Geodesic PCA in the Wasserstein space by Convex PCA

Jérémie Bigot[1], Raúl Gouet[2], Thierry Klein[3] & Alfredo López[4]

Institut de Mathématiques de Bordeaux et CNRS (UMR 5251)[1]
Université de Bordeaux

Depto. de Ingeniería Matemática and CMM (CNRS, UMI 2807)[2]
Universidad de Chile

Institut de Mathématiques de Toulouse et CNRS (UMR 5219)[3]
Université de Toulouse

CSIRO Chile International Centre of Excellence[4]

July 17, 2015

Abstract

We introduce the method of Geodesic Principal Component Analysis (GPCA) on the space of probability measures on the line, with finite second moment, endowed with the Wasserstein metric. We discuss the advantages of this approach, over a standard functional PCA of probability densities in the Hilbert space of square-integrable functions. We establish the consistency of the method by showing that the empirical GPCA converges to its population counterpart, as the sample size tends to infinity. A key property in the study of GPCA is the isometry between the Wasserstein space and a closed convex subset of the space of square-integrable functions, with respect to an appropriate measure. Therefore, we consider the general problem of PCA in a closed convex subset of a separable Hilbert space, which serves as basis for the analysis of GPCA and also has interest in its own right. We provide illustrative examples on simple statistical models, to show the benefits of this approach for data analysis. The method is also applied to a real dataset of population pyramids.

Keywords: Wasserstein space, Geodesic and Convex Principal Component Analysis, Fréchet mean, Functional data analysis, Geodesic space, Inference for family of densities.
AMS classifications: Primary 62G08; secondary 62G20.

Acknowledgments

Associate Editor for their constructive criticism, comments and remarks that resulted in a major revision of the original manuscript.

1 Introduction

1.1 Main goal of this paper

The main goal of this paper is to define a notion of principal component analysis (PCA) of a family of probability measures $\nu_1, \ldots, \nu_n$, defined on the real line $\mathbb{R}$. In the case where the measures admit square-integrable densities $f_1, \ldots, f_n$, the standard approach is to use functional PCA (FPCA) (see e.g. [DPR82, RS05, Sil96]) on the Hilbert space $L^2(\mathbb{R})$, of square-integrable functions, endowed with its usual inner product. This method has already been applied in [Del11, KU01] for analysing the main modes of variability of a set of densities.

We briefly introduce elements of standard PCA in a separable Hilbert space H, endowed with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. A PCA of the data $x_1, \ldots, x_n$ in H is carried out by diagonalizing the empirical covariance operator $Kx = \frac{1}{n} \sum_{i=1}^{n} \langle x_i - \bar{x}_n, x \rangle (x_i - \bar{x}_n)$, $x \in H$, where $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the Euclidean mean.

The eigenvectors of K, associated to the largest eigenvalues, describe the principal modes of data variability around $\bar{x}_n$. The first principal mode of linear variation of the data is defined by the H-valued curve $g : \mathbb{R} \to H$ given by

$$g_t = \bar{x}_n + t\sigma_1 w_1, \ t \in \mathbb{R}, \tag{1.1}$$

where $w_1 \in H$ is the eigenvector corresponding to the largest eigenvalue $\sigma_1 \geq 0$ of K. On the other hand, it is well known that PCA can be formulated as the problem of finding a sequence of nested affine subspaces, minimizing the sum of norms of projection residuals. In particular, w_1 is a solution of

$$\min_{v \in H, \, \|v\|=1} \sum_{i=1}^{n} d^2(x_i, S_v) = \min_{v \in H, \, \|v\|=1} \sum_{i=1}^{n} \|x_i - \bar{x}_n - \langle x_i - \bar{x}_n, v \rangle v\|^2, \tag{1.2}$$

where $S_v = \{\bar{x}_n + tv, \ t \in \mathbb{R}\}$ is the affine subspace through $\bar{x}_n$, with direction $v \in H$, and $d(x, S) = \inf_{x' \in S} \|x - x'\|$ denotes the distance from $x \in H$ to $S \subset H$.

We illustrate the strategy discussed above on the set of Gaussian densities $f_1, \ldots, f_4$, shown in Figure 1. These densities are sampled from the following location-scale model, to be used throughout the paper as illustrative example: let f_0 be a density in $L^2(\mathbb{R})$ and, for $(a_i, b_i) \in (0, \infty) \times \mathbb{R}$, $i = 1, \ldots, n$, we define ν_i as the probability measure with density

$$f_i(x) := a_i^{-1} f_0 \left(a_i^{-1}(x - b_i) \right), \ x \in \mathbb{R}. \tag{1.3}$$

This model is appropriate in many applications such as curve registration and signal warping, see e.g. [BK12] and [GLM13]. The main sources of variability in these densities are the variation in location along the x-axis, and the scaling variation. One of the purposes of this paper is to develop a notion of PCA, that has desirable and coherent properties with respect to this variability and the model. A first requirement is that the principal modes of variation be

densities. Moreover, they should reflect the fact that the data vary in location and scale around f_0.

The densities displayed in Figure 1 represent an example of realizations of this model, with f_0 the standard normal density and $n = 4$. Let us first consider the FPCA of this dataset. To that end we compute the Euclidean mean $\bar{f}_4$, shown in Figure 1(e), a bi-modal density which is not a "satisfactory" average of the uni-modal densities $f_1, \ldots, f_4$. In Figure 2 we display the first mode of linear variation g, given by (1.1), and observe that it is not a "meaningful" descriptor of the variability in the data. Indeed, for $|t|$ sufficiently large, g_t may take negative values and does not integrate to one, as illustrated in Figure 2(a),(e),(f). Moreover, even for small values of $|t|$, g_t does not represent the typical shape of the observed densities, as shown by Figure 2(c),(d). Therefore, the FPCA of densities in $L^2(\mathbb{R})$ is not always appropriate as it may lead to principal modes of linear variation that are not coherent with the sources of variability observed in the data (e.g., sampled from a location-scale model). To overcome some of these issues, one could constrain the first mode of variation to lie in the set of positive functions, integrating to one. However, such a constrained PCA would be computed via the $L^2(\mathbb{R})$ norm, so the Euclidean mean $\bar{f}_4$ would stay unchanged and still not be satisfactory. We believe these drawbacks of FPCA are mainly due to the fact that the Euclidean distance in $L^2(\mathbb{R})$ is not appropriate to perform PCA for densities.
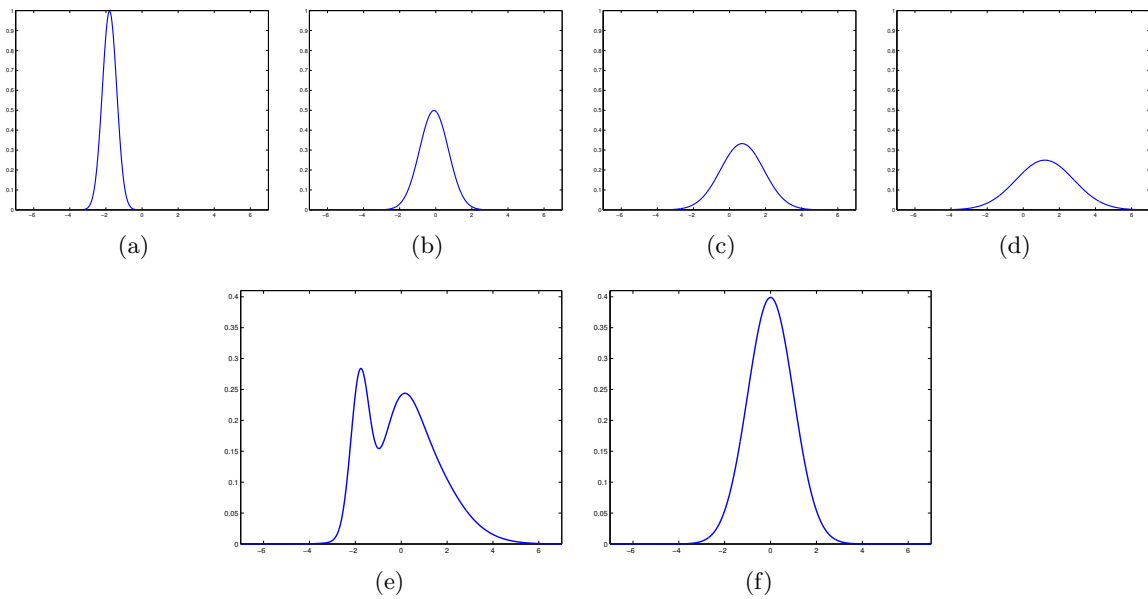

Figure 1: (a,b,c,d) Graphs of Gaussian densities $f_1, \ldots, f_4$, with different means and variances sampled from a location-scale model. (e) Euclidean mean of $f_1, \ldots, f_4$ in $L^2(\mathbb{R})$. (f) Density of the barycenter $\bar{\nu}_4$ of $\nu_1, \ldots, \nu_4$ in the Wasserstein space $W_2(\mathbb{R})$.
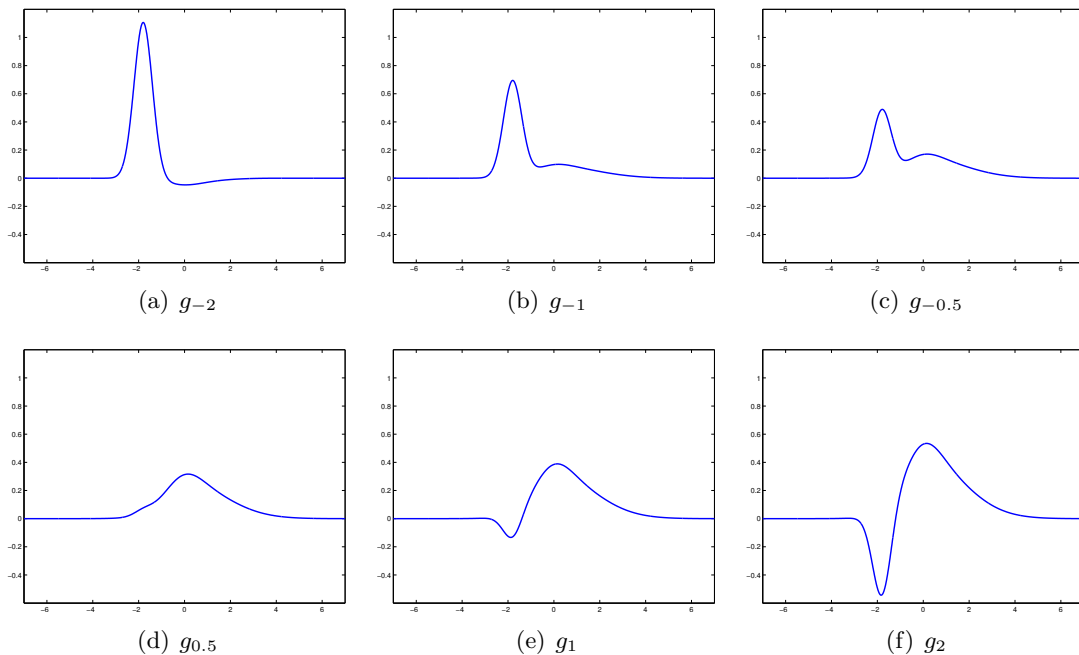
3

Figure 2: An example of functional PCA of densities. First principal mode of linear variation g_t in $L^2(\mathbb{R})$, for $-2 \leq t \leq 2$, of the densities displayed in Figure 1; see equation (1.1).

1.2 Main contributions and organization of the paper

In this paper we suggest to rather consider that $\nu_1, \ldots, \nu_n$ belong to the Wasserstein space $W_2(\Omega)$ of probability measures over Ω, with finite second order moment, where Ω is $\mathbb{R}$ or a closed interval of $\mathbb{R}$. This space is endowed with the Wasserstein distance, associated to the quadratic cost; see [Vil03] for an overview of Wasserstein spaces. In this setting it is not possible to define a notion of PCA in the usual sense as $W_2(\Omega)$ is not a linear space. Nevertheless, we show how to define a proper notion of Geodesic PCA (GPCA), by relying on the formal Riemannian structure of $W_2(\Omega)$, developed in [AGS04], that we describe in Section 2.1. A first idea in that direction is related to the mean of the data, which is an essential ingredient in any notion of PCA. We propose to use the Fréchet mean (also called barycenter) as introduced in [AC11], with asymptotic properties studied in [BK12]. It is significant that the barycenter of $\nu_1, \ldots, \nu_4$, in our example above, preserves the shapes of the densities; see Figure 1(f).

Before precisely defining GPCA in $W_2(\Omega)$, we display $\tilde{g}$ in Figure 3, the first principal mode of geodesic variation in $W_2(\Omega)$, of the data displayed in Figure 1; see equation (5.1). GPCA clearly gives a better description of the variability in the data, compared to the results in Figure 2, that correspond to the first principal mode of linear variation g in $L^2(\mathbb{R})$, given by (1.1).

Our approach shares similarities with analogs of PCA for data belonging to a Riemannian manifold. There is currently a growing interest in the statistical literature on the development of nonlinear analogs of PCA, for the analysis of data belonging to curved Riemannian man-
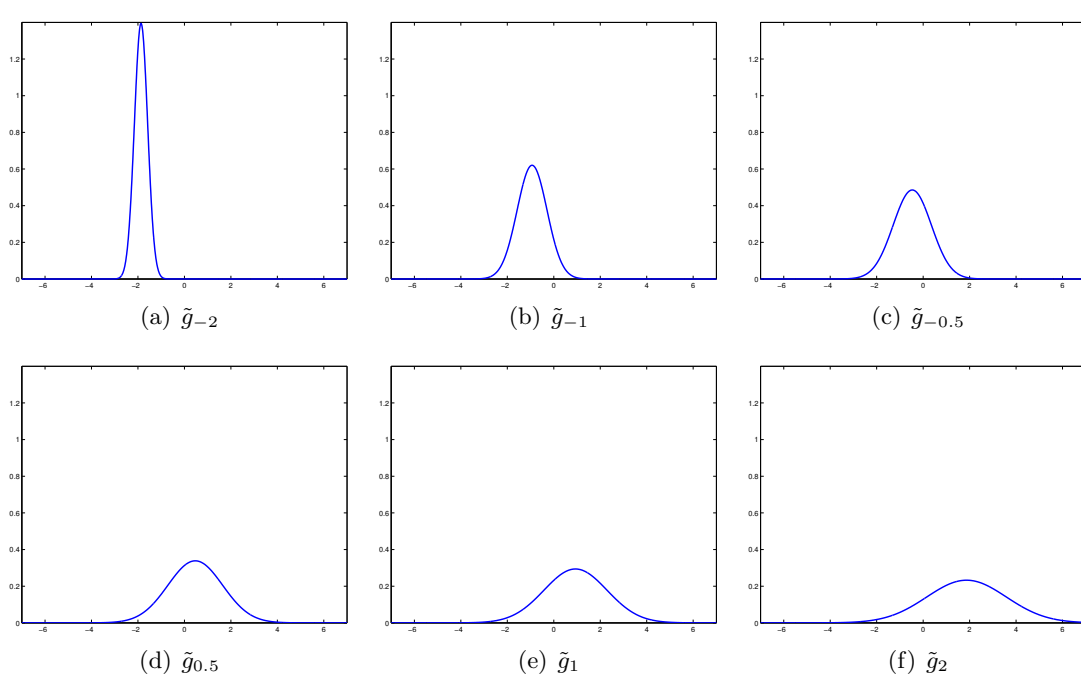
4

Figure 3: An example of GPCA of densities. First principal mode of geodesic variation $\tilde{g}_t$ in $W_2(\mathbb{R})$, for $-2 \leq t \leq 2$, of the densities displayed in Figure 1; see (5.1).

ifolds; see e.g. [FLPJ04, HHM10, SLHN10] and references therein. These methods, generally referred to as Principal Geodesic Analysis (PGA), extend the notion of classical PCA in Hilbert spaces. Nevertheless, as the Wasserstein space is not a Riemannian manifold, existing methods to perform a PGA cannot be directly applied to the setting of this paper.

