Fast PCA in 1-D Wasserstein Spaces via B-splines Representation and Metric Projection

Matteo Pegoraro[1] **and Mario Beraha**[2, 3]
[1] MOX - Department of Mathematics, Politecnico di Milano
[2] Department of Mathematics, Politecnico di Milano
[3] Department of Computer Science, Università di Bologna
{matteo.pegoraro, mario.beraha}@polimi.it

Abstract

We address the problem of performing Principal Component Analysis over a family of probability measures on the real line, using the Wasserstein geometry. We present a novel representation of the 2-Wasserstein space, based on a well known isometric bijection and a B-spline expansion. Thanks to this representation, we are able to reinterpret previous work and derive more efficient optimization routines for existing approaches. As shown in our simulations, the solution of these optimization problems can be costly in practice and thus pose a limit to their usage. We propose a novel definition of Principal Component Analysis in the Wasserstein space that, when used in combination with the B-spline representation, yields a straightforward optimization problem that is extremely fast to compute. Through extensive simulation studies, we show how our PCA performs similarly to the ones already proposed in the literature while retaining a much smaller computational cost. We apply our method to a real dataset of mortality rates due to Covid-19 in the US, concluding that our analyses are consistent with the current scientific consensus on the disease.

1 Introduction

In many fields of machine learning and statistics, comparing, averaging, and, broadly speaking, performing inference on a set of distributions is an ubiquitous but arduous task. In some cases, the single datum itself can be seen as a distribution, as in the analysis of images (Cuturi and Doucet 2014), census data (Cazelles et al. 2017), econometric surveys (Potter et al. 2017) and process monitoring (Hron et al. 2014). In others, distributions could be the output of different machine learning models, and one wants to aggregate the outputs as in (Srivastava et al. 2015).

The possibility to carry out meaningful statistical analysis on distributions depends on the tools that can be employed, with a focus on the ability to scale to big dataset and the interpretability of the obtained results. Along this line, among the statistical tools which can be defined on a space, Principal Components Analysis (PCA) is one of the richest ones. One of the key features of PCA, which contributes to its popularity among practitioners, is its interpretability, since it provides insights into the variability of the data. Moreover, by projecting the data into the Euclidean space of the scores relative to the principal components, one can apply well established machine learning techniques, defined on multivariate data, also to more complex data such as distributions.

Optimal Transport (OT) defines a natural framework to compare different probability measures, as testified by the huge interest sparkled on different topics related to OT in recent years, especially in the field of machine learning (Cuturi and Doucet 2014; Genevay, Peyré, and Cuturi 2018; Cuturi, Teboul, and Vert 2019). Closely related to OT, lies the definition of Wasserstein distances and, henceforth, Wasserstein spaces (Villani 2008; Ambrosio, Gigli, and Savaré 2008).

PCA in Wasserstein spaces has already been investigated in (Bigot et al. 2017; Cazelles et al. 2017), where the authors define the concepts of *geodesic* and *log* PCA, in analogy to the definitions for Riemannian manifolds. In their work, it is clearly shown how the *log* PCA loses its effectiveness as soon as data is not well concentrated around the barycenter. On the other hand, the *geodesic* PCA is less sensitive to this kind of issues, but can become impractical due to the complex optimization problems that need to be solved.

The contribution of this work is two-folded. First we show how by employing a suitable B-spline representation of the space of quantile functions, the optimization problems required for the *geodesic* PCA can be solved in a much more efficient manner. Despite the improvements obtained with this representation, the computational cost to perform *geodesic* PCA on a real dataset can still be burdensome. This motivates our second contribution, that is an alternative definition of PCA in Wasserstein spaces, that we call *projected* PCA. Our methodology is substantially different from both the *geodesic* and *log* PCA, while being connected to both. In particular, our insight is that our *projected* PCA shares the same spirit of the *geodesic* approach in preserving the metric structure, while leading to a much simpler optimization problem required to find the principal components, as in the case of *log* PCA.

2 Related Works

As already mentioned, at least two definitions of PCA on distributions are available using the geometry of the Wasserstein space. However, several other works have proposed different PCAs on distributions, outside the domain of Wasserstein spaces. Here, we make a brief review of these methods.

In the context of histogram data, symbolic data analysis (SDA) has been employed to perform PCA in (Nagabhushan and Kumar 2007; Rodrıguez, Diday, and Winsberg 2000; Le-Rademacher and Billard 2017). Moreover, in (Verde, Irpino, and Balzanella 2015) some of these attempts have also been extended to work with generic distributional data using Wasserstein metrics. However, the components of PCA in SDA are difficult even to be represented, providing an efficient framework but with low interpretability.

When working with absolutely continuous distributions, hence endowed with a probability density function (pdf), the techniques generally employed derive from functional data analysis (FDA), as in (Delicado 2011; Kneip and Utikal 2001), by considering, for instance, the pdfs as functions in Hilbert spaces like L_2. Despite its simplicity (FDA techniques are well established and readily available in various software packages), it is clear that this approach has a major drawback in that it completely disregards the structure of the space of distributions. This is due to the fact that the linear structure of the traditionally employed functional spaces cannot capture the constrained nature of probability densities. Moreover, we also notice that functional PCA under the L_2 metric is not always easy to interpret as shown in (Colosimo and Pacella 2007).

In an attempt to overcome this drawback, while continuing to use FDA tools, (Hron et al. 2014) employs a transformation that maps the space of continuous distributions endowed with the Aitchison geometry (called Bayes space), into L_2 through an isometric bijection. In such a way, the transformed densities are analyzed using the L_2 metric and the results are mapped back to the original Bayes space via the inverse map. However Bayes spaces are defined only for densities, and not for discrete measures, moreover all the densities must share the same support (the smallest closed set with probability 1), which must also be a compact interval in order for the maps employed to be well defined. Of course, this assumption is hardly verified in practice.

