of two-dimensional histograms.

1.3 Structure of the paper

In Section 2, we provide some background on GPCA in the Wasserstein space $W_2(\Omega)$, borrowing material from previous work in [BGKL15]. Section 3 describes log-PCA in $W_2(\Omega)$, and some of its limitations are discussed. Section 4 contains the main results of the paper, namely two algorithms for computing GPCA. In Section 5, we provide a comparison between GPCA and log-PCA using statistical analysis of real datasets of histograms. Section 6 contains a discussion on a comparison of GPCA and log-PCA of histograms supported on $\mathbb{R}^2$. Some perspectives on this work are also given. Finally, various details on the implementation of the algorithms are deferred to technical Appendices.

2 Background on Geodesic PCA in the Wasserstein space

2.1 Definitions and notations

Let Ω be a (possibly unbounded) interval in $\mathbb{R}$. Let ν be a probability measure (also called distribution) over $(\Omega, \mathcal{B}(\Omega))$ where $\mathcal{B}(\Omega)$ is the σ-algebra of Borel subsets of Ω. For a mapping $T : \Omega \to \Omega$, the push-forward measure $T\#\nu$ is a probability measure on Ω defined by $(T\#\nu)(A) = \nu\{x \in \Omega | T(x) \in A\}$, for any $A \in \mathcal{B}(\Omega)$. The cumulative distribution function (cdf) and the (generalized) quantile function of ν are denoted respectively by F_ν and F_ν^-. The Wasserstein space $W_2(\Omega)$ is the set of probability measures with support included in Ω and having a finite second moment, that is endowed with the quadratic Wasserstein distance d_W defined by

$$d_W^2(\mu, \nu) := \int_0^1 (F_\mu^-(\alpha) - F_\nu^-(\alpha))^2 d\alpha, \ \mu, \nu \in W_2(\Omega). \tag{2.1}$$

We also denote by $W_2^{ac}(\Omega)$ the set of measures $\nu \in W_2(\Omega)$ that are absolutely continuous with respect to the Lebesgue measure dx on $\mathbb{R}$. If $\mu \in W_2^{ac}(\Omega)$ then $T^* = F_\nu^- \circ F_\mu$ will be referred to as the optimal mapping to push-forward μ onto ν and in this case $d_W^2(\mu, \nu) = \int_\Omega (T^*(x) - x)^2 d\mu(x)$. For a detailed analysis of $W_2(\Omega)$ and its connection with optimal transport theory, we refer to [Vil03].

2.2 The pseudo Riemannian structure of the Wasserstein space

In what follows, μ_r denotes a reference measure in $W_2^{ac}(\Omega)$, whose choice will be discussed later on. The space $W_2(\Omega)$ has a formal Riemannian structure described, for example, in [AGS04]. The tangent space at μ_r is defined as the Hilbert space $L^2_{\mu_r}(\Omega)$ of real-valued, μ_r-square-integrable functions on Ω, equipped with the inner product $\langle \cdot, \cdot \rangle_{\mu_r}$ defined by $\langle u, v \rangle_{\mu_r} = \int_\Omega u(x)v(x)d\mu_r(x), \ u, v$ $L^2_{\mu_r}(\Omega)$, and associated norm $\| \cdot \|_{\mu_r}$. We define the exponential and the logarithmic maps at μ_r, as follows.

Definition 2.1. Let $\mathrm{id} : \Omega \to \Omega$ be the identity mapping. The exponential $\exp_{\mu_r} : L^2_{\mu_r}(\Omega) \to W_2(\mathbb{R})$ and logarithmic $\log_{\mu_r} : W_2(\Omega) \to L^2_{\mu_r}(\Omega)$ maps are defined respectively as

$$\exp_{\mu_r}(v) = (\mathrm{id} + v)\#\mu_r \quad \text{and} \quad \log_{\mu_r}(\nu) = F_\nu^- \circ F_{\mu_r} - \mathrm{id}. \tag{2.2}$$

Contrary to the setting of Riemannian manifolds, the "exponential map" $\exp_{\mu_r}$ defined above is not a local homeomorphism from a neighborhood of the origin in the "tangent space" $L^2_{\mu_r}(\Omega)$ to the space $W_2(\Omega)$, see e.g. [AGS04]. Nevertheless, it is shown in [BGKL15] that $\exp_{\mu_r}$ is an isometry when restricted to the following specific set of functions

$$V_{\mu_r}(\Omega) := \log_{\mu_r}(W_2(\Omega)) = \left\{ \log_{\mu_r}(\nu) \ ; \ \nu \in W_2(\Omega) \right\} \subset L^2_{\mu_r}(\Omega),$$

and that the following results hold (see [BGKL15]).

Proposition 2.1. *The subspace $V_{\mu_r}(\Omega)$ satisfies the following properties :*

(P1) *the exponential map $\exp_{\mu_r}$ restricted to $V_{\mu_r}(\Omega)$ is an isometric homeomorphism, with inverse $\log_{\mu_r}$. We have hence $W_2(\nu, \eta) = ||\log_{\mu_r}(\nu) - \log_{\mu_r}(\eta)||_{L^2_{\mu_r}(\Omega)}$.*

(P2) *the set* $V_{\mu_r}(\Omega) := \log_\mu(W_2(\Omega))$ *is closed and convex in* $L^2_{\mu_r}(\Omega)$.

(P3) *the space* $V_{\mu_r}(\Omega)$ *is the set of functions* $v \in L^2_{\mu_r}(\Omega)$ *such that* $T := \mathrm{id} + v$ *is* μ_r*-almost everywhere non decreasing and that* $T(x) \in \Omega$*, for* $x \in \Omega$.

Moreover, it follows, from [BGKL15], that geodesics in $W_2(\Omega)$ are exactly the image under $\exp_{\mu_r}$ of straight lines in $V_{\mu_r}(\Omega)$. This property is stated in the following lemma.

Lemma 2.1. *Let* $\gamma : [0,1] \to W_2(\Omega)$ *be a curve and let* $v_0 := \log_{\mu_r}(\gamma(0))$, $v_1 := \log_{\mu_r}(\gamma(1))$. *Then* $\gamma = (\gamma_t)_{t\in[0,1]}$ *is a geodesic if and only if* $\gamma_t = \exp_{\mu_r}((1-t)v_0 + tv_1)$, *for all* $t \in [0,1]$.

2.3 GPCA for probability measures

Let $\nu_1, \ldots, \nu_n$ be a set of probability measures in $W_2^{ac}(\Omega)$. Assuming that each ν_i is absolutely continuous simplify the following presentation, and it is in line with the purpose of statistical analysis of histograms. We define now the notion of (empirical) GPCA of this set of probability measures by following the approach in [BGKL15]. The first step is to choose the reference measure μ_r. To this end, let us introduce the Wasserstein barycenter [AC11] or Fréchet mean of the ν_i's, that is defined as the probability measure $\bar{\boldsymbol{\nu}}$,

$$\bar{\boldsymbol{\nu}} = \mathrm{argmin}_{\mu\in W_2(\Omega)} \frac{1}{n}\sum_{i=1}^{n} d_W^2(\nu_i, \mu).$$

Note that it immediately follows from results in [AC11] that $\bar{\boldsymbol{\nu}} \in W_2^{ac}(\Omega)$, and that its cdf satisfies

$$F^-_{\bar{\boldsymbol{\nu}}} = \frac{1}{n}\sum_{i=1}^{n} F^-_{\nu_i}. \tag{2.3}$$

A typical choice for the reference measure is to take $\mu_r = \bar{\boldsymbol{\nu}}$ which represents an average location in the data around which can be computed the principal sources of geodesic variability. To introduce the notion of a principal geodesic subspace of the measures $\nu_1, \ldots, \nu_n$, we need to introduce further notation and definitions. Let G be a subset of $W_2(\Omega)$. The distance between $\mu \in W_2(\Omega)$ and the set G is $d_W(\nu, G) = \inf_{\lambda\in G} d_W(\nu, \lambda)$, and the average distance between the data and G is taken as

$$\mathrm{D}_W(G) := \frac{1}{n}\sum_{i=1}^{n} d_W^2(\nu_i, G). \tag{2.4}$$

Definition 2.2. Let K be some positive integer. A subset $G \subset W_2(\Omega)$ is said to be a geodesic set of dimension $\dim(G) = K$ if $\log_{\mu_r}(G)$ is a convex set such that the dimension of the smallest affine subspace of $L^2_{\mu_r}(\Omega)$ containing $\log_{\mu_r}(G)$ is of dimension K.

The notion of principal geodesic subspace (PGS) with respect to the reference measure $\mu_r = \bar{\boldsymbol{\nu}}$ can now be presented below.

Definition 2.3. Let $\mathrm{CL}(W)$ be the metric space of nonempty, closed subsets of $W_2(\Omega)$, endowed with the Hausdorff distance, and

$$\mathrm{CG}_{\bar{\boldsymbol{\nu}},K}(W) = \{G \in \mathrm{CL}(W) \mid \bar{\boldsymbol{\nu}} \in G,\ G \text{ is a geodesic set and } \dim(G) \leq K\}\,,\ K \geq 1.$$

A principal geodesic subspace (PGS) of $\boldsymbol{\nu}$ of dimension K with respect to $\bar{\boldsymbol{\nu}}$ is a set

$$G_K \in \operatorname*{argmin}_{G \in \mathrm{CG}_{\bar{\boldsymbol{\nu}},K}(W)} \mathrm{D}_W(G). \tag{2.5}$$

When $K = 1$, searching for the first PGS of $\boldsymbol{\nu}$ simply amounts to search for a geodesic curve $\gamma^{(1)}$ that is a solution of the following optimization problem:

$$\tilde{\gamma}^{(1)} := \operatorname*{argmin}_{\gamma} \left\{ \frac{1}{n} \sum_{i=1}^{n} d_W^2(\nu_i, \gamma) \mid \ \gamma \text{ is a geodesic in } W_2(\Omega) \text{ passing through } \mu_r = \bar{\boldsymbol{\nu}}. \right\}.$$

We remark that this definition of $\tilde{\gamma}^{(1)}$ as the first principal geodesic curve of variation in $W_2(\Omega)$ is consistent with the usual concept of PCA in a Hilbert space in which geodesic are straight lines.

For a given dimension k, the GPCA problem consists in finding a nonempty closed geodesic subset of dimension k which contains the reference measure μ_r and minimizes Eq. (2.4). We describe in the next section how we can parameterize such sets G.

2.4 Geodesic PCA parameterization

GPCA can be formulated as an optimization problem in the Hilbert space $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$. To this end, let us define the functions $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$ for $1 \leq i \leq n$ that corresponds to the data mapped in the tangent space. It can be easily checked that this set of functions is centered in the sense that $\frac{1}{n}\sum_{i=1}^{n} \omega_i = 0$. Note that, in a one-dimensional setting, computing ω_i (mapping of the data to the tangent space) is straightforward since the optimal maps $T_i^* = F_{\nu_i}^- \circ F_{\bar{\boldsymbol{\nu}}}$ between the data and their Fréchet mean are available in a simple and closed form.

For $\mathcal{U} = \{u_1, \ldots, u_K\}$ a collection of $K \geq 1$ functions belonging to $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$, we denote by $\mathrm{Sp}(\mathcal{U})$ the subspace spanned by $u_1, \ldots, u_K$. Defining $\mathit{\Pi}_{\mathrm{Sp}(\mathcal{U})} v$ as the projection of $v \in L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ onto $\mathrm{Sp}(\mathcal{U})$, and $\mathit{\Pi}_{\mathrm{Sp}(\mathcal{U}) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)} v$ as the projection of v onto the closed convex set $\mathrm{Sp}(\mathcal{U}) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)$, then we have

Proposition 2.2. *Let $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$ for $1 \leq i \leq n$, and $\mathcal{U}^* = \{u_1^*, \ldots, u_k^*\}$ be a minimizer of*

$$\frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \mathit{\Pi}_{\mathrm{Sp}(\mathcal{U}) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)} \omega_i\|^2_{\bar{\boldsymbol{\nu}}}, \tag{2.6}$$

over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_K\}$ of functions in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ of dimension K (namely such that $\langle u_j, u_{j'} \rangle_{\bar{\boldsymbol{\nu}}} = 0$ if $j \neq j'$). If we let

$$G_{\mathcal{U}^*} := \exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\mathcal{U}^*) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)),$$

then $G_{\mathcal{U}^*}$ is a principal geodesic subset (PGS) of dimension k of the measures $\nu_1, \ldots, \nu_n$, meaning that $G_{\mathcal{U}^*}$ belongs to the set of minimizers of the optimization problem (2.5).