The key property that we use to develop a notion of GPCA in the Wasserstein space is the isometry between $W_2(\Omega)$ and a closed convex subset of the Hilbert space of square-integrable functions $L_\mu^2(\Omega)$, with respect to an appropriate measure μ; see Theorem 2.2. In this paper we thus consider the statement of the general problem of PCA in a closed convex subset of a Hilbert space, which not only serves as basis for the analysis of GPCA in $W_2(\Omega)$, but may also have interest in its own right, for further developments. For example, the notion of convex PCA introduced in this paper could be of interest, when probability distributions are characterized by observed parameters, belonging to some convex subset of an Euclidean space.

Throughout the paper, various notions from Riemannian geometry such as geodesic, tangent space, exponential and logarithmic maps, are used to illustrate the connection between our approach and PGA. However, the important issue here is not the geometry of $W_2(\Omega)$ but rather the use of these notions to state precisely the isometry between $W_2(\Omega)$ and a closed convex set of $L_\mu^2(\Omega)$. The GPCA in the Wasserstein space is then an application of these results.

The rest of the paper is organized as follows. In Section 2, we present the isometry between $W_2(\Omega)$ and a closed convex subset of $L_\mu^2(\Omega)$. We also recall basic definitions such as tangent space, geodesic, exponential and logarithmic maps in the Wasserstein space framework, having

their analogs in the Riemannian setting. Section 3 is devoted to the definition and analysis of Convex PCA (CPCA) in a general framework. The main results on GPCA are gathered in Section 4. In Section 5 we describe some numerical aspects of GPCA on simulated examples, using simple statistical models. We also analyze a real dataset of population pyramids of 223 countries, for the year 2000. Section 6 is dedicated to the consistency of the empirical CPCA and GPCA, as the number of random data points tends to infinity. We conclude the paper in Section 7, discussing the differences between GPCA and existing PGA methods on Riemannian manifolds. We also mention potential extensions of this work. Finally, to make the paper self-contained, we collect in the Appendix some technical results about quantiles, geodesic spaces, Kuratowski convergence and Γ-convergence.

Remark 1.1. In this paper we assume that the input data consist of probabilities $\nu_1, \ldots, \nu_n$ belonging to $W_2(\Omega)$. However, in many applications we may have access, only to random observations from each of these probabilities. A natural strategy to address this issue is to estimate the associated densities by means of kernel estimators, for instance, and then a GPCA could be applied to the estimations. Possibly, more efficient estimators of principal geodesics could be obtained by adapting ideas from [KU01] which would, however, require a simple representation of principal geodesics in terms of densities.

2 Convexity of the Wasserstein space $W_2(\Omega)$ up to an isometry

2.1 The pseudo-Riemannian structure of $W_2(\Omega)$

Let Ω be either the real line $\mathbb{R}$ or a closed interval of $\mathbb{R}$ and let $W_2(\Omega)$ be the set of probability measures over $(\Omega, \mathcal{B}(\Omega))$, with finite second moment, where $\mathcal{B}(\Omega)$ is the σ-algebra of Borel subsets of Ω. For $\nu \in W_2(\Omega)$ and $T : \Omega \to \Omega$ (always assumed measurable), we recall that the push-forward measure $T\#\nu$ is defined by $(T\#\nu)(A) = \nu\{x \in \Omega | T(x) \in A\}$, for $A \in \mathcal{B}(\Omega)$. The cumulative distribution function (cdf) and the quantile function of ν are denoted respectively by F_ν and F_ν^-; see Definition A.2. If ν is absolutely continuous (a.c.), its density is denoted by f_ν.

Definition 2.1. The quadratic Wasserstein distance d_W in $W_2(\Omega)$ is defined by

$$d_W^2(\nu_1, \nu_2) := \inf_{\pi \in \Pi(\nu_1, \nu_2)} \int |x - y|^2 d\pi(x, y), \ \nu_1, \nu_2 \in W_2(\Omega),$$

where $\Pi(\nu_1, \nu_2)$ is the set of probability measures on $\Omega \times \Omega$, with marginals ν_1 and ν_2.

It can be shown that $W_2(\Omega)$ endowed with d_W is a metric space, usually called Wasserstein space. For a detailed analysis of $W_2(\Omega)$, we refer to [Vil03]. In particular, the following formula, from Theorem 2.18 in [Vil03], is important in the sequel:

$$d_W^2(\nu_1, \nu_2) = \int_0^1 (F_{\nu_2}^-(y) - F_{\nu_1}^-(y))^2 dy. \tag{2.1}$$

Also important is the following celebrated theorem (stated for measures on $\mathbb{R}^d$), from optimal transportation theory, due to Brenier [Bre91].

Theorem 2.1. *Let $\mu, \nu \in W_2(\mathbb{R}^d)$ such that μ gives no mass to small sets, then*

$$d_W^2(\mu, \nu) = \inf_{T \in \mathrm{MP}(\mu,\nu)} \int_\Omega |T(x) - x|^2 d\mu(x), \tag{2.2}$$

where $\mathrm{MP}(\mu, \nu) = \{T : \mathbb{R}^d \to \mathbb{R}^d \mid \nu = T\#\mu\}$. Moreover, there exists $T^* \in \mathrm{MP}(\mu, \nu)$ such that $d_W^2(\mu, \nu) = \int_\Omega |T^*(x) - x|^2 d\mu(x)$, characterized as the unique (up to a μ-negligible set) element in $\mathrm{MP}(\mu, \nu)$ that can be represented, μ-almost everywhere (a.e.), as the gradient of a convex function.

Since we are in dimension $d = 1$, T^* in Theorem 2.1, being the gradient of a convex function, is increasing. Observe also that T^* may possibly be defined and be increasing only in a set of μ measure 1, but still $T^*\#\mu$ makes sense; see [Vil03], page 67. Finally note that in $\mathbb{R}$ it suffices to assume μ atomless, that is, F_μ continuous. Under the above stated conditions it is well known that $T^* = F_\nu^- \circ F_\mu$ and

$$d_W^2(\mu, \nu) = \int_\Omega (F_\nu^- \circ F_\mu(x) - x)^2 d\mu(x), \tag{2.3}$$

with $F_\nu^- \circ F_\mu$ defined on the full μ-measure set $A_\mu := \{x \in \Omega | F_\mu(x) \in (0, 1)\}$.

The $W_2(\Omega)$ space has a formal Riemannian structure described, for example, in [AGS04]. We provide some basic definitions, having their analogs in the Riemannian manifold setting.

From here onwards we consider that $\mu \in W_2(\Omega)$ is a reference measure, with continuous cdf F_μ. Following [AGS04], we define the tangent space at μ as the Hilbert space $L_\mu^2(\Omega)$ of real-valued, μ-square-integrable functions on Ω, equipped with the standard inner product $\langle \cdot, \cdot \rangle_\mu$ and norm $\| \cdot \|_\mu$. Furthermore, we define the exponential and the logarithmic maps at μ, as follows.

Definition 2.2. Let id be the identity on Ω. The exponential $\exp_\mu : L_\mu^2(\Omega) \to W_2(\Omega)$ and logarithmic $\log_\mu : W_2(\Omega) \to L_\mu^2(\Omega)$ maps are defined respectively as

$$\exp_\mu(v) = (v + \mathrm{id})\#\mu \quad \text{and} \quad \log_\mu(\nu) = F_\nu^- \circ F_\mu - \mathrm{id}. \tag{2.4}$$

Remark 2.1. (a) $\exp_\mu(v) \in W_2(\Omega)$, for any $v \in L_\mu^2(\Omega)$, since

$$\int x^2 d\exp_\mu(v)(x) = \int (x + v(x))^2 d\mu(x) \leq 2 \int x^2 d\mu(x) + 2 \int v^2(x) d\mu(x) < +\infty.$$

(b) By Theorem 2.1 and (2.3), $\log_\mu(\nu)$ is unique (μ-a.e.) and belongs to $L_\mu^2(\Omega)$ since $\| \log_\mu(\nu) \|_\mu^2 = d_W^2(\mu, \nu) < +\infty$, for all $\nu \in W_2(\Omega)$. But, as commented after (2.3), $\log_\mu(\nu)$ is only defined on A_μ. Finally, the continuity of F_μ implies $\exp_\mu(\log_\mu(\nu)) = \nu$.

Example 2.1. We illustrate the notions of exponential and logarithmic maps, using again the location-scale model. For $\mu_0 \in W_2(\mathbb{R})$ a.c. and $(a, b) \in (0, \infty) \times \mathbb{R}$, let $\nu^{(a,b)}$ be the probability measure, with cdf and density respectively given by

$$F^{(a,b)}(x) := F_{\mu_0}((x - b)/a), \quad f^{(a,b)}(x) := f_{\mu_0}((x - b)/a)/a, \ x \in \mathbb{R}. \tag{2.5}$$

From (2.4), $\log_\mu(\nu^{(a,b)})(x) = [F^{(a,b)}]^- \circ F_\mu(x) - x$ and $\log_{\mu_0}(\nu^{(a,b)})(x) = (a - 1)x + b$. Therefore, letting $v(x) = (a - 1)x + b$, we have

$$\exp_{\mu_0}(v) = \nu^{(a,b)}. \tag{2.6}$$

7

In the setting of Riemannian manifolds, the exponential map at a given point is a local homeomorphism from a neighborhood of the origin in the tangent space to the manifold. However, this is not the case for $\exp_\mu$ defined above, as it is possible to find two arbitrarily small functions in $L^2_\mu(\Omega)$, with equal exponentials, see e.g. [AGS04]. On the other hand, we show that $\exp_\mu$ is an isometry when restricted to a specific set of functions defined below.

2.2 Isometry between $W_2(\Omega)$ and a closed convex subset of $L^2_\mu(\Omega)$

We consider below the image of $W_2(\Omega)$ under the logarithmic map, denoted $V_\mu(\Omega)$, which is shown to be a closed convex subset of $L^2_\mu(\Omega)$. We also prove that $\exp_\mu$, restricted to $V_\mu(\Omega)$, is an isometry. These are crucial properties needed to define and to compute the GPCA in $W_2(\Omega)$.

Theorem 2.2. *The exponential map $\exp_\mu$ restricted to $V_\mu(\Omega) := \log_\mu(W_2(\Omega))$ is an isometric homeomorphism, with inverse $\log_\mu$.*

Proof. Let $\nu \in W_2(\Omega)$ then, from Theorem 2.1, $F_\nu^- \circ F_\mu$ is the unique μ-a.e. increasing map (see Definition A.1), such that $(F_\nu^- \circ F_\mu)\#\mu = \nu$. In other words, $v := \log_\mu(\nu) = F_\nu^- \circ F_\mu - \mathrm{id}$ is the unique element in $V_\mu(\Omega)$ such that $\exp_\mu(v) = \nu$. The isometry property follows from (2.1) because $d_W^2(\nu_1, \nu_2) = \int_0^1 (F_{\nu_2}^-(y) - F_{\nu_1}^-(y))^2 dy = \|F_{\nu_1}^- \circ F_\mu - F_{\nu_2}^- \circ F_\mu\|_\mu^2 = \|\log_\mu(\nu_1) - \log_\mu(\nu_2)\|_\mu^2$, for any $\nu_1, \nu_2 \in W_2(\Omega)$. $\qquad\square$

Proposition 2.1. *The set $V_\mu(\Omega) := \log_\mu(W_2(\Omega))$ is closed and convex in $L^2_\mu(\Omega)$.*

Proof. Let (ν_n) be a sequence in $W_2(\Omega)$, such that $\log_\mu(\nu_n) \to v \in L^2_\mu(\Omega)$. Then $F_{\nu_n}^- \circ F_\mu \to v + \mathrm{id}$ and, because F_μ is continuous, we have $F_{\nu_n}^- \to w \in L^2(0,1)$ (the space of square-integrable functions with respect to the Lebesgue measure on $(0,1)$). From Proposition A.2, there exists $\nu \in W_2(\Omega)$ such that $w = F_\nu^{-1}$ a.e. and so, $F_{\nu_n}^- \circ F_\mu \to F_\nu^- \circ F_\mu$ in $L^2_\mu(\Omega)$, that is, $\log_\mu(\nu_n) \to \log_\mu(\nu) \in V_\mu(\Omega)$. Convexity follows also from Proposition A.2 because, for $\lambda \in [0,1]$, there exists $\nu_\lambda \in W_2(\Omega)$ such that $\lambda \log_\mu(\nu_1) + (1-\lambda)\log_\mu(\nu_2) = (\lambda F_{\nu_1}^- + (1-\lambda)F_{\nu_2}^-) \circ F_\mu - \mathrm{id} = F_{\nu_\lambda}^- \circ F_\mu - \mathrm{id} \in V_\mu(\Omega)$. $\qquad\square$

Remark 2.2. The space $V_\mu(\Omega)$ can be characterized as the set of functions $v \in L^2_\mu(\Omega)$ such that $T := \mathrm{id} + v$ is μ-a.e. increasing (see Definition A.1) and that $T(x) \in \Omega$, for $x \in \Omega$.

2.3 Geodesics in $W_2(\Omega)$

A general overview of geodesics in a metric space is given in the Appendix. In this section, we consider the notion of geodesic in $W_2(\Omega)$, as given in Definition A.4. A direct consequence of Corollary A.1, Proposition 2.1 and Theorem 2.2 is that geodesics in $W_2(\Omega)$ are exactly the image under $\exp_\mu$ of straight lines in $V_\mu(\Omega)$. In particular, given two measures in $W_2(\Omega)$, there exists a unique shortest path connecting them. This property is stated in the following lemma.

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

Example 2.2. To illustrate Lemma 2.1, let us consider again the location-scale model (2.5). Then one has $v_0(x) := \log_{\mu_0}(\nu^{(1,0)}) = 0$ and $v_1(x) := \log_{\mu_0}(\nu^{(a,b)}) = (a-1)x + b$, $x \in \mathbb{R}$. From Lemma 2.1, the curve $\gamma : [0,1] \to W_2(\Omega)$, defined by

$$\gamma(t) = \exp_{\mu_0}((1-t)v_0 + tv_1) = \exp_{\mu_0}(t(a-1)x + tb) = \nu^{(a_t,b_t)}, \ t \in [0,1],$$

is a geodesic such that $\gamma(0) = \mu_0 = \nu^{(1,0)}$ and $\gamma(1) = \nu^{(a,b)}$, where $a_t = 1 - t + ta$ and $b_t = tb$. Moreover, for each time $t \in [0,1]$, the measure $\gamma(t)$ admits the density

$$f^{(a_t,b_t)}(x) = a_t^{-1} f_0\left(a_t^{-1}(x - b_t)\right), \ x \in \mathbb{R}. \tag{2.7}$$

In Figure 4 we display the densities $f^{(a_t,b_t)}$ for some values of $t \in [0,1]$, with μ_0 the standard Gaussian measure, $a = 0.5$ and $b = 2$.