On the other hand, there have already been proposals to define a PCA in Wasserstein spaces that avoid all the drawbacks of the more traditional approaches discussed above. In particular, the authors in (Bigot et al. 2017; Cazelles et al. 2017) have defined *geodesic* and *log* PCA, translating the corresponding techniques available for Riemannian Manifolds into the Wasserstein setting. As shown in (Bigot et al. 2017), the *geodesic* PCA enjoys several nice theoretical properties, but the optimization problems to be solved in order to find the principal components is highly nonlinear and requires complex and time consuming optimization routines to be carried out. Moreover, as discussed in (Cazelles et al. 2017), despite the computational advantages of the *log* PCA, the fact that the Wasserstein space is not a manifold causes some limitations. In particular, the *log* principal components are not always accurate and may sometimes be hard to interpret.

3 Preliminaries

In the remaining of the paper, we will denote with $\mathcal{P}(\mathbb{R})$ the space of probability measures on $\mathbb{R}$ and with μ and ν generic probability measures in $\mathcal{P}(\mathbb{R})$. For any fixed probability measure μ let F_μ be its cumulative distribution function (cdf) and let F_μ^- be the associated quantile function, i.e. the pseudo inverse of the cdf. We also denote with p_1 and p_2 the canonical projection operators, $p_1 : (x, y) \mapsto x$ and $p_2 : (x, y) \mapsto y$.

Given a probability measure μ and a function $f : \mathbb{R} \to \mathbb{R}$, we denote with $f\#\mu$ the pushforward of μ through f, defined by the identity $(f\#\mu)(B) = \mu(f^{-1}(B))$ for every measurable B, where f^{-1} denotes the preimage of f.

3.1 Wasserstein Spaces

We give a brief summary of the main mathematical definitions and results needed to develop our methodology.

Definition 1. *(Adapted from (Villani 2008) Def. 6.1) Let $\Gamma(\mu, \nu)$ be the set of probability measures on $\mathbb{R} \times \mathbb{R}$ with marginals μ and ν, i.e. for every $\gamma \in \Gamma$, $p_1\#\gamma = \mu$ and $p_2\#\gamma = \nu$. The squared 2–Wasserstein distance between μ and ν is:*

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x - y|^2 d\gamma(x, y) \qquad (1)$$

Observe that (1) can be unbounded. In order for W_2 to define a metric, we shall restrict ourselves to the space of probability measures with finite second moment, as in the following definition.

Definition 2. *The Wasserstein space $\mathcal{W}_2(\mathbb{R})$ is defined as*

$$\mathcal{W}_2(\mathbb{R}) = \{\mu \in \mathcal{P}(\mathbb{R}) : \mathbb{E}_\mu[X^2] < \infty\} \qquad (2)$$

Equation (1) is due to Kantorovich and is the weak formulation of Monge's optimal transport problem:

$$\inf_{T:T\#\mu=\nu} \int_{\mathbb{R}} |x - T(x)|^2 d\mu(x) \qquad (3)$$

which can be ill posed.

The following theorems are the base tools needed to define our PCA.

Theorem 1. *((Ambrosio, Gigli, and Savaré 2008) Theorem 6.0.2) Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$.*

1. *We have*

$$\min_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 d\gamma(x, y) = \int_0^1 |F_\mu^-(s) - F_\nu^-(s)|^2 ds$$

2. *If μ has not atom, i.e. F_μ is continuous, then $T_\mu^\nu := F_\nu^- \circ F_\mu$ is the unique optimal transport map in* (3)*.*

Following (Panaretos and Zemel 2020) Sec 2.3.2, it is possible to embed $\mathcal{W}_2(\mathbb{R})$ into $L_2([0, 1])$ through the isometric map $\mu \mapsto F_\mu^-$. Indeed, one has that F_μ^- is square integrable on $[0, 1]$ if and only if $\mathbb{E}_\mu[X^2] < +\infty$. The image of this map, $L_2([0, 1])^\uparrow \subset L_2([0, 1]$ is the closed convex set given by (equivalence classes of) left-continuous non decreasing functions on $[0, 1]$.

The geodesics in $\mathcal{W}_2(\mathbb{R})$ are obtained by pushing forward straight lines between functions in $L_2([0, 1])$, see (Panaretos and Zemel 2020), i.e. the line between F_μ^- and F_ν^- gives the unique geodesic between μ and ν:

$$\beta(t) = (tF_\mu^- + (1 - t)F_\nu^-)\#\mathcal{U}([0, 1]),$$

where $\mathcal{U}([0,1])$ denotes the uniform distribution on $[0,1]$.

Note that inverse of the map $\mu \mapsto F_\mu^-$ is given by $F_\mu^- \mapsto F_\mu^- \# \mathcal{U}([0,1])$. For coherence of notation with (Bigot et al. 2017), we call $log : \mathcal{W}_2(\mathbb{R}) \to L_2([0,1])$ such that $log(\mu) = F_\mu^-$ and $exp : L_2([0,1]) \to \mathcal{W}_2(\mathbb{R})$, with $exp(f) = f \# \mathcal{U}([0,1])$.

Hence, through the use of these maps, we can carry out statistical inference in the space $L_2([0,1])^\uparrow$ and retrieve the corresponding results in $\mathcal{W}_2(\mathbb{R})$ through the exp map.

3.2 Quadratic B-Splines

In this section we give a very handy representation of the space $L_2([0,1])^\uparrow$ of non decreasing functions on $[0,1]$.