Proof. For $v \in L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ and a subset $C \in L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$, we define $d_{\bar{\nu}}(v, C) = \inf_{u \in C} \|v - u\|_{\bar{\nu}}$. Remark that $\sum_i \omega_i = 0$. Hence by Proposition 3.3 in [BGKL15], if $\mathcal{U}^*$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n} d^2_{\bar{\nu}}(\omega_i, \mathrm{Sp}(\mathcal{U}^*) \cap V_{\bar{\nu}}(\Omega)) = \frac{1}{n}\sum_{i=1}^{n} \|\omega_i - \Pi_{\mathrm{Sp}(\mathcal{U}^*)\cap V_{\bar{\nu}}(\Omega)}\omega_i\|^2_{\bar{\boldsymbol{\nu}}},$$

then $\mathrm{Sp}(\mathcal{U}^*) \cap V_{\bar{\nu}}(\Omega) \in \mathrm{argmin}_C \frac{1}{n}\sum_{i=1}^{n} d^2_{\bar{\nu}}(\omega_i, C)$, where C is taken over all nonempty, closed, convex set of $V_{\bar{\nu}}(\Omega)$ such that $dim(C) \leq K$ and $0 \in C$. By Proposition 4.3 in [BGKL15], and since $\log_{\bar{\nu}}(\bar{\nu}) = 0$, we can conclude that G^* is a geodesic subset of dimension K which minimizes (2.4). □

Thanks to Proposition 2.2, it follows that GPCA in $W_2(\Omega)$ corresponds to a mapping of the data into the Hilbert space $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ which is followed by a PCA in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ that is constrained to lie in the convex and closed subset $V_{\bar{\boldsymbol{\nu}}}(\Omega)$. This has to be interpreted as a geodesicity constraint coming from the definition of a PGS in $W_2(\Omega)$.

3 The log-PCA approach

For data in a Riemannian manifold, we recall that log-PCA consists in solving a linearized version of the PGA problem by mapping the whole data set to the tangent space at the Fréchet mean through the logarithmic map [FLPJ04]. This approach is computationally attractive since it boils down to computing a standard PCA. [WSB$^+$13] used this idea to linearize geodesic PCA in the Wasserstein space $W_2(\mathbb{R}^d)$, by defining the logarithmic map of a probability measure as the barycentric projection of the transport plan with respect to a template measure. This approach has the two drawbacks (1) and (2) of log-PCA mentioned in Section 1.1. A third limitation inherent to the Wasserstein space is that when this template or mean probability measure is discrete, the logarithmic map cannot be defined anywhere. This is why the authors of [WSB$^+$13] had to apply the barycentric projection, which is a lossy process, to the data set before being able to map it to the tangent space.

We consider as usual a subset $\Omega \subset \mathbb{R}$. In this setting, $W_2(\Omega)$ is a flat space as shown by the isometry property (P1) of Proposition 2.1. Moreover, if the Wasserstein barycenter $\bar{\boldsymbol{\nu}}$ is assumed to be absolutely continuous, then Definition 2.1 shows that the logarithmic map at $\bar{\boldsymbol{\nu}}$ is well defined everywhere. Under such an assumption, log-PCA in $W_2(\Omega)$ corresponds to the following steps:

1. compute the log maps (see Definition 2.1) $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, $i = 1, \ldots, n$,
2. perform the PCA of the projected data $\omega_1, \cdots, \omega_n$ in the Hilbert space $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ to obtain K orthogonal directions $\tilde{u}_1, \ldots, \tilde{u}_K$ in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ of principal variations,
3. recover a principal subspace of variation in $W_2(\Omega)$ with the exponential map $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}))$ of the principal eigenspace $\mathrm{Sp}(\tilde{\mathcal{U}})$ in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ spanned by $\tilde{u}_1, \ldots, \tilde{u}_K$.

For specific datasets, log-PCA in $W_2(\Omega)$ may be equivalent to GPCA, in the sense that the set $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega))$ is a principal geodesic subset of dimension K of the measures

$\nu_1, \ldots, \nu_n$, as defined by (2.5). Informally, this case corresponds to the setting where the data are sufficiently concentrated around their Wasserstein barycenter $\bar{\boldsymbol{\nu}}$ (we refer to Remark 3.5 in [BGKL15] for further details). However, carrying out a PCA in the tangent space of $W_2(\mathbb{R})$ at $\bar{\boldsymbol{\nu}}$ is a relaxation of the convex-constrained GPCA problem (2.6), where the elements of the sought principal subspace do not need to be in $V_{\bar{\boldsymbol{\nu}}}$. Indeed, standard PCA in the Hilbert space $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$, amounts to find $\tilde{\mathcal{U}} = \{\tilde{u}_1, \ldots, \tilde{u}_K\}$ minimizing,

$$\frac{1}{n}\sum_{i=1}^{n} \|\omega_i - \Pi_{\mathrm{Sp}(\mathcal{U})}\omega_i\|^2_{\bar{\boldsymbol{\nu}}},, \tag{3.1}$$

over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_k\}$ of functions in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$. It is worth noting that the three steps of log-PCA in $W_2(\Omega)$ are simple to implement and fast to compute, but that performing log-PCA or GPCA (2.6) in $W_2(\Omega)$ is not necessarily equivalent.

Log-PCA is generally used for two main purposes. The first one is to obtain a low dimensional representation of each data measure $\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ through the coefficients $\langle \omega_i, \tilde{u}_k \rangle_{L^2_{\bar{\boldsymbol{\nu}}}}$. From this low dimensional representation, the measure $\nu_i \in W_2(\Omega)$ can be approximated through the exponential mapping $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\mathcal{U})}\omega_i)$. The second one is to visualize each mode of variation in the dataset, by considering the evolution of the curve $t \mapsto \exp_{\bar{\boldsymbol{\nu}}}(t\tilde{u}_k)$ for each $\tilde{u}_k \in \tilde{\mathcal{U}}$.

However, relaxing the convex-constrained GPCA problem (2.6) when using log-PCA results in several issues. Indeed, as shown in the following paragraphs, not taking into account this geodesicity constraint makes difficult the computation and interpretation of $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}))$ as a principal subspace of variation, which may limit its use for data analysis.

Numerical implementation of pushforward operators A first downside to the log-PCA approach is the difficulty of the numerical implementation of the pushforward operator in the exponential map $\exp_{\bar{\boldsymbol{\nu}}}(v) = (\mathrm{id} + v)\#\bar{\boldsymbol{\nu}}$ when the mapping $\mathrm{id} + v$ is not a strictly increasing function for a given vector $v \in \mathrm{Sp}(\tilde{\mathcal{U}})$. This can be shown with the following proposition, which provides a formula for computing the density of a pushforward operator.

Proposition 3.1. *(Density of the pushforward) Let $\mu \in W_2(\mathbb{R})$ be an absolutely continuous measure with density ρ (that is possibly supported on an interval $\Omega \subset \mathbb{R}$). Let $T : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $|T'(x)| > 0$ for almost every $x \in \mathbb{R}$, and define $\nu = T\#\mu$. Then, ν admits a density g given by,*

$$g(y) = \sum_{x \in T^{-1}(y)} \frac{\rho(x)}{|T'(x)|},\ y \in \mathbb{R}. \tag{3.2}$$

When T is injective, this simplifies to,

$$g(y) = \frac{\rho(T^{-1}(y))}{|T'(T^{-1}(y))|}. \tag{3.3}$$

Proof. Under the assumptions made on T, the coarea formula (which is a more general form of Fubini's theorem, see e.g. [KP08] Corollary 5.2.6 or [EG15] Section 3.4.3) states that, for any

measurable function $h : \mathbb{R} \to \mathbb{R}$, one has

$$\int_{\mathbb{R}} h(x)|T'(x)|dx = \int_{\mathbb{R}} \sum_{x \in T^{-1}(y)} h(x)dy. \tag{3.4}$$

Let B a Borel set and choose $h(x) = \frac{\rho(x)}{|T'|}\mathbf{1}_{T^{-1}(B)}$, x. Hence, using (3.4), one obtains that

$$\int_{T^{-1}(B)} \rho(x)dx = \int_{\mathbb{R}} \sum_{x \in T^{-1}(y)} \frac{\rho(x)}{|T'(x)|}\mathbf{1}_{T^{-1}(B)}(x)dy = \int_{B} \sum_{x \in T^{-1}(y)} \frac{\rho(x)}{|T'(x)|}dy.$$

The definition of the pushforward $\nu(B) = \mu(T^{-1}(B))$ then completes the proof. □

The numerical computation of formula (3.2) or (3.3) is not straightforward. When T is not injective, computation of the formula (3.2) must be done carefully by partitioning the domain of T in sets on which T is injective. Such a partitioning depends on the method of interpolation for estimating a continuous density ρ from a finite set of its values on a grid of reals. More importantly, when $T'(x)$ is very small, $\frac{\rho(x)}{T'(x)}$ may become very irregular and the density of $\nu = T\#\mu$ may exhibit large peaks, see Figure 2 for an illustrative example.

Pushforward of the barycenter outside the support Ω. A second downside of log-PCA in $W_2(\Omega)$ is that the range of the mapping $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i$ may be larger than the interval Ω. This implies that the density of the pushforward of the Wassertain barycenter $\bar{\boldsymbol{\nu}}$ by this mapping, namely $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)$, may have a support which is not included in Ω. This issue may be critical when trying to estimate the measure $\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ by its projected measure $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)$. For example, in a dataset of histograms with bins necessarily containing only positive reals, a projected distribution with positive mass on negative reals would be hard to interpret.

A higher Wasserstein reconstruction error. Finally, relaxing the geodesicity constraint (2.6) may actually increase the Wasserstein reconstruction error with respect to the Wasserstein distance. To state this issue more clearly, we define the reconstruction error of log-PCA as

$$\tilde{r}(\tilde{\mathcal{U}}) = \frac{1}{n}\sum_{i=1}^{n} d_W^2\big(\nu_i, \exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)\big). \tag{3.5}$$

and the reconstruction error of GPCA as

$$r(\mathcal{U}^*) = \frac{1}{n}\sum_{i=1}^{n} d_W^2\big(\nu_i, \exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\mathcal{U}^*)\cap V_{\bar{\boldsymbol{\nu}}}(\Omega)}\omega_i)\big). \tag{3.6}$$

where $\mathcal{U}^*$ is a minimiser of (2.6). Note that in (3.5), the projected measures $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)$ might have a support that lie outside Ω. Hence, the Wasserstein distance d_W in (3.5) has to be understood for measures supported on $\mathbb{R}$ (with the obvious extension to zero of ν_i outside Ω).

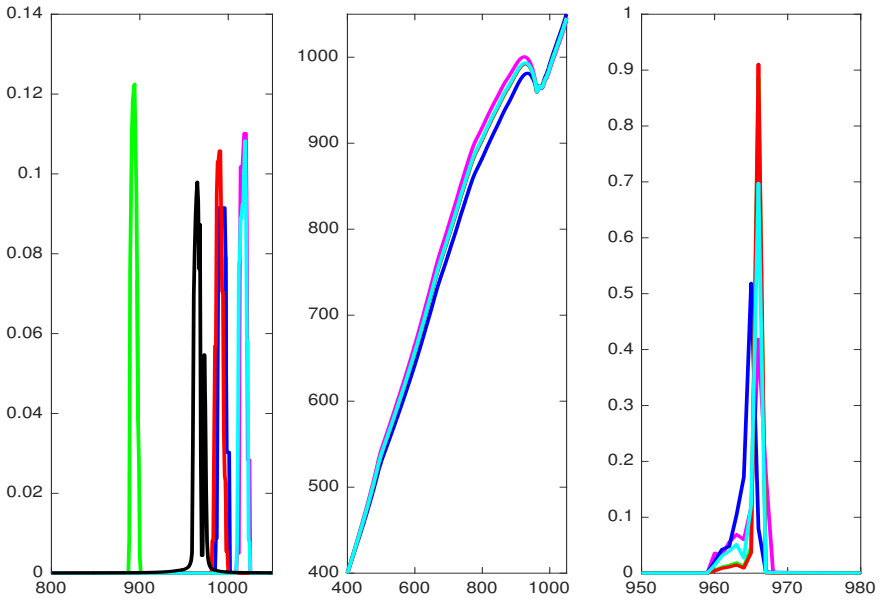

Figure 2: (Left) Distribution of the total precipitation (mm) collected in a year in $1 \leq i \leq 5$ stations among 60 in China - Source : Climate Data Bases of the People's Republic of China 1841-1988 downloaded from http://cdiac.ornl.gov/ndps/tr055.html. The black curve is the density of the Wasserstein barycenter of the 60 stations. (Middle) Mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\tilde{u}_2)}\omega_i$ obtained from the projections of these 5 distributions onto the second eigenvector $\tilde{u}_2$ given by log-PCA of the whole dataset. (Right) Pushforward $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_2)}\omega_i) = T_i \# \bar{\boldsymbol{\nu}}$ of the Wasserstein barycenter $\bar{\boldsymbol{\nu}}$ for each $1 \leq i \leq 5$. As the derivative T_i' take very small values, the densities of the pushforward barycenter $T_i \# \bar{\boldsymbol{\nu}}$ for $1 \leq i \leq 5$ exhibit large peaks (between 0.4 and 0.9) whose amplitude is beyond the largest values in the original data set (between 0.08 and 0.12).