(a) $f^{(a_t,b_t)}$ for $t = 0$ (b) $f^{(a_t,b_t)}$ for $t = 0.2$ (c) $f^{(a_t,b_t)}$ for $t = 0.4$

(d) $f^{(a_t,b_t)}$ for $t = 0.6$ (e) $f^{(a_t,b_t)}$ for $t = 0.8$ (f) $f^{(a_t,b_t)}$ for $t = 1$
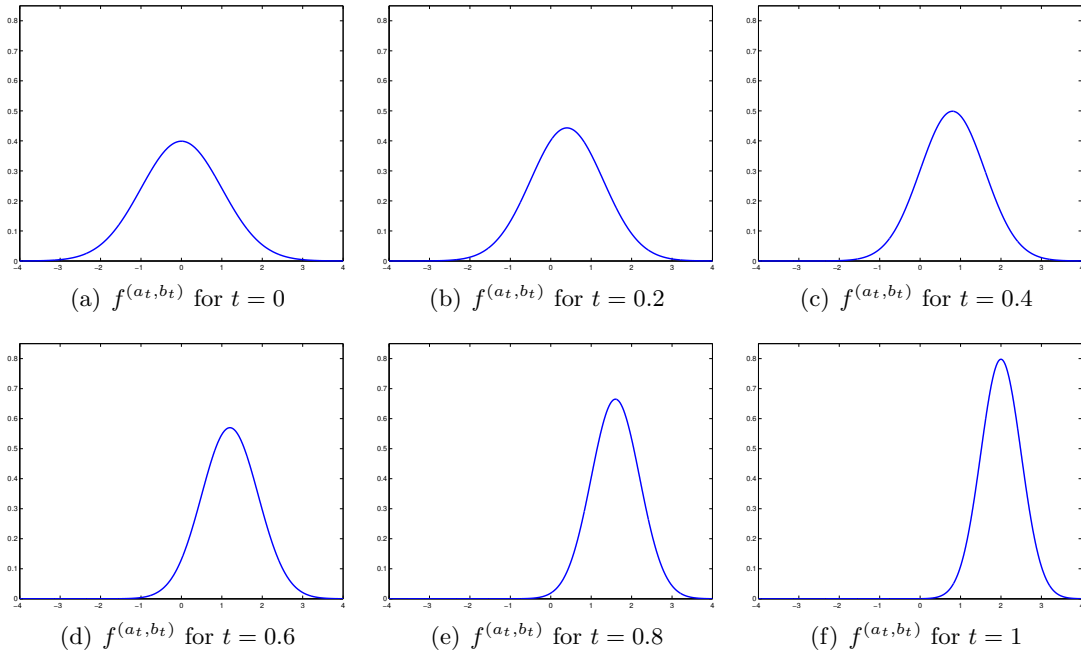
Figure 4: Visualization of the densities $f^{(a_t,b_t)}$ associated to the geodesic curve $\gamma(t) = \nu^{(a_t,b_t)}$ in W_2, described in Example 2.2, with $a = 0.5$ and $b = 2$, in the case where $\mu = \mu_0$ is the standard Gaussian measure.

Remark 2.3. By Lemma 2.1, $W_2(\Omega)$ endowed with the Wasserstein distance d_W is a geodesic space. Moreover, we have the following corollary.

Corollary 2.1. *A set $G \subset W_2(\Omega)$ is geodesic (in the sense of Definition A.5) if and only if $\log_\mu(G)$ is convex.*

Definition 2.3. Let $G \subseteq W_2(\Omega)$ be geodesic. The dimension of G, denoted $\dim(G)$, is defined as the dimension of the smallest affine subspace of $L_\mu^2(\Omega)$ containing $\log_\mu(G)$.

9

Remark 2.4. $\dim(G)$ does not depend on the the reference measure μ. Indeed, $\mu' \in W_2(\Omega)$ (atomless) and E an affine subspace of $L^2_\mu(\Omega)$, such that $\log_\mu(G) \subseteq E$. It is easy to see that $\log_{\mu'} \circ \exp_\mu : L^2_\mu(\Omega) \to L^2_{\mu'}(\Omega)$ is affine, therefore $\log_{\mu'} \circ \exp_\mu(E)$ is an affine subspace of $L^2_{\mu'}(\Omega)$ containing $\log_{\mu'}(G)$ and $\dim(E) = \dim(\log_{\mu'} \circ \exp_\mu(E))$. Observe also that, if $\gamma : [0,1] \to W_2(\Omega)$ is a geodesic, then $\gamma([0,1])$ is a geodesic space of dimension 1.

3 Convex PCA

We have shown in Section 2 that $W_2(\Omega)$ is isometric to the closed convex subset $V_\mu(\Omega)$, of the Hilbert space $L^2_\mu(\Omega)$. As can be seen in Section 4, the notion of GPCA in $W_2(\Omega)$ is strongly linked to a PCA constrained to $V_\mu(\Omega)$. It is then natural to develop a general strategy of convex-constrained PCA, in a general Hilbert space. This method, which we call Convex PCA (CPCA), could be applicable beyond the GPCA in $W_2(\Omega)$. We introduce the following notation:
- H is a separable Hilbert space, with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$.
- $d(x,y) := \|x - y\|$ and $d(x,E) := \inf_{z \in E} d(x,z)$, for $x, y \in H, E \subset H$.
- X is a closed convex subset of H, equipped with its Borel σ-algebra $\mathcal{B}(X)$.
- $\boldsymbol{x}$ is an X-valued random element, assumed square-integrable, in the sense that $\mathbb{E}\|\boldsymbol{x}\|^2 < +\infty$, with expected value $\mathbb{E}\boldsymbol{x}$.
- $x_0 \in X$ is a reference element and $k \geq 1$ an integer.

Remark 3.1. $(\mathbb{E}\|\boldsymbol{x}\|)^2 \leq \mathbb{E}\|\boldsymbol{x}\|^2 < +\infty$ and so, $\mathbb{E}\boldsymbol{x} \in H$. It is well-known that $\mathbb{E}\boldsymbol{x}$ is characterized as the unique element in H satisfying $\langle \mathbb{E}\boldsymbol{x}, x \rangle = \mathbb{E}\langle \boldsymbol{x}, x \rangle$, for all $x \in H$, and also, as the unique element in $\arg\min_{y \in X} \mathbb{E}d^2(\boldsymbol{x}, y)$. Hence $\mathbb{E}\boldsymbol{x}$ can be seen as a natural notion of average in X.

3.1 Principal convex components

Definition 3.1. For $C \subset X$, let $\mathrm{K}_X(C) = \mathbb{E}d^2(\boldsymbol{x}, C)$.

Remark 3.2. Note that $\mathrm{K}_X(C)$ is the expected value of the squared residual of $\boldsymbol{x}$ projected onto C, necessarily finite since $\boldsymbol{x}$ is assumed square-integrable. Observe also that K_X is monotone, in the sense that $\mathrm{K}_X(C) \geq \mathrm{K}_X(B)$, if $C \subset B$.

Definition 3.2. Let
(a) $\mathrm{CL}(X)$ be the metric space of nonempty, closed subsets of X, endowed with the Hausdorff distance h (see Definitions A.7, A.8),
(b) $\mathrm{CC}_k(X)$ be the family of convex sets $C \in \mathrm{CL}(X)$, such that $\dim(C) \leq k$, where $\dim(C)$ is the dimension of the smallest affine subspace of H containing C, and
(c) $\mathrm{CC}_{x_0,k}(X)$ be the family of sets $C \in \mathrm{CC}_k(X)$, such that $x_0 \in C$.

Proposition 3.1. *If X is compact, then K_X is continuous on $\mathrm{CL}(X)$.*

Proof. Let $C_n, C \in \mathrm{CL}(X), n \geq 1$, such that $h(C_n, C) \to 0$, and observe that $d^2(\boldsymbol{x}, C_n)$ is a.s. bounded by the diameter of X. Then, by Proposition A.3 and the dominated convergence theorem, $\mathrm{K}_X(C_n) \to \mathrm{K}_X(C)$. $\square$

Proposition 3.2. *If X is compact, then $\mathrm{CL}(X), \mathrm{CC}_k(X)$ and $\mathrm{CC}_{x_0,k}(X)$ are compact.*

Proof. The compactness of $\mathrm{CL}(X)$ is proved in [Pri40] and [HT13] and so we proceed with $\mathrm{CC}_k(X)$ and $\mathrm{CC}_{x_0,k}(X)$. Let $C_n \in \mathrm{CC}_k(X), n \geq 1$, and $C \in \mathrm{CL}(X)$, such that $h(C_n, C) \to 0$. Then, from Blaschke's selection theorem in Banach spaces (see [Pri40], [HT13]), C is convex.

Let us check by contradiction that $\dim(C) \leq k$. Assume that that $\dim(C) > k$, then there exists linearly independent elements $x_1, \ldots, x_{k+1} \in C$ or, equivalently, with Gram determinant $\det(GM) \neq 0$ (the Gram matrix GM has elements $GM_{i,j} = \langle x_i, x_j \rangle$, $i, j = 1, \ldots, k+1$). Observe that $h(C_n, C) \to 0$ implies that $C_n \to C$ in the sense of Kuratowski (see Remark A.2). By Definition A.6(i), there exist $x_{1,n}, \ldots, x_{k+1,n} \in C_n$, for every $n \geq 1$, such that $x_{j,n} \to x_j$, for $j = 1, \ldots, k+1$. But as $\dim(C_n) \leq k$, the Gram determinant $\det(GM_n)$ of $x_{1,n}, \ldots, x_{k+1,n}$ is zero. Also, it is easy to see that $\det(GM_n) \to \det(GM)$, which implies that $\det(GM) = 0$, a contradiction. We conclude that $\mathrm{CC}_k(X)$ is closed, hence compact, as it is a subset of the compact space $\mathrm{CL}(X)$. Finally, observe that if $x_0 \in C_n$, for all $n \geq 1$, then $x_0 \in C$, by Definition A.6(ii). So $\mathrm{CC}_{x_0,k}(X)$ is also closed, thus compact. $\qquad\square$

We define two notions of principal convex component (PCC), nested and global, and prove their existence. In the nested case, the definition is inductive and is motivated by the usual characterization of PCA, in terms of a nested sequence of optimal linear subspaces.

Definition 3.3. (a) A (k, x_0)-global principal convex component (GPCC) of $\boldsymbol{x}$ is a set $C_k \in \mathcal{G}_{x_0,k}(X) := \underset{C \in \mathrm{CC}_{x_0,k}(X)}{\arg\min} \ \mathrm{K}_X(C)$.

(b) Let $\mathcal{N}_{x_0,1}(X) = \mathcal{G}_{x_0,1}(X)$ and $C_1 \in \mathcal{G}_{x_0,1}(X)$. For $k \geq 2$, a (k, x_0)-nested principal convex component (NPCC) of $\boldsymbol{x}$ is a set

$$C_k \in \mathcal{N}_{x_0,k}(X) := \underset{C \in \mathrm{CC}_{x_0,k}(X), C \supset C_{k-1} \in \mathcal{N}_{x_0,k-1}(X)}{\arg\min} \ \mathrm{K}_X(C).$$

Theorem 3.1. *If X is compact, then $\mathcal{G}_{x_0,k}(X)$ and $\mathcal{N}_{x_0,k}(X)$ are nonempty.*

Proof. The result for $\mathcal{G}_{x_0,k}(X)$ is a direct consequence of Propositions 3.1 and 3.2. We show that $\mathcal{N}_{x_0,k}(X) \neq \emptyset$ by induction on k: first observe that $\mathcal{N}_{x_0,1}(X) = \mathcal{G}_{x_0,1}(X) \neq \emptyset$ and suppose that $C_{k-1} \in \mathcal{N}_{x_0,k-1}(X) \neq \emptyset, k \geq 2$. Furthermore, let $B_n \in \mathrm{CL}(X)$, such that $C_{k-1} \subset B_n, n \geq 1$, and $K\text{-}\lim B_n = B \in \mathrm{CL}(X)$ (the notation $K\text{-}\lim$ denotes convergence in the sense of Kuratowski, see Appendix A.3 for a precise definition, where it is also recalled, that since X is compact, the convergence with respect to the Hausdorff distance is equivalent to convergence in the sense of Kuratowski). It is clear that $C_{k-1} \subset B$, hence $\mathcal{C}_{k-1} := \{C \in \mathrm{CL}(X) \mid C \supset C_{k-1}\}$ is closed and so, by Proposition 3.2, $\{C \in \mathrm{CC}_{x_0,k}(X) \mid C \supset C_{k-1}\} = \mathcal{C}_{k-1} \cap \mathrm{CC}_{x_0,k}(X)$ is closed, thus compact. Finally, Proposition 3.1 implies $\mathcal{N}_{x_0,k}(X) \neq \emptyset$. $\qquad\square$

Remark 3.3. For $k = 1$ the notions of GPCC and NPCC coincide. However, this might not be the case for $k \geq 2$.

Definition 3.4. Given $x_1, \ldots, x_n \in X$, we denote by $\boldsymbol{x}^{(n)}$ the (square-integrable) X-valued random element such that $\mathbb{P}(\boldsymbol{x}^{(n)} \in A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(x_i)$, for any $A \in \mathcal{B}(\mathcal{X})$, where $\mathbb{1}_A$ is the indicator function of A.

Definition 3.5. The empirical GPCC and NPCC are defined as in Definition 3.3, with $\boldsymbol{x}$ replaced by $\boldsymbol{x}^{(n)}$. The empirical version of K_X is $\mathrm{K}_X^{(n)}(C) := \mathbb{E}d^2(\boldsymbol{x}^{(n)}, C) = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, C)$.

11

3.2 Formulation of CPCA as an optimization problem in H

Definition 3.6. For $\mathcal{U} = \{u_1, \ldots, u_k\} \subset H$, let
(a) $\mathrm{Sp}(\mathcal{U})$ be the subspace spanned by $u_1, \ldots, u_k$,
(b) $C_{\mathcal{U}} = (x_0 + \mathrm{Sp}(\mathcal{U})) \cap X \in \mathrm{CC}_{x_0,k}(X)$ and
(c) $\mathrm{H}_X(\mathcal{U}) := \mathrm{K}_X(C_{\mathcal{U}})$.

To simplify notations in Definition 3.6, we write $\mathrm{Sp}(u)$, $\mathrm{H}_X(u)$ or C_u whenever $\mathcal{U} = \{u\}$. We show below that finding a GPCC can be formulated as an optimization problem in H^k.

Proposition 3.3. *Let $\mathcal{U}^* = \{u_1^*, \ldots, u_k^*\}$ be a minimizer of H_X over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_k\} \subset H$, then $C_{\mathcal{U}^*} \in \mathcal{G}_{x_0,k}(X)$.*

Proof. For any $C \in \mathrm{CC}_{x_0,k}(X)$, there exists an orthonormal set $\mathcal{U} = \{u_1, \ldots, u_k\} \subset H$, such that $C \subset C_{\mathcal{U}}$. Thus, as K_X is monotone, $\mathrm{K}_X(C) \geq \mathrm{K}_X(C_{\mathcal{U}}) = \mathrm{H}_X(\mathcal{U}) \geq \mathrm{H}_X(\mathcal{U}^*) = \mathrm{K}_X(C_{\mathcal{U}^*})$, and the conclusion follows. $\square$

The analogous result for NPCC is stated below. The proof, similar to that of Proposition 3.3, is omitted.