Proposition 1. *Let $\{\psi_j^k\}_{j=1}^J$ be a basis of B-splines of order k defined over the knots $x_1, \dots, x_{J+k+2}$. Let $f(x) = \sum_{j=1}^J a_j \psi_j^k(x)$, then:*

1. *If the coefficients $\{a_j\}$ are monotonically increasing (decreasing) f is monotonically increasing (decreasing).*
2. *If $k = 2$, then the statement in Item 1. holds with an "if and only if".*

This result implies that any finite sum of quadratic B-splines which is in $L_2([0,1])^\uparrow$, lies in the convex polytope defined by the linear equations which constrain the coefficients to be non decreasing. Moreover any such truncation is obviously left-continuous and so it represents the quantile function of some probability distribution. Being monotone splines dense in $L_2([0,1])^\uparrow$ we can use them to approximate any quantile function. In what follows we will focus on quadratic splines, i.e. fix $k = 2$ and omit the superscript k when referring to spline functions.

Having fixed $J \in \mathbb{N} > 0$ and a quadratic B-spline basis $\{\psi_j\}_{j=1}^J$, we call $L_2([0,1])^{J\uparrow}$ the space of monotone splines spanned by such basis, that is, the space of quantile functions of the kind:

$$F^-(x) = \sum_{j=1}^J a_j \psi_j(x) \quad \text{s.t. } a_i - a_{i-1} \geq 0$$

for every $i = 2, \dots, J$. These constraints can be equivalently written as:

$$\{G \cdot [a_1, \cdots, a_J]^T\}_i \geq 0 \quad i = 2, \dots, J$$

where G is the $(J-1) \times J$ matrix with entries g_{ij} such that $\sum_j g_{ij} \cdot a_j = a_i - a_{i-1}$.

Remark 1. *Define $\mathbb{R}^{J\uparrow} \subset \mathbb{R}^J$ the convex polytope given by all vectors with non decreasing coefficients. By fixing $J \in \mathbb{N} > 0$ and a quadratic B-spline basis $\{\psi_j\}_{j=1}^J$ the space $L_2([0,1])^{J\uparrow}$ is isomorphic to $\mathbb{R}^{J\uparrow}$ endowed with the metric induced by $E = \{e_{lm}\}$ where $e_{lm} =< \psi_l, \psi_m >_{L_2}$.*

4 Principal Component Analysis

As already mentioned, (Bigot et al. 2017) exploits exp and log maps to give different definitions of PCA: a *geodesic* PCA and a *log* PCA, both translating into the Wasserstein space tools developed for Riemannian manifolds (Huckemann, Hotzand, and Munk 2010; Patrangenaru and Ellingson 2015). We refer to their papers for a comprehensive and detailed description.

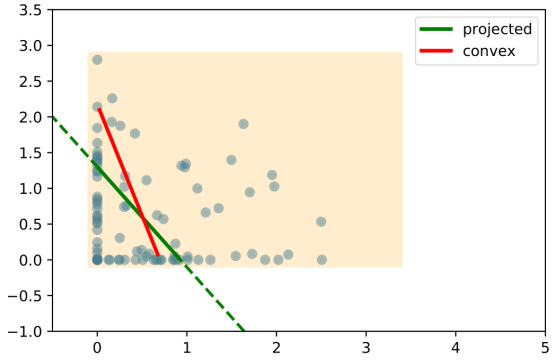

Figure 1: An example in which solving problem (5) differs from problem (6), with the constraint's polytope being the colored rectangle. Notice that the difference is caused by the amount of points on the borders of the politope.

In this section, we summarize the definitions of *global* and *nested* PCA given in (Bigot et al. 2017; Cazelles et al. 2017) and reformulate them in light of the spline basis we introduced above, deriving alternative optimization problems. Furthermore, we propose another definition of PCA, called *projected* PCA, and show how this definition yields a straightforward optimization problem.

4.1 Geodesic PCA

The *geodesic* PCA in $\mathcal{W}_2(\mathbb{R})$ amounts to finding closed convex subsets in $L_2([0,1])^\uparrow$ minimizing the average distance from the data (mapped in $L_2([0,1])$ with the log map).

Remark 2. *The choice of restricting the search to convex sets is due to the fact that geodesic subsets (i.e. sets which contains all the geodesics connecting any couple of points in the set) of $\mathcal{W}_2(\mathbb{R})$ are mapped to convex subsets of $L_2([0,1])^\uparrow$ through the* log *map (Bigot et al. 2017).*

In this context, the dimension of a closed convex subset is defined as the dimension of the smallest affine subset of $L_2([0,1])$ containing it and the distance of a point x from a closed convex set C is the minimum distance from x to a point in C. Moreover, a k dimensional principal component (PC) refers to a closed convex set in $L_2([0,1])^\uparrow$ of dimension k and not to a set of k orthonormal elements, as instead is the case for standard PCA in Euclidean spaces.

As in (Cazelles et al. 2017), we distinguish between a *global* approach and an *iterative* (or *nested*) one to the problem. In the following let $F_1^-, \cdots, F_n^-$ denote a set of n quantile functions, each one associated to one data point, and $F_0^- \in L_2([0,1])^\uparrow$ be another quantile function to be considered as the "center" of the PCA. We report the definitions from (Bigot et al. 2017).

Definition 3. *(Global geodesic PCA) A (k, F_0^-)-global geodesic PC is a set C^* minimizing $n^{-1} \sum_{i=1}^n d(F_i^-, C)$ over the closed convex sets $C \subset L_2([0,1])^\uparrow$ such that $F_0^- \in C$ and $dim(C) \leq k$*

Definition 4. *(Nested geodesic PCA) A (k, F_0^-)-nested geodesic PC is a set C_k^* such that C_k^* is a minimizer*

of $n^{-1}\sum_{i=1}^{n} d(F_i^-, C)$ *over the closed convex sets* $C \subset L_2([0,1])^{\uparrow}$ *such that* $F_0^- \in C$, $dim(C) \leq k$, *and* $C \supset C^*_{k-1}$, *where* C^*_{k-1} *is a* $(k-1, F_0^-)$-nested *geodesic PC. When* $k = 0$, *the PC coincides with* $\{F_0^-\}$.

