Distribution's template estimate with Wasserstein metrics

Emmanuel Boissard, Thibaut Le Gouic, Jean-Michel Loubes

Distribution's template estimate with Wasserstein metrics

Emmanuel Boissard, Thibaut Le Gouic, Jean-Michel Loubes

Institut de Mathématiques de Toulouse, Université Toulouse Paul Sabatier

Abstract

In this paper we tackle the problem of comparing distributions of random variables and defining a mean pattern between a sample of random events. Using barycenters of measures in the Wasserstein space, we propose an iterative version as an estimation of the mean distribution. Moreover, when the distributions are a common measure warped by a centered random operator, then the barycenter enables to recover this distribution template.

Keywords: Wasserstein Distance; Template estimation; Clustering, Fréchet mean.

e-mail: boissard,tlegouic,loubes@math.univ-toulouse.fr

1 Introduction

Giving a sense to the notion of *mean behaviour* may be counted among the very early activities of statisticians. When confronted to a large sample of high dimensional data, the usual notion of Euclidean mean is not usually enough since the information conveyed by the data possesses an inner geometry far from the Euclidean one. Indeed, deformations on the data such as translations, scale location models for instance or more general warping procedures prevent the use of the usual methods in data analysis. The mere issue of defining the mean of the data becomes a difficult task. This problem arises naturally for a wide range of statistical research fields such as functional data analysis for instance in [16], [24], [7] and references therein, image analysis in [26] or [5], shape analysis in [20] or [17] with many applications ranging from biology in [9] to pattern recognition [25] just to name a few.

Without any additional knowledge, this problem is too difficult to solve. Hence to tackle this issue, two main directions have been investigated. On the one hand, some assumptions are made on the deformations. Models governed by parameters have been proposed, involving for instance scale location parameters, rotations, actions of parameters of Lie groups as in [8] or in a more general way deformations parametrized by their coefficients on a given basis or in an RKHS set [2]. Adding structure on the deformations enables to define the *mean behaviour* as the data warped by the *mean deformation*, i.e the deformation parametrized by the mean of the parameters. Bayesian or semi-parametric

statistics enable to provide sharp estimation of these parameters. However, the consistency of the estimator remains a theoretical issue for many cases.

On the other hand, another direction consists in finding an adequate distance between the data which reveals the information which is conveyed. Actually, the chosen distance depends on the nature of the set where the observations belong, whose estimation is a hard task. We refer for instance to [23] for some examples. Once an appropriate distance has been chosen, difficulties arise when trying to define the mean as the minimum of the square distance since both existence and uniqueness rely on assumptions on the geometry of the data sets. This will be the framework of our work.

Assume that we observe $j = 1, \dots, J$ samples of $i = 1, \dots, n$ independent random variables $X_{i,j} \in \mathbb{R}^d$ with distribution μ_j. We aim at defining the *mean* behaviour of these observations, i.e their *mean* distribution. For this we will extend the notion of barycenter of the distributions with respect to the Wasserstein distance defined in [1] to the empirical measures and prove the consistency of its estimate. Actually, Wasserstein distance is a powerful tool to compute distance between distributions, with application in statistics pioneered in [13], [3] or [12] for instance. Moreover, we will tackle the case where the distributions are the images of an unknown original distribution by random operators under some suitable assumptions. In this case, we prove that an iterative version of the barycenter of the empirical distributions provides an estimate which enables to recover the original template distribution when the number of replications J is large enough.

The paper falls into the following parts. Section 2 is devoted to the extension of the notion of Barycenter in the Wasserstein space for empirical measures. In Section 3.2, we consider a modification of the notion of barycenter by considering iterative barycenters, which have the advantage to enable to recover the distribution pattern as proved in Section 4. Finally, some data applications are outlined in Section 5.

