Machine learning and optimal transport: some statistical and algorithmic tools

Elsa Cazelles

ESAIM: PROCEEDINGS AND SURVEYS, November 2023, Vol. 74, p. 158-168
Editors: Marguerite Zani, Clément Marteau - Journées MAS 2020, Fédération Siméon Poisson (Orléans-Tours),

MACHINE LEARNING AND OPTIMAL TRANSPORT: SOME STATISTICAL AND ALGORITHMIC TOOLS. *

Elsa Cazelles[1]

Abstract. In this paper, we focus on the analysis of data that can be described by probability measures supported on a Euclidean space, by way of optimal transport. Our main objective is to present a first and second order statistical analyses in the space of distributions in a concise manner, as a first approach to understand the general modes of variation of a set of observations. In the context of optimal transport, these studies correspond to the barycenter and the decomposition into geodesic principal components in the Wasserstein space. In particular, we aim attention at a regularised estimator of the barycenter, in order to handle the noise coming from the observations. Additionally, we leverage these tools for time series analysis, whose spectral informations are compared using optimal transport.

Résumé. Dans cet article, nous nous concentrons sur l'analyse de données pouvant être décrites par des mesures de probabilité supportées sur un espace Euclidien, au moyen du transport optimal. Notre objectif principal est de présenter de manière concise l'analyse statistique de premier et second ordre dans l'espace des distributions comme une première approche pour comprendre les tendances générales d'un ensemble d'observations. Dans le contexte du transport optimal, ces études correspondent au barycentre et à la décomposition en composantes géodésiques principales dans l'espace de Wasserstein. Notamment, nous nous intéressons à un estimateur régularisé du barycentre, afin de gérer le bruit provenant des observations. Par ailleurs, nous exploitons ces outils pour l'analyse des séries temporelles, dont les informations spectrales sont comparées à l'aide du transport optimal.

The Wasserstein distance in brief

The Wasserstein distance between probability distributions is a special case of optimal transport introduced by Monge (1781) and generalised by Kantorovich (1940'). It was designed to find the most efficient way, i.e. requiring the least possible effort, to transport a pile of sand into a hole of the same volume. In other words, in mathematical language, Monge's problem comes to minimising the cost of transferring mass from one probability measure to another, that is, for probability measures μ and ν supported respectively on $\mathcal{X}$ and $\mathcal{Y}$:

$$\inf_{T:T\#\mu=\nu} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\mu(x) \qquad \text{for an arbitrary cost } c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}, \tag{1}$$

* *I would like to thank my co-authors Jérémie Bigot, Marco Cuturi, Nicolas Papadakis, Arnaud Robert, Vivien Seguy and Felipe Tobar.*

[1] CNRS, IRIT, Université de Toulouse.

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202374158

where T belongs to the set of measurable functions $T : \mathcal{X} \to \mathcal{Y}$ such that $T\#\mu = \nu$. The *pushforward measure* $T\#\mu$ is a probability measure on $\mathcal{Y}$ defined by $T\#\mu(B) = \mu\{x \in \mathcal{X} \mid T(x) \in B\}$, for any measurable set $B \subset \mathcal{Y}$ (see Figure 1). However, such a map T with $T\#\mu = \nu$ does not always exists.

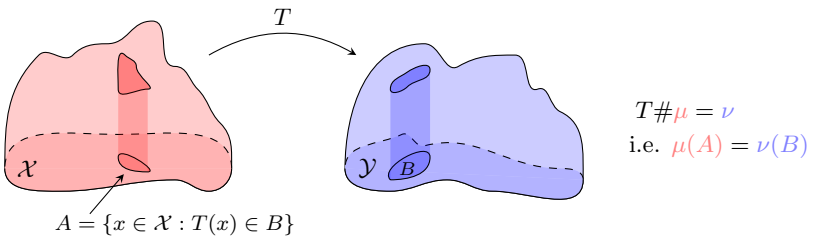

FIGURE 1. Mass transfer through the pushforward operator and the map T between two probability distributions μ (in red) and ν (in blue). Figure adapted from Thorpe's book [Thorpe, 2019].

In its modern formulation, Kantorovich's problem is a relaxed version of Monge's that consists in finding a transport plan between a source measure μ and a target measure ν, which minimises the global effort (see e.g. [Villani, 2008] and [Ambrosio et al., 2004]). In this article, we focus on measures with support included in $\mathbb{R}^d$ and on a Euclidean cost, thus defining the Wasserstein p-distance introduced by Leonid Wasserstein (1969), given for two probability distributions μ, ν of finite p-momentum by

$$W_p(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y) \right)^{1/p}, \text{ for } p \geq 1, \tag{2}$$

where π contains the behaviour of the mass transfer. More precisely $\Pi(\mu,\nu)$ is the set of measures supported on $\mathbb{R}^d \times \mathbb{R}^d$ of respective marginals μ and ν (see Figure 2). This distance has in particular the advantage of characterising the weak convergence of measures on the metric space $(\mathcal{P}_p(\mathcal{X}), W_p)$ of probabilities admitting a moment of order p (see e.g. Chapter 7 of [Villani, 2003]).