Proposition 3.4. *Let $u_1^*, \ldots, u_k^* \in H$ such that $u_1^* \in \underset{u \in H, \|u\|=1}{\arg\min}\ \mathrm{H}_X(u)$ and, for $j = 2, \ldots, k$, let $u_j^* \in \underset{u \in \mathrm{Sp}(u_1^*, \ldots, u_{j-1}^*)^{\perp}, \|u\|=1}{\arg\min}\ \mathrm{H}_X(u)$, where $\perp$ denotes orthogonal. Then $C_{\{u_1^*, \ldots, u_k^*\}} \in \mathcal{N}_{x_0,k}(X)$.*

Remark 3.4. The empirical version of H_X is $\mathrm{H}_X^{(n)}(\mathcal{U}) := \mathrm{K}_X^{(n)}(C_{\mathcal{U}}) = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, C_{\mathcal{U}})$. A minimizer $\mathcal{U}^* = \{u_1^*, \ldots, u_k^*\}$ of $\mathrm{H}_X^{(n)}(\mathcal{U})$ leads to the construction of the empirical GPCC.

In the following proposition we give a sufficient condition for the standard PCA on H to be a solution of the CPCA problem. For the sake of simplicity, we state the result only for GPCC. Given $x \in H$ and C a closed convex subset of H, we denote by $\Pi_C x$ the projection of x onto C.

Proposition 3.5. *Let $\tilde{\mathcal{U}} = \{\tilde{u}_1, \ldots, \tilde{u}_k\} \subset H$ be a set of orthonormal eigenvectors, associated to the k largest eigenvalues of the covariance operator $Ky = \mathbb{E}\langle \boldsymbol{x} - x_0, y\rangle(\boldsymbol{x} - x_0)$, $y \in H$. If $\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})}\boldsymbol{x} \in X$ a.s., then $C_{\tilde{\mathcal{U}}} \in \mathcal{G}_{x_0,k}(X)$.*

Proof. It is well known that $\tilde{\mathcal{U}}$ is minimizer of $\mathrm{K}_X(x_0 + \mathrm{Sp}(\mathcal{U})) = \mathbb{E}\left(\|\boldsymbol{x} - \Pi_{x_0 + \mathrm{Sp}(\mathcal{U})}\boldsymbol{x}\|^2\right)$, over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_k\} \subset H$. Further, $\mathrm{H}_X(\mathcal{U}) = \mathbb{E}\left(\|\boldsymbol{x} - \Pi_{(x_0 + \mathrm{Sp}(\mathcal{U})) \cap X}\boldsymbol{x}\|^2\right)$ and, since by hypothesis $\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})}\boldsymbol{x} \in X$, we have $\mathrm{H}_X(\tilde{\mathcal{U}}) = \mathrm{K}_X(x_0 + \mathrm{Sp}(\tilde{\mathcal{U}}))$.

Also, the monotonicity of K_X implies $\mathrm{K}_X(x_0 + \mathrm{Sp}(\mathcal{U})) \leq \mathrm{K}_X((x_0 + \mathrm{Sp}(\mathcal{U})) \cap X) = \mathrm{H}_X(\mathcal{U})$. Finally, from the relations above, we get

$$\mathrm{H}_X(\tilde{\mathcal{U}}) = \mathrm{K}_X(x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})) \leq \mathrm{K}_X(x_0 + \mathrm{Sp}(\mathcal{U})) \leq \mathrm{H}_X(\mathcal{U}),$$

which means that $\tilde{\mathcal{U}}$ is a minimizer of $\mathrm{H}_X(\mathcal{U})$ over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_k\} \subset H$. Finally, from Proposition 3.3 we obtain the result. $\square$

Remark 3.5. (a) We can informally say that, if the data is sufficiently concentrated around the reference element x_0, then the CPCA in X is simply obtained from the standard PCA in H. In particular, if there exists a ball $B(x_0, r)$, with center x_0 and radius $r > 0$, such that $\boldsymbol{x} \in B(x_0, r) \subset X$, a.s., then the hypothesis of Proposition 3.5 is satisfied. Indeed, $\|\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})} \boldsymbol{x} - x_0\| = \|\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})}(\boldsymbol{x} - x_0)\| \le \|\boldsymbol{x} - x_0\| \le r$ and so $\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})} \boldsymbol{x} \in X$.

(b) The previous condition of data concentration is quite strong. However, obtaining weaker conditions ensuring $\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})} \boldsymbol{x} \in X$ a.s., seems to be a difficult problem.

(c) If we replace $\boldsymbol{x}$ by $\boldsymbol{x}^{(n)}$, we obtain the empirical version of Proposition 3.5. In this case, if $\tilde{\mathcal{U}} = \{\tilde{u}_1, \ldots, \tilde{u}_k\} \subset H$ are orthonormal eigenvectors associated to the k largest eigenvalues of the empirical covariance operator $Ky = \frac{1}{n} \sum_{i=1}^{n} \langle x_i - x_0, y \rangle (x_i - x_0)$, $y \in H$, and if $\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})} x_i \in X$, for $i = 1, \ldots, n$, then $G_{\tilde{\mathcal{U}}}$ is an empirical GPCC.

(d) In this section we have used an arbitrary reference element $x_0 \in X$. However, a natural choice for x_0 would be $\mathbb{E}\boldsymbol{x}$ or $\bar{x}_n := \mathbb{E}\boldsymbol{x}^{(n)}$, in the empirical case.

4 Geodesic PCA

We consider $W_2(\Omega)$ equipped with the Borel σ-algebra $\mathcal{B}(W)$, relative to the Wasserstein metric. Also, $\boldsymbol{\nu}$ denotes a $W_2(\Omega)$-valued random element, assumed square-integrable in the sense that $\mathbb{E}d_W^2(\boldsymbol{\nu}, \lambda) < +\infty$, for some (thus for all) $\lambda \in W_2(\Omega)$. As in Section 2, we assume that $\mu \in W_2(\Omega)$ is atomless, thus F_μ is continuous.

4.1 Fréchet Mean

A natural notion of average in $W_2(\Omega)$ is the Fréchet mean, studied in [BK12] in a more general setting. In what follows we define and give some properties of the population Fréchet mean ν^* of $\boldsymbol{\nu}$. Our results are stated in dimension one, that is, in $W_2(\mathbb{R})$. The higher dimensional case is more involved and we refer to [AC11, BK12] for further details.

Observe that if $\boldsymbol{u}$ is a $L_\mu^2(\Omega)$-valued random element, such that $\mathbb{E}\|\boldsymbol{u}\|_\mu < +\infty$, then its expectation $\mathbb{E}\boldsymbol{u}$ is given by $(\mathbb{E}\boldsymbol{u})(x) = \mathbb{E}\boldsymbol{u}(x)$, for all $x \in \mathbb{R}$. Clearly $\|\mathbb{E}\boldsymbol{u}\|_\mu \le \mathbb{E}\|\boldsymbol{u}\|_\mu < \infty$, hence $\mathbb{E}\boldsymbol{u} \in L_\mu^2(\Omega)$. Also, if $\mathbb{P}(\boldsymbol{u} \in V_\mu(\Omega)) = 1$, then $\mathbb{E}\boldsymbol{u} \in V_\mu(\Omega)$.

Proposition 4.1.
(i) There exists a unique $\nu^* \in \mathcal{W} := \arg\min_{\nu \in W_2(\Omega)} \mathbb{E}d_W^2(\boldsymbol{\nu}, \nu)$, called the Fréchet mean of $\boldsymbol{\nu}$.
(ii) $\nu^* = \exp_\mu(\mathbb{E}\boldsymbol{v})$, where $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$.
(iii) $F_{\nu^*}^- = \mathbb{E}(F_{\boldsymbol{\nu}}^-)$, where $F_{\boldsymbol{\nu}}$ is the (random) cdf of $\boldsymbol{\nu}$.
(iv) If $F_{\boldsymbol{\nu}}$ is continuous a.s., then F_{ν^*} is continuous.

Proof. (*i, ii*) Let $\mathcal{L} = \arg\min_{u \in L_\mu^2(\Omega)} \mathbb{E}\|\boldsymbol{v} - u\|_\mu^2$, $\mathcal{V} = \arg\min_{u \in V_\mu(\Omega)} \mathbb{E}\|\boldsymbol{v} - u\|_\mu^2$. From Theorem 2.2, $\mathbb{E}\|\boldsymbol{v} - u\|_\mu^2 = \mathbb{E}d_W^2(\boldsymbol{\nu}, \exp_\mu(u))$, for all $u \in L_\mu^2(\Omega)$. Therefore $\inf_{\nu \in W_2(\Omega)} \mathbb{E}d_W^2(\boldsymbol{\nu}, \nu) = \inf_{u \in V_\mu(\Omega)} \mathbb{E}\|\boldsymbol{v} - u\|_\mu^2$, and $u^* \in \mathcal{V}$ if and only if $\exp_\nu(u^*) \in \mathcal{W}$.

On the other hand, $\mathbb{E}\boldsymbol{v} \in V_\mu(\Omega)$ is the unique element of $\mathcal{L}$, hence the unique element in $\mathcal{V}$. Thus, by Theorem 2.2, $\nu^* = \exp_\mu(\mathbb{E}\boldsymbol{v})$ is the unique element of $\mathcal{W}$.

(*iii*) From (*ii*) and (2.4), we have the chain of equalities $F_{\nu^*}^- \circ F_\mu = \log_\mu(\nu^*) + \mathrm{id} = \mathbb{E}\boldsymbol{v} + \mathrm{id} =$

$\mathbb{E}(\boldsymbol{v} + \mathrm{id}) = \mathbb{E}(\log_\mu(\boldsymbol{\nu}) + \mathrm{id}) = \mathbb{E}(F_{\boldsymbol{\nu}}^- \circ F_\mu) = \mathbb{E}(F_{\boldsymbol{\nu}}^-) \circ F_\mu$, which implies $F_{\nu^*}^- = \mathbb{E}(F_{\boldsymbol{\nu}}^-)$ because F_μ is continuous.

(iv) Observe that F_ν is continuous if and only if F_ν^- is strictly increasing. So, if $F_{\boldsymbol{\nu}}^-(y) - F_{\boldsymbol{\nu}}^-(x) > 0$ a.s., for $x < y$, then $F_{\nu^*}^-(y) - F_{\nu^*}^-(x) = \mathbb{E}(F_{\boldsymbol{\nu}}^-(y) - F_{\boldsymbol{\nu}}^-(x)) > 0$, that is, F_{ν^*} is continuous. $\qquad\square$

Remark 4.1. It is interesting to see, from Proposition 4.1(ii), that $\exp_\mu(\mathbb{E}(\log_\mu(\boldsymbol{\nu})))$ does not depend on μ.

4.2 Principal geodesics

In this section we present definitions and results similar to those of Section 3; k denotes a positive integer and $\nu_0 \in W_2(\Omega)$ is a reference measure.

Definition 4.1. For $\nu \in W_2(\Omega)$, $G \subset W_2(\Omega)$, let $d_W(\nu, G) = \inf_{\lambda \in G} d_W(\nu, \lambda)$ and $\mathrm{K}_W(G) := \mathbb{E}d_W^2(\boldsymbol{\nu}, G)$.

Definition 4.2. Let
(a) $\mathrm{CL}(W)$ be the metric space of nonempty, closed subsets of $W_2(\Omega)$, endowed with the Hausdorff distance h_{W_2}, and
(b) $\mathrm{CG}_{\nu_0,k}(W) = \{G \in \mathrm{CL}(W) \mid \nu_0 \in G, \ G \text{ is a geodesic set and } \dim(G) \leq k\}$, $k \geq 1$.

The notions of global and nested principal geodesics of $\boldsymbol{\nu}$ with respect to ν_0, are presented below, followed by the main existence result. In the case of the nested geodesics, the definition is inductive. The proof depends on the relation between GPCA and CPCA in $V_\mu(\Omega)$.

Definition 4.3. (a) A (k, ν_0)-global principal geodesic (GPG) of $\boldsymbol{\nu}$ is a set $G_k \in \mathcal{G}_{\nu_0,k}(W) := \arg\min_{G \in \mathrm{CG}_{\nu_0,k}(W)} \mathrm{K}_W(G)$.

(b) A (k, ν_0)-nested principal geodesic (NPG) of $\boldsymbol{\nu}$ is a set $G_k \in \arg\min_{G \in \mathrm{CG}_{\nu_0,k}(W), G \supset G_{k-1}} \mathrm{K}_W(G), k \geq 2$, with $G_1 \in \mathcal{G}_{\nu_0,k}(W)$.

Theorem 4.1. *If Ω is compact, then $\mathcal{G}_{\nu_0,k}(W)$ and $\mathcal{N}_{\nu_0,k}(W)$ are nonempty.*

Proof. As Ω is compact, $W_2(\Omega)$ is also compact (see [Vil03]) and so is $V_\mu(\Omega)$, by Theorem 2.2. Then, Theorem 3.1 and Propositions 4.3, 4.4 ensure the existence of GPG and NPG. $\qquad\square$

Remark 4.2. As commented in Remark 3.3 for CPCA, the notions of GPG and NPG are not equivalent, except obviously for $k = 1$.

4.3 Empirical Fréchet mean and principal geodesics

Definition 4.4. Given $\nu_1, \ldots, \nu_n \in W_2(\Omega)$, we denote by $\boldsymbol{\nu}^{(n)}$ the $W_2(\Omega)$-valued random element, such that $\mathbb{P}(\boldsymbol{\nu}^{(n)} \in A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(\nu_i)$, for any $A \in \mathcal{B}(W)$.

Definition 4.5. The empirical Fréchet mean of $\nu_1, \ldots, \nu_n \in W_2(\Omega)$, denoted by ν_n^*, is defined, following Proposition 4.1, as the Fréchet mean of $\boldsymbol{\nu}^{(n)}$ defined above. Equivalently, ν_n^* is the unique element of $\arg\min_{\nu \in W_2(\Omega)} \frac{1}{n} \sum_{i=1}^n d_W^2(\nu_i, \nu)$.

Proposition 4.2. *Let* $\nu_1, \ldots, \nu_n \in W_2(\Omega)$. *Then, the following formula holds*

$$F_{\nu_n^*}^- = \frac{1}{n} \sum_{i=1}^n F_{\nu_i}^-. \tag{4.1}$$

Proof. The result is a direct consequence of Proposition 4.1(iii). □

Remark 4.3. Formula (4.1) is known in statistics as quantile averaging; see [ZM11, GLM13]. A detailed characterization of ν_n^* can be found in [AC11], for measures supported on $\mathbb{R}^d$, $d \geq 1$.

Definition 4.6. The empirical GPG and NPG are defined as in Definition 4.3, with $\boldsymbol{\nu}$ replaced by $\boldsymbol{\nu}^{(n)}$.

Remark 4.4. (a) A natural choice for the reference measure ν_0 is the Fréchet mean ν^*, which is atomless thanks to Proposition 4.1(iv).
(b) In the empirical case K_W is given by $\mathrm{K}_W^{(n)}(G) := \mathbb{E}d_W^2(\boldsymbol{\nu}^{(n)}, G) = \frac{1}{n} \sum_{i=1}^n d_W^2(\nu_i, G)$.