2 Barycenters in the Wasserstein space: Notations and general results

Let (E, d, Ω) denote a metric measurable space. The set of probability measures over E is denoted by $\mathcal{P}(E)$. Given a collection of probability measures $\mu_1, \dots, \mu_J$ over E, and weights $\lambda_1, \dots, \lambda_J \in \mathbb{R}$, $\lambda_j \geq 0$, $1 \leq j \leq J$, $\sum_{j=1}^J \lambda_j = 1$, there are several natural ways to define a weighted average of these measures. Perhaps the most straightforward is to take the convex combination of these measures

$$\mu_c = \sum_{j=1}^{J} \lambda_j \mu_j,$$

using the fact that probability measures form a convex subset of the linear space of finite measures. However, if we provide $\mathcal{P}(E)$ with some metric structure, the definition above is not really appropriate.

We denote by $\mathcal{P}_2(E)$ the set of all probability measures over E with a finite second-order moment. Given two measures μ, ν in $\mathcal{P}(E)$, we denote by $\mathcal{P}(\mu, \nu)$ the set of all

probability measures π over the product set $E \times E$ with first, resp. second, marginal μ, resp. ν.

The transportation cost with quadratic cost function, or quadratic transportation cost, between two measures μ, ν in $\mathcal{P}_2(E)$, is defined as

$$\mathcal{T}_2(\mu, \nu) = \inf_{\pi \in \mathcal{P}(\mu,\nu)} \int d(x, y)^2 d\pi.$$

The quadratic transportation cost allows to endow the set of probability measures (with finite second-order moment) with a metric by setting

$$W_2(\mu, \nu) = \mathcal{T}_2(\mu, \nu)^{1/2}.$$

This metric is known under the name of 2-Wasserstein distance.

In Euclidean space, the barycenter of the points $x_1, \ldots, x_J$ with weights $\lambda_1, \ldots, \lambda_J$, $\lambda_j \geq 0$, $\sum_{j=1}^J \lambda_j = 1$, is defined as

$$b = \sum_{j=1}^J \lambda_j x_j.$$

It is also the unique minimizer of the functional

$$y \mapsto E(y) = \sum_{j=1}^J \lambda_j |x_j - y|^2.$$

By analogy with the Euclidean case, we give the following definition for Wasserstein barycenter, introduced by M. Agueh and G. Carlier in [1].

Definition 2.1. We say that the measure $\mu \in \mathcal{P}_2(E)$ is a Wasserstein barycenter for the measures $\mu_1, \ldots, \mu_J \in \mathcal{P}_2(E)$ endowed with weights $\lambda_1, \ldots, \lambda_J$, where $\lambda_j \geq 0$, $\leq j \leq J$, and $\sum_{j=1}^J \lambda_j = 1$, if μ minimizes

$$E(\nu) = \sum_{j=1}^J \lambda_j W_2^2(\nu, \mu_j).$$

We will write

$$\mu_B(\lambda) = \mathrm{Bar}((\mu_j, \lambda_j)_{1 \leq j \leq J}).$$

In other words, the barycenter is the weighted Fréchet mean in the Wasserstein space. In [1], the authors prove that when $E = \mathbb{R}^d$ the barycenter exists. They also provide suitable assumptions on the measures μ_j, $1 \leq j \leq J$ to ensure that the barycenter is unique. For example, a sufficient condition is that one of the measures μ_j admits a density with respect to the Lebesgue measure. They also provide a problem that is the dual of the minimization of the functional E defined above, as well as characterizations of the barycenter.

Next, we recall a version of Brenier's theorem on the characterization of quadratic optimal transport in $\mathbb{R}^d$. Throughout all the paper we will use the following notation.

Definition 2.2. Let E, F be measurable spaces and $\mu \in \mathcal{P}(E)$. Let $T : E \to F$ be a measurable map. The push-forward of μ by T is the probability measure $T_{\#}\mu \in \mathcal{P}(F)$ defined by the relations

$$T_{\#}\mu(A) = \mu(T^{-1}(A)), \quad A \subset F \text{ measurable.}$$

Hence Brenier's theorem can be stated as follows.

Theorem 2.1 (Brenier's theorem, see [10])**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be measures, with μ absolutely continuous w.r.t. Lebesgue measure. Then there exists a μ-a.e. unique map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that*

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

Moreover, there exists a lower semi-continuous convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $T = \nabla\varphi$ μ-a.e., and T is the only map of this type pushing forward μ to ν, up to a μ-negligible modification. The map T is called the Brenier map from μ to ν.