The Wasserstein reconstruction error $\tilde{r}(\tilde{\mathcal{U}})$ of log-PCA is the sum of the Wasserstein distances of each data point ν_i to a point on the surface $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}))$ which is given by the decomposition of ω_i on the orthonormal basis $\tilde{\mathcal{U}}$. However, by Proposition 2.1, the isometry property (P1) only holds between $W_2(\mathbb{R})$ and the convex subset $V_{\bar{\boldsymbol{\nu}}} \subset L^2_{\bar{\boldsymbol{\nu}}}(\mathbb{R})$. Therefore, we may not have $d^2_W\big(\nu_i, \exp_{\bar{\boldsymbol{\nu}}}\left(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i\right)\big) = \|\omega_i - \Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i\|^2_{\bar{\boldsymbol{\nu}}}$ as $\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i$ is a function belonging to $L^2_{\bar{\boldsymbol{\nu}}}(\mathbb{R})$ which may not necessarily be in $V_{\bar{\boldsymbol{\nu}}}$. In this case, the minimal Wasserstein distance between ν_i and the surface $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\mathcal{U}^*))$ is not equal to $\|\omega_i - \Pi_{\mathrm{Sp}(\mathcal{U})}\omega_i\|_{\bar{\boldsymbol{\nu}}}$, and this leads to situations where $\tilde{r}(\tilde{\mathcal{U}}) > r(\mathcal{U}^*)$ as illustrated in Figure 3.

4 Two algorithmic approaches for GPCA in $W_2(\Omega)$, for $\Omega \subset \mathbb{R}$

In this section, we introduce two algorithms which solve some of the issues of log-PCA that have been raised in Section 3. First, the output of the proposed algorithms guarantees that the computation of mappings to pushforward the Wassertein barycenter to approximate elements in

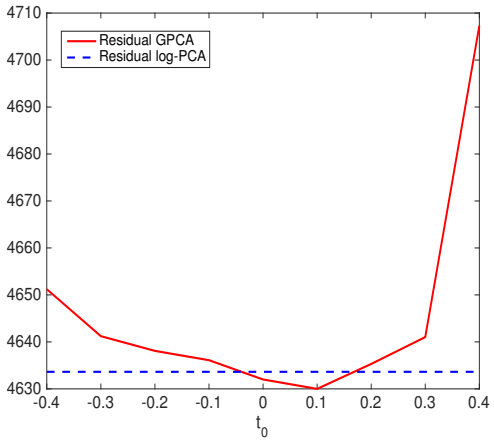

Figure 3: Comparison of the Wasserstein reconstruction error between GPCA and log-PCA on the synthetic dataset displayed in Figure 1 for the first component, with an illustration of the role of the parameter t_0 in (4.2).

the datasets are strictly increasing (that is they are optimal). As a consequence, the resulting pushforward density behaves numerically much better. Secondly, the geodesic curve or surface are constrained to lie in $W_2(\Omega)$, implying that the projection of the data are distributions whose supports do not lie outside Ω.

4.1 Iterative geodesic approach

In this section, we propose an algorithm to solve a variant of the convex-constrained GPCA problem (2.6). Rather than looking for a geodesic subset of a given dimension which fits well the data, we find iteratively orthogonal principal geodesics (i.e. geodesic set of dimension one). Assuming that that we already know a subset $\mathcal{U}^{k-1} \subset L^2_{\boldsymbol{\nu}}(\Omega)$ containing $k-1$ orthogonal principal directions $\{u_l\}_{l=1}^{k-1}$ (with $\mathcal{U}^0 = \emptyset$), our goal is to find a new direction $u_k \in L^2_{\boldsymbol{\nu}}(\Omega)$ of principal variation by solving the optimisation problem:

$$u_k \in \underset{v \perp \mathcal{U}^{k-1}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \Pi_{\mathrm{Sp}(v) \cap V_{\boldsymbol{\nu}}(\Omega)} \omega_i\|^2_{\boldsymbol{\nu}}, \tag{4.1}$$

where the infimum above is taken over all $v \in L^2_{\boldsymbol{\nu}}(\Omega)$ belonging to the orthogonal of $\mathcal{U}^{k-1}$. This iterative process is not equivalent to the GPCA problem (2.6), with the exception of the first principal geodesic ($k = 1$). Nevertheless, it computes principal subsets $\mathcal{U}^k$ of dimension k such that the projections of the data onto every direction of principal variation lie in the convex set $V_{\boldsymbol{\nu}}$.

The following proposition is the key result to derive an algorithm to solve (4.1) on real data.

Proposition 4.1. *Introducing the characteristic function of the convex set $V_{\boldsymbol{\nu}}(\Omega)$ as:*

$$\chi_{V_{\boldsymbol{\nu}}(\Omega)}(v) = \begin{cases} 0 & if\ v \in V_{\boldsymbol{\nu}}(\Omega) \\ +\infty & otherwise \end{cases}$$

the optimisation problem (4.1) *is equivalent to*

$$u_k = \underset{v \perp \mathcal{U}^{k-1}}{\operatorname{argmin}} \min_{t_0 \in [-1;1]} H(t_0, v), \tag{4.2}$$

where

$$H(t_0, v) = \frac{1}{n} \sum_{i=1}^{n} \min_{t_i \in [-1;1]} \|\omega_i - (t_0 + t_i)v\|^2_{\bar{\boldsymbol{\nu}}} + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0 - 1)v) + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0 + 1)v). \tag{4.3}$$

Proof. We first observe that $\Pi_{\mathrm{Sp}(u) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)} \omega_i = \beta_i u$, with $\beta_i \in \mathbb{R}$ and $\beta_i u \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$. Hence, for u_k solution of (4.1), we have:

$$\frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \Pi_{\mathrm{Sp}(u_k) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)} \omega_i\|^2_{\bar{\boldsymbol{\nu}}} = \frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \beta_i u_k\|^2_{\bar{\boldsymbol{\nu}}}.$$

such that $\beta_i \in \mathbb{R}$ and $\beta_i u_k \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$ for all $i \in \{1, \ldots, n\}$. We take $M \in \operatorname{argmax}_{1 \leq i \leq n} \beta_i$ and $m \in \operatorname{argmin}_{1 \leq i \leq n} \beta_i$. Without loss of generality, we can assume that $\beta_M > 0$ and $\beta_m < 0$. We then define $v = (\beta_M - \beta_m)u_k/2$ and $t_0 = (\beta_M + \beta_m)/(\beta_M - \beta_m)$, that checks $|t_0| < 1$. Hence, for all $i = 1, \ldots, n$, there exists $t_i \in [-1; 1]$ such that: $\beta_i u_k = (t_0 + t_i)v \in V_{\bar{\boldsymbol{\nu}}}$. In particular, one has $t_M = 1$ and $t_m = -1$, which means that $(t_0 \pm 1)v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$. Reciprocally, $(t_0 \pm 1)v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$ insure us by convexity of $V_{\bar{\boldsymbol{\nu}}}(\Omega)$ that for all $t_i \in [-1; 1]$, $(t_0 + t_i)v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$. □

Proposition 4.1 may be interpreted as follows. For a given $t_0 \in [-1; 1]$, let $v \in \perp\, \mathcal{U}^{k-1}$ satisfying $(t_0 - 1)v \in V_{\bar{\boldsymbol{\nu}}}$ and $(t_0 + 1)v \in V_{\bar{\boldsymbol{\nu}}}$. Then, if one defines the curve

$$g_t(t_0, v) = (id + (t_0 + t)v) \# \bar{\boldsymbol{\nu}} \text{ for } t \in [-1; 1], \tag{4.4}$$

it follows, from Lemma 2.1, that $(g_t(t_0, v))_{t \in [-1;1]}$ is a geodesic since it can be written as $g_t(t_0, v) = \exp_{\bar{\boldsymbol{\nu}}}((1 - u)w_0 + uw_1), u \in [0, 1]$ with $w_0 = (t_0 - 1)v$, $w_1 = (t_0 + 1)v$, $u = (t + 1)/2$, and with w_0 and w_1 belonging to $V_{\bar{\boldsymbol{\nu}}}$ for $|t_0| < 1$. From the isometry property (P1) in Proposition 2.1, one has

$$\min_{t_i \in [-1;1]} \|\omega_i - (t_0 + t_i)v\|^2_{\bar{\boldsymbol{\nu}}} = \min_{t_i \in [-1;1]} d^2_W(\nu_i, g_{t_i}(v)), \tag{4.5}$$

and thus the objective function $H(t_0, v)$ in (4.2) is equal to the sum of the squared Wasserstein distances between the dataset to the geodesic curve $(g_t(t_0, v))_{t \in [-1;1]}$.

The choice of the parameter t_0 corresponds to the location of the mid-point of the geodesic $g_t(t_0, v)$, and it plays a crucial role. Indeed, the minimisation of $H(t_0, v)$ over $t_0 \in [-1; 1]$ in (4.2) cannot be avoided to obtain an optimal Wasserstein reconstruction error. This is illustrated by the Figure 3, where the Wasserstein reconstruction error $\tilde{r}(\tilde{\mathcal{U}})$ of log-PCA (see relation (3.5)) is compared with the ones of GPCA, for different t_0, obtained for $k = 1$ as

$$t_0 \in [-1; 1] \mapsto H(t_0, u_1^{t_0})$$

with $u_1^{t_0} = \operatorname{argmin}_v H(t_0, v)$.We therefore exhibit that GPCA can lead to a better low dimensional data representation than log-PCA in term of Wasserstein residuals.

4.2 Geodesic surface approach

Once a family of vectors $(v_1, \cdots, v_k)$ has been found through the minisation of problem (4.1), one can recover a geodesic subset of dimension k by considering all convex combinations of the vectors $((t_0^1+1)v_1, (t_0^1-1)v_1, \cdots, (t_0^k+1)v_k, (t_0^k-1)v_k)$. However, this subset may not be a solution of (2.6) since we have no guarantee that a data point ν_i is actually close to this geodesic subset. This discussion suggests that we may consider solving the GPCA problem (2.6) over geodesic set parameterized as in Proposition 4.1. In order to find principal geodesic subsets which are close to the dataset, we consider a family $V^K = (v_1, \cdots, v_K)$ of linearly independant vectors and $\mathbf{t}_0^K = (t_0^1, \cdots, t_0^K) \in [-1,1]^K$ such that $(t_0^1-1)v_1, (t_0^1+1)v_1, \cdots, (t_0^K-1)v_K, (t_0^K+1)v_K$ are all in $V_{\bar{\boldsymbol{\nu}}}$. Convex combinations of the latter family provide a parameterization of a geodesic set of dimension K by taking the exponential map $\exp_{\bar{\boldsymbol{\nu}}}$ of

$$\hat{V}_{\bar{\boldsymbol{\nu}}}(V^K, \mathbf{t}_0^K) = \{\sum_{k=1}^{K} (\alpha_k^+(t_0^k+1) + \alpha_k^-(t_0^k-1))v_k, \alpha^{\pm} \in A\} \tag{4.6}$$

where A is a simplex constraint: $\alpha^{\pm} \in A \Leftrightarrow \alpha_k^+, \alpha_k^- \geq 0$ and $\sum_{k=1}^K (\alpha_k^+ + \alpha_k^-) \leq 1$. We hence substitute the general sets $\mathrm{Sp}(\mathcal{U}) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)$ in the definition of the GPCA problem (2.6) to obtain,

$$\begin{aligned}
(u_1, \cdots, u_K) &= \underset{V^K, \mathbf{t}_0^K}{\operatorname{argmin}} \quad \frac{1}{n}\sum_{i=1}^{n} \|\omega_i - \Pi_{\hat{V}_{\bar{\boldsymbol{\nu}}}(V^K, t^K)}\omega_i\|_{\bar{\boldsymbol{\nu}}}^2, \\
&= \underset{v_1, \cdots, v_K}{\operatorname{argmin}} \quad \min_{\mathbf{t}_0^K \in [-1,1]^K} \frac{1}{n}\sum_{i=1}^{n} \min_{\alpha_i^{\pm} \in A} \|\omega_i - \sum_{k=1}^{K} (\alpha_{ik}^+(t_0^k+1) + \alpha_{ik}^-(t_0^k-1))v_k\|_{\bar{\boldsymbol{\nu}}}^2 \\
&\qquad + \sum_{k=1}^{K} \Big(\chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0^k+1)v_k) + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0^k-1)v_k)\Big) + \sum_{i=1}^{n} \chi_A(\alpha_i^{\pm}).
\end{aligned} \tag{4.7}$$

4.3 Discretization and Optimization

In this section we follow the framework of the iterative geodesic algorithm. We provide additional details When the optimization procedure of the geodesic surface approach differs from the iterative one.