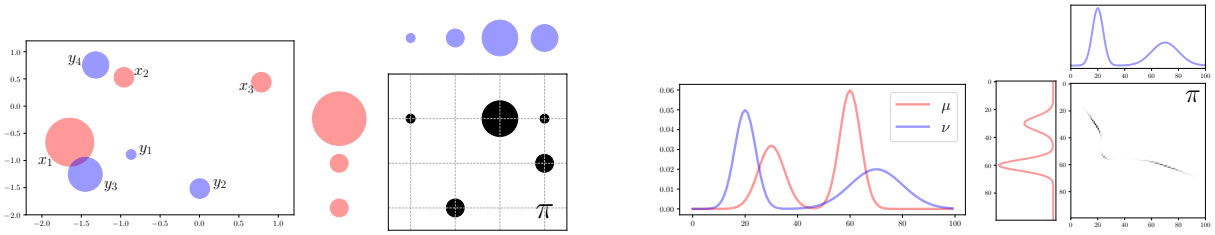

FIGURE 2. Representation of an optimal transport plan π in eq. (2) between two probability distributions μ (in red) and ν (in blue) that are discrete (left) and absolutely continuous with respect to Lebesgue measure (right). Figure adapted from Peyré and Cuturi's book [Peyré et al., 2012].

Additionally, on the real line, the Wasserstein distance is closed-form : let Ω be a (possibly unbounded) interval in $\mathbb{R}$ and let ν be a probability measure over $(\Omega, \mathcal{B}(\Omega))$ where $\mathcal{B}(\Omega)$ is the σ-algebra of Borel subsets of Ω. The cumulative distribution function (cdf) and the (generalized) quantile function of ν are denoted

respectively by F_ν and F_ν^-. Then, the Wasserstein distance W_p is defined for probability measures μ and ν in $\mathcal{P}_p(\Omega)$ by

$$W_p(\mu,\nu) := \left(\int_0^1 (F_\mu^-(\alpha) - F_\nu^-(\alpha))^p \mathrm{d}\alpha\right)^{1/p}. \tag{3}$$

Note that if $\mu \in \mathcal{P}_p(\Omega)$ is absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x$, then $T^* = F_\nu^- \circ F_\mu$ will be referred to as the optimal mapping to pushforward μ onto ν in the Monge's problem (1). For a detailed analysis of $\mathcal{P}_p(\Omega)$ and its connection with optimal transport theory, we refer to [Villani, 2003]. For a computational point of view, including applications, we refer to [Peyré and Cuturi, 2019].

In the following, we discuss the first order statistical analysis of a set of probability distributions, namely the barycenter in the Wasserstein space. In particular, we present an entropy regularised estimator of the barycenter and study its variance. Next, we outline the difficulties of conducting a principal component analysis for a set of probability measures, and present a PCA based on the geodesics in the Wasserstein space. The final section addresses the analysis of time series using the tools presented here.

1. First order statistical analysis : entropy regularised barycenter

1.1. Wasserstein barycenters

A statistical analysis of order one requires an object equivalent to the Euclidean mean, adapted to non-linear spaces, in this case the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. In this regard, the Fréchet mean [Fréchet, 1948] for the W_2 metric is a natural tool. As introduced by [Agueh and Carlier, 2011], an empirical Wasserstein barycenter $\hat{\nu}_n$ of a set of n probability measures $\nu_1, \ldots, \nu_n$ in $\mathcal{P}_2(\mathbb{R}^d)$ is given by

$$\hat{\nu}_n \in \underset{\mu\in\mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n W_2^2(\mu,\nu_i). \tag{4}$$

A detailed characterisation of these barycenters in terms of existence, uniqueness and regularity for probability measures whose support is included in $\mathbb{R}^d$ is available in [Agueh and Carlier, 2011].

The notion of Wasserstein barycenter was first generalized in [Le Gouic and Loubes, 2017] for random probability measures (see also [Álvarez-Esteban et al., 2015] for similar concepts). A probability measure $\boldsymbol{\nu}$ in $\mathcal{P}_2(\mathbb{R}^d)$ is said to be random if it has distribution $\mathbb{P}$ on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{B}\left(\mathcal{P}_2(\mathbb{R}^d)\right))$, where $\mathcal{B}\left(\mathcal{P}_2(\mathbb{R}^d)\right)$ is the σ-Borel algebra generated by the topology induced by the distance W_2. In other words, when well defined, the Wasserstein barycenter of a random probability measure of law $\mathbb{P}$ supported on the space of distributions $\mathcal{P}_2(\mathbb{R}^d)$ is given by

$$\nu_{\mathbb{P}} \in \underset{\mu\in\mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu,\boldsymbol{\nu})\mathrm{d}\mathbb{P}(\boldsymbol{\nu}). \tag{5}$$

In the case where $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are independent and identically distributed random probability measures (*iid*) of law $\mathbb{P}$, the barycenter $\nu_{\mathbb{P}}$ is referred to as the population counterpart of $\hat{\boldsymbol{\nu}}_n$.