4.2 Projected PCA

In this section, we propose an alternative way to use the log map to define our *projected* PCA. The idea is quite natural, and is to exploit the Hilbert structure of $L_2([0,1])$, where the PCA is well defined and easily available (see Ramsay 2004), and the fact that the metric projection on a convex set always exists and is unique. For a given set of quantile functions $F_1^-, \ldots, F_n^-$, call $U_k = \{u_1, \ldots, u_k\}$ the principal components of the $L_2([0,1])$-PCA. Given the PCs in $L_2([0,1])$, we project their span $Sp(\{u_1, \ldots, u_k\})$ onto $L_2([0,1])^{\uparrow}$ and then find the biggest closed convex subset of this projection. We formalize the definition as follows:

Definition 5. *(Projected PCA) A* (k, F_0^-)-projected *PC is the biggest closed convex subset* C^* *of* $L_2([0,1])^{\uparrow}$ *such that: (i)* $F_0^- \in C^*$, *(ii)* $dim(C^*) = k$ *and (iii)* $C^* \subseteq \Pi(Sp(U_k))$, *where* Π *denotes the metric projection onto* $L_2([0,1])^{\uparrow}$.

We observe that simply considering the projection $\Pi(Sp(U_k))$ cannot define a valid PCA. Indeed, the projection of $Sp(U_k)$ might have a dimension greater than k and, moreover, it is not guaranteed to be a convex subset. To see this, think for example of projecting the line $y = -x$ onto the first quadrant in $\mathbb{R}^2$.

It is clear from the definition that, in general, the *projected* PCA is less respectful of the metric structure of the Wasserstein space than the *geodesic* ones. We will comment more on such differences in the next section.

Remark 3. *We point out that* projected *PCA is very different from* log *PCA defined in (Bigot et al. 2017; Cazelles et al. 2017) since the composition* $exp \circ log : L_2([0,1]) \to L_2([0,1])^{\uparrow}$ *is by no means a projection. As a consequence, the* log*-principal directions might even end up not being geodesics, as opposed to the* projected *ones which are surely geodesics thanks to Remark 2. See also (Pegoraro and Beraha 2021).*

4.3 Computing the PCAs in Practice

In their work, (Cazelles et al. 2017) show how the definitions of *global* and *nested* PCA can be translated into numerical optimization problems, which however are hard to solve. Here, we firstly show how our spline representation yields more tractable optimization routines for both the *global* and *nested* PCA and secondly demonstrate that the *projected* PCA instead can be solved in a straightforward way.

Let $\{\psi_j\}_{j=1}^J$ be a fixed quadratic B-spline basis and denote by $\boldsymbol{a}_i = \{a_{ij}\}_j$ and $\boldsymbol{a}_0 = \{a_{0j}\}_j$ the coefficients associated to F_i^- and F_0^- respectively, i.e. $F_i^- = \sum_{j=}^{J} a_{ij}\psi_j$.

Thanks to Remark 1, we can develop our methodologies in $\mathbb{R}^J$, considering the metric induced by E instead of the usual one. Indeed, given a vector $\boldsymbol{w} \in \mathbb{R}^J$, we can identify the corresponding function in L_2 by the map $\boldsymbol{w} \mapsto \sum_{j=}^{J} w_j\psi_j$.

The next three propositions formalize the optimization problems that need to be solved. Proofs are given in the Supplementary Material (Pegoraro and Beraha 2020).

Proposition 2. *(Global geodesic PCA) A* k *dimensional* global *geodesic PC centered in* $\boldsymbol{a}_0$ *is the subset of* $\mathbb{R}^{J\uparrow}$ *spanned by* $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k\}$, *linearly independent, which solve:*

$$\underset{\{\boldsymbol{\lambda}_i\}_1^n, \{\boldsymbol{w}_j\}_1^k}{\arg\min} \sum_{i=1}^{n} ||\boldsymbol{a_i} - \boldsymbol{a_0} - \sum_{j=1}^{k} \lambda_{ij} \cdot \boldsymbol{w}_j||_E^2 \qquad (4)$$

subject to: $G\left(\sum_j \lambda_{ij}\boldsymbol{w}_j + \boldsymbol{a}_0\right) \geq 0$.

Proposition 3. *(Nested geodesic PCA) With the same notation as above, a* k *dimensional* nested *geodesic PC, centered in* $\boldsymbol{a}_0$ *is the set spanned by* $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k\}$ *in* $\mathbb{R}^{J\uparrow}$, *where the* $\boldsymbol{w}_i$*s are found recursively from* $\boldsymbol{w}_1$ *to* $\boldsymbol{w}_k$, *such that* $\boldsymbol{w}_h$, *for every* h, *is a solution of:*

$$\underset{\{\boldsymbol{\lambda}_{ih}\}_{i=1}^n, \boldsymbol{w}_h}{\arg\min} \sum_{i=1}^{n} ||\boldsymbol{a_i} - \boldsymbol{a_0} - \lambda_{ih}\boldsymbol{w}_h||_E^2 \qquad (5)$$

subject to: $||\boldsymbol{w}_h||_E = 1$, $< \boldsymbol{w}_j, \boldsymbol{w}_h >_E = 0$ *for all* $j = 1, \ldots, h-1$ *and* $G\left(\lambda_{ih}\boldsymbol{w}_h + \boldsymbol{a_0}\right) \geq 0$.

Proposition 4. *(Projected PCA) A* k *dimensional* projected *PC, centered in* $\boldsymbol{a}_0$, *is the set spanned by* $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k\}$ *in* $\mathbb{R}^{J\uparrow}$, *where the* $\boldsymbol{w}_i$*s are found recursively from* $\boldsymbol{w}_1$ *to* $\boldsymbol{w}_k$, *such that* $\boldsymbol{w}_h$ *is a solution, for every* h, *of:*

$$\underset{\boldsymbol{w_h}:||\boldsymbol{w}_h||_E=1}{\arg\max} \sum_{i=1}^{n} | < \boldsymbol{a}_i - \boldsymbol{a}_0, \boldsymbol{w}_h >_E |^2 \qquad (6)$$

subject to: $< \boldsymbol{w}_i, \boldsymbol{w}_h >_E = 0$ *for all* $i = 1, \ldots, h-1$.