Remark. The theorem above is commonly referred to as Brenier's theorem and originated from Y. Brenier's work in the analysis and mechanics literature. Much of the current interest in transportation problems emanates from this area of mathematics. We conform to the common use of the name. However, it is worthwile pointing out that a similar statement was established earlier independently in a probabilistic framework by J.A. Cuesta-Albertos and C. Matrán [11] : they show existence of an optimal transport map for quadratic cost over Euclidean and Hilbert spaces, and prove monotonicity of the optimal map in some sense (Zarantarello monotonicity).

As observed in [1], the barycenter of two measures is the interpolant of these two measures in the sense of McCann.

Proposition 2.2 (See [1], Section 6.2)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous w.r.t. Lebesgue measure. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ denote the Brenier map from μ to ν. The barycenter of (μ, λ) and $(\nu, 1-\lambda)$ is*

$$\mu_\lambda = (\lambda Id + (1-\lambda)T)_{\#}\,\mu.$$

This provides a natural expression for the barycenter of measures as a convex combination of measures.

3 Estimation of Barycenters of empirical measures

Assume we do not observe the distributions μ_j's but approximations of these distributions. Let $\mu_j^n \in \mathcal{P}_2(\mathbb{R}^d)$ for $1 \leq j \leq J$ be these approximations in the sense that they converge with respect to Wasserstein distance, i.e $W_2(\mu_j^n, \mu_j) \to 0$ when $n \to +\infty$. Our aim is to study the asymptotic behaviour of the barycenter of the μ_j^n's when n goes to infinity.

3.1 Consistency of the approximated barycenter

We are interested here in statistical properties of the barycenter of the $\mu_1^n, \ldots, \mu_J^n$. We begin by establishing a consistency result.

Theorem 3.1. *Let $J \geq 1$, and for every $n \geq 0$, let $\mu_j^n \in \mathcal{P}_2(\mathbb{R}^d)$, $1 \leq j \leq J$, be probability measures converging in Wasserstein topology to some probability measure μ_j for $1 \leq j \leq J$. Let $\lambda_1, \ldots, \lambda_J$ be positive weights. Let μ_B^n be a barycenter of the (μ_j^n, λ_j). The sequence $(\mu_B^n)_{n\geq 1}$ is compact and any of its limit points lies in $Bar((\mu_j, \lambda_j)_{1\leq j\leq J})$.*

Note that If any of the μ_j^n is absolutely continuous with respect to the Lebesgue measure, then μ_B^n is unique and our theorem states that it converges to a barycenter of the limit measures (μ_j, λ_j). Likewise, if any of the μ_j is absolutely continuous with respect to the Lebesgue measure, any μ_B^n is converging to the unique barycenter of (μ_j, λ_j).

The proof of this theorem relies on the following lemma which provides a characterization of a barycenter of measures. The proof of the lemma is inspired by the proof of Proposition 4.2 in [1] and is postponed to the Appendix.

Lemma 3.2. *Let $\Gamma(\mu_1, \ldots, \mu_J)$ be the set of probability measures on $(\mathbb{R}^d)^J$ with marginals $\mu_1, \ldots, \mu_J$, respectively and $T(x_1, \ldots, x_J) = \sum_{j=1}^J \lambda_j x_j$ with weights $\lambda_j \geq 0$ such that $\sum_{j=1}^J \lambda_j = 1$. A probability measure ν is a barycenter of $\mu_1, ..., \mu_J$ with weights $(\lambda_j)_{\leq j\leq J}$ if and only if $\nu = T_{\#}\gamma$ where $\gamma \in \Gamma(\mu_1, ..., \mu_J)$ minimizes*

$$\int \sum_{1\leq j\leq J} \lambda_j \|T(x_1, ..., x_J) - x_j\|^2 d\gamma(x_1, ..., x_J). \tag{1}$$

3.2 An Iterative version of barycenters of measures

Barycenters in Euclidean spaces enjoy the *associativity property* : the barycenter of x_1, x_2, x_3 with weights $\lambda_1, \lambda_2, \lambda_3$ coincides with the barycenter of x_{12}, x_3 with weights $\lambda_1 + \lambda_2, \lambda_3$ when x_{12} is the barycenter of x_1, x_2 with weights λ_1, λ_2. This property, as we will see, no longer holds when considering barycenters in Wasserstein spaces over Euclidean spaces, with the notable exception of dimension 1.