We can then construct regularised versions of these barycenters, and conduct their statistical analysis. The motivation is twofold: adding an entropy term to the Wasserstein distance not only allows us to take advantage of a fast algorithm to compute the barycenter but also to obtain a smoother estimator. Moreover, we present the first bound on the variance of the proposed regularised estimator, which will allow to choose the regularisation parameters appropriately.

1.2. Entropy regularised barycenters

We consider a dataset composed of n random discrete measures $\boldsymbol{\nu}_{p_1}, \ldots, \boldsymbol{\nu}_{p_n}$ obtained from random observations $\boldsymbol{X} = (\boldsymbol{X}_{i,j})_{1\le i\le n;\, 1\le j\le p_i}$ organised as n subjects (or experimental units), such that $\boldsymbol{\nu}_{p_i}$ is defined

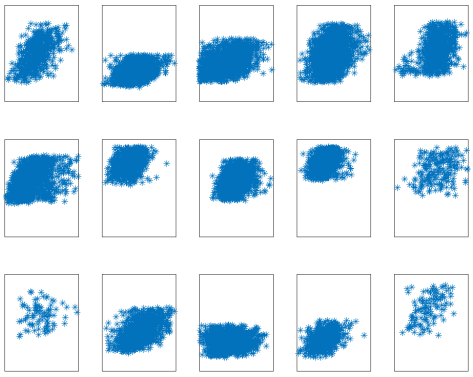

FIGURE 3. Cytometry dataset from the Immune Tolerance Network. Each point cloud represents the FSC and SSC marker values for a set of cells from a patient.

by

$$\boldsymbol{\nu}_{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}. \tag{6}$$

To better understand these objects, we can lean on the example of the cytometry dataset (Immune Tolerance Network `http://bioconductor.org/packages/release/bioc/html/flowStats.html`) in Figure 3, which shows the FSC (*forward-scattered light*, x-axis, ranging from 200 to 600) and SSC (*side-scattered light*, y-axis, ranging from 0 to 250) marker values of human cells. In that specific example, the population barycenter would be seen as the cell measurements of a *super patient* from which $\boldsymbol{\nu}_{p_1}, \dots, \boldsymbol{\nu}_{p_n}$ patients are sampled, such that for each of them we have only a finite number p_i of observations (i.e. cell's measurements). From such a sample, we are interested in finding the underlying two-dimensional distribution, a priori absolutely continuous, of the patients' FSC and SSC markers. A regularised version (in the form of a entropy penalty term) of the Wasserstein barycenter can be of great advantages since one usually only has access to a dataset of observations $\boldsymbol{X}$. Therefore the resulting barycenter may suffer from irregularities due to outliers or a lack of observations per measurement.

In the following, we will consider the discrete setting, meaning that the measures are supported on a fixed finite number of points $\mathcal{X} := \{x_1, \dots, x_N\}$. A probability measure is then identified by a vector of positive weights summing to 1, that is an element of the N-dimensional simplex denoted Σ_N. In this framework, the optimal transport corresponds to a linear optimisation problem on the space of transport matrices of size N, with constraints on the marginals. However, the excessive cost of computing such an optimal mass transfer —of the order of $\mathcal{O}(N^3 \log N)$— is clearly prohibitive. To alleviate this computational cost, [Cuturi, 2013] proposed to add an entropy regularisation term to the classical linear transport problem, leading to the notion of entropy regularised optimal transport, or Sinkhorn divergence, between discrete probability measures. The Sinkhorn divergence is then defined for $a, b \in \Sigma_N$ and regularisation parameter $\varepsilon > 0$ by

$$W^p_{p,\varepsilon}(a, b) = \min_{U \in U(a,b)} \langle U, C \rangle - \varepsilon h(U), \tag{7}$$

where $h(U) = -\sum_{i,j} U_{ij} \log U_{ij}$ is the negative entropy of the transport matrix $U \in U(a, b) := \{U \in \mathbb{R}_+^{N \times N}$ such that $U \mathbb{1}_N = a, U^T \mathbb{1}_N = b\}$, and C is the cost matrix between the points of the support $\mathcal{X}$, i.e. $C_{ij} = \|x_i - x_j\|^2$ for $i, j \in \{1, \dots, N\}$.