4.3.1 Discrete optimization problem

Let $\Omega = [a; b]$ be a compact interval, and consider its discretization over N points $a = x_1 < x_2 < \cdots < x_N = b$, $\Delta_j = x_{j+1} - x_j$, $j = 1, \ldots, N-1$. We recall that the functions $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$ for $1 \leq i \leq n$ are elements of $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ which correspond to the mapping of the data to the tangent space at the Wasserstein barycenter $\bar{\boldsymbol{\nu}}$. In what follows, for each $1 \leq i \leq n$, the discretization of the function ω_i over the grid reads $\mathbf{w}_i = (w_i^j)_{j=1}^N \in \mathbb{R}^N$. We also recall that $\chi_A(u)$ is the characteristic function of a given set A, namely $\chi_A(u) = 0$ if $u \in A$ and $+\infty$ otherwise. Finally, the space $\mathbb{R}^N$ is understood to be endowed with the following inner product and norm $\langle \mathbf{u}, \mathbf{v} \rangle_{\bar{\boldsymbol{\nu}}} = \sum_{j=1}^N \bar{\boldsymbol{f}}(x_j) u_j v_j$ and $\|\mathbf{v}\|_{\bar{\boldsymbol{\nu}}}^2 = \langle \mathbf{v}, \mathbf{v} \rangle_{\bar{\boldsymbol{\nu}}}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, where $\bar{\boldsymbol{f}}$ denotes the density of the measure $\bar{\boldsymbol{\nu}}$. Let us now suppose that we have already computed $k-1$ orthogonal (in the

sense $\langle \mathbf{u}, \mathbf{v} \rangle_{\bar{\boldsymbol{\nu}}} = 0$) vectors $\mathbf{u}_1, \cdots \mathbf{u}_{k-1}$ in $\mathbb{R}^N$ which stand for the discretization of orthonormal functions $u_1, \ldots, u_{k-1}$ in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ over the grid $(x_j)_{j=1}^N$.

Discretizing problem (4.2) for a fixed $t_0 \in]-1;1[$, our goal is to find a new direction $\mathbf{u}_k \in \mathbb{R}^N$ of principal variations by solving the following problem over all $\mathbf{v} = \{v_j\}_{j=1}^N \in \mathbb{R}^N$:

$$\mathbf{u}_k \in \underset{\mathbf{v}\in\mathbb{R}^N}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(\min_{t_i \in [-1;1]} \|\mathbf{w}_i - (t_0 + t_i)\mathbf{v}\|^2_{\bar{\boldsymbol{\nu}}} \right) + \chi_S(\mathbf{v}) + \chi_V((t_0-1)\mathbf{v}) + \chi_V((t_0+1)\mathbf{v}), \quad (4.8)$$

where $S = \{\mathbf{v} \in \mathbb{R}^N \text{ s.t. } \langle \mathbf{v}, \mathbf{u}_l \rangle_{\bar{\boldsymbol{\nu}}} = 0, \, l = 1 \cdots k-1\}$,is a convex set that deals with the orthogonality constraint $\mathbf{v} \perp \mathcal{U}^{k-1}$ and V corresponds to the discretization of the constraints contained in $V_{\bar{\boldsymbol{\nu}}}(\Omega)$. From Proposition 2.1 (P3), we have that $\forall v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$, $T := id + v$ is non decreasing and $T(x) \in \Omega$ for all $x \in \Omega$. Hence the discrete convex set V is defined as

$$V = \{\mathbf{v} \in \mathbb{R}^N \text{ s.t. } x_{j+1} + v_{j+1} \geq x_j + v_j, \, j = 1 \cdots N-1 \text{ and } x_j + v_j \in [a;b], \, j = 1 \cdots N\}$$

and can be rewritten as the intersection of two convex sets dealing with each constraint separately.

Proposition 4.2. *One has*

$$\chi_V((t_0 - 1)\mathbf{v}) + \chi_V((t_0 + 1)\mathbf{v}) = \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}),$$

where the convex sets D and E respectively deal with the domain constraints $x_j + (t_0+1)v_j \in [a;b]$ and $x_j + (t_0 - 1)v_j \in [a;b]$, i.e.:

$$D = \{\mathbf{v} \in \mathbb{R}^N, \, s.t. \; m_j \leq v_j \leq M_j\}, \quad (4.9)$$

with $m_j = \max\left(\frac{a-x_j}{t_0+1}, \frac{b-x_j}{t_0-1}\right)$ and $M_j = \min\left(\frac{a-x_j}{t_0-1}, \frac{b-x_j}{t_0+1}\right)$, and the non decreasing constraint of $id + (t_0 \pm 1)\mathbf{v}$:

$$E = \{\mathbf{z} \in \mathbb{R}^N \, s.t. \; -1/(t_0+1) \leq z_j \leq 1/(1-t_0)\}. \quad (4.10)$$

with the differential operator $K : \mathbb{R}^N \to \mathbb{R}^N$ computing the discrete derivative of $\mathbf{v} \in \mathbb{R}^n$ as

$$(K\mathbf{v})_j = \begin{cases} (v_{j+1} - v_j)/(x_{j+1} - x_j) & \text{if } 1 \leq j < N \\ 0 & \text{if } j = N, \end{cases} \quad (4.11)$$

Having D and E both depending on t_0 is not an issue since problem (4.8) is solved for fixed t_0.

Introducing $\mathbf{t} = \{t_i\}_{i=1}^n \in \mathbb{R}^n$, problem (4.8) can be reformulated as:

$$\min_{\mathbf{v}\in\mathbb{R}^N} \min_{\mathbf{t}\in\mathbb{R}^n} J(\mathbf{v}, \mathbf{t}) := \underbrace{\sum_{i=1}^{n} \|\mathbf{w}_i - (t_0 + t_i)\mathbf{v}\|^2_{\bar{\boldsymbol{\nu}}}}_{F(\mathbf{v},\mathbf{t})} + \underbrace{\chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) + \chi_{B_1^n}(\mathbf{t})}_{G(\mathbf{v},\mathbf{t})}. \quad (4.12)$$

where B_1^n is the L^∞ ball of $\mathbb{R}^n$ with radius 1 dealing with the constraint $t_i \in [-1;1]$. Notice that F is differentiable but non-convex in $(\mathbf{v}, \mathbf{t})$ and G is non-smooth and convex.

Geodesic surface approach For fixed $(t_0^1, \dots, t_0^K) \in \mathbb{R}^K$ and $\mathbf{ff}^{\pm} = \{\alpha_k^+, \alpha_k^-\}_{k=1}^K$, the discretized version of (4.7) is then

$$\min_{\mathbf{v_1},\dots,\mathbf{v_K}\in\mathbb{R}^N} \min_{\mathbf{ff_1^{\pm}},\dots,\mathbf{ff_n^{\pm}}\in\mathbb{R}^{2K}} F'(\mathbf{v},\mathbf{t}) + G'(\mathbf{v},\mathbf{t}), \tag{4.13}$$

where $F'(\mathbf{v},\mathbf{t}) = \sum_{i=1}^n \|\mathbf{w}_i - \sum_{k=1}^K (\alpha_{ik}^+(t_0^k+1) + \alpha_{ik}^-(t_0^k-1))\mathbf{v_k}\|_{\boldsymbol{\nu}}^2$ is still non-convex and differentiable, $G'(\mathbf{v},\mathbf{t}) = \sum_{k=1}^K \left(\chi_E(K\mathbf{v_k}) + \chi_{D_k}(\mathbf{v_k})\right) + \sum_{i=1}^n \chi_A(\mathbf{ff_i^{\pm}})^2$ is convex and non smooth, A is the simplex of $\mathbb{R}^{2K}$ and D_k is defined as in (4.9), depending on t_0^k. We recall that the orthogonality between vectors $\mathbf{v_k}$ is not taken into account in the geodesic surface approach.

4.3.2 Optimization through the Forward-Backward Algorithm

Following [ABS13], in order to compute a critical point of problem (4.12), one can consider the Forward-Backward algorithm (see also [OCBP14] for an acceleration using inertial terms). Denoting as $X = (\mathbf{v},\mathbf{t}) \in \mathbb{R}^{N+n}$, taking $\tau > 0$ and $X^{(0)} \in \mathbb{R}^{N+n}$, it reads:

$$X^{(\ell+1)} = \mathrm{Prox}_{\tau G}(X^{(\ell)} - \tau \nabla F(X^{(\ell)})), \tag{4.14}$$

where $\mathrm{Prox}_{\tau G}(\tilde{X}) = \mathrm{argmin}_X \frac{1}{2\tau}||X - \tilde{X}||^2 + G(X)$ with the Euclidean norm $||\cdot||$. In order to guarantee the convergence of this algorithm, the gradient of F has to be Lipschitz continuous with parameter $M > 0$ and the time step should be taken as $\tau < 1/M$. The details of computation of ∇F and $\mathrm{Prox}_{\tau G}$ for the two algorithms are given in Appendix A.

5 Statistical comparison between log-PCA and GPCA on synthetic and real data

5.1 Synthetic example - Iterative versus geodesic surface approaches

First, for the synthetic example displayed in Figure 1, we compare the two algorithms (iterative and geodesic surface approaches) described in Section 4. The results are reported in Figure 4 by comparing the projection of the data onto the first and second geodesics computed with each approach. We also also display the projection of the data onto the two-dimensional surface generated by each method. It should be recalled that the principal surface for the iterative geodesic algorithm is not necessarily a geodesic surface but each $g_t(t_0^k, u_k)_{t\in[-1;1]}$ defined by (4.4) for $k = 1, 2$ is a geodesic curve for $\mathcal{U} = \{u_1, u_2\}$. For data generated from a location-scale family of Gaussian distributions, it appears that each algorithm provides a satisfactory reconstruction of the data set. The main divergence concerns the first and second principal geodesic. Indeed enforcing the orthogonality between components in the iterative approach enables to clearly separate the modes of variation in location and scaling, whereas searching directly a geodesic surface in the second algorithm implies a mixing of these two types of variation.

Note that the barycenter of Gaussian distributions $\mathcal{N}(m_i, \sigma_i^2)$ can be shown to be Gaussian with mean $\sum m_i$ and variance $(\sum \sigma_i)^2$.

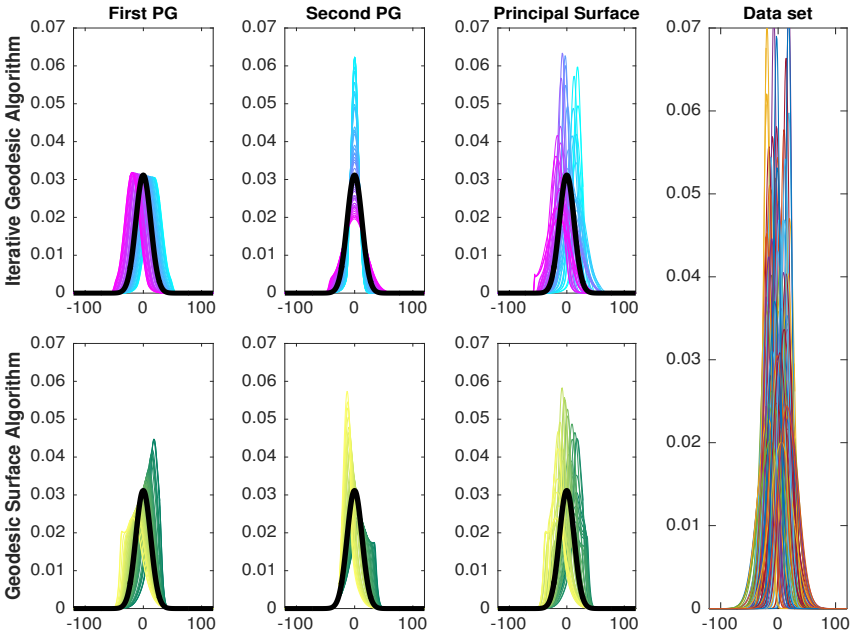

Figure 4: Synthetic example - Data sampled from a location-scale family of Gaussian distributions. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $W_2(\Omega)$.