Therefore we introduce a notion of *iterated barycenter* as the point obtained by successively taking two-measures barycenters with appropriate weights. This does not in general coincide with the ordinary barycenter. However, we will identify cases where the two notions match.

Definition 3.1. Let $\mu_i \in \mathcal{P}_2(E)$, $1 \leq i \leq n$, and $\lambda_i > 0$, $1 \leq i \leq n$ with $\sum_{i=1}^n \lambda_i = 1$. The iterated barycenter of the measures $\mu_1, \ldots, \mu_n$ with weights $\lambda_1, \ldots, \lambda_n$ is denoted by $\mathrm{IB}((\mu_i, \lambda_i)_{1\leq i\leq n})$ and is defined as follows :

- $\mathrm{IB}((\mu_1, \lambda_1)) = \mu_1$,
- $\mathrm{IB}((\mu_i, \lambda_i)_{1\leq i\leq n}) = \mathrm{Bar}\ [(\mathrm{IB}((\mu_i, \lambda_i)_{1\leq i\leq n-1}), \lambda_1 + \ldots + \lambda_{n-1}), (\mu_n, \lambda_n)]$

The next proposition establishes consistency of iterated barycenters of approximated measures μ_j^n, for $j = 1, \ldots, J$.

Theorem 3.3. *The iterated barycenter is consistent : if $\mu_j^n \to \mu_j$ in W_2 distance for $j = 1, \dots, J$, then*

$$IB((\mu_j^n, \lambda_j)_{1 \leq j \leq J}) \to IB((\mu_j, \lambda_j)_{1 \leq j \leq J})$$

in W_2 distance.

Remark. Iterated barycenters as well as barycenters are well-suited to computations, since there exist efficient numerical methods to compute McCann's interpolant, see e.g. [6, 18]. The purpose of introducing the iterative barycenters is that, as shown in the next Section 4.2, the resulting measure has an expression as the image of a measure by a linear combination of maps. This will be helpful when considering a warping setting. Moreover, as we will see later, in some cases of interest the iterated barycenter does not depend on the order in which two-measures barycenters are taken, allowing for parallel computation schemes.

4 Deformations of a template measure

We now would like to use Wasserstein barycenters or iterated barycenters in the following framework : let (E, d, Ω) denotes a metric measurable space and assume that we observe probability measures in $\mathcal{P}(E)$, $\mu_1, \dots, \mu_J$ that are deformed versions, in some sense, of an original measure μ. We would like to recover μ from the observations. Here, we propose to study the relevance of the barycenter as an estimator of the template measure, when the deformed measures are of the type $\mu_j = T_{j\#}\mu$ for suitable push-forward maps T_j.

Our aim here is to extend the results of J.F. Dupuy, J.M. Loubes and E. Maza in [14]. They study the problem of *curve registration*, that we can describe as follows : given an unknown increasing function $F : [a, b] \mapsto [0, 1]$, and a random variable H with values in the set of continuous increasing functions $h : [a, b] \mapsto [a, b]$, we observe $F \circ h_1^{-1}, \dots, F \circ h_n^{-1}$ where h_i are i.i.d. versions of H (randomly warped versions of F). Let $\mu \in \mathcal{P}(\mathbb{R})$ denote the probability measure that admits F as its c.d.f. : then the above amounts to saying that we observe $h_{i\#}\mu$, $1 \leq i \leq n$. The authors build an estimator by using quantile functions that turns out to be the Wasserstein barycenter of the observed measures. They show that the estimator converges to $(\mathbb{E}H)_{\#}\mu$.

Hereafter, we first define a class of deformations for distributions, which are modeled by a push forward action by a family of measurable maps T_j, $j = 1, \dots, J$ undergoing the following restrictions. Such deformations will be called *admissible*.

4.1 Admissible deformations

Definition 4.1. The set $GCF(\Omega)$ is the set of all gradients of convex functions, that is to say the set of all maps $T : \Omega \to \mathbb{R}^n$ such that there exists a proper convex l.s.c. function $\phi : \Omega \to \mathbb{R}$ with $T = \nabla\phi$.