Note that entropy regularised transport has considerably gained popularity in machine learning and statistics, as it makes it possible to use an approximation of transport distances for high-dimensional data analysis; in particular for generative models, multi-label learning, dictionary learning or image processing, see *e.g.* [Cuturi and Peyré, 2016, Rabin and Papadakis, 2015], text extraction by keyword comparison and in the averaging of

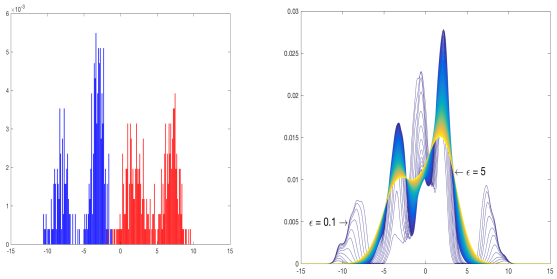

FIGURE 4. A simulated example of $n = 2$ distributions constructed with $p_1 = p_2 = 300$ observations generated from mixtures of Gaussians of random means and variances. (Left) The blue and red graphs are histograms of equal and small bins. (Right) 400 Sinkhorn barycenters $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ for ε ranging from 0.1 to 5. The colours encode the variation of ε.

neuroimaging data. Peyré and Cuturi's book [Peyré et al., 2012] presents a large part of the applications specific to optimal transport, and in particular to regularised transport.

As previously explained, the initial purpose of this entropy regularisation was to efficiently compute the Wasserstein distance, by way of an iterative algorithm for which each iteration costs $\mathcal{O}(N^2)$. Singularly, such a regularised transport can be instrumental in handling outliers or smoothing an estimator of barycenter, beyond the purely computational advantage. This approach leads to the Sinkhorn barycenter [Cuturi and Doucet, 2014, Cuturi and Peyré, 2016, Carlier et al., 2017, Benamou et al., 2015].

The following presents the first statistical study of this regularised barycenter. We consider n discrete random measures $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n \in \Sigma_N$ generated from a distribution $\mathbb{P} \in \Sigma_N$. Additionally, for each i, we assume that the observations $(\boldsymbol{X}_{i,j})_{1 \leq j \leq p_i}$ are random variables with distribution $\boldsymbol{q}_i$. We then define for $\varepsilon > 0$ the empirical Sinkhorn barycenter $\hat{r}^{\varepsilon}_{n,p}$, where p depends on $(p_1, \ldots, p_n)$, and its equivalent in population r^{ε}

$$\begin{aligned} \hat{r}^{\varepsilon}_{n,p} &= \arg\min_{r} \frac{1}{n} \sum_{i=1}^{n} W^2_{2,\varepsilon} \left(r, \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{X_{ij}} \right) \\ r^{\varepsilon} &= \arg\min_{r} \mathbb{E}_{q \sim \mathbb{P}}[W^2_{2,\varepsilon}(r, q)] \end{aligned} \tag{8}$$

which correspond to Fréchet averages with respect to the Sinkhorn divergence $W^2_{2,\varepsilon}$, and depend on the regularisation parameter ε that informs on the amount of entropy within the transport problem. We can notice (see Figure 4) that the parameter ε has a smoothing effect on the barycenter $\hat{r}^{\varepsilon}_{n,p}$: the larger the parameter, the more the mass spreads. Thus the entropy penalty is no longer only of computational interest (in order to speed up the calculation time of a transport distance), but becomes a real regularisation tool.

We proved the strong convexity of the Sinkhorn divergence [Bigot et al., 2019], which allowed us to obtain a bound on the variance of the estimator $\hat{r}^{\varepsilon}_{n,p}$ of the Sinkhorn barycenter. For this, it is necessary to restrict the analysis to discrete measures belonging to the space

$$\Sigma^{\rho}_N = \left\{ r \in \Sigma_N \ : \ \min_{1 \leq \ell \leq N} r_\ell \geq \rho \right\},$$

as well as constraining the barycenter to belong to this space. This amounts to imposing a constraint on the support of the Sinkhorn barycenter. The bound is given by the following theorem, where all constants are explicit:

Theorem 1.1 (Bigot, C., Papadakis, 2018)**.** *Let* $p = \min_{1\leq i\leq n} p_i$ *et* $\varepsilon > 0$*. So*

$$\mathbb{E}(|r^{\varepsilon} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}|^2) \leq \frac{32L^2_{\rho,\varepsilon}}{\varepsilon^2 n} + \frac{2L_{\rho,\varepsilon}}{\varepsilon}\sqrt{\frac{N}{p}},$$

with

$$L_{\rho,\varepsilon} = \left(\sum_{1\leq m\leq N}\left(2\varepsilon\log(N) + \inf_{1\leq k\leq N}\sup_{1\leq \ell\leq N}|C_{m\ell} - C_{k\ell}| - 2\varepsilon\log(\rho)\right)^2\right)^{1/2}, \tag{9}$$

where we recall that C *is the cost matrix in the entropy regularised optimal transport problem and* N *the number of support points of the distributions.*