In the context of *nested* and *projected* PCA, we will call the $\boldsymbol{w}_i$ *principal directions* of the PCA. Observe that the *global* PCA produces only a principal component and not a set of preferential directions.

As in (Cazelles et al. 2017), the optimization problems in (4) and (5) are non-convex and hence, to solve them, we employed the well known solver Ipopt, which implements an interior point method. Note that the worst-case computational cost associated to (4) is $O((nk + Jk)^{3.5})$ and the one for (5) is $O((n + J)^{3.5})$, cf. (Nocedal and Wright 2006).

As a rough comparison, we measured the execution times needed to to find the 2-*nested* PC on a dataset of $n = 100$ Gaussian distributions, with our implementation (using $J = 20$ spline basis) and the one in (Cazelles et al. 2017) [1]. We observed a massive speedup (approx. 30 – 35 times faster) gained using our formulation.

On the other hand, problem (6) can be solved trivially by recognizing it as a Rayleigh quotient. Indeed, by letting A the $n \times J$ matrix with rows $\boldsymbol{a}_i - \boldsymbol{a}_0$, (6) is equivalent to

$$\underset{\boldsymbol{w} \in \mathbb{R}^J}{\arg\max} \frac{\boldsymbol{w}^t (AE)^T (AE) \boldsymbol{w}}{\boldsymbol{w}^t E \boldsymbol{w}}$$

whose solution is the generalized eigenvalue problem $(E^T A^T AE)$ which may be solved by the eigenvalue problem for $A^T AE$. The computational cost associated to this operation is $O(J^3)$.

[1] The code is publicly available at https://github.com/ecazelles/2017-GPCA-vs-LogPCA-Wasserstein

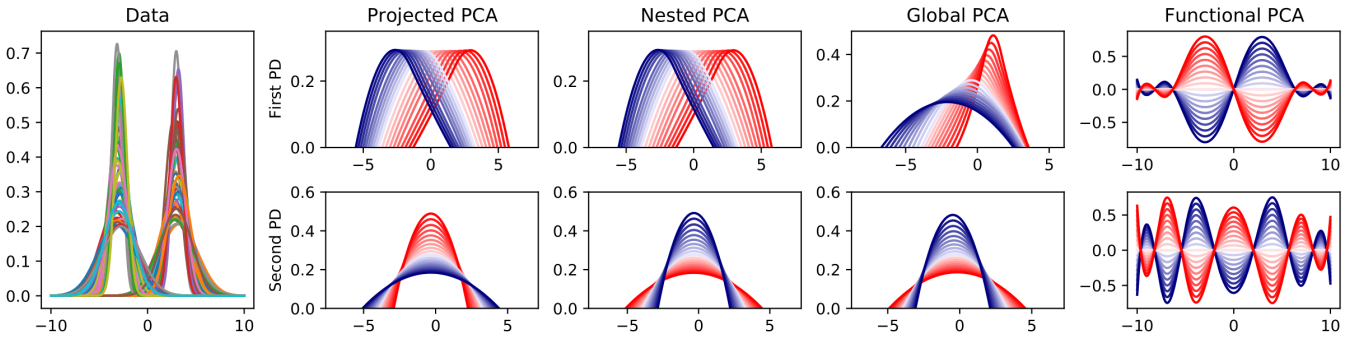

Figure 2: From left to right: dataset of pdfs, top row: first principal directions (PD) for different methods, bottom row: second PD. For the Projected, Nested and Global panels, each plot displays the pdfs associated to $\lambda \boldsymbol{w}_i$ where i is the index of the component (either $i = 1$ or $i = 2$), $\boldsymbol{w}_i$ is the PC and λ ranges from -2 (darkest blue) to $+2$ (darkest red). For the Functional panels, each plot displays the functions given by $\lambda \boldsymbol{w}_i$ where $\boldsymbol{w}_i$ is the i–th PC found by functional PCA.

In all the three different PCAs, once the PC is found, to project a (possibly new) datum represented by the spline coefficients $\boldsymbol{a}^*$ on the component, we solve

$$\arg\min_{\boldsymbol{\lambda}} ||\boldsymbol{a}^* - \boldsymbol{a_0} - \textstyle\sum_{j=1}^{k} \lambda_j \cdot \boldsymbol{w}_j||_E^2 \qquad (7)$$

subject to: $G\left(\sum_j \lambda_j \boldsymbol{w}_j + \boldsymbol{a}_0\right) \geq 0$.

Observe that problem (7) has the same form of (4), while keeping the vectors $\boldsymbol{w}_j$ fixed, which extremely simplifies the problem. Indeed the projection of a point on a PC is equivalent to minimizing the norm in a polytope, which in $\mathbb{R}^J$ is a well studied problem (see Sekitani and Yamamoto 1993) and can be solved using an interior point method.

The formulation of the optimization problems presented in this section allows a straightforward description of the differences between the *projected* PCA and the *geodesic* ones. Indeed, the main difference lies in the the monotonicity constraints: in (5) these are part of the optimization problems, while in (6) they are not present. In fact, in the *projected* PCA the constraints are used only to restrict the span of $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ to $\mathbb{R}^{J\uparrow}$ in (7). Figure 1 shows a toy example where the two definitions yield different principal directions.

This, a priori, might result in a loss of performance in fitting the data. However, in Section 5 we show that the loss is marginal in several simulations, especially when compared to the computational advantages provided by our definition.