Definition 4.2. We say that the family $(T_i)_{i \in I}$ of maps on Ω is an *admissible family of deformations* if the following requirements are satisfied :

1. there exists $i_0 \in I$ with $T_{i_0} = \mathrm{Id}$,

2. the maps $T_i : \Omega \to \Omega$ are one-to-one and onto,

3. for $i, j \in I$ we have $T_i \circ T_j^{-1} \in GCF(\Omega)$.

The following Proposition provides examples are of such deformations.

Proposition 4.1. *The following are admissible families of deformations on domains of* $\mathbb{R}^n$.

- *The set of all product continuous increasing maps on* $\mathbb{R}^n$*, i.e. the set of all maps*

$$T : x \mapsto (F_1(x_1), \ldots, F_n(x_n))$$

where the functions $F_i : \mathbb{R} \to \mathbb{R}$ *are continuous increasing functions with* $F_i \to_{-\infty} -\infty$, $F_i \to_{+\infty} +\infty$.

In particular, this includes the family of scale-location transformations, i.e. maps of the type $x \mapsto ax + b$, $a > 0$, $b \in \mathbb{R}^n$.

- *The set of* radial distorsion *transformations, i.e. the set of maps*

$$T : \mathbb{R}^n \to \mathbb{R}^n, \quad x \mapsto F(|x|)\frac{x}{|x|}$$

where $F : \mathbb{R}^+ \mapsto \mathbb{R}^+$ *is a continuous increasing function such that* $F(0) = 0$.

- *The maps* ${}^tG \circ T_i \circ G$ *where* $(T_i)_{i \in I}$ *is an admissible family of deformations on* Ω *and* $G \in \mathcal{O}_n$ *is a fixed orthogonal matrix. This family has* ${}^tG(\Omega)$ *as its domain.*

Proof of Proposition 4.1

Proof. Let us consider the first family. Checking the two first requirements is straightforward and we only take care of the last one. Let $S : x \mapsto (F_1(x_1), \ldots, F_n(x_n))$ and $T : x \mapsto (G_1(x_1), \ldots, G_n(x_n))$. The map $S \circ T^{-1}$ is given by

$$S \circ T^{-1}(x) = \left(F_1 \circ G_1^{-1}(x_1), \ldots, F_n \circ G_n^{-1}(x_n)\right),$$

and this is the gradient of the function

$$x \mapsto \int_0^{x_1} F_1 \circ G_1^{-1}(z)dz + \ldots + \int_0^{x_n} F_n \circ G_n^{-1}(z)dz.$$

The functions $F_i \circ G_i^{-1}$ are increasing, so that their primitives are convex functions, which makes the function above convex.

Second point : observe that radial distortion transformations form a group, so that we only need show that each such transformation is the gradient of a convex function. And indeed, $T : x \mapsto F(|x|)\frac{x}{|x|}$ is the gradient of the function

$$x \mapsto \int_0^{|x|} F(r)dr$$

and this is a convex function because F is increasing.

The final item is a simple consequence of the observation that if $G \in \mathrm{GL}_n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then $\nabla(f \circ G) =^t G \circ \nabla f \circ G$. □

4.2 Barycenter of measures warped using admissible deformations

We are interested in recovering a template measure from deformed observations. The unknown template is a probability measure μ on the domain $\Omega \subset \mathbb{R}^d$, absolutely continuous w.r.t. the Lebesgue measure λ. We represent the deformed observations as push-forwards of μ by maps $T : \Omega \to \Omega$, i.e. we observe $(T_j)_{\#}\mu$, $j = 1, \dots, J$.

Theorem 4.2 states that when T_j belongs to an admissible family of deformations, taking the iterated barycenter of the observations corresponds to averaging the deformations. With this explicit expression at hand, we can check that in the case described above, the iterated barycenter coincides with the usual notion of barycenter.