1.3. **Application and choice of the regularisation parameter** ε

Histogram registration problems have applications in many fields. In bioinformatics, for example, researchers aim to automatically normalise large datasets to compare and analyse characteristics within a single population of cells, taking into account phase variability (see the previous cytometry example in Figure 3). Unfortunately, the acquired information is often noisy due to misalignment, caused by technical variations in the environment. The need to take phase variability into account in the statistical analysis of such datasets is a known problem. Examples can be found in the one-dimensional case ($d = 1$) with biodemographic and genomic studies [Zhang and Müller, 2011], economic studies [Kneip and Utikal, 2001], analysis of neuronal activity in neuroscience [Wu and Srivastava, 2011] or the functional connectivity between brain regions [Petersen et al., 2016]. In higher dimension, $d \geq 2$, the data registration problem comes for instance from the statistical analysis of spatial point processes [Gervini, 2016, Panaretos and Zemel, 2017] or from flow cytometry data [Hahne et al., 2010, Pyne et al., 2014].

Optimal transport allows to correct the effects of misalignments within a dataset, however its usefulness has only been exploited by few authors. Additionally, the noise can be dealt with a smoothing step, that in our case is included in the computation of the regularised Wasserstein barycenter defined in (8). The estimator then fulfils its role in efficiently recovering the structure of a dataset from a small number of observations, as shown in Figure 5.

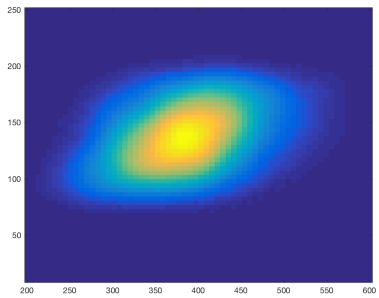

FIGURE 5. Sinkhorn barycenter associated to the cytometry dataset in Figure 3.

In order to automatically choose the regularisation parameter, the Goldenshluger-Lepski method suggests a solution based on the variance of the estimators, for a choice of parameters guided by the dataset. Therefore the theoretical results on the upper bound of the variance of the estimator in Theorem 1.1 allow us to tackle the problem of histogram registration and especially the automatic choice of the regularisation parameter ε of the barycenter estimator in (8). In Figure 6, we present a toy example for $n = 15$ mixtures of Gaussian distributions,

each with $p = 50$ observations. The GL bias-variance trade-off function associated to the Sinkhorn barycenters and plotted on the left suggests to choose $\varepsilon = 2.55$ as the optimal regularisation parameter. In Figure 6 (right), we display the associated Sinkhorn barycenter $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$. This work [Bigot et al., 2019] is the first to propose an automatic choice of parameters in the context of regularisation related to optimal transport.

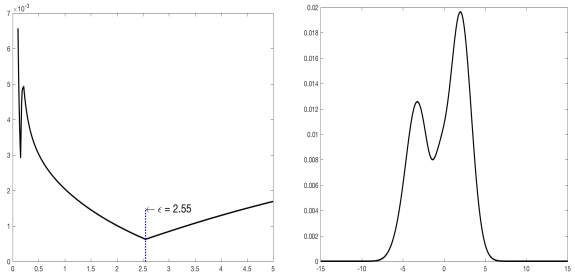

FIGURE 6. (Left) Bias-variance trade-off function given by the Goldenshluger-Lepski method, for $n = 15$ mixtures of random Gaussians. (Right) Optimal Sinkhorn barycenter associated with $\varepsilon = 2.55$.

2. Second order statistical analysis : Geodesic PCA

We present in this section a second-order statistical analysis, which is naturally expressed through a principal component analysis (PCA). In the same way as a usual PCA, the aim is to calculate the main modes of variation of one-dimensional histograms around their average element in order to better summarise and represent the information of a dataset. However, as the number, size or locations of significant bins in the histograms of interest may vary from one histogram to another, using standard PCA on histograms (with respect to the Euclidean metric) is bound to fail. The usual (functional) PCA of a set of probability densities $(f_i)_{i=1,\dots,n}$ seen as functions of $\mathbb{L}_2(\mathbb{R})$ consists in diagonalizing the covariance operator Cov. The eigenvectors of Cov associated with the largest eigenvalues describe the main modes of variability of the data around the Euclidean mean $\bar{f}_n$. The functional PCA results are very unsatisfactory for several reasons. Firstly, the functions obtained are not probability densities, in particular they take negative values. Secondly, the $\mathbb{L}_2$ metric only takes into account variations in the amplitude of the data.