5 Numerical Illustrations

There are several situations in which one may want to perform PCA on a generic dataset, that mainly boil down to two different purposes: descriptive analysis of the main sources of variability in a set of data and dimensionality reduction. Finally, a related but different purpose, in the context of complex data, is to map structured data (as distributions) in an Euclidean space and perform inference, such as regression or classification, using standard techniques.

The interpretability of the analysis (being it descriptive, inferential, etc...) and the effectiveness of the dimensionality reduction depend on the metric chosen and on how the PCA is able to fit the data according to such metric.

The first issue can be checked by visual inspection, while the second one can be measured in terms of the *reconstruction error*: the average distance between each datum and its projection onto the principal components, i.e. for a datum $F^* = \sum_{j=1}^{J} a_j^* \psi_j$, the quantity $||\boldsymbol{a}^* - \sum_{j=1}^{k} \lambda_j \boldsymbol{w}_j||_E$ where λ_js are the solution of (7) and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are solution of the optimization problems defining the different PCAs.

We design three different simulations to address all these points. In the following, we will always center the PCA in the barycenter of the data, i.e. $\boldsymbol{a}_0 = n^{-1} \sum_{i=1}^{n} \boldsymbol{a}_i$. Moreover, we consider the spline basis $\{\psi_j\}_{j=1}^{J}$ with $J = 20$ and equispaced knots in $[0, 1]$. The average error measured in terms of Wasserstein distance between the original data and the spline representation is roughly ten times smaller than the average pairwise distance between measures in each dataset. Of course, a larger number of elements in the basis would help reduce even more this approximation error, but would greatly increase the computational cost required to find the *geodesic* PCAs.

All experiments were performed on a laptop equipped with a 8-core Intel i7-7700HQ CPU 2.80GHz and 16Gb of RAM. The main numerical libraries employed consist of the Python packages numpy, scipy and qpsolvers (v 1.1) and of the optimization library Ipopt (v 3.12.12) interfaced with the Python package pyomo.

5.1 Intepretability: Population of Gaussians

We consider a sample of $n = 100$ Gaussian measures $\{\mu_i\}_{i=1}^{100}$, simulated as follows

$$\begin{aligned} \mu_i(dx) &= \mathcal{N}(dx \mid m_i, \sigma_i^2) \\ m_i &\sim 0.5\mathcal{N}(-3, (0.5)^2) + 0.5\mathcal{N}(3, (0.5)^2) \\ \sigma_i &\sim \mathcal{U}([0.5, 2]) \end{aligned}$$

It is clear that there are two sources of variability in this dataset: the location of the maximum of the probability density functions as well as the width of each pdf. Thus, a well behaved PCA should be able to detangle these two main "directions".

Figure 2 displays the results obtained using the three different methods, as well as functional PCA (FPCA). In the

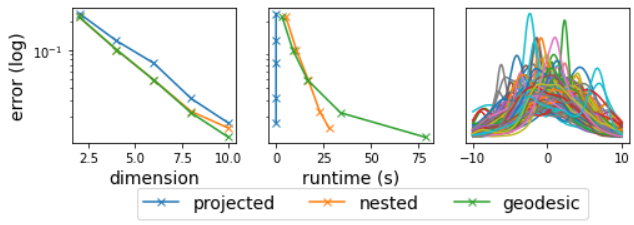

Figure 3: Left: reconstruction error as a function of the number of the number of components. Center: reconstruction error as a function of runtime. Right: example of dataset.

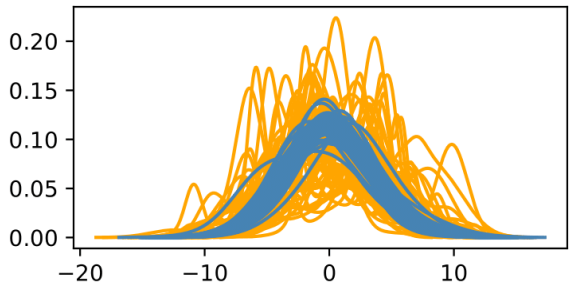

Figure 4: Example of dataset for the classification problem, each curve represents a pdf and the colors correspond to the classes to be predicted.

latter case, the pdfs are considered as functions in L_2, and the PCA is performed using the Python package scikit-fda. It is obvious how the PCAs based on the Wasserstein metric are able to separate the variability given by the location (first direction) and scale (second direction), whereas the FPCA captures only the pointwise amplitude variability in the pdfs. In addition to that, the output of FPCA are clearly not pdfs (being negative in some part of the domain).

Furthermore, we can appreciate that the principal directions found by the *projected* and *nested* PCA are extremely similar, hence showing that, at least in this example, the projection does not alter much the results. Observe how the $\boldsymbol{w}_i$s found by the *global* PCA (especially the first one) are quite different from the others, and in fact they are not *principal directions* (see Section 4.3). For this reason we believe one should prefer either the *projected* or *nested* PCA for visualization purposes.

In summary, this example shows how the *projected* PCA in Wasserstein spaces can provide a useful tool to gain insights into the sources variability of a collection of probability measures.

5.2 Dimensionality Reduction: Dirichlet Process Mixtures

In this example, we compare the ability of the different PCAs to efficiently compress the information present in a dataset of distributions, especially when no particular structure is recognizable among the data, unlike in the previous example. This example is intended to quantify the amount of information lost in the projection step.

For this reason, we simulate data from a Dirichlet Process Mixture (DPM, see Ferguson 1983), which is a well known workhorse in the Bayesian literature. We argue that this is a sensible choice for generating a sample of distributions with little (recognizable) structure (see the right panel in Figure 3). We sample $n = 100$ random probability density functions from:

$$p_i(\cdot) = \int_\Theta k(\cdot \mid \theta) G_i(d\theta), \quad G_i \sim \mathrm{DP}(\alpha H_0) \qquad (8)$$

where $k(\cdot \mid \theta)$ is the Gaussian kernel with parameters $\theta = (m, \sigma^2)$, $H_0(dm, d\sigma) = \mathcal{N}(dm \mid 0, 4)\mathcal{U}(d\sigma \mid [0.5, 2])$, $\alpha = 50$ and DP denotes the Dirichlet measure with base measure αG_0. Additional details on how we simulate from (8) are given in the Supplementary Material.