4.4 Formulation of GPCA as CPCA in $V_\mu(\Omega)$

Recall that geodesic sets in $W_2(\Omega)$ are the image under the exponential map $\exp_\mu$, of convex sets in $V_\mu(\Omega)$ (see Corollary 2.1). Thus, the GPCA in $W_2(\Omega)$ can be formulated as a CPCA in $V_\mu(\Omega)$, as shown in this section. CPCA is applied to $H = L_\mu^2(\Omega)$, $X = V_\mu(\Omega)$, $x_0 = \log_\mu(\nu_0)$ and $\boldsymbol{x} = \log_\mu(\boldsymbol{\nu})$. In this setting $\mathrm{K}_X(C) = \mathbb{E}d_\mu^2(\boldsymbol{x}, C)$, $C \subset V_\mu(\Omega)$.

The following proposition shows that the search of GPG in $W_2(\Omega)$ is equivalent to the search of GPCC in $V_\mu(\Omega)$. The same principle applies to NPG.

Proposition 4.3. *Let* $\mathcal{G}_{\nu_0, k}(W)$ *be the set of GPG of* $\boldsymbol{\nu}$ *and* $\mathcal{G}_{x_0, k}(V_\mu(\Omega))$ *be the set of GPCC of* $\boldsymbol{x} = \log_\mu(\boldsymbol{\nu})$. *Then* $\mathcal{G}_{\nu_0, k}(W) = \exp_\mu(\mathcal{G}_{x_0, k}(V_\mu(\Omega)))$.

Proof. From Corollary 2.1 we have $\mathrm{CG}_{\nu_0, k}(W) = \exp_\mu(\mathrm{CC}_{x_0, k}(V_\mu(\Omega)))$. On the other hand, from Theorem 2.2 and Definition 4.1, $\mathrm{K}_W(G) = \mathbb{E}d_W^2(\boldsymbol{\nu}, G) = \mathbb{E}d_\mu^2(\boldsymbol{x}, \log_\mu(G))$. Therefore, $\mathrm{K}_W(G) = \mathrm{K}_X(\log_\mu(G))$, for $G \subset W_2(\Omega)$. The result follows from Theorem 2.2. □

Proposition 4.4. *Let* $\mathcal{N}_{\nu_0, k}(W)$ *be the set of NPG of* $\boldsymbol{\nu}$ *and* $\mathcal{N}_{x_0, k}(V_\mu(\Omega))$ *the set of NPCC of* $\boldsymbol{x} = \log_\mu(\boldsymbol{\nu})$. *Then* $\mathcal{N}_{\nu_0, k}(W) = \exp_\mu(\mathcal{N}_{x_0, k}(V_\mu(\Omega)))$.

5 Numerical examples of GPCA in $W_2(\mathbb{R})$

In Section 5.1 we show an example of concentrated data, such that Proposition 3.5 can be applied and the problem of finding GPG is reduced to standard PCA on the logarithms; see Remark 3.5(a). In Section 5.2 we exhibit "spread-out data", where the GPG cannot be obtained from standard PCA.

i	1	2	3	4
a_i	0.4	0.8	1.2	1.6
b_i	−1.8	−0.1	0.7	1.2

Table 1: Values of parameters for concentrated data.

5.1 Concentrated data

We consider the set of probabilities $\nu_1, \ldots, \nu_4$, with densities $f_1, \ldots, f_4$, displayed in Figure 1. These measures satisfy the location-scale model (2.5), with μ_0 being the standard Gaussian measure and the values of a_i and b_i given in Table 1.

The Fréchet mean ν_4^* of $\nu_1, \ldots, \nu_4$ is computed using the quantile average formula (4.1), from which we obtain the density g_4^* of ν_4^* (Figure 1(f)), given by

$$g_4^*(x) = f^{(\bar{a}_4, \bar{b}_4)}(x) = f_{\mu_0}\left((x - \bar{b}_4)/\bar{a}_4\right)/\bar{a}_4 = f_{\mu_0}(x), \ x \in \mathbb{R},$$

where $\bar{a}_4 = 1$ and $\bar{b}_4 = 0$ are the arithmetic means of the parameters a_i, b_i, and so, $\nu_4^* = \mu_0$. Observe that the measures $\nu_1, \ldots, \nu_4$ are concentrated around their Fréchet mean, in the sense that their expectations and variances are not too far from those of ν_4^* (see Figure 1).

We apply Propositions 3.5 and 4.3 to compute an empirical first GPG, with both μ and ν_0 equal to μ_0. Let w_1 be the eigenvector associated to the largest eigenvalue of the empirical covariance operator $Kv = \sum_{i=1}^4 \langle v_i, v \rangle v_i / 4$, $v \in L^2_{\mu_0}(\mathbb{R})$, where

$$v_i(x) = \log_{\mu_0}(\nu_i)(x) = (a_i - 1)\,x + b_i, \ i = 1, \ldots, 4; x \in \mathbb{R}.$$

Given that the $v_i \in A \subseteq L^2_{\mu_0}(\mathbb{R})$, the subspace of affine functions (generated by the identity and the constant function 1, which are orthonormal in $L^2_{\mu_0}(\mathbb{R})$), the operator K can be identified with the 2×2 matrix $M = \sum_{i=1}^4 V_i' V_i / 4$ with $V_i = (a_i - 1, b_i)' \in \mathbb{R}^2$, $1 \leq i \leq 4$. Therefore, $w_1 \in A$ and $w_1(x) = \alpha_1 x + \beta_1$, where $W_1 := (\alpha_1, \beta_1)' = (0.36, 0.93)' \in \mathbb{R}^2$ is the eigenvector associated to the largest eigenvalue of M. In other words, computing w_1 simply amounts to calculating the first eigenvector associated to the standard PCA of the $V_i \in \mathbb{R}^2$, which represent the slope and intercept parameters of the functions v_i. In Figure 5 we display the vectors V_i (circles), together with the linear space spanned by W_1 (dash-dot line), which corresponds to the first principal direction of variation of this dataset.

Affine functions $u(x) = \alpha x + \beta$ in $V_{\mu_0}(\mathbb{R})$ are represented by points $(\alpha, \beta)' \in \mathbb{R}^2$, with $\alpha \geq -1$, which is the region to the right of the vertical dashed line in Figure 5. Hence, it can be seen from the projections of the V_i onto the space spanned by W_1, that $\Pi_{\mathrm{Sp}(w_1)} v_i \in V_{\mu_0}(\mathbb{R})$, for $1 \leq i \leq 4$. Therefore, from Propositions 3.5 and 4.3, the set of probability measures

$$G_1 = \left\{ \nu_{1,t} := \exp_{\mu_0}(tw_1) \mid t \in \mathbb{R}, 1 + t\alpha_1 \geq 0 \right\},$$

is a first empirical GPG. From (2.5) and (2.6), each $\nu_{1,t} \in G_1$ admits the density

$$g_{1,t}(x) = f_{\mu_0}\left((x - b_{1,t})/a_{1,t}\right)/a_{1,t}, \ x \in \mathbb{R}, \tag{5.1}$$

with $a_{1,t} = 1 + t\alpha_1$ and $b_{1,t} = t\beta_1$. In Figure 3, we display the first principal mode of geodesic variation $g_{1,t}$, for $-2 \leq t \leq 2$, of the densities displayed in Figure 1. As already mentioned, the

16

GPCA in $W_2(\Omega)$ gives a better interpretation of the data variability, when compared to results from the first principal mode of linear variation of the densities in $L^2_{\mu_0}(\mathbb{R})$, displayed in Figure 2.
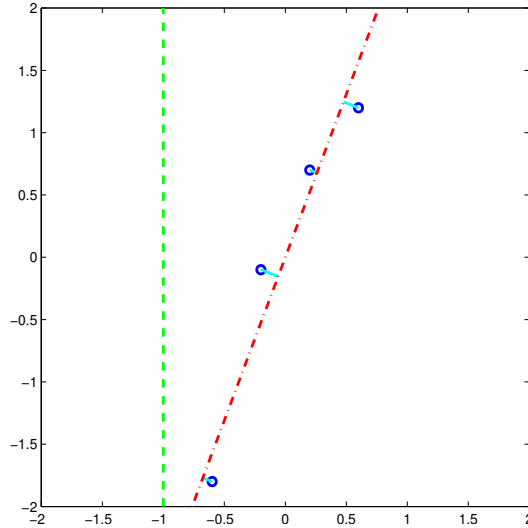

Figure 5: Two-dimensional representation of the affine functions $u(x) = \alpha x + \beta$ in $L^2_{\mu_0}(\mathbb{R})$. The horizontal and vertical axes represent the slope and the intercept parameters α, β, respectively. Points to the right of the vertical dashed line at $\alpha = -1$, correspond to affine functions in $V_{\mu_0}(\mathbb{R})$. Circles represent the vectors $V_i = (a_i - 1, b_i)'$, associated to the functions $v_i(x) = (a_i - 1)x + b_i$, for $1 \le i \le 4$, corresponding to the measures with densities displayed in Figure 1. The dash-dot line is the linear space spanned by the first eigenvector W_1 from the standard PCA of $V_1, \ldots, V_4$.

5.2 The case of spread-out data

We exhibit a case where standard PCA of logs in $L^2_\mu(\mathbb{R})$ does not lead to a solution of GPCA in $W_2(\Omega)$. Measures $\nu_1, \ldots, \nu_4$ are as in Section 5.1, with parameters a_i, b_i given in Table 2. We have again $\bar{a}_4 = 1$, $\bar{b}_4 = 0$ and $\nu_4^* = \mu_0$. From Figure 6, we see that $\nu_1, \ldots, \nu_4$ are less concentrated around ν_4^*, compared to the foregoing example (see Figure 1).

i	1	2	3	4
a_i	0.2	0.2	0.2	3.4
b_i	-3	-1	1	3

Table 2: Values of parameters for spread-out data.

As for concentrated data, we first perform a standard PCA on the logarithms in $V_{\mu_0}(\mathbb{R})$. In what follows, we keep the same notation as in Section 5.1. In Figure 7 we display the vectors
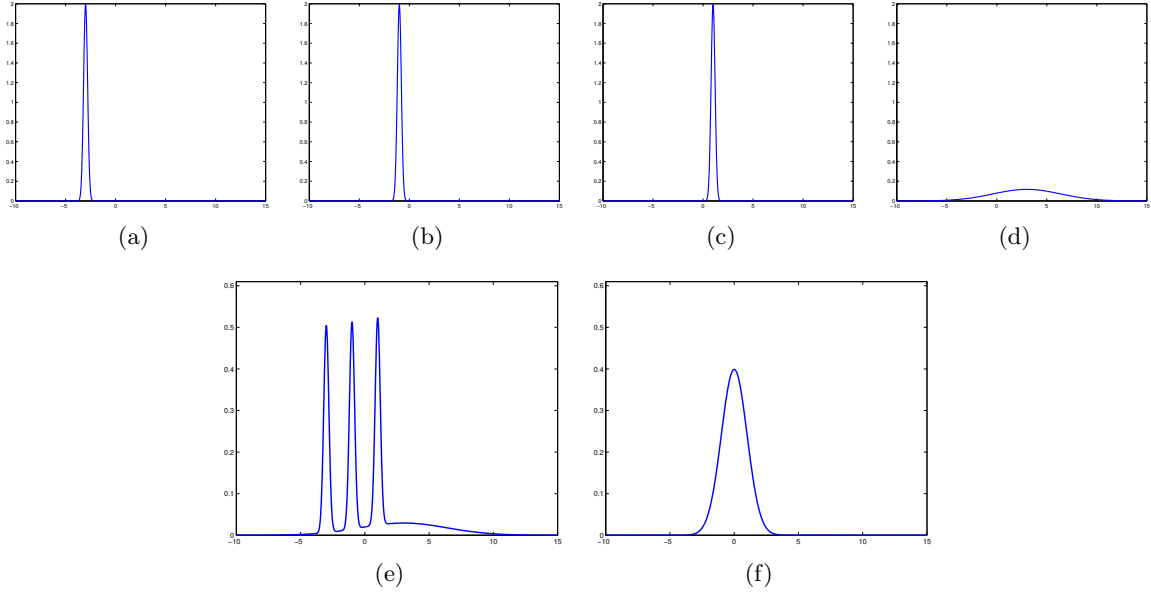
Figure 6: Gaussian densities $f_1, \ldots, f_4$ from the location-scale model (2.5), with means and variances given in Table 2. (e) Euclidean mean of the densities in $L^2(\mathbb{R})$. (f) Density of the barycenter $\nu_4^* \in W_2(\Omega)$ of $\nu_1, \ldots, \nu_4$, with densities $f_1, \ldots, f_4$.

V_i and the linear space $\mathrm{Sp}(W_1)$. From the projections of the vectors V_i onto the space spanned by W_1, it can be seen that $\Pi_{\mathrm{Sp}(w_1)} v_1 \notin V_{\mu_0}(\mathbb{R})$. Therefore, the condition $\Pi_{x_0 + \mathrm{Sp}(\tilde{\mathcal{U}})} \boldsymbol{x} \in X$ a.s. in Proposition 3.5 is not satisfied. Thus, one cannot conclude that G_1 is a first empirical GPG.

Now, in order to show that G_1 is not a GPG, it suffices to find $G_1^* \in \mathrm{CG}_{\mu_0,1}(W)$, such that $\mathrm{K}_W^{(n)}(G^*) < \mathrm{K}_W^{(n)}(G_1^*)$. To that end we perform a CPCA of $v_1, \ldots, v_4$, with $X = A \cap V_{\mu_0}(\mathbb{R})$ and reference element $x_0 = 0$. By Proposition 3.3, this amounts to solving

$$\min_{u \in A, \|u\|_{\mu_0} = 1} \mathrm{H}_X^{(n)}(u) = \frac{1}{n} \sum_{i=1}^n d_{\mu_0}^2(v_i, \mathrm{Sp}(u) \cap X). \tag{5.2}$$

On the other hand, letting $Y = \{(\alpha, \beta)' \in \mathbb{R}^2 \mid \alpha \geq -1\}$, (5.2) is equivalent to

$$\min_{U \in \mathbb{R}^2, \|U\| = 1} \mathrm{H}_Y^{(n)}(U) := \frac{1}{n} \sum_{i=1}^n d^2(V_i, \mathrm{Sp}(U) \cap Y), \tag{5.3}$$

where $d, \|\cdot\|$ are the Euclidean distance and norm in $\mathbb{R}^2$ and $V_i = (a_i - 1, b_i)' \in \mathbb{R}^2$, $1 \leq i \leq 4$. We have numerically found a unique minimizer $W_1^* = (\alpha_1^*, \beta_1^*)$ of (5.3) and so, $w_1^*(x) = \alpha_1^* x + \beta_1^*$ is the unique minimizer of (5.2).

Letting $G_1^* := \{\nu_{1,t}^* := \exp_{\mu_0}(t w_1^*) \mid t \in \mathbb{R}, 1 + t\alpha_1^* \geq 0\} \in \mathrm{CG}_{\mu_0,1}(W)$, we find that $G_1^* \neq G_1$ and $\mathrm{K}_W^{(n)}(G_1^*) < \mathrm{K}_W^{(n)}(G_1)$. Indeed, from Figure 7 it can be seen that $W_1^* \neq W_1$ and also that $\mathrm{H}_Y^{(n)}(W_1^*) < \mathrm{H}_Y^{(n)}(W_1)$.