5.2 Population pyramids

As a first real example, we consider a real dataset whose elements are histograms representing the population pyramids of $n = 217$ countries for the year 2000 (this dataset is produced by the International Programs Center, US Census Bureau (IPC, 2000), available at `http://www.census.gov/ipc/www/idb/region.php`). Each histogram in the database represents the relative frequency by age, of people living in a given country. Each bin in a histogram is an interval of one year, and the last interval corresponds to people older than 85 years. The histograms are normalized so that their area is equal to one, and thus they represent a set of pdf. In Figure 5, we display the population pyramids of 4 countries, and the whole dataset. Along the interval $\Omega = [0, 84]$, the variability in this dataset can be considered as being small.

For $K = 2$, log-PCA and the iterative GPCA algorithm lead to the same principal orthogonal directions in $L^2_{\boldsymbol{\nu}}(\Omega)$, namely that $\tilde{u}_1 = u^*_1$ and $\tilde{u}_2 = u^*_2$ where $(\tilde{u}_1, \tilde{u}_2)$ minimizes (3.1) and (u^*_1, u^*_2) are minimizers of (4.2). In this case, all projections of data $\omega_i = \log_{\boldsymbol{\nu}}(\nu_i)$ for $i = 1, \ldots, n$ onto $\mathrm{Sp}(\{\tilde{u}_1, \tilde{u}_2\})$ lie in $V_{\boldsymbol{\nu}}(\Omega)$, which means that log-PCA and the iterative geodesic algorithm lead exactly the same principal geodesics. Therefore, population pyramids is an example of data that are sufficiently concentrated around their Wasserstein barycenter so that log-PCA and GPCA are equivalent approaches (see Remark 3.5 in [BGKL15] for further details). Hence, we only display in Figure 6 the results of the iterative and geodesic surface algorithms.

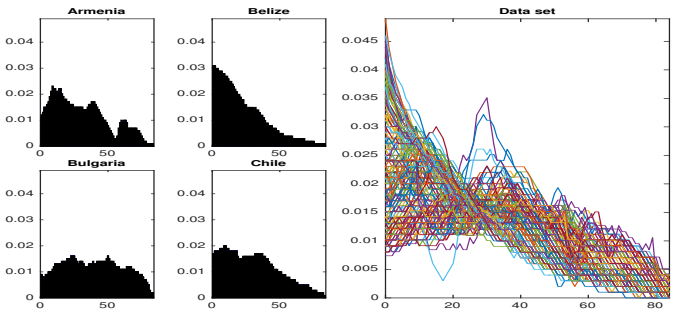

Figure 5: Population pyramids. A subset of population pyramids for 4 countries (left) for the year 2000, and the whole dataset of $n = 217$ population pyramids (right) displayed as pdf over the interval $[0, 84]$.

In the iterative case, the projection onto the first geodesic exhibits the difference between less developed countries (where the population is mostly young) and more developed countries (with an older population structure). The second geodesic captures more subtle divergences concentrated on the middle age population. It can be observed that the geodesic surface algorithm gives different results since the orthogonality constraint on the two principal geodesics is not required. In particular, the principal surface mainly exhibit differences between countries with a young population with countries having an older population structure, but the difference between its first and second principal geodesic is less contrasted.

5.3 Children's first name at birth

In a second example, we consider a dataset of histograms which represent, for a list of $n = 1060$ first names, the distribution of children born with that name per year in France between years 1900 and 2013. In Figure 7, we display the histograms of four different names, as well as the whole dataset. Along the interval $\Omega = [1900, 2013]$, the variability in this dataset is much larger than the one observed for population pyramids. This dataset has been provided by the INSEE (French Institute of Statistics and Economic Studies).

This is an example of real data where log-PCA and GPCA are not equivalent procedures for $K = 2$ principal components. We recall that log-PCA leads to the computation of principal orthogonal directions $\tilde{u}_1, \tilde{u}_2$ in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ minimizing (3.1). First observe that in the left column of Figure 8, for some data $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, the mappings $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ are decreasing, and their range is larger than the interval Ω (that is, for some $x \in \Omega$, one has that $\tilde{T}_i(x) \notin \Omega$). Hence, such $\tilde{T}_i$ are not optimal mappings. Therefore, the condition $\Pi_{\mathrm{Sp}(\tilde{U})}\omega_i \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$ for all $1 \leq i \leq n$ (with $\tilde{U} = \{\tilde{u}_1, \tilde{u}_2\}$) is not satisfied, implying that log-PCA does not lead to a solution of GPCA thanks to Proposition 3.5 in [BGKL15].

Hence, for log-PCA, the corresponding histograms displayed in the right column of Figure 8 are such that $\Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i \notin V_{\bar{\boldsymbol{\nu}}}(\Omega)$. This implies that the densities of the projected measures $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ have a support outside $\Omega = [1900, 2013]$. Hence, the estimation of the measure

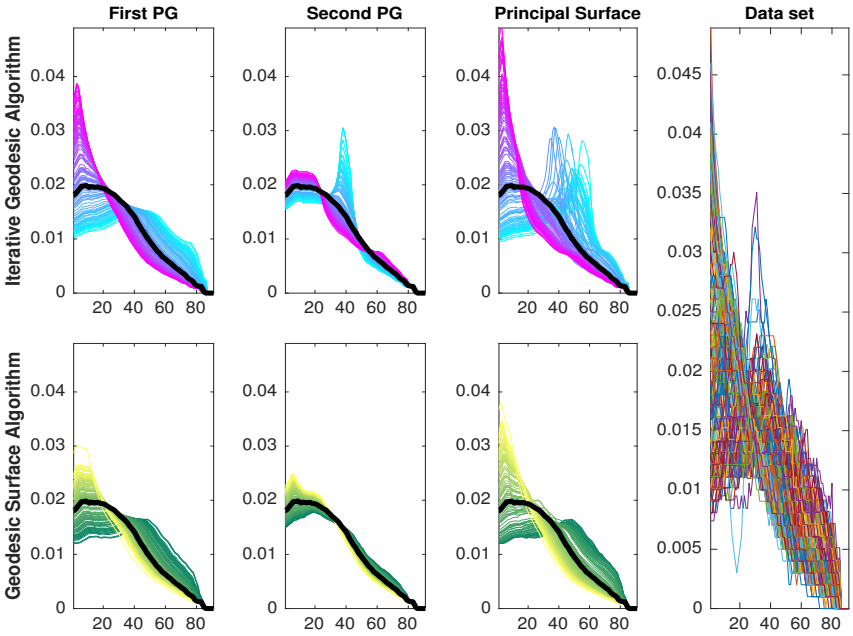

Figure 6: Population pyramids. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The first (resp. second) column is the projection of the data into the first (resp. second) principal direction. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $W_2(\Omega)$.

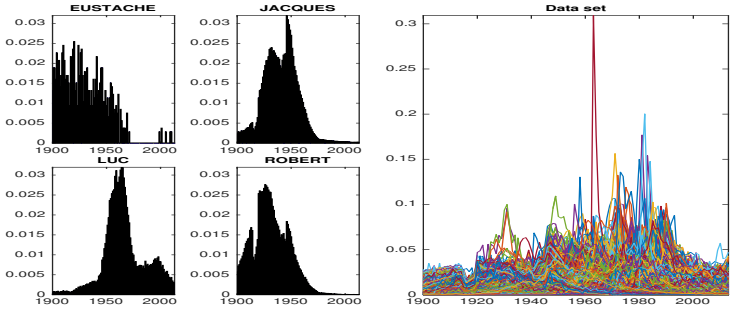

Figure 7: Children's first name at birth. An subet of 4 histograms representing the distribution of children born with that name per year in France, and the whole dataset of $n = 1060$ histograms (right), displayed as pdf over the interval $[1900, 2013]$

$\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ by its projection onto the first mode of variation obtained with log-PCA is not satisfactory.

In Figure 8, we also display the results given by the iterative geodesic algorithm, leading to orthogonal directions u_1^*, u_2^* in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ that are minimizers of (4.2). Contrary to the results obtained with log-PCA, one observes in Figure 8 that all the mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{u_1^*\})}\omega_i$ are non-decreasing, and such that $T_i(x) \in \Omega$ for all $x \in \Omega$. Nevertheless, by enforcing these

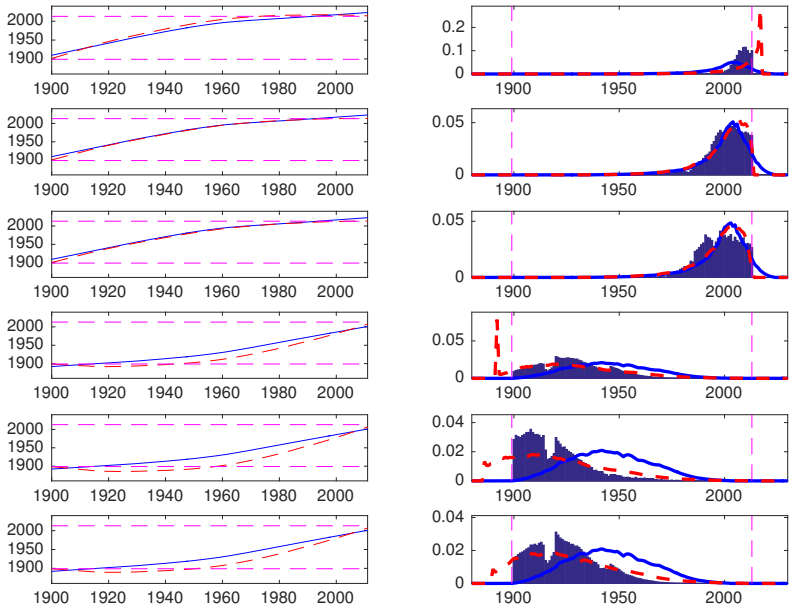

Figure 8: Children's first name at birth with support $\Omega = [1900, 2013]$. (Left) The dashed red curves represent the mapping $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ where $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, and $\tilde{u}_1$ is the first principal direction in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via log-PCA. The blue curves are the mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{u_1^*\})}\omega_i$, where u_1^* is the first principal direction in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via the iterative algorithm. (Right) The histogram stands for the pdf of measures ν_i that have a large Wasserstein distance with respect to the barycenter $\bar{\boldsymbol{\nu}}$. The red curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ with log-PCA, while the blue curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ with GPCA.

two conditions, one has that a good estimation of the measure $\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ by its projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ is made difficult as most of the mass of ν_i is located at either the right or left side of the interval Ω which is not the case for its projection. The histograms displayed in the right clumn of Figure 8 correspond to the elements in the dataset that have a large Wasserstein distance with respect to the barycenter $\bar{\boldsymbol{\nu}}$. This explains why it is difficult to have good projected measures with GPCA. For elements in the dataset that are closest to $\bar{\boldsymbol{\nu}}$, the projected measures $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ and $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ are much closer to ν_i and for such elements, log-PCA and the iterative geodesic algorithm lead to similar results in terms of data projection.

To better estimate the extremal data in Figure 8, a solution is to increase the support of the data to the interval $\Omega_0 = [1850, 2050]$, and to perform log-PCA and GPCA in the Wasserstein space $W_2(\Omega_0)$. The results are reported in Figure 9. In that case, it can be observed that both algorithms lead to similar results, and that a better projection is obtained for the extremal data. Notice that with this extended support, all the mappings $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ obtained with log-PCA are optimal in the sense that they are non-decreasing with a range inside Ω_0.