Theorem 4.2. *Assume that $(T_i)_{i\in I}$ is an admissible family of deformations on a domain $\Omega \subset \mathbb{R}^n$, and let $\mu \in \mathcal{P}_2(\Omega)$, $\mu << \lambda$. Let $\mu_j = (T_j)_{\#}\mu$. The following holds :*

$$IB((\mu_j, \lambda_j)_{1\leq j\leq J}) = (\sum_{j=1}^{J} \lambda_j T_j)_{\#}\mu. \tag{2}$$

Moreover

$$IB((\mu_j, \lambda_j)_{1\leq j\leq J}) = Bar((\mu_j, \lambda_j)_{1\leq j\leq J}). \tag{3}$$

Remark.

1. The special case of the dimension 1

   In dimension 1, the set of *all* continuous increasing maps is an admissible family of deformations. The previous theorem applies for this very large class of deformations. Results in this case are known from [14] or [15]: the only new part here is that the estimator can be computed iteratively.

2. Barycenters and iterated barycenters do not match in general.

   The fact that the two notions of barycenter introduced above coincide no longer holds as soon as the dimension is larger than 2. For a counterexample, consider the case of non-degenerate centered Gaussian measures $\gamma_1, \dots, \gamma_J$ on $\mathbb{R}^n$, defined by their covariances matrices $S_1, \dots S_J \in \mathcal{S}_n^{++}$.

   According e.g. to [22], Example 1.7, the optimal transport map from $\mathcal{N}(0, S)$ to $\mathcal{N}(0, T)$ is given by

   $$x \mapsto T^{1/2}(T^{1/2}ST^{1/2})^{-1/2}T^{1/2}x.$$

   From this result, it is possible to give an explicit expression of the iterated barycenter.

   On the other hand, according to Theorem 6.1 in [1], the barycenter of the μ_j with weights $1/J$ is the Gaussian measure with covariance matrix the unique positive definite solution of the fixed point equation

$$M = \frac{1}{J}\sum_{j=1}^{J}\left(M^{1/2}S_jM^{1/2}\right)^{1/2}.$$

One may check that these two covariance matrices do not match in general.

4.3 Template Estimation from admissible deformations

Thanks to Theorem 4.2, we can study the asymptotic behaviour of the barycenter when the number of replications of the warped distributions J increases. Actually, we prove that the barycenter is an estimator of the template distribution.

Let T be a process with values in some admissible family of deformations acting on a subset $\mathcal{I} \subset \mathbb{R}^d$.

$$\begin{array}{rccc} T: & \Omega & \to & \mathcal{T}(\mathcal{I}) \\ & w & \mapsto & T(w,\cdot), \end{array}$$

where $(\Omega, \mathcal{A}, \mathbb{P})$ is an unknown probability space, Assume that T is bounded and has a finite moment $\varphi(.) = \mathbb{E}(T(.))$. Let T_j for $j = 1, \ldots, J$ be a random sample of realizations of the process T. Then, we observe measures μ_j which are warped by T_j in the sense that for all , $\mu_j = T_{j\#}\mu$.

Theorem 4.3. *Assume that μ is compactly supported. As soon as $\varphi = \mathrm{id}$, μ_B the barycenter of the μ_j's with weights $1/J$ is a consistent estimate of μ when J tends to infinity in the sense that a.s*

$$W_2^2(\mu_B, \mu) \overset{J\to\infty}{\longrightarrow} 0.$$

Moreover, assuming that $\|T - id\|_{L^2} \leq M$ a.s., we get the following error bound :

$$\mathbb{P}(W_2(\mu_B, \mu) \geq \varepsilon) \leq 2\exp\left(-J\frac{\varepsilon^2}{M^2(1 + c\varepsilon/M)}\right).$$

Note that when the warping process is not centered, the problem of estimating the original measure μ is not identifiable and we can only estimate by the barycenter μ_B the original measure transported by the mean of the deformation process, namely $\varphi_{\#}\mu$.

The proof of this theorem relies on the following proposition.

Proposition 4.4. *Let $(T_i)_{i\in I}$ be an admissible family of deformations on a domain $\Omega \subset \mathbb{R}^n$, and let $\mu \in \mathcal{P}_2(\Omega)$, absolutely continuous with respect to the n-dimensional Lebesgue measure. Let $\mu_j = (T_j)_{\#}\mu$. Denote by μ_B the barycenter with equal weights $1/J$. For every ν in $\mathcal{P}