Figure 3 reports the results obtained, averaged over 20 repetitions. All the three approaches present an exponentially decreasing reconstruction error; as expected, the *nested* and *global* PCA seem to perform a little better than our *projected* PCA, for any fixed choice of the number of components. However, when comparing the reconstruction error against the runtime, it is clear how the *projected* PCA has an almost negligible runtime. On the other hand, both the *nested* and *global* PCA show that a much greater computational effort is needed to match the performance of the *projected* PCA, hence validating the choice of *projected* PCA as a fast black-box tool for dimensionality reduction. Although not shown here, the same conclusions can be drawn when comparing the runtimes needed to obtain any PC for a varying number of basis functions.

5.3 Classification on Distributions

Finally, we consider a binary classification problem, where the goal is to predict a class label given a distribution. Data $\{(p_i, y_i)\}$, where $y_i \in \{0, 1\}$ is a class label to be predicted, are generated again from Dirichlet Process Mixtures, with different parameters depending on the class. In particular $n = 100$ data are generated as in (8), with the difference that $G_i \sim DP(\alpha H_{y_i})$ where

$$H_j(dm, d\sigma) = \mathcal{N}(dm \mid 0, \eta_j^2)\, \mathcal{U}(d\sigma \mid [l_j, u_j]) \quad j = 0, 1$$

and $\eta_0 = 4$, $\eta_1 = 2$, $(l_0, u_0) = (0.5, 2.0)$, $(l_1, u_1) = (2.0, 4.0)$, see Figure 4. In all the simulations, the two classes are equally balanced.

We compare on this classification task three classifiers based the different PCAs presented. In particular, after performing a PCA, a Support Vector Machine (SVM) classifier is fit, with parameters $C = 1.0$, radial basis function kernel and default value for the parameter γ (see the Python package scikit-learn (Pedregosa et al. 2011)). Moreover, we compare the results obtained with FDA techniques, namely functional logistic regression and functional nearest neighbor classifier. Both methods work as the corresponding multivariate ones, only substituting the L_2 inner product and the L_2 norm in place of the inner product and norm in the finite dimensional Euclidean space.

Results are shown in Table 1. We can notice how the three classifiers based on the presented PCAs perform very similarly and consistently outperform the techniques based on

Method	Accuracy
ProjectedPCA + SVM	0.86 ± 0.11
NestedPCA + SVM	0.86 ± 0.11
GlobalPCA + SVM	0.85 ± 0.10
Functional Logistic Regression	0.71 ± 0.15
Functional Nearest Neighbor	0.75 ± 0.13

Table 1: Accuracy of classification for different algorithms. Each result displays the average 10-fold cross validation accuracy, averaged again over 20 repetitions $\pm$ one standard deviation.

FDA, hence again providing empirical evidence of: (i) the advantages of using the Wasserstein metric to analyze distributional data and (ii) the similarity in performance between the *projected* PCA and the *geodesic* ones. Finally, similar considerations with respect to computational time hold as discussed in the previous examples.

5.4 Covid-19 Mortality Data in the USA

We illustrate the *projected* PCA using data from the US Centers for Disease Control and Prevention. In particular, we examine the Covid-19 mortality in 53 among US states and inhabited territories. Further, we divide the data between males and females, as there is a scientific consensus that mortality due to Covid-19 is higher among males. Data consists of unnormalized histograms, counting the number of deaths subdivided in 11 age bins.[2]

We apply the *projected* PCA after having normalized the histograms and computed the spline basis representation of the quantile functions; using $J = 20$ basis we obtain an average approximation error of 2×10^{-4}. We selected the dimension of the PC by looking at the reconstruction error: using the 2 dimensional PC, we obtained an error, weighted by the L_2 norm of the quantile function, as low as 0.01. Therefore, we stick to this choice for subsequent analyses.

Figure 5 reports a summary of the inference. We observe how the first principal direction highlights differences in the mortality of the elders, while the second one discriminates the mortality in the age range $50 - 75$. In particular, negative values on the first principal direction correspond to higher mortality rates among the elders while positive values in the second principal direction correspond to higher mortality rates in the age range $50 - 75$.

By looking at the scores of the projections on these directions, we notice two slightly overlapping point clouds: in orange the women and in blue the men. The fact that the blue dots tend to have higher values on the y-axis is in accordance to the fact that Covid-19 mortality affects more males in their $50 - 70$s than women in the same age range.

Finally we report also the plots for the two populations having the most extreme values on the first direction, namely woman in Massachussets (green) and men in Hawaii (red). In the first case, the mortality is concentrated in the last bin

[2]Data are freely available at https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku.

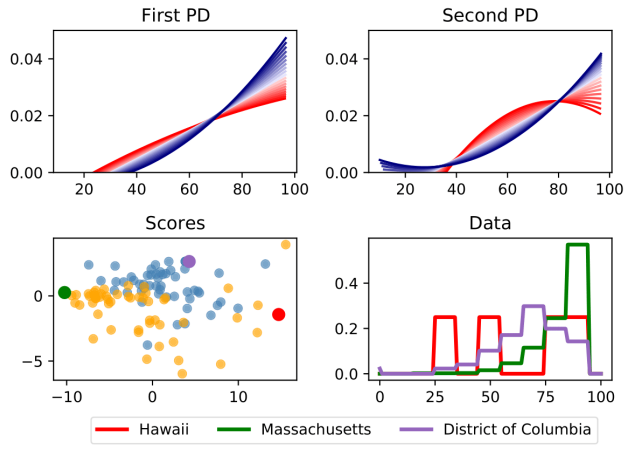

Figure 5: Top left: first principal direction. Top right: second principal direction. Each plot displays the pdfs associated to $\lambda \boldsymbol{w}_i$ (i is the index of the direction), $\boldsymbol{w}_i$ is the principal direction and λ ranges from -4 (darkest blue) to $+4$ (darkest red). Bottom left: scores of the projections on the 2-*projected* PC, in blue: male populations, in orange: female populations, the other colors match the one of the bottom right plot. Bottom right: data for three particular populations.

of the histogram, while in the latter one there are numerous deaths of people in their 30s and 50s.