Finally, we display in Figure 10 and Figure 11 the results of the iterative and geodesic surface algorithms with either $\Omega = [1900, 2013]$ or with data supported on the extended support $\Omega_0 = [1850, 2050]$. The projection of the data onto the first principal geodesic suggests that

the distribution of a name is deeply dependent on the part of the century. The second geodesic express a popular trend through a spike effect. In Figure 10, the artefacts in the principal surface that are obtained with the iterative algorithm at the end of the century, correspond to the fact that the projection of the data ω_i onto the surface spanned by the first two components is not ensured to belong to the set $V_{\bar{\boldsymbol{\nu}}}(\Omega)$.

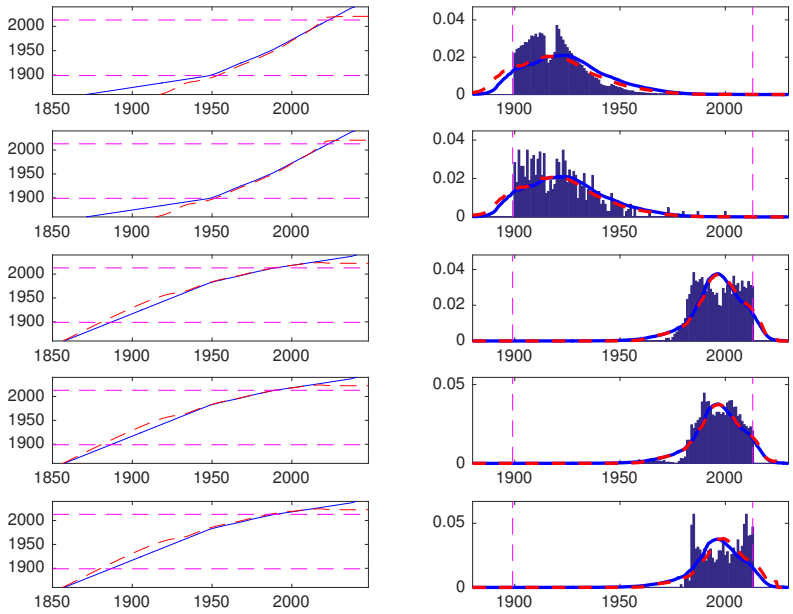

Figure 9: Children's first name at birth with extended support $\Omega_0 = [1850, 2050]$. (Left) The dashed red curves represent the mapping $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ where $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, and $\tilde{u}_1$ is the first principal direction in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via log-PCA. The blue curves are the mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{u_1^*\})}\omega_i$, where u_1^* is the first principal direction in $L^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via the iterative algorithm. (Right) The histogram stands for the pdf of measures ν_i that have a large Wasserstein distance with respect to the barycenter $\bar{\boldsymbol{\nu}}$. The red curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ with log-PCA, while the blue curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ with GPCA.

6 Extensions beyond $d > 1$ and some perspectives.

We now briefly show that our iterative algorithm for finding principal geodesics can be adapted to the general case $d > 1$. This requires to take into account two differences with the one-dimensional case. First, the definition of the space $V_{\mu_r}(\Omega)$ in 2.1 relies on the explicit close-form formulae (2.2) for the computation of a solution to the optimal transport problem which is specific to the one-dimensional case. We must hence provide a more general definition of $V_{\mu_r}(\Omega)$. Second, the isometry property $(P1)$ does not hold for $d > 1$, so that Wasserstein distances cannot

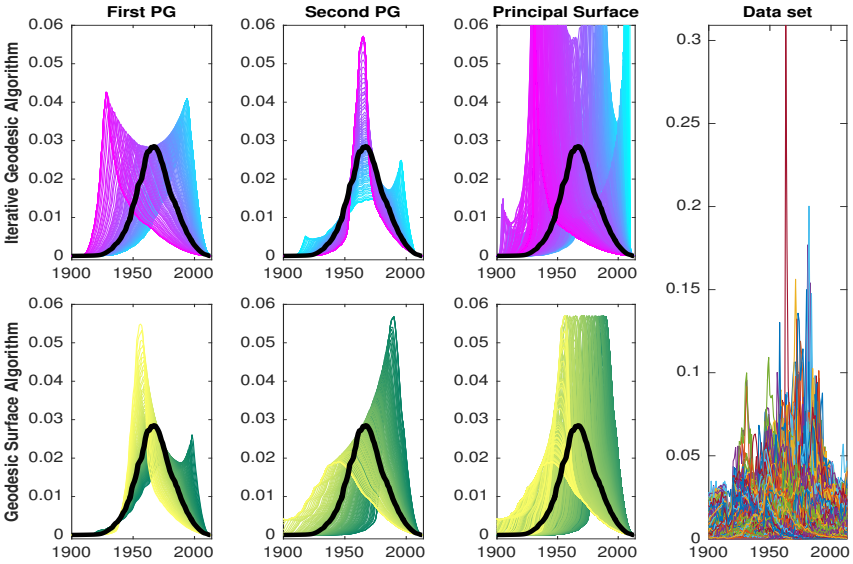

Figure 10: Children's first name at birth with support $\Omega = [1900, 2013]$. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The first (resp. second) column is the projection of the data into the first (resp. second) principal direction. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $W_2(\Omega)$.

be replaced by the $L^2_{\tilde{\nu}}$ norm between log maps as in (4.5) and must be explicitly computed and differentiated.

Definition of $V_{\mu_r}(\Omega)$ in the general case In the one dimensional case, $V_{\mu_r}(\Omega)$ is characterized in Proposition 2.1 (P3) as the set of functions $v \in L^2_{\mu_r}(\Omega)$ such that $T := \mathrm{id} + v$ is μ_r-almost everywhere non decreasing. A striking result by Brenier [Bre91] is that, in any dimension, if μ_r does not give mass to small set, there exists an optimal mapping $T \in L^2_{\mu_r}(\Omega)$ between μ_r and any probability measure ν, and T is equal to the gradient of a convex function u ie. $T = \nabla u$. Hence, we define the set $V_{\mu_r}(\Omega)$ as the set of functions $v \in L^2_{\mu_r}(\Omega)$ such that $id + v = \nabla u$ for an arbitrary convex function u.

In order to deal with the latter constraint, we note that this space of functions is equal to the space of functions $v \in L^2_{\mu_r}(\Omega)$ such that $div(v) \geq -1$. Indeed, assuming that $x + v = \nabla u$, then u being a convex potential involves $div(\nabla u) \geq 0$ which is equivalent to $div(x+v) = div(v)+1 \geq 0$.

General objective function Without the isometry property (P1), the objective function $H(t_0, v)$ in (4.3) must be written with the explicit Wasserstein distance d_W,

$$H(t_0, v) = \frac{1}{n} \sum_{i=1}^{n} \min_{t_i \in [-1;1]} d^2_W(\nu_i, g_{t_i}(t_0, v)) + \chi_{V_{\tilde{\nu}}(\Omega)}((t_0 - 1)v) + \chi_{V_{\tilde{\nu}}(\Omega)}((t_0 + 1)v), \qquad (6.1)$$

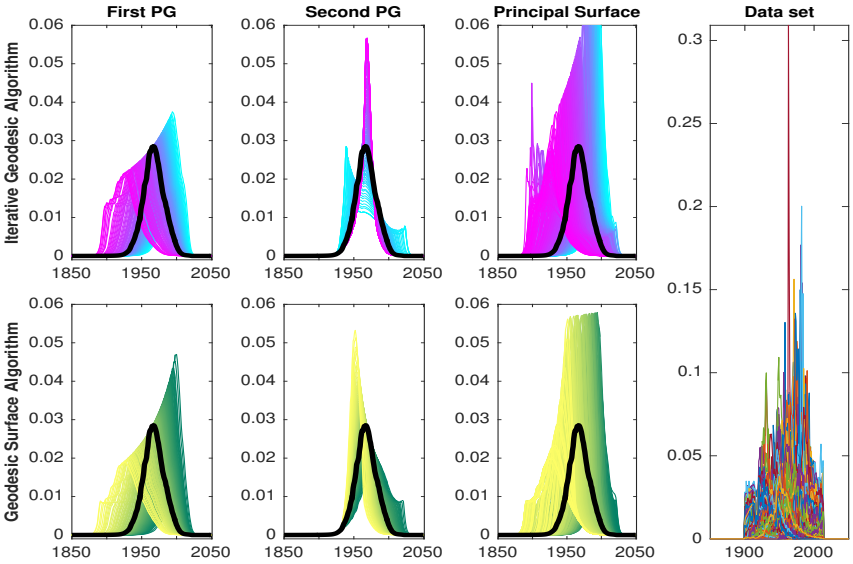

Figure 11: Children's first name at birth with extended support $\Omega_0 = [1850, 2050]$. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The first (resp. second) column is the projection of the data into the first (resp. second) principal direction. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $W_2(\Omega)$.

where $g_t(t_0, v) = (\mathrm{id} + (t_0 + t)v)\#\bar{\boldsymbol{\nu}}$ for $t \in [-1; 1]$ as defined in (4.4). Optimizing over both the functions $\mathbf{v} \in (\mathbb{R}^d)^N$ and the projection times $\mathbf{t}$, the discretized objective function to minimize is,

$$\min_{\mathbf{v}\in\mathbb{R}^N} \min_{\mathbf{t}\in\mathbb{R}^n} J(\mathbf{v}, \mathbf{t}) := \underbrace{\sum_{i=1}^{n} d_W^2(\nu_i, g_{t_i}(t_0, \mathbf{v}))}_{F(\mathbf{v},\mathbf{t})} + \underbrace{\chi_S(v) + \chi_E(K\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_{B_1^n}(\mathbf{t})}_{G(\mathbf{v},\mathbf{t})}. \tag{6.2}$$

where K is a discretized divergence operator, and $E = \{\mathbf{z} \in \mathbb{R}^N : \frac{-1}{t_0+1} \leqslant \mathbf{z} \leqslant \frac{1}{1-t_0}\}$, $D = \{\mathbf{v} : \mathrm{id}+(t_0\pm 1)\mathbf{v} \in \Omega\}$ deals with the domain constraint and S deals with the orthogonality constraint w.r.t. to the preceding principal components. As for the one-dimensional case, we minimize J through the Forward-Backward algorithmas detailed in the appendix B.

Extension to higher dimensions is straightforward. However, considering that we have to discretize the support of the Wasserstein mean $\bar{\boldsymbol{\nu}}$, the approach becomes intractable for $d > 3$.

6.1 Application to grayscale images

We consider the MNIST dataset [LeC98] which contains grayscale images of handwritten digits. All the images have identical size 28×28 pixels. Each grayscale image, once normalized so that the sum of pixel grayscape values sum to one, can be interpreted as a discrete probability

measure, which is supported on the 2D grid of size 28×28. The ground metric for the Wasserstein distance is then the 2D square Euclidean distance between the locations of the pixels of the two-dimensional grid. We compute the first principal components on 1000 images of each digit. Wasserstein barycenters, which are required as input to our algorithm, are approximated efficiently through iterative bregman projections as proposed in [BCC$^+$15]. We use the network simplex algorithm[1] to compute Wasserstein distances.

Figure 12 displays the results obtained with our Forward-Backward algorithm (with t_0 set to 0 for simplicity), and the ones given by Log-PCA as described in section 3. These two figures are obtained by sampling the first principal components. We then use kernel smoothing to display the discrete probability measures back to the original grid and present the resulting grayscale image with an appropriate colormap.

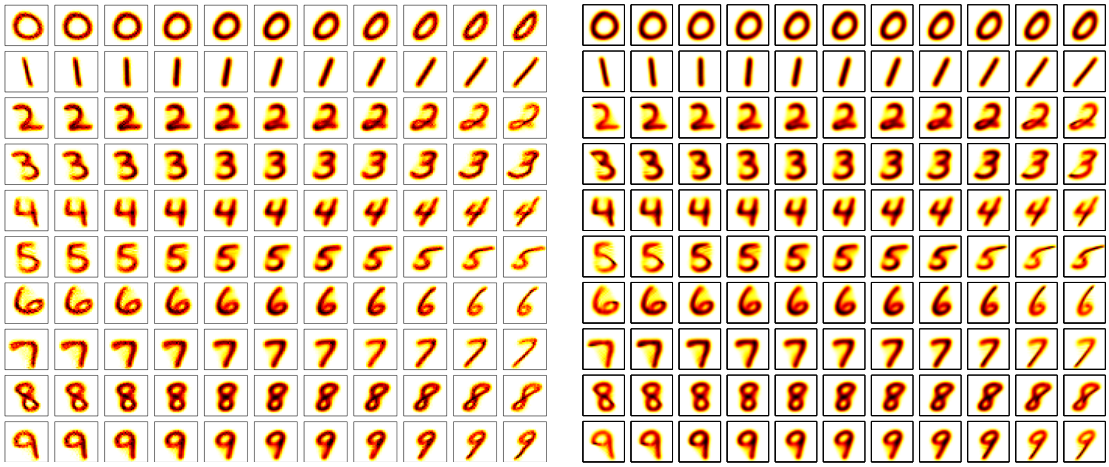

Figure 12: First principal geodesics for 1000 images of each digit from the MNIST dataset, computed through the proposed Forward-Backward algorithm (left) and log-PCA (right).