18

Remark 5.1. For this example of spread-out data, it should be noted that G_1^* is not necessarily the first empirical GPG. Indeed, G_1^* is a minimizer of $\mathrm{K}_W^{(n)}(G)$ over the sets $G \in \mathrm{CG}_{\mu_0,1}(W)$ such that $G \subset \{\nu^{(a,b)} \mid (a,b) \in (0,\infty) \times \mathbb{R}\}$. Whether or not G_1^* is a first empirical GPC remains as an open issue. Problem (5.3) is locally convex in a neighborhood of the (unique) optimum, after parametrization of the unitary sphere of $\mathbb{R}^2$ using polar coordinates. In the general case of the optimization problems in Propositions 3.3 and 3.4, issues such as convexity, uniqueness of solution and optimality conditions, need to be addressed.
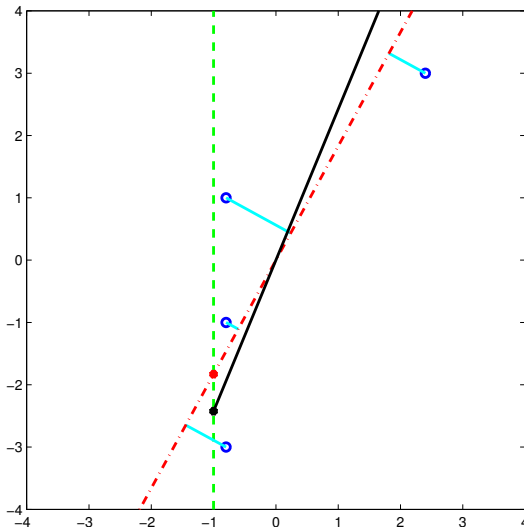

Figure 7: Same interpretation as Figure 5. The dash-dot line is the linear space spanned by the first eigenvector W_1 from the standard PCA of $V_1, \ldots, V_4$. The solid line is the convex set $\mathrm{Sp}(W_1^*) \cap Y$ where $W_1^* = (\alpha_1^*, \beta_1^*)$ is the minimizer of (5.3). The dot on the solid line is the projection of $V_1 = (0.2, -3)$ onto $\mathrm{Sp}(W_1^*) \cap Y$, while the dot on the dash-dot line is the projection of V_1 onto $\mathrm{Sp}(W_1) \cap Y$.

5.3 Real data example: statistical analysis of population pyramids

We analyze a real dataset consisting of histograms that represent the population pyramids of 223 countries for the year 2000. This dataset has been studied in [Del11] using FPCA of densities. The data are available from the International Data Base (IDB), produced by the International Programs Center, US Census Bureau (IPC, 2000), and they can be downloaded from the URL `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 probability density functions. In Figure 8, we display the population pyramids of five countries.

For the purpose of summarizing this dataset in an efficient way, we propose to compare the results obtained using either FPCA or GPCA. Note that FPCA of histograms amounts to a
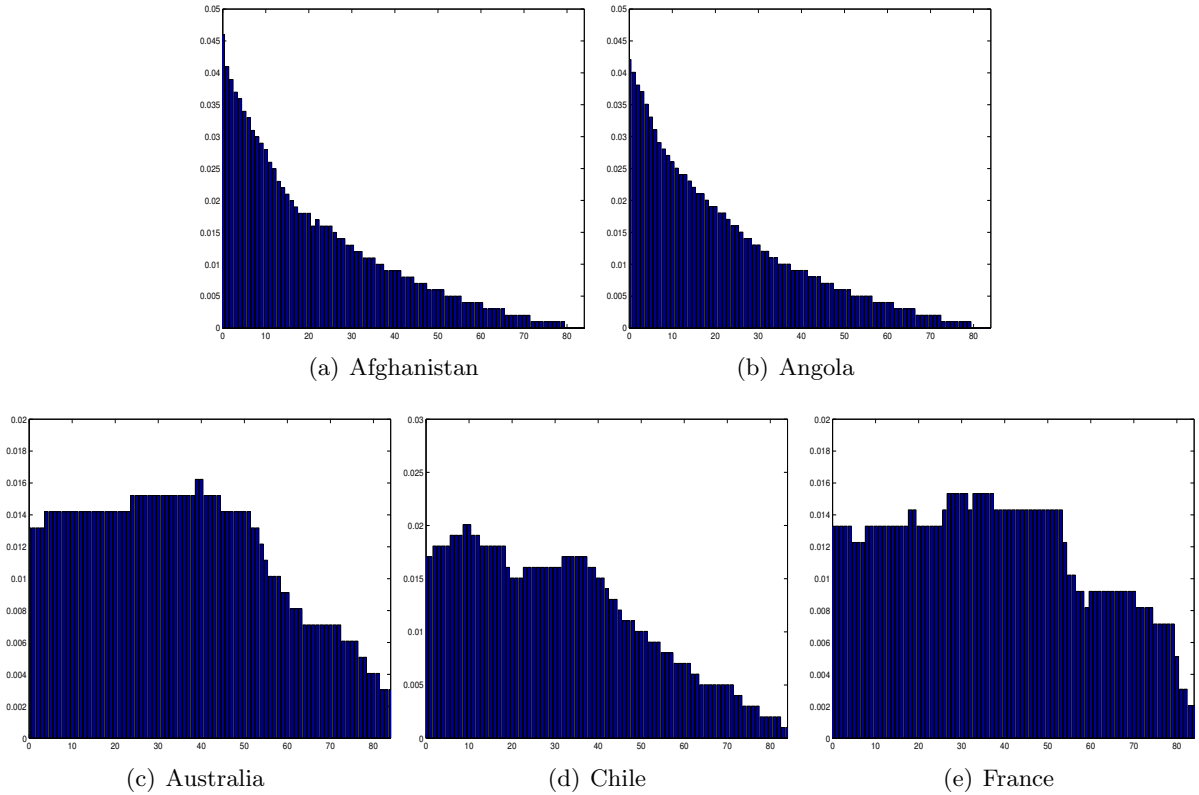
Figure 8: Population pyramids of 5 countries for the year 2000.

standard multivariate PCA in the Euclidean space $\mathbb{R}^p$ with $p = 85$. In Figure 9(a), we display the projection of the data onto the first two principal components (PC) when performing FPCA. Note that 81 % (resp. 8 %) of variability is explained by the first PC (resp. the second PC).

To perform GPCA we proceed as follows. First we compute the cdf of each histogram, which allows, from (2.4), the computation of $v_i = \log_{\nu_n^*}(\nu_i), i = 1, \ldots, n = 223$, where ν_i is the probability associated to the i-th histogram and ν_n^* is the Fréchet mean of these probability measures in $W_2(\Omega)$. Then, we perform the FPCA of the v_i in $L^2_{\nu_n^*}(\mathbb{R})$ to compute the first two PC that we denote by w_1 and w_2. For this dataset, we notice that the conditions $\Pi_{\mathrm{Sp}(\{w_1,w_2\})}v_i \in V_{\nu_n^*}$, for all $i = 1, \ldots, n = 223$, are satisfied. Therefore, Propositions 3.5 and 4.3, the FCPA of data-logarithms in $L^2_{\nu_n^*}(\mathbb{R})$ leads to a solution of GPCA in $W_2(\Omega)$. In Figure 9(b), we display the projection in $L^2_{\nu_n^*}(\mathbb{R})$ of $v_i, i = 1, \ldots, n$, onto the first two PC w_1 and w_2. Note that, when using GPCA, 96 % (resp. 2 %) of variability is explained by the first PC (resp. the second PC). Hence, we achieve a better reduction of dimensionality by the use of GPCA. Using the representation in the Wasserstein space, one may conclude that this dataset is essentially one dimensional, in terms of variability around its Fréchet mean in $W_2(\Omega)$. In particular, this fact can be observed in Figure 10, where we plot the projections of the five histograms displayed in Figure 8.
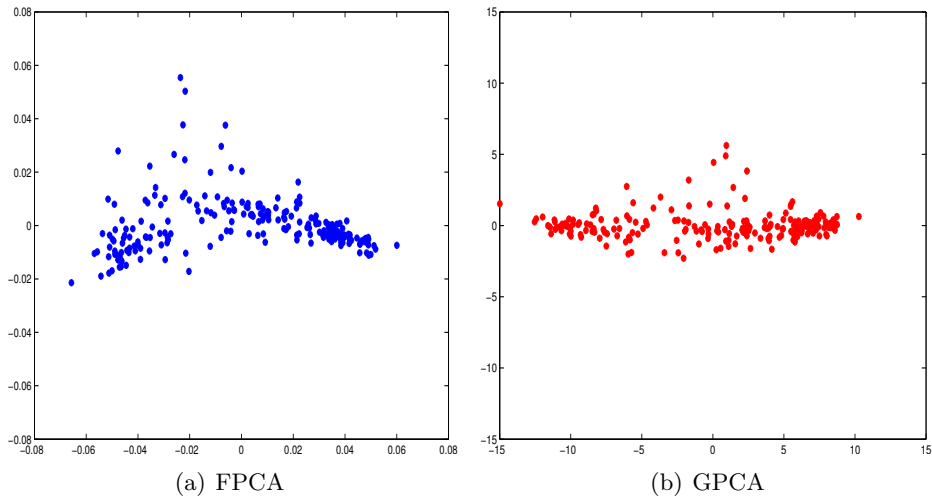
(a) FPCA

(b) GPCA

Figure 9: Population pyramids of 223 countries. (a) Standard PCA: projection of the data on the first PC (81 %) and second PC (8 %). (b) GPCA: projection of the data on the first PC (96 %) and second PC (2 %)
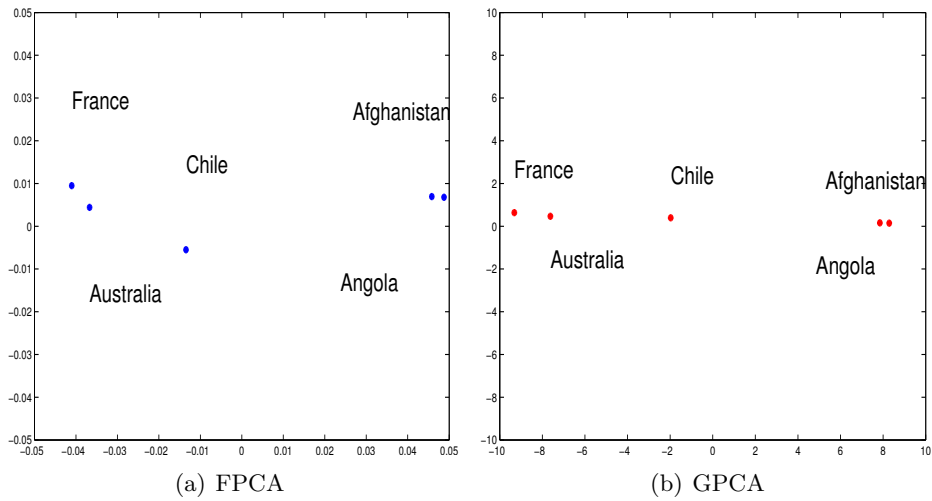

(a) FPCA

(b) GPCA

Figure 10: Projection of the five histograms displayed in Figure 8, using the whole dataset, on the first two PC, from (a) standard PCA and (b) GPCA.

6 Analysis of consistency

6.1 Consistency of the empirical CPCA

Throughout this section we use the notation of Section 3; limits are understood as $n \to \infty$. Let $x_0 = \mathbb{E}\boldsymbol{x}$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent, identically distributed (iid) copies of $\boldsymbol{x}$. Denote by $\bar{\boldsymbol{x}}_n := \sum_{i=1}^{n} \boldsymbol{x}_i / n$ their arithmetic mean and observe that $\bar{\boldsymbol{x}}_n \to x_0$ a.s., by the strong law of

large numbers (SLLN) in a Hilbert space (see [LT11]). Let also $\mathbf{K}_X^{(n)}(C) = \frac{1}{n} \sum_{i=1}^{n} d^2(\boldsymbol{x}_i, C)$ be the random version of $\mathrm{K}_X^{(n)}$.

We prove in Theorem 6.1 that empirical GPCC based on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ converge, in a sense defined below, to GPCC of $\boldsymbol{x}$. The analogous result for NPCC is omitted.

Following Definition 3.5, let $\mathcal{G}_{x_0,k}(X)$ be the set of GPCC of $\boldsymbol{x}$, with reference point $x_0 = \mathbb{E}\boldsymbol{x}$, and $\mathcal{G}_{n,k}(X) := \underset{C \in \mathrm{CC}_{\bar{\boldsymbol{x}}_n,k}(X)}{\arg\min} \mathbf{K}_X^{(n)}(C)$ the (random) set of empirical GPCC of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with $\bar{\boldsymbol{x}}_n$ as reference point.

Definition 6.1. The empirical GPCC are consistent, denoted $\mathcal{G}_{n,k}(X) \to \mathcal{G}_{x_0,k}(X)$ a.s., if for every measurable selection of $C_n \in \mathcal{G}_{n,k}(X), n \geq 1$, and $C \in \mathcal{G}_{x_0,k}(X)$,

(a) $\mathbf{K}_X^{(n)}(C_n) \to \mathrm{K}_X(C)$ a.s., and
(b) the accumulation points of (C_n) belong to $\mathcal{G}_{x_0,k}(X)$ a.s.

In the following lemma we show that the indicators of $\mathrm{CC}_{x_n,k}(X)$ (denoted $\chi_{n,k}$) Γ-converge to the indicator of $\mathrm{CC}_{x,k}(X)$ (denoted χ_k) when $x_n \to x \in X$. We refer to Section A.4 for the definitions of Γ-convergence and indicator.

Lemma 6.1. *Let* $x_n \in X, n \geq 1$, *with* $x_n \to x \in X$. *If* X *is compact then* $\Gamma\text{-}\lim_{n\to\infty} \chi_{n,k} = \chi_k$.

Proof. Recall that under compactness of X, $h(C_n, C) \to 0$ is equivalent to $K\text{-}\lim C_n = C$ (see Section A.3). By Lemma A.5, it is sufficient to show that $\mathrm{CC}_{x_n,k}(X)$ converges to $\mathrm{CC}_{x,k}(X)$ in the sense of Kuratowski. That is, we have to show that:

(a) for every $C \in \mathrm{CC}_{x,k}(X)$ there exist $C_n \in \mathrm{CC}_{x_n,k}(X), n \geq 1$, with $h(C_n, C) \to 0$, and
(b) if C is an accumulation point of $C_n \in \mathrm{CC}_{x_n,k}(X), n \geq 1$, then $C \in \mathrm{CC}_{x,k}(X)$.

For (a) take $C \in \mathrm{CC}_{x,k}(X)$ and let $C_n := C + x_n - x \in \mathrm{CC}_{x_n,k}(X)$, $n \geq 1$. After some calculation we find that the deviations $d(C, C_n)$ and $d(C_n, C)$ (see Definition A.7) are bounded above by $\|x - x_n\|$. Therefore, $h(C, C_n) \leq \|x - x_n\| \to 0$.