In order to overcome these two drawbacks, it is essential to work directly on the space of probability measures $\mathcal{P}_2(\mathbb{R})$ endowed with the 2-Wasserstein distance. However, this space is not Hilbertian. Consequently, standard PCA, which involves the calculation of a covariance matrix, cannot be applied directly to compute the principal modes of variation in the Wasserstein sense. Nevertheless, a meaningful notion of PCA can still be defined based on the pseudo-Riemannian structure of the Wasserstein space, which has been extensively studied in [Ambrosio et al., 2004] and [Ambrosio et al., 2005]. Following this principle, a structure for the geodesic principal component analysis (GPCA) of probabilities measures supported on an interval $\Omega \subset \mathbb{R}$ has been introduced in [Bigot et al., 2017]. GPCA is defined as the problem of estimating a principal geodesic subspace (of a given dimension) that maximises the variance of the projection of the data into this subspace. In this approach, the base point of the subspace is the Wasserstein barycenter $\hat{f}_n$ of the data f_i as mentioned in (4). The existence, consistency and a detailed characterisation of the GPCA in $\mathcal{P}_2(\Omega)$ have been studied in [Bigot et al., 2017]. In particular, the authors showed that this approach is equivalent to projecting the data into the tangent space of $\mathcal{P}_2(\Omega)$ at the Fréchet mean, and then performing a PCA in this Hilbert space, while constraining the problem to a convex and closed subset of functions. Projecting the data into this tangent space is not difficult in the one-dimensional case since it boils down to computing a set of optimal mappings, or Monge maps, between the data and their Wasserstein barycenter, for which an explicit form is available (see eq. (3)). However, the authors of [Bigot et al., 2017] did not construct an algorithm to solve the GPCA problem, only a numerical

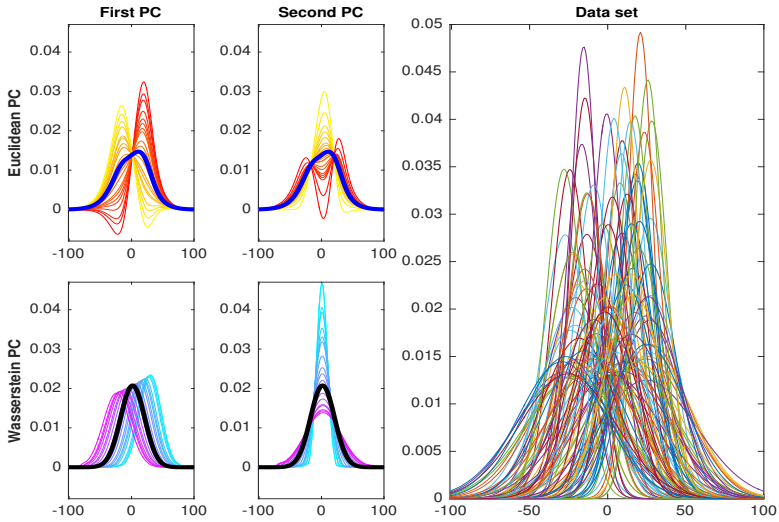

FIGURE 7. (Right) Data set of $n = 100$ randomly translated and expanded Gaussian histograms. (Top-left) Usual PCA via the Euclidean metric. The Euclidean barycentre is shown in blue. (Bottom-left) Geodesic PCA with respect to the Wasserstein distance. The black curve represents the density of the Wasserstein barycentre. The colours encode the progression of the densities along the geodesic principal components.

approximation of the computation of the geodesic principal components has been proposed. This last approach consists in applying a log-PCA, i.e. a standard PCA of the dataset previously projected in the tangent space of $\mathcal{P}_2(\Omega)$ to its Wasserstein barycenter $\hat{f}_n$.

In our paper [Cazelles et al., 2018], we proposed to compare the log-ACP and GPCA methods as introduced in [Bigot et al., 2017, Seguy and Cuturi, 2015]. In this setting, histograms are seen as piecewise constant probability densities supported on a given interval Ω. Therefore, the modes of variation of a set of histograms can be studied through the notion of geodesic PCA of probability measures in the Wasserstein space $\mathcal{P}_2(\Omega)$ admitting these histograms for density. The results are presented in Figure 7. The components recover well the translation effects (first component, left) and the amplitude effects (second component, right) of the dataset, when a so-called Euclidean functional PCA is not able to do so. However, the computation of the GPCA remains complicated even in the simplest case of $\mathbb{R}$ supported probability densities.

We have therefore provided a novel algorithm *forward-backward* to perform geodesic principal component analysis of measures defined on the real line, by solving the non-convex GPCA optimization problem exactly. This allowed us to present a detailed comparison between log-PCA and geodesic PCA of one-dimensional histograms, for different datasets. We have also extended the results for two-dimensional measures.The codes are available online at `https://github.com/ecazelles/2017-GPCA-vs-LogPCA-Wasserstein`.