6 Discussion and Future Work

In this paper, we have introduced a simple and handy representation of the 2-Wasserstein space, based on the isometric bijection that associates to a measure on the real line the corresponding quantile function. By considering a suitable B-spline representation of this space, and truncating the number of basis, we were able to reinterpret the definition of *nested* and *global* PCA, defined in (Bigot et al. 2017), in a simpler way, thus leading to easier optimization problems.

Furthermore, we introduced a novel notion of PCA in Wasserstein spaces, based on the projection of a standard PCA onto a convex polytope. Although our method is different from the *geodesic* approach and shares some similarities with the less accurate *log* approach, we showed empirically how in several different tasks our *projected* PCA has a similar performance, while requiring an almost null computational effort, when compared to the *geodesic* PCAs.

We believe that this work is a first step towards the development of fast and easy statistical methodologies in Wasserstein spaces. Nonetheless, there is broad room for improvements: many novel methodologies can be defined starting from our spline representation, such as linear models in Wasserstein spaces. Moreover, we still need to clarify the impact that the truncation of the B-spline basis produces over the inference.

Acknowledgments

We thank Federico Bassetti for his encouragements and helpful comments on an earlier version of the manuscript, and Alessandra Guglielmi and Piercesare Secchi for their advices.

References

Ambrosio, L.; Gigli, N.; and Savaré, G. 2008. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Bigot, J.; Gouet, R.; Klein, T.; and López, A. 2017. Geodesic PCA in the Wasserstein space by convex PCA. *Annales De L Institut Henri Poincare-probabilites Et Statistiques* 53: 1–26.

Cazelles, E.; Seguy, V.; Bigot, J.; Cuturi, M.; and Papadakis, N. 2017. Log-PCA versus Geodesic PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing* 40.

Colosimo, B.; and Pacella, M. 2007. On the use of principal component analysis to identify systematic patterns in roundness profiles. *Quality and Reliability Engineering International* 23: 707 – 725.

Cuturi, M.; and Doucet, A. 2014. Fast Computation of Wasserstein Barycenters. In *International Conference on Machine Learning*, 685–693.

Cuturi, M.; Teboul, O.; and Vert, J.-P. 2019. Differentiable ranking and sorting using optimal transport. In *Advances in Neural Information Processing Systems*, 6861–6871.

Delicado, P. 2011. Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis* 55: 401–420.

Ferguson, T. S. 1983. Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, 287–302. Elsevier.

Genevay, A.; Peyré, G.; and Cuturi, M. 2018. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 1608–1617. PMLR.

Hron, K.; Menafoglio, A.; Templ, M.; Hrzov, K.; and Filzmoser, P. 2014. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis* 94.

Huckemann, S.; Hotzand, T.; and Munk, A. 2010. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica* 1–58.

Kneip, A.; and Utikal, K. J. 2001. Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association* 96(454): 519–532.

Le-Rademacher, J.; and Billard, L. 2017. Principal component analysis for histogram-valued data. *Advances in Data Analysis and Classification* 11(2): 327–351.

Nagabhushan, P.; and Kumar, R. P. 2007. Histogram pca. In *International Symposium on Neural Networks*, 1012–1021. Springer.

Nocedal, J.; and Wright, S. 2006. *Numerical optimization*. Springer Science & Business Media.

Panaretos, V. M.; and Zemel, Y. 2020. *An Invitation to Statistics in Wasserstein Space*. Springer Nature.

Patrangenaru, V.; and Ellingson, L. 2015. *Nonparametric statistics on manifolds and their applications to object data analysis*. CRC Press.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Pegoraro, M.; and Beraha, M. 2020. Supplementary Materials for Fast PCA in 1-D Wasserstein Spaces via B-splines Representation and Metric Projection. URL https://drive.google.com/file/d/1fZ9PQTUh7l5ludtHFhE-3BtzaVLFNeEt/view.

Pegoraro, M.; and Beraha, M. 2021. Projected Statistical Methods for Distributional Data on the Real Line with the Wasserstein Metric. *arXiv preprint arXiv:2101.09039* .

Potter, S.; Del Negro, M.; Topa, G.; and Van der Klaauw, W. 2017. The advantages of probabilistic survey questions. *Review of Economic Analysis* 9(1): 1–32.

Ramsay, J. O. 2004. Functional data analysis. *Encyclopedia of Statistical Sciences* 4.

Rodrıguez, O.; Diday, E.; and Winsberg, S. 2000. Generalization of the principal components analysis to histogram data. In *Workshop on simbolic data analysis of the 4th European Conference on principles and practice of knowledge discovery in data bases, Setiembre*, 12–16.

Sekitani, K.; and Yamamoto, Y. 1993. A recursive algorithm for finding the minimum norm point in a polytope and a pair of closest points in two polytopes. *Mathematical Programming* 61: 233–249.

Srivastava, S.; Cevher, V.; Dinh, Q.; and Dunson, D. 2015. WASP: Scalable Bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, 912–920.

Verde, R.; Irpino, A.; and Balzanella, A. 2015. Dimension reduction techniques for distributional symbolic data. *IEEE transactions on cybernetics* 46(2): 344–355.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.