Visually, both the Log-PCA and GPCA approaches capture well the main source of variability of each set of grayscape images ie each number. We observe variations in the slant of the handwritten digits for all digits, the most obvious case being digit '1'. As a principal component is parameterized by a whole velocity field on the support of the Wasserstein mean of the data, single principal components can capture more interesting patterns, such as changes in the shape of the '0' or the presence or absence of the lower loop of the '2'. From purely visual inspection, it is difficult to tell which approach, Log-PCA or GPCA, provides a "better" principal component. For this purpose we compute the reconstruction error of each digit. This reconstruction error is computed in the same way for both Log-PCA and GPCA principal components: We sample the principal components at many times t and find for each image in a given dataset, the time

[1] `http://liris.cnrs.fr/~nbonneel/FastTransport/`

at which the geodesic is the closest to the image sample. This provides an approximation of $\min_{t\in[-1;1]} d_W^2(\nu_i, g_t(v))$ for each image $i = 1, \dots, n$, where $(g_t)_{t\in[-1,1]}$ is the principal component. For the Log-PCA principal component, we take $\tilde{g}_t = (\mathrm{id} + t1.25\lambda v)\#\bar{\boldsymbol{\nu}}$, where λ is the eigenvalue corresponding to the first principal component. The 1.25 factor is useful to consider a principal curve which goes through the whole range of the dataset. For the GPCA principal geodesic, we have $g_t^* = (\mathrm{id} + tv)\#\bar{\boldsymbol{\nu}}$. The reconstruction errors are shown in Table 1. We see that, for each digit, we obtain a better, i.e. smaller, reconstruction error when using the proposed Forward-Backward algorithm. This result is not surprising, since the reconstruction error is explicitly minimized through the Forward-Backward algorithm. As previously mentioned, Log-PCA rather computes linearized Wasserstein distances. In one-dimension, the isometry property (P1) states that these quantities are equal. In dimension two or larger, that property does not hold.

MNIST digit	Log-PCA RE $(\cdot 10^3)$	GPCA RE $(\cdot 10^3)$
0	2.0355	**1.9414**
1	3.1426	**1.0289**
2	3.4221	**3.3575**
3	2.6528	**2.5869**
4	2.8792	**2.8204**
5	2.9391	**2.9076**
6	2.1311	**1.9864**
7	4.7471	**2.8205**
8	2.0741	**2.0222**
9	1.9303	**1.8728**

Table 1: Reconstruction Errors (RE) computed on 1000 sample images of each digit of the MNIST dataset. (center) Reconstruction error w.r.t. the first principal component computed with the Log-PCA algorithm. (right) Reconstruction error w.r.t. the first principal geodesic computed with the proposed Forward-Backward algorithm.

6.2 Discussion

The proposed Forward-Backward algorithm minimizes the same objective function as defined in [SC15]. The first difference with the algorithm provided [SC15] is that we take gradient steps with respect to both $\mathbf{v}$ and $\mathbf{t}$, while the latter first attempts to find the optimal t (by sampling the geodesics at many time t), before taking a gradient step of $\mathbf{v}$. Our approach reduces the cost of computing a gradient step by one order of magnitude. Secondly, [SC15] relied on barycentric projections to preserve the geodesicity of the principal curves in between gradient steps. That heuristic does not guarantee a decrease in the objective after a gradient step. Morever, the method in [SC15] considered two velocity fields $\mathbf{v}_1, \mathbf{v}_2$ rather than a single $\mathbf{v}$ since the optimality of both $\mathbf{v}$ and $-\mathbf{v}$ could not be preserved through the barycentric projection.

When considering probability measures over high dimensional space ($d > 3$), our algorithm becomes intractable since we need to discretize the support of the Wasserstein mean of the data with a regular grid, while the approach of [SC15] is still tractable since an arbitrary support for the Wasserstein mean is used. A remaining challenge for computing principal geodesics in the

Wasserstein space is then to propose an algorithm for GPCA which is still tractable in higher dimensions while not relying on barycentric projections.

A Dimension $d = 1$

We here detail the application of Algorithm (4.14) to the iterative GPCA procedure that consists in solving the problem (4.12):

$$\min_{\mathbf{v}\in\mathbb{R}^N} \min_{\mathbf{t}\in\mathbb{R}^n} J(\mathbf{v},\mathbf{t}) := \underbrace{\sum_{i=1}^{n} \|\mathbf{w}_i - (t_0+t_i)\mathbf{v}\|_{\boldsymbol{\nu}}^2}_{F(\mathbf{v},\mathbf{t})} + \underbrace{\chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) + \chi_{B_1^n}(\mathbf{t})}_{G(\mathbf{v},\mathbf{t})}.$$

A.1 Lispschitz constant of ∇F

Let us now look at the Lipschitz constant of $\nabla F(\mathbf{v},\mathbf{t})$ on the restricted acceptable set $D \times B_1^n$. We first denote as $\mathcal{H}$ the hessian matrix (of size $(N+n)\times(N+n)$) of the $\mathcal{C}^2$ function $F(X)$. We know that if the spectral radius of $\mathcal{H}$ is bounded by a scalar value M, i.e. $\rho(\mathcal{H}) \leq M$, then ∇F is a Lipschitz continuous function with constant M. Hence, we look at the eigenvalues of the Hessian matrix of $F = \sum_{i=1}^{n}\sum_{j=1}^{N} \bar{\boldsymbol{f}}_n(x_j)(w_i^j - (t_0+t_i)v_j)^2$ that is

$$\frac{\partial^2 F}{\partial t_i^2} = \sum_{j=1}^{N} 2v_j^2 \bar{\boldsymbol{f}}_n(x_j), \quad \frac{\partial^2 F}{\partial v_j^2} = \sum_{i=1}^{n} 2(t_0+t_i)^2 \bar{\boldsymbol{f}}_n(x_j), \quad \frac{\partial^2 F}{\partial t_i \partial v_j} = 2\bar{\boldsymbol{f}}_n(x_j)(2(t_0+t_i)v_j - w_i^j)$$

and $\frac{\partial^2 F}{\partial t_i \partial t_{i'}} = \frac{\partial^2 F}{\partial v_j \partial v_{j'}} = 0$, for all $i \neq i'$ or $j \neq j'$. Being $\{\mu_k\}_{k=1}^{n+N}$ the eigenvalues of $\mathcal{H}$, we have $\rho(\mathcal{H}) = \max_k |\mu_k| \leq \max_k \sum_l |\mathcal{H}_{kl}|$. We denote as $f_\infty = \max_j |\bar{\boldsymbol{f}}_n(x_j)|$ and likewise $w_\infty = \max_{i,j} |w_i^j|$. Since $|t_0| < 1$, $t_i^2 \leq 1$, $\forall \mathbf{t} \in B_1^n$ and $v_j^2 \leq \alpha^2 = (b-a)^2$, $\forall \mathbf{v} \in D$, by defining $\gamma = 2(1+|t_0|)\alpha + w_\infty$, we thus have

$$\rho(\mathcal{H}) \leq 2f_\infty \max\left\{n\alpha^2 + N\gamma, n\gamma + N(1+|t_0|)^2\right\} := M. \tag{A.1}$$

A.2 Computing $\mathbf{Prox}_{\tau G}$

In order to implement the algorithm (4.14), we finally need to compute the proximity operator of G defined as:

$$(\mathbf{v}^*,\mathbf{t}^*) = \mathrm{Prox}_{\tau G}(\tilde{\mathbf{v}},\tilde{\mathbf{t}}) = \operatorname*{argmin}_{\mathbf{v},\mathbf{t}} \frac{1}{2\tau}(||\mathbf{v}-\tilde{\mathbf{v}}||^2 + ||\mathbf{t}-\tilde{\mathbf{t}}||^2) + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) + \chi_{B_1^n}(\mathbf{t}).$$

This problem can be solved independently on $\mathbf{v}$ and $\mathbf{t}$. For $\mathbf{t}$, it can be done pointwise as $t_i^* = \mathrm{argmin}_{t_i} \frac{1}{2\tau}||t_i - \tilde{t}_i||^2 + \chi_{B_1^1}(t_i) = \mathrm{Proj}_{[-1;1]}(\tilde{t}_i)$. Unfortunately, there is no closed form expression of the proximity operator for the component $\mathbf{v}$. It requires to solve the following intern optimization problem at each extern iteration (ℓ) of the algorithm (4.14):

$$\mathbf{v}^* = \operatorname*{argmin}_{\mathbf{v}} \frac{1}{2\tau}||\mathbf{v}-\tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}), \tag{A.2}$$

where, to avoid confusions, we denote by $\mathbf{v}$ the variable that is optimized within the intern optimization problem (A.2).

Remark A.1. The Lipschitz constant of $\nabla F(\mathbf{v}, \mathbf{t})$ in (A.1) relies independantly on $\mathbf{v}$ and $|t_0|$, thus we can choose the optimal gradient descent step τ for $\mathbf{v}^*$ and $\mathbf{t}^*$.

Primal-Dual reformulation Using duality (through Fenchel transform), one has:

$$\begin{aligned} &\min_{\mathbf{v}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v}-\tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) \\ =& \min_{\mathbf{v}\in\mathbb{R}^N} \max_{\mathbf{z}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v}-\tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \langle K\mathbf{v}, \mathbf{z}\rangle - \chi_E^*(\mathbf{z}), \end{aligned} \tag{A.3}$$

where $\mathbf{z} = \{\mathrm{z}_j\}_{j=1}^N \in \mathbb{R}^N$ is a dual variable and $\chi_E^* = \sup_{\mathbf{v}}\langle \mathbf{v}, \mathbf{z}\rangle - \chi_E(\mathbf{v})$ is the convex conjugate of χ_E that reads:

$$(\chi_E^*(\mathbf{z}))_j = \begin{cases} -\mathrm{z}_j/(1+t_0) & \text{if } \mathrm{z}_j \leq 0, \\ \mathrm{z}_j/(1-t_0) & \text{if } \mathrm{z}_j > 0. \end{cases}$$

Hence, one can use the Primal-Dual algorithm proposed in [CP14] to solve the problem (A.3). For two parameters $\sigma, \theta > 0$ such that $||K||^2 \leq \frac{1}{\sigma}(\frac{1}{\theta} - \frac{1}{\tau})$ and given $\mathbf{v}^0, \bar{\mathbf{v}}^0, \mathbf{z}^0 \in \mathbb{R}^N$, the algorithm is:

$$\begin{cases} \mathbf{z}^{(m+1)} &= \mathrm{Prox}_{\sigma\chi_E^*}(\mathbf{z}^{(m)} + \sigma K\bar{\mathbf{v}}^{(m)}) \\ \mathbf{v}^{(m+1)} &= \mathrm{Prox}_{\theta(\chi_D+\chi_S)}(\mathbf{v}^{(m)} - \theta(K^*\mathbf{z}^{(m+1)} + \frac{1}{\tau}(\mathbf{v}^{(m)} - \tilde{\mathbf{v}})) \\ \bar{\mathbf{v}}^{(m+1)} &= 2\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)} \end{cases} \tag{A.4}$$

where K^* is defined as $\langle K\mathbf{v}, \mathbf{z}\rangle = \langle \mathbf{v}, K^*\mathbf{z}\rangle$. Using the operator K defined in (4.11), we thus have:

$$(K^*\mathbf{z})_j = \begin{cases} -\mathrm{z}_1/\Delta_1 & \text{if } j = 1 \\ \mathrm{z}_{j-1}/\Delta_{j-1} - \mathrm{z}_j/\Delta_j & \text{if } 1 < j < N. \\ \mathrm{z}_{N-1}/\Delta_{N-1} & \text{if } j = N, \end{cases} \tag{A.5}$$

where $\Delta_j = x_{j+1} - x_j$. We have that $||K||^2 = \rho(K^*K)$, the largest eigenvalue of K^*K. With the discrete operators (4.11) and (A.5), $\rho(K^*K)$ can be bounded by

$$\delta^2 = 2\max_j(1/\Delta_j^2 + 1/\Delta_{j+1}^2). \tag{A.6}$$

One can therefore for instance take $\sigma = \frac{1}{\delta}$ and $\theta = \tau/(1+\delta\tau)$.