For (b) let C be an accumulation point of (C_n). Then, since $x_n \in C_n$ and $x_n \to x$, it follows that $x \in C$, by (ii) in Definition A.6. On the other hand, recall that $\mathrm{CC}_k(X)$ is compact, thanks to Proposition 3.2. Then, as $C_n \in \mathrm{CC}_k(X), n \geq 1$, we have $C \in \mathrm{CC}_k(X)$ and, since $x \in C$, we conclude that $C \in \mathrm{CC}_{x,k}(X)$. $\qquad\square$

Theorem 6.1. *If* X *is compact then* $\mathcal{G}_{n,k}(X) \to \mathcal{G}_{x_0,k}(X)$ *a.s.*

Proof. Let $\chi_{0,k}, \chi_{n,k}$ be the indicators of $\mathrm{CC}_{x_0,k}(X), \mathrm{CC}_{\bar{\boldsymbol{x}}_n,k}(X)$ respectively. Note that

$$\mathcal{G}_{x_0,k}(X) = \underset{C \in \mathrm{CL}(X)}{\arg\min} \mathrm{K}_X(C) + \chi_{0,k}(C) \quad \text{and} \quad \mathcal{G}_{n,k}(X) = \underset{C \in \mathrm{CL}(X)}{\arg\min} \mathbf{K}_X^{(n)}(C) + \chi_{n,k}(C). \quad (6.1)$$

From Lemma 6.1, we have

$$\Gamma\text{-}\lim_{n\to\infty} \chi_{n,k} = \chi_{0,k}, \text{ a.s.}, \tag{6.2}$$

where the Γ-convergence takes place in the space $\mathrm{CL}(X)$. From Proposition A.3 and recalling that X is compact, we have that $d^2(x, C)$ is separately continuous in $x \in X$ and $C \in \mathrm{CL}(X)$.

22

Hence, $d^2(x, C)$ is measurable on the product space $X \times \mathrm{CL}(X)$; see [Joh69] or [Rud81]. Thus, from Theorem 2.3 in [AW95], we have the following Γ-convergence in $\mathrm{CL}(X)$,

$$\Gamma\text{-}\lim_{n\to\infty} \mathbf{K}_X^{(n)}(\cdot) = \mathrm{K}_X(\cdot) \text{ a.s.} \tag{6.3}$$

On the other hand, as X is compact, there exists a constant $R > 0$ such that $d^2(x, C) \le R$, for all $x \in X$ and $C \in \mathrm{CL}(X)$. Also, by Proposition 3.2 , $\mathrm{CL}(X)$ is a compact set. Therefore, by the uniform strong law of large number (see Lemma 2.4 in [NM94]), $\mathbf{K}_X^{(n)}(C) \to \mathrm{K}_X(C)$ uniformly in $\mathrm{CL}(X)$ a.s., that is,

$$\lim_{n\to\infty} \sup_{C \in \mathrm{CL}(X)} |\mathrm{K}_X(C) - \mathbf{K}_X^{(n)}(C)| = 0 \text{ a.s.} \tag{6.4}$$

From (6.2) to (6.4) and by Proposition 6.24 in [DM93], we obtain

$$\Gamma\text{-}\lim_{n\to\infty} \mathbf{K}_X^{(n)} + \chi_{n,k} = \mathrm{K}_X + \chi_{0,k} \text{ a.s.} \tag{6.5}$$

Therefore, from (6.1), (6.5), the compactness of $\mathrm{CL}(X)$ and Theorem A.1, the conclusion follows. $\qquad\square$

6.2   Consistency of the empirical GPCA

In this section we use the notation of Section 4, with $\nu_0 = \nu^*$, the Fréchet mean of $\boldsymbol{\nu}$. Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ be iid copies of $\boldsymbol{\nu}$ and let $\boldsymbol{\nu}_n^*$ be their empirical Fréchet mean. Let $\boldsymbol{x} = \log_\mu(\boldsymbol{\nu}), \boldsymbol{x}_i = \log_\mu(\boldsymbol{\nu}_i), i = 1, \ldots, n$ and $x_0 = \log_\mu(\nu_0)$. Let also $\mathbf{K}_W^{(n)}(G) = \frac{1}{n}\sum_{i=1}^n d_{W_2}^2(\boldsymbol{\nu}_i, G)$ be the random version of $\mathrm{K}_W^{(n)}$. We show the convergence of $\boldsymbol{\nu}_n^*$ and of the empirical GPG to their population counterparts.

Proposition 6.1. $d_W(\boldsymbol{\nu}_n^*, \nu_0) \to 0$ *a.s.*

Proof. From Proposition 4.1(*ii*), $x_0 = \log_\mu(\nu_0) = \mathbb{E}\boldsymbol{x}$ and $\log_\mu(\boldsymbol{\nu}_n^*) = \bar{\boldsymbol{x}}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$. By Theorem 2.2 and the SLLN in a Hilbert space (see [LT11]), $d_W^2(\boldsymbol{\nu}_n^*, \nu_0) = \|\bar{\boldsymbol{x}}_n - \mathbb{E}\boldsymbol{x}\|_\mu^2 \to 0$, a.s. $\qquad\square$

Remark 6.1. As pointed out by an anonymous reviewer, the result in Proposition 6.1 follows from Ziezold's strong law of large number [Zie77].

Recall that if Ω is compact then $W_2(\Omega)$ is compact. In this case $\mathrm{CL}(W)$ is also compact, as can be easily shown from Theorem 2.2 and Proposition 3.2. Therefore, if Ω is compact, then every sequence $G_n \in \mathcal{G}_{n,k}(W), n \ge 1$, has a convergent subsequence in $\mathrm{CL}(W)$.

Let $\mathcal{G}_{\nu_0,k}(W)$ be the set of GPG of $\boldsymbol{\nu}$, with reference measure $\nu_0 = \nu^*$. Let also $\mathcal{G}_{n,k}(W) :=$ $\underset{G \in \mathrm{CG}_{\boldsymbol{\nu}_n^*,k}(W)}{\arg\min} \mathbf{K}_W^{(n)}(G)$ the (random) set of empirical GPG of $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$, with reference measure $\boldsymbol{\nu}_n^*$, and $\mathcal{G}_{n,k}(V_\mu(\Omega)) := \underset{C \in \mathrm{CC}_{\bar{\boldsymbol{x}}_n,k}(V_\mu(\Omega))}{\arg\min} \mathbf{K}_X^{(n)}(C)$.

Definition 6.2. The empirical GPG are consistent, denoted $\mathcal{G}_{n,k}(W) \to \mathcal{G}_{\nu_0,k}(W)$ a.s., if for every $G_n \in \mathcal{G}_{n,k}(W), n \geq 1$, and $G \in \mathcal{G}_{\nu_0,k}(W)$,

(a) $\mathbf{K}_W^{(n)}(G_n) \to \mathrm{K}_W(G)$ a.s., and

(b) the accumulation points of (G_n) belong to $\mathcal{G}_{\nu_0,k}(W)$ a.s.

Theorem 6.2. *If Ω is compact then $\mathcal{G}_{n,k}(W) \to \mathcal{G}_{\nu_0,k}(W)$ a.s.*

Proof. From Proposition 4.3

$$\mathcal{G}_{x_0,k}(V_\mu(\Omega)) = \log_\mu\left(\mathcal{G}_{\nu_0,k}(W)\right) \quad \text{and} \quad \mathcal{G}_{n,k}(V_\mu(\Omega)) = \log_\mu\left(\mathcal{G}_{n,k}(W)\right). \tag{6.6}$$

On the other hand, from Theorem 2.2, it can be shown that $\log_\mu$ is an isometric bijection for the Hausdorff distance, between $\mathrm{CL}(W)$ and $\mathrm{CL}(V_\mu(\Omega))$. Let $G_n \in \mathcal{G}_{n,k}(W), n \geq 1$, with a subsequence $(G_{n'})$ converging to $G \in \mathrm{CL}(W)$. Then, by the continuity of $\log_\mu$, $C_{n'} := \log_\mu(G_{n'}) \to C := \log_\mu(G)$.

From (6.6), we have $C_n \in \mathcal{G}_{n,k}(V_\mu(\Omega)), n \geq 1$. Therefore, Theorem 6.1 implies that $C \in \mathcal{G}_{x_0,k}(V_\mu(\Omega))$ and $\mathbf{K}_X^{(n)}(C_n) \to \mathrm{K}_X(C)$ a.s. Finally, the result follows from the equalities $\mathrm{K}_X(C) = \mathrm{K}_W(G), \mathbf{K}_X^{(n)}(C_n) = \mathbf{K}_W^{(n)}(G_n), n \geq 1$, and (6.6). $\qquad\square$

Remark 6.2. From the proof of Theorem 6.2 one can see that the sequence of reference measures $\boldsymbol{\nu}_n^*$ (the empirical Fréchet means) can be replaced by any other sequence of atomless measures in $W_2(\Omega)$, converging to an atomless limit, say, μ_0. Of course, the limiting GPG would have μ_0 as reference measure.

Remark 6.3. As commented in Remark 1.1, in practical situations we often have access only to, say, n_i random observations from each ν_i. In this context, consistency has to be redefined as n and the n_i tend to infinity.

7   GPCA in $W_2(\Omega)$ and PCA in a Riemannian manifold

As already mentioned in the introduction, nonlinear analogs of PCA have been proposed in the literature for the analysis of data belonging to curved Riemannian manifolds [FLPJ04, HHM10]. To perform a PCA-like analysis two popular approaches are:

1. Standard PCA of the data projected onto the tangent space at their Fréchet mean, with back projection onto the manifold and

2. Principal Geodesic Analysis (PGA), that is, a PCA along geodesics.

Below, we briefly recall the main ideas of these two approaches that generally lead to different directions of geodesic variability in a curved manifold [SLHN10].

Consider $y_1, \ldots, y_n$ belonging to a complete Riemannian manifold $\mathcal{M}$ admitting a geodesic distance $d_\mathcal{M}$. In order to define a PCA like analysis in $\mathcal{M}$, one needs a notion of average. It has been suggested [FLPJ04] that the appropriate notion is the Fréchet mean, defined as an element $z \in \mathcal{M}$ (not necessarily unique) minimizing the sum of squared distances to the data, namely

$$z \in \arg\min_{y \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^n d_\mathcal{M}^2(y, y_i).$$

We refer to [BP03] for details and properties of the Fréchet mean in Riemannian manifolds.

Let $T_z\mathcal{M}$ be the tangent space to $\mathcal{M}$ at z. If v denotes a tangent vector in $T_z\mathcal{M}$, there exists a unique geodesic $\gamma_v(t)$ having v as its initial velocity, where $t \in \mathbb{R}$ is a time parameter. The Riemannian exponential map $\exp_z : T_z\mathcal{M} \to \mathcal{M}$, defined by $\exp_z(v) = \gamma_v(1)$ is a diffeomorphism on a neighborhood of zero and its inverse is the Riemannian log map, denoted by $\log_z$.

(1) PCA via linearization in the tangent space: in this approach the data $y_1, \ldots, y_n$ is first projected on $T_z\mathcal{M}$ by means of the $\log_z$ map, thus obtaining $x_i = \log_z(y_i)$, $i = 1, \ldots, n$. Next, a standard PCA of $x_1, \ldots, x_n$ is performed in the linear space $(T_z\mathcal{M}, \langle \cdot, \cdot \rangle, \| \cdot \|)$, which leads to computing the first principal component v^{lin}, the eigenvector associated with the largest eigenvalue of the covariance operator

$$Kv = \frac{1}{n}\sum_{i=1}^{n}\langle x_i - \bar{x}_n, v\rangle(x_i - \bar{x}_n), \; v \in T_z\mathcal{M},$$

where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Finally, v^{lin} is projected back onto $\mathcal{M}$ by means of the $\exp_z$ map to obtain $w^{lin} = \exp_z(v^{lin})$, which represents a first notion of principal direction of geodesic variability. The main drawback of PGA via linearization is the fact that distances are generally not preserved by the projection step, that is, $\|x_i - x_j\| \neq d_\mathcal{M}(y_i, y_j)$.

(2) PGA on $\mathcal{M}$: the notion of PCA along geodesics on $\mathcal{M}$ is motivated by formulation (1.2), which characterizes standard PCA. In a first step, one computes

$$v^{geo} = \underset{v \in T_z\mathcal{M}, \|v\|=1}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} d_\mathcal{M}^2\left(y_i, G_v\right),$$

where $G_v = \{\exp_z(tv), \; t \in \mathbb{R}\}$ and $d_\mathcal{M}(y, G) = \inf_{y' \in G} d_\mathcal{M}(y, y')$ for $y \in \mathcal{M}$ and $G \subset \mathcal{M}$. Then, in a second (and final) step, one projects the element $v^{geo} \in T_z\mathcal{M}$ onto $\mathcal{M}$, by computing $w^{geo} = \exp_z(v^{geo})$. This yields another notion of principal direction of geodesic variability of the data and generally one has that $w^{lin} \neq w^{geo}$, except if $\mathcal{M}$ is a Hilbert space. Therefore, PCA via linearization on the tangent space and PCA along geodesics may lead to different directions of geodesic variability in a curved manifold. A detailed analysis of the differences between these methods can be found in [SLHN10]. In both methods it is also possible to define subsequent principal directions (second, third, and so on) of geodesic variability in a recursive manner, and we refer to [FLPJ04] for further details.

In this paper, we have considered the analysis of data in the Wasserstein space $W_2(\Omega)$, which is not a Riemannian manifold but has pseudo-Riemannian structure, rich enough to allow the definition of a notion of geodesic PCA. By means of the analogs of the logarithmic and of the exponential maps, we also introduce the corresponding version of the standard PCA in the tangent space, with back projection onto $W_2(\Omega)$, thus establishing a parallel to the methodological duality available for data in Riemannian manifolds, as presented above. Also, as could be expected, these two approaches yield, in general, different forms of geodesic variability.

There is however a significant distinguishing feature of our methodology, namely the possibility of performing a PCA in the tangent space under convexity restrictions, which is equivalent

(after projection) to the geodesic PCA in $W_2(\Omega)$. This motivates the definition of Convex PCA (see Section 3), a general PCA-like method for analyzing data on a closed convex subset of a Hilbert space, which can be of interest beyond its specific application in the context of GPCA. The CPCA applied to the logarithms of the data measures is interesting because it is formally simpler than the geodesic PCA in $W_2(\Omega)$ although more complex than standard PCA. In this respect it is also worth noticing that if the data are "sufficiently concentrated", the standard and the restricted PCA in the tangent space yield the same results.

It should be mentioned that the terminology geodesic PCA (GPCA) was used previously by Huckemann et al. in [HHM10] to denote a Riemannian manifold generalization of linear PCA. Their approach shares similarities with the PGA method introduced in [FLPJ04] but optimizes additionally for the placement of the center point (not necessarily equal to the Fréchet mean). Furthermore, it does not use a linear approximation of the manifold and is only suited for Riemannian manifolds, where explicit formulas for geodesics exist. However, it is difficult to compare our approach to the GPCA in [HHM10] since the notion of principal geodesic, that we propose in this paper, is defined with respect to a given reference measure ν_0 (chosen to be either the population or the empirical Fréchet mean). For a precise comparison it would be necessary to carry out the optimization in Definition 4.3(a), with respect to the reference measure ν_0, a task which is beyond the scope of this paper.