3. Application to time series analysis

As an object allowing displacements on the support of the measures, the Wasserstein distance can be instrumental for signal processing problems and stationary time series analysis. The previous theoretical studies then naturally led to comparing time series in terms of their normalised Power Spectral Density (PSD) (i.e. of mass 1) for three main reasons:

(i) comparing two signals in terms of their PSD is possible even when they do not share the same sampling rate, length, magnitude or phase;

(ii) in the very simple case of a cosine of frequency ω, its PSD is given by the sum of Dirac mass at $-\omega$ and ω. It is thus reasonable to use the Wasserstein metric to emphasise the location of the support of the PSD, which contains all the information of the cosine;
(iii) in order to leverage the vast literature on the Wasserstein distance, which deals with positive functions of mass 1.

Going into details, the PSD of a signal is given by the modulus of its Fourier transform. After normalisation, it is possible to characterise an equivalence class for signals of the same Normalised PSD (NPSD). Once we have these objects in hand, we can define the so-called Wasserstein-Fourier distance between two classes of signals as the Wasserstein distance between their NPSDs, as proposed in [Cazelles et al., 2021]. From this construction we can easily deduce basic properties of this distance, on time and frequency translations for example. Similarly, we validated the proposed distance as a measure of interest for time series by relating convergence results between the time and frequency representations when the number of observations tends to infinity.

Once this framework is well established, we can apply the known statistical tools defined in the Wasserstein space. We emphasise that for many applications, differentiating and quantifying information across the spectrum of a signal, in the frequency domain, is more coherent then in the time domain. In particular, we have focused on the following applications. The code to reproduce the experiments is available in Python at `https://github.com/GAMES-UChile/Wasserstein-Fourier`.

Interpolation. As in the intuitive example (ii), an interpolation between two cosines of frequency $\omega_1 \leq \omega_2$ is interpreted as a cosine of frequency ω_γ with $\omega_1 \leq \omega_\gamma \leq \omega_2$. More generally, an interpolation boils down to computing the geodesic in the Wasserstein space between the normalised PSDs associated with the signals. This is summarised in Figure 8 (left). We also present in Figure 8 (right) the results when the signals are Gaussian processes with different kernels, indeed Bochner's Theorem directly relates the PSD and the kernel of a Gaussian process. In this special case, the interpolation appears to be a natural way of encoding the deformation of one signal onto another. As a consequence, interpolation allows us to generate new data along the geodesic whose dynamic content is close to the source and target data, thus performing data augmentation. Note that a GPCA procedure could also be applied to a set of signals for data augmentation. However, in order to properly capture the diversity of the dataset, the data must have a strong common geometric structure.

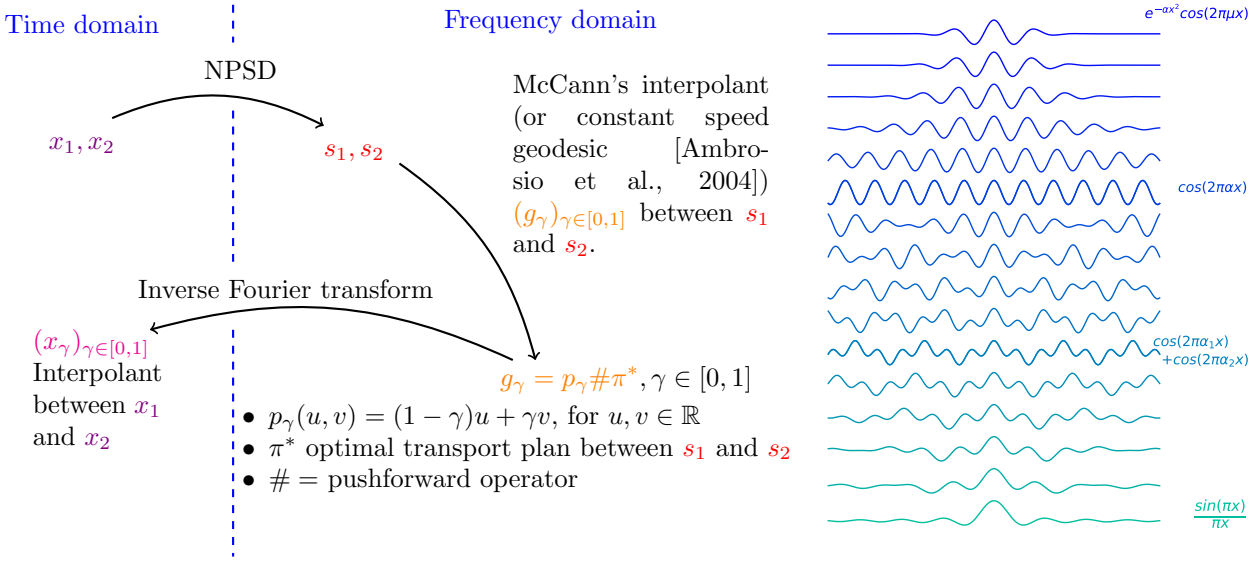

FIGURE 8. (Left) Interpolation principle with respect to Wasserstein-Fourier distance. (Right) Interpolation between Gaussian processes.