Proximity operators in (A.4) The proximity operator of $\chi_D + \chi_S$ is obtained as:

$$(\mathrm{Prox}_{\theta(\chi_D+\chi_S)}(\mathbf{v}))_j = (\mathrm{Proj}_{D\cap S}(\mathbf{v}))_j = \mathrm{Proj}_{[m_j;M_j]}\left(\left(\mathbf{v} - \sum_{l=1}^{k-1}\frac{\langle \mathbf{u}_l, \mathbf{v}\rangle_{\bar{\boldsymbol{\nu}}}}{||\mathbf{u}_l||_{\bar{\boldsymbol{\nu}}}^2}\mathbf{u}_l\right)_j\right), \tag{A.7}$$

since projecting onto $D \cap S$ is equivalent to first project onto the orthogonal of $\mathrm{Sp}(\mathcal{U}^{k-1})$ and then onto D. One can finally show that the proximity operator of χ_E^* can be computed pointwise as:

$$(\mathrm{Prox}_{\sigma\chi_E^*}(\mathbf{z}))_j = \begin{cases} \mathrm{z}_j - \sigma/(1-t_0) & \text{if } \mathrm{z}_j > \sigma/(1-t_0) \\ \mathrm{z}_j + \sigma/(1+t_0) & \text{if } \mathrm{z}_j < -\sigma/(1+t_0) \\ 0 & \text{otherwise.} \end{cases} \tag{A.8}$$

A.3 Algorithms for GPCA

Gathering all the previous elements, we can finally find a critical point of the non-convex problem (4.12) using the Forward-Backward (FB) framework (4.14), as detailed in Algorithm 1.

Algorithm 1 Resolution with FB of problem (4.12): $\min_{\mathbf{v},\mathbf{t}} F(\mathbf{v},\mathbf{t}) + G(\mathbf{v},\mathbf{t})$

Require: $\boldsymbol{w}_i \in \mathbb{R}^N$ for $i = 1 \cdots n$, $\mathbf{u}_1, \cdots \mathbf{u}_{k-1}$, $t_0 \in]-1;1[$, $\alpha = (b-a) > 0$, $\eta > 0$, $\delta > 0$ (defined in (A.6)) and $M > 0$ (defined in (A.1)).
Set $(\mathbf{v}^{(0)}, \mathbf{t}^{(0)}) \in D \times B_1^n$
Set $\tau < 1/M$, $\sigma = 1/\delta$ and $\theta = \tau/(1+\delta\tau)$.
%Extern loop:
while $||\mathbf{v}^{(\ell)} - \mathbf{v}^{(\ell-1)}||/||\mathbf{v}^{(\ell-1)}|| > \eta$ **do**
% FB on $\mathbf{t}$ with $\mathbf{t}^{(\ell+1)} = Prox_{\tau G}(\mathbf{t}^{(\ell)} - \tau\nabla F(\mathbf{v}^{(\ell)}, \mathbf{t}^{(\ell)}))$:
$t_i{}^{(\ell+1)} = \mathrm{Proj}_{[-1;1]}\left(t_i{}^{(\ell)} - \tau\sum_{j=1}^N v_j{}^{(\ell)} \bar{\boldsymbol{f}}_n(x_j)\left((t_0 + t_i{}^{(\ell)}) v_j{}^{(\ell)} - w_i^j\right)\right)$
% Gradient descent on $\mathbf{v}$ with $\tilde{\mathbf{v}} = \mathbf{v}^{(\ell)} - \tau\nabla F(\mathbf{v}^{(\ell)}, \mathbf{t}^{(\ell)})$:
$\tilde{v}_j = v_j{}^{(\ell)} - \tau\bar{\boldsymbol{f}}_n(x_j)\sum_{i=1}^n (t_0 + t_i{}^{(\ell)})\left((t_0 + t_i{}^{(\ell)})\, v_j{}^{(\ell)} - w_i^j\right)$
%Intern loop for $\mathbf{v}^{(\ell+1)} = Prox_{\tau G}(\tilde{\mathbf{v}})$:
Set $\mathbf{z}^{(0)} \in E$, $\mathbf{v}^{(0)} = \tilde{\mathbf{v}}$, $\bar{\mathbf{v}}^{(0)} = \tilde{\mathbf{v}}$
while $||\mathbf{v}^{(m)} - \mathbf{v}^{(m-1)}||/||\mathbf{v}^{(m-1)}|| > \eta$ **do**
$\mathbf{z}^{(m+1)} = \mathrm{Prox}_{\sigma\chi_E^*}\left(\mathbf{z}^{(m)} + \sigma K\bar{\mathbf{v}}^{(m)}\right)$ (using (A.8))
$\mathbf{v}^{(m+1)} = \mathrm{Prox}_{\theta(\chi_D+\chi_S)}\left(\mathbf{v}^{(m)} - \theta(K^*\mathbf{z}^{(m+1)} + \frac{1}{\tau}(\mathbf{v}^{(m)} - \tilde{\mathbf{v}})\right)$ (using (A.7))
$\bar{\mathbf{v}}^{(m+1)} = 2\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)}$
$m := m + 1$
end while
$\mathbf{v}^{(\ell+1)} = \mathbf{v}^{(m)}$
$\ell := \ell + 1$
end while
return $\mathbf{u}_k = \mathbf{v}^{(\ell)}$

Geodesic surface approach In order to solve the problem (4.13), we follow the same steps as in the section A.1-A.2. First we obtain the Lipchitz constant of the function $\tilde{F}$ by the same tricks as in the iterative algorithm. Then, since the constraints' problem in G' are separable, we can compute each component $\mathbf{v_k}$ and each $\mathbf{ff_i^{\pm}}$ independantly. The only difference with the

iterative algorithm concerns the proximal operator of the function χ_A, which is the projection into the simplex of $\mathbb{R}^{2K}$.

B Dimension $d = 2$

We now show how to generalize the algorithm to the two-dimensional case.

Gradients of F. We write $X = (x_1, \cdots, x_N) \in (\mathbb{R}^2)^N$ the discretized support of $\bar{\boldsymbol{\nu}}$, $Z_t = (x_1 + (t_0 + t)v_1, \cdots, x_N + (t_0 + t)v_N)$ the support $g_t(t_0, \mathbf{v})$, the geodesic sampled at time t. Let P^* be an optimal transport plan between $\bar{\boldsymbol{\nu}}$ and $g_t(t_0, \mathbf{v})$. The function $F(\mathbf{v}, \mathbf{t})$ is differentiable almost everywhere. Gradients can be computed in the same fashion as [SC15] to obtain,

$$\nabla_{\mathbf{v}} F = 2\sum_{i=1}^{n}(t_0+t_i)(Z_{t_i} - XP^{*T}\mathrm{diag}(1/\bar{\boldsymbol{f}}_n)), \quad \nabla_{t_i} F = 2\langle Z_{t_i}\mathrm{diag}(\bar{\boldsymbol{f}}_n), \mathbf{v}\,\rangle - 2\langle P^*, \mathbf{v}^T X\,\rangle, \tag{B.1}$$

Proximal operator of G. The only difference between the one-dimensional case and the two-dimensional case considered here concerns the projection step of $\mathbf{v}$,

$$\mathbf{v}^* = \operatorname*{argmin}_{\mathbf{v}} \frac{1}{2\tau}||\mathbf{v} - \tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}), \tag{B.2}$$

Primal-Dual reformulation As for the on-dimensional case, one has,

$$\begin{aligned} &\min_{\mathbf{v}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v} - \tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) \\ &= \min_{\mathbf{v}\in\mathbb{R}^N} \max_{\mathbf{z}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v} - \tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \langle K\mathbf{v}, \mathbf{z}\rangle - \chi_E^*(\mathbf{z}), \end{aligned} \tag{B.3}$$

where $\mathbf{z} = \{\mathrm{z}_j\}_{j=1}^N \in \mathbb{R}^N$ is a dual variable and $\chi_E^* = \sup_{\mathbf{v}}\langle \mathbf{v}, \mathbf{z}\rangle - \chi_E(\mathbf{v})$ is the convex conjugate of χ_E. This can be solve with the same iterative steps as described in A.2,

$$\left\{ \begin{array}{rcl} \mathbf{z}^{(m+1)} & = & \mathrm{Prox}_{\sigma\chi_E^*}(\mathbf{z}^{(m)} + \sigma K\bar{\mathbf{v}}^{(m)}) \\ \mathbf{v}^{(m+1)} & = & \mathrm{Prox}_{\theta(\chi_D+\chi_S)}(\mathbf{v}^{(m)} - \theta(K^*\mathbf{z}^{(m+1)} + \frac{1}{\tau}(\mathbf{v}^{(m)} - \tilde{\mathbf{v}})) \\ \bar{\mathbf{v}}^{(m+1)} & = & 2\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)} \end{array} \right. \tag{B.4}$$

Here the definition of the divergence operator K and the transpose of the divergence operator K^* are specific to the dimension. For $d = 2$, with a regular grid discretizing Ω in $M \times N$ points, we take

$$K^T\mathbf{z} = -\nabla\mathbf{z} = -\begin{bmatrix}\partial_x^+\mathbf{z} \\ \partial_y^+\mathbf{z}\end{bmatrix},$$

with

$$\partial_x^+\mathbf{z}(i,j) = \left\{ \begin{array}{ll} \mathbf{z}(i+1,j) - \mathbf{z}(i,j) & if\, i < M \\ 0 & otherwise, \end{array} \right.$$

$$\partial_y^+ \mathbf{z}(i,j) = \begin{cases} \mathbf{z}(i,j+1) - \mathbf{z}(i,j) & if\, j < N \\ 0 & otherwise \end{cases}$$

so that

$$K\mathbf{u} = K \begin{bmatrix} \mathbf{u}_x \\ \mathbf{u}_y \end{bmatrix} = \partial_x^- \mathbf{u}_x + \partial_y^- \mathbf{u}_y,$$

with

$$\partial_x^- \mathbf{u}(i,j) = \begin{cases} \mathbf{u}(i,j) - \mathbf{u}(i-1,j) & if\, 1 < i < M \\ \mathbf{u}(i,j) & if\, i = 1 \\ -\mathbf{u}(i-1,j) & if\, i = M. \end{cases}$$

To ensure convergence of B.4, one can take $1/\sigma.(1/\theta - 1/\tau) = ||K||^2$. See [CP15, LP15] for more details. Since we have $||K||^2 = 8$, the parameters can be taken as $\sigma = 1/4$ and $\theta = \tau/(1+2\tau)$.

References

[ABS13] H. Attouch, J. Bolte, and B. Svaiter. Convergence of descent methods for semi-algebraic and tame problems. *Mathematical Programming*, 137(1-2):91–129, 2013.

[AC11] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

[AGS04] L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl.*, 15(3-4):327–343, 2004.

[AGS06] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2006.

[BCC+15] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[BGKL15] J. Bigot, R. Gouet, T. Klein, and A. Lopez. Geodesic PCA in the Wasserstein space by Convex PCA. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, To be published, 2015.

[Bre91] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[CP14] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal-dual algorithm. *Preprint*, 2014.

[CP15] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal,Äìdual algorithm. *Mathematical Programming*, pages 1–35, 2015.

[EG15] Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.

[FLPJ04] P. T. Fletcher, C. Lu, Stephen M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004.

[KP08] Steven G Krantz and Harold R Parks. *Geometric integration theory*. Springer Science & Business Media, 2008.

[LeC98] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[LP15] D. A. Lorenz and T. Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015.

[OCBP14] P. Ochs, Y. Chen, T. Brox, and T. Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[PM16] A. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

[SC15] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015.

[SLHN10] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision ECCV 2010*, volume 6316 of *Lecture Notes in Computer Science*, pages 43–56. Springer Berlin Heidelberg, 2010.

[VIB15] R. Verde, A. Irpino, and A. Balzanella. Dimension reduction techniques for distributional symbolic data. *IEEE Transactions on Cybernetics*, 2(46):344–355, 2015.

[Vil03] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[WSB+13] W. Wang, D. Slepcev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013.