Finally observe that, from Theorem 2.2, one can interpret $W_2(\Omega)$ as a space with no curvature, and hence the pseudo-Riemannian formalism, used in Section 2.1, is not essential for our development. However, such a framework allows making a connection between our approach and PCA methods adapted to Riemannian manifolds.

A Appendix

A.1 Increasing functions and quantiles

We present some useful, well-known results about increasing functions and quantiles. For additional information, see [EH13], [RR14]. In this section μ, ν denote probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, F_ν denotes the (right-continuous) cdf of ν and $L^2(0,1)$ is the space of square-integrable functions, with respect to the Lebesgue measure on $(0,1)$.

Definition A.1. Let $A \subseteq \mathbb{R}$ and $T : A \to \mathbb{R}$.
(a) T is increasing on $B \subseteq A$ if $\forall x, y \in B$, $x < y$ implies $T(x) \leq T(y)$.
(b) T is μ-a.e. increasing if there exists $B_\mu \in \mathcal{B}(\Omega)$, with $B_\mu \subseteq A$, $\mu(B_\mu) = 1$ and T increasing on B_μ.

Remark A.1. A μ-a.e. increasing function $T : A \to \mathbb{R}$ needs not to have a version increasing on A. A version of T is a function $\tilde{T} : A \to \mathbb{R}$ such that $T = \tilde{T}$, μ-a.e.

Definition A.2. The quantile function of ν is defined as $F_\nu^-(y) = \inf\{x \in \mathbb{R} | F_\nu(x) \geq y\}$, $y \in (0,1)$.

Proposition A.1.
(a) F_ν^- *is left-continuous and increasing on* $(0,1)$.

(b) Any left-continuous and increasing $T : (0, 1) \to \mathbb{R}$ is the quantile of some probability ν.
(c) ν has finite second moment if and only if $\int_0^1 (F_\nu^-(x))^2 dx < \infty$.

Proof. See [EH13], [RR14]. $\square$

Lemma A.1. *Let $T \in L^2(0, 1)$ a.e. increasing. Then there exists $\nu \in W_2(\mathbb{R})$ such that $T = F_\nu^-$ a.e.*

Proof. Suppose T is increasing on a full measure set $B \subseteq (0, 1)$ (that is the Lebesgue measure of B is one). Let $\tilde{T} : (0, 1) \to \mathbb{R}$ be defined as $\tilde{T}(x) = T(x)$, for $x \in B$, and $\tilde{T}(x) = \inf_{y \in B, x < y} T(y)$, for $x \notin B$. Then $\tilde{T}$ is increasing in $(0, 1)$ and $\tilde{T} = T$ a.e. Finally, let $\hat{T}$ be the left-continuous version of $\tilde{T}$, that is, $\hat{T}(x) := \lim_{t \to x^-} \tilde{T}(t)$. So, as $\hat{T}$ is left-continuous and increasing on $(0, 1)$, from Proposition A.1(b,c) there exists a probability $\nu \in W_2(\mathbb{R})$, such that $F_\nu^- = \hat{T}$. Finally, since the number of discontinuities of any increasing function is countable, we have $\hat{T} = \tilde{T}$ a.e. $\square$

Proposition A.2. *Let Ω be an interval of real numbers (not necessarily bounded). Then, the set of quantile functions $\{F_\nu^- | \nu \in W_2(\Omega)\}$ is closed and convex in $L^2(0, 1)$.*

Proof. For convexity let $\alpha \in (0, 1)$ and $\nu_1, \nu_2 \in W_2(\Omega)$. Then $T_\alpha := \alpha F_{\nu_1}^- + (1 - \alpha) F_{\nu_2}^-$ is increasing, left-continuous and square integrable. Hence, by Proposition A.1(b,c), T_α is the quantile of some $\nu_\alpha \in W_2(\Omega)$. For closedness consider a sequence (ν_n) in $W_2(\Omega)$, such that $\int_0^1 (F_{\nu_n}^-(x) - T(x))^2 dx \to 0$, as $n \to \infty$. Then, there exists a subsequence (ν_{k_n}) of (ν_n) such that $F_{\nu_{k_n}}^- \to T$ a.e. and hence, T is square-integrable and a.e. increasing. So, by Lemma A.1, T is a quantile. As usual, the elements of $L^2(0, 1)$ are understood as equivalence classes. $\square$

A.2 Geodesics in metric spaces

We introduce the concept of geodesic in metric spaces. For notations, definitions and results, we follow [Cho11] and references therein. For convenience, without loss of generality, we consider I such that $[0, 1] \subset I$.

Definition A.3. A curve in a metric space (X, d) is a continuous function $\gamma : I \to X$, where $I \subset \mathbb{R}$ is a closed (not necessarily bounded) interval. Also
(i) γ is said to pass through $z \in X$ if $\gamma(t) = z$, for some $t \in I$;
(ii) γ joins $x, y \in X$ if there exists $a, b \in I$, such that $\gamma(a) = x$ and $\gamma(b) = y$ and
(iii) γ is rectifiable if its length $L(\gamma)$ is finite.

Definition A.4. A metric space (X, d) is said to be geodesic if for every $x, y \in X$, there exists a rectifiable curve γ joining x and y, such that $d(x, y) = L(\gamma)$. Such minimum length curve γ is called a shortest path between x and y. A curve $\gamma : I \to X$ is a geodesic if for every $t \in I$, there exist $a, b \in I, a < b, a \le t \le b$ such that the restriction of γ to $[a, b]$ is a shortest path between $\gamma(a)$ and $\gamma(b)$.

The following is a useful characterization of shortest path (See [Cho11] for a proof).

Lemma A.2. *For any shortest path, there exists a continuous reparametrization γ on $[0, 1]$ such that*

$$d(\gamma(s), \gamma(t)) = |t - s| d(\gamma(0), \gamma(1)) \text{ for all } s, t \in [0, 1].$$

Lemma A.3. *Let H be a Hilbert space and $x, y \in H$. Then γ is a shortest path joining x and y if and only if $\gamma(t) = (1 - t)x + ty$, for all $t \in [0, 1]$, up to a continuous reparametrization.*

Proof. Denote the inner product and the induced norm in H by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively. Let γ be a shortest path between x and y, and $t \in [0, 1]$. After a reparametrization such that $\gamma(0) = x$ and $\gamma(1) = y$, from Lemma A.2 we have $\|x - \gamma(t)\| = t\|x - y\|$ and $\|\gamma(t) - y\| = (1 - t)\|x - y\|$, then $\|x - \gamma(t)\| + \|\gamma(t) - y\| = \|x - y\|$.

Squaring and simplifying the former expression above, we obtain $\|x - \gamma(t)\|\|\gamma(t) - y\| = \langle x - \gamma(t), \gamma(t) - y \rangle$. Hence, by the Cauchy-Schwartz inequality, there exists $\lambda \geq 0$ such that $x - \gamma(t) = \lambda(\gamma(t) - y)$. Finally, taking norm we find $\lambda = \frac{t}{1-t}$ and the result follows. The other implication is direct. $\qquad\square$

From the previous lemma we deduce that, in Hilbert spaces, any geodesic is locally a segment and so, geodesics are straight lines. We state this in the following corollary.

Corollary A.1. *Let H be a Hilbert space and $\gamma : I \to H$ a curve, such that $\gamma(0) = x$ and $\gamma(1) = y$. Then γ is a geodesic if and only if $\gamma(t) = (1 - t)x + ty$, for all $t \in I$, up to a continuous reparametrization.*

Definition A.5. *Let (X, d) be a geodesic space and $Y \subset X$. We say that Y is geodesic if the induced metric space (Y, d) is geodesic. In other words, if for any $x, y \in Y$, there exists a shortest path joining x and y, totally contained in Y.*

Note from Lemma A.3 that a Hilbert space H is geodesic and $C \subset H$ is geodesic if and only if C is convex.

A.3 K-convergence

In this section we present definitions and results that we use for proving the existence of principal geodesics (see Section 4.2). In particular, we define an appropriate concept of convergence for sequences of convex sets in a metric space (X, d).

Definition A.6. *Let $C, C_n \subset X, n \geq 1$. We say that the sequence (C_n) converges to C in the sense of Kuratowski, denoted by $K\text{-}\lim_{n \to \infty} C_n = C$, if*
(i) for all $x \in C$, there exist $x_n \in C_n, n \geq 1$, such that $x_n \to x$ and
(ii) for all $x_n \in C_n, n \geq 1$, and for any accumulation point x of (x_n), $x \in C$.

Definition A.7. *The deviation from $x \in X$ to $B \subseteq X$ is defined by $d(x, B) := \inf_{x' \in B} d(x, x')$; the deviation from $A \subseteq X$ to B is $d(A, B) := \sup_{x \in A} d(x, B)$ and the Hausdorff distance between the sets A and B is*

$$h(A, B) := \max\{d(A, B), d(B, A)\}. \tag{A.1}$$

Remark A.2. It is well known (see [Pri40, Bee85] and references therein) that convergence with respect to the Hausdorff distance is stronger than convergence in the sense of Kuratowski. Moreover, if X is compact both notions of convergence coincide.

Definition A.8. We define the metric space $\mathrm{CL}(X)$ as the set of nonempty, closed subsets of X, endowed with the Hausdorff distance h.

Proposition A.3. *For all $x \in X$, then $d(x, \cdot)$ is continuous on $\mathrm{CL}(X)$.*

Proof. Observe that, for all $x \in X$ and $A, B \in \mathrm{CL}(X)$, $d(x, A) \leq d(x, B) + h(A, B)$. Then $|d(x, A) - d(x, B)| \leq h(A, B)$ and the conclusion follows. $\square$

Lemma A.4. *Let $B, C, B_n, C_n \subset X$, with $B_n \subset C_n$, $n \geq 1$, such that $K\text{-}\lim_{n\to\infty} B_n = B$ and $K\text{-}\lim_{n\to\infty} C_n = C$. Then $B \subset C$.*

Proof. By Definition A.6(i), for any $x \in B$ there exist $x_n \in B_n, n \geq 1$, such that $x_n \to x$. As $x_n \in B_n \subset C_n$, $n \geq 1$, from Definition A.6(ii) we have $x \in C$. $\square$

A.4 Γ-convergence

The notion of Γ-convergence in a metric space (X, d) ([Att84, DM93]) is used in the proof of Theorem 6.2.

Definition A.9. Let $F, F_n : X \mapsto \overline{\overline{\mathbb{R}}} := \mathbb{R} \cup \{+\infty, -\infty\}, n \geq 1$, a sequence of functions. We say that (F_n) Γ-converges to F, denoted $\Gamma\text{-}\lim_{n\to\infty} F_n = F$, if, for every $x \in X$,
(i) $F(x) \leq \liminf_{n\to\infty} F_n(x_n)$, for any $x_n \in X, n \geq 1$, with $x_n \to x$, and
(ii) there exist $x_n \in X, n \geq 1$, with $x_n \to x$, such that $F(x) = \lim_{n\to\infty} F_n(x_n)$.

Definition A.10. For $F : X \to \overline{\overline{\mathbb{R}}}$, let $M(F) := \{x \in X : F(x) = \inf_{y \in X} F(y)\}$.

The following result (see [DM93], Theorems 7.8 and 7.23) shows that Γ-convergence together with compactness (or more generally equicoercivity) implies convergence of minimum values and minimizers.

Theorem A.1. *Assume that X is compact and let $F, F_n : X \mapsto \overline{\overline{\mathbb{R}}}, n \geq 1$, such that $\Gamma\text{-}\lim_{n\to\infty} F_n = F$. Then $M(F)$ is nonempty and*

$$\lim_{n\to\infty} \inf_{x \in X} F_n(x) = \min_{x \in X} F(x).$$

Moreover, if $x_n \in M(F_n), n \geq 1$, then the accumulation points of (x_n) belong to $M(F)$.

Definition A.11. The indicator of $A \subset X$ is the function $\chi_A : X \to \mathbb{R} \cup \{+\infty\}$ defined by $\chi_A(x) = 0$, if $x \in A$, and $\chi_A(x) = +\infty$, if $x \notin A$.

The following Proposition (see [Att84], Proposition 4.15.) shows the relation between K-convergence (see Definition A.6) and Γ-convergence.

Lemma A.5. *Let $A, A_n \subset X$, $n \geq 1$. Then $K\text{-}\lim_{n\to\infty} A_n = A$ if and only if $\Gamma\text{-}\lim_{n\to\infty} \chi_{A_n} = \chi_A$.*

References

[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.

[Att84]   H. Attouch. *Variational Convergence for Functions and Operators.* Applicable Mathematic series. Pitman, London, 1984.

[AW95]   Z. Artstein and R. J. B. Wets. Consistency of minimizers and the SLLN for stochastic programs. *J. Convex Anal.*, 2(1-2):1–17, 1995.

[Bee85]   G. Beer. On convergence of closed sets in a metric space and distance functions. *Bulletin of the Australian Mathematical Society*, 31:421–432, 6 1985.

[BK12]   J. Bigot and T. Klein. Consistent estimation of a population barycenter in the Wasserstein space. *Preprint, ArXiv/1212.2562*, 2012.

[BP03]   R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of Statistics*, 31(1):1–29, 2003.

[Bre91]   Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[Cho11]   O. Chodosh. Optimal transport and Ricci curvature: Wasserstein space over the interval. *Preprint, ArXiv/1105.2883*, 2011.

[Del11]   P. Delicado. Dimensionality reduction when data are density functions. *Comput. Statist. Data Anal.*, 55(1):401–420, 2011.

[DM93]   G. Dal Maso. *An Introduction to Γ-convergence*, volume 8 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser, Boston, MA, 1993.

[DPR82]   J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.*, 12(1):136–154, 1982.

[EH13]   P. Embrechts and M. Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77:423–432, 2013.

[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.

[GLM13]   S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical biosciences*, 242(2):129–142, 2013.

[HHM10] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, 20:1–100, 2010.

[HT13] N. N. Hai and An P. T. A generalization of Blaschke's convergence theorem in metric spaces. *Journal of Convex Analysis*, to be published, 2013.

[Joh69] B. E. Johnson. Separate continuity and measurability. *Proceedings of the American Mathematical Society*, 20(2):420–422, 1969.

[KU01] A. Kneip and K. J. Utikal. Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.*, 96(454):519–542, 2001. With comments and a rejoinder by the authors.

[LT11] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[NM94] W. K. Newey and D. McFadden. Large Sample Estimation and Hypothesis Testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam, 1994.

[Pri40] B. Price. On the completeness of a certain metric space with an application to Blaschke's selection theorem. *Bull. Amer. Math. Soc.*, 46:278–280, 1940.

[RR14] R.T. Rockafellar and J.O. Royset. Random variables, monotone relations, and convex analysis. *Math. Program.*, pages 1–35, 2014.

[RS05] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[Rud81] W. Rudin. Lebesgue's first theorem. In *Mathematical analysis and applications, Part B*, volume 7b of *Advances in Math. Suppl. Stud*, pages 741–747. Academic Press, 1981.

[Sil96] B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1):1–24, 1996.

[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.

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

[Zie77] Herbert Ziezold. On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Trans. 7th Prague Conf. Inf. Theory, Stat. Dec. Func., Random Processes*, volume A, pages 591–602. Reidel, Dordrecht, 1977.

[ZM11]   Z. Zhang and H.-G. Müller. Functional density synchronization. *Comput. Stat. Data Anal.*, 55(7):2234–2249, 2011.