PCA. The counterpart of classical PCA in the space of probability distributions, presented in Section 2, can be directly applied to the NPSDs space. That allows to visualise and identify data that have similar dynamic content as soon as their projections are close on the principal components, and then to deduce groups of signals sharing some behaviour in frequency.

Classification. We propose a simple logistic regression framework, as well as a classifier based on nearest neighbours to classify time series based on their NPSD, which we compare through Wasserstein and Euclidean distances and Kullback-Leibler divergence.

Conclusion

In this paper, we present a small sample of the statistical analyses that can be performed using optimal transport when dealing with a dataset of objects described by probability distributions. The main message is that using an adequate space and metric to process data can be critical in some applications.

References

[Agueh and Carlier, 2011] Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

[Álvarez-Esteban et al., 2015] Álvarez-Esteban, P., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2015). Wide consensus for parallelized inference. *ArXiv e-prints*, 1511.05350.

[Ambrosio et al., 2004] Ambrosio, L., Gigli, N., and Savaré, G. (2004). Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei (9), Matematica e Applicazioni*, 15(3-4).

[Ambrosio et al., 2005] Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

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

[Bigot et al., 2019] Bigot, J., Cazelles, E., and Papadakis, N. (2019). Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755.

[Bigot et al., 2017] Bigot, J., Gouet, R., Klein, T., and López, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut H. Poincaré, Probabilités et Statistiques*, 53(1).

[Carlier et al., 2017] Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Analysis*, 49(2):1385–1418.

[Cazelles et al., 2021] Cazelles, E., Robert, A., and Tobar, F. (2021). The Wasserstein-Fourier distance for stationary time series. *IEEE Transactions on Signal Processing*, 69, 709-721.

[Cazelles et al., 2018] Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., and Papadakis, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc.

[Cuturi and Doucet, 2014] Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning 2014, PMLR W&CP*, volume 32, pages 685–693.

[Cuturi and Peyré, 2016] Cuturi, M. and Peyré, G. (2016). A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343.

[Fréchet, 1948] Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut H.Poincaré, Sect. B, Probabilités et Statistiques*, 10:235–310.

[Gervini, 2016] Gervini, D. (2016). Independent component models for replicated point processes. *Spatial Statistics*, 18:474 – 488.

[Hahne et al., 2010] Hahne, F., Khodabakhshi, A., Bashashati, A., Wong, C.-J., Gascoyne, R., Weng, A., Seyfert-Margolis, V., Bourcier, K., Asare, A., Lumley, T., Gentleman, R., and Brinkman, R. (2010). Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, 77(2):121–131.

[Kneip and Utikal, 2001] Kneip, A. and Utikal, K. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96(454):519–542.

[Le Gouic and Loubes, 2017] Le Gouic, T. and Loubes, J.-M. (2017). Existence and Consistency of Wasserstein Barycenters. *Probability Theory and Related Fields*, 168(3):901–917.

[Panaretos and Zemel, 2017] Panaretos, V. M. and Zemel, Y. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2), 932-976.

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

[Peyré et al., 2012] Peyré, G., Fadili, J., and Rabin, J. (2012). Wasserstein active contours. In *IEEE International Conference on Image Processing (ICIP)*.

[Peyré and Cuturi, 2019] Peyré, G., Cuturi, M. (2019). Computational optimal transport: With applications to data science. In *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

[Pyne et al., 2014] Pyne, S., Lee, S., Wang, K., Irish, J., Tamayo, P., Nazaire, M.-D., Duong, T., Ng, S.-K., Hafler, D., Levy, R., Nolan, G., Mesirov, J., and McLachlan, G. (2014). Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PloS one*, 9(7).

[Rabin and Papadakis, 2015] Rabin, J. and Papadakis, N. (2015). Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer.

[Seguy and Cuturi, 2015] Seguy, V. and Cuturi, M. (2015). Principal geodesic analysis for probability measures under the optimal transport metric. *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc.

[Villani, 2003] Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society.

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

[Wu and Srivastava, 2011] Wu, W. and Srivastava, A. (2011). An information-geometric framework for statistical inferences in the neural spike train space. *Journal of Computational Neuroscience*, 31(3):725–748.

[Zhang and Müller, 2011] Zhang, Z. and Müller, H.-G. (2011). Functional density synchronization. *Computational Statistics & Data Analysis*, 55(7):2234–2249.

[Thorpe, 2019] Thorpe, M. (2019). *Introduction to optimal transport*. Lecture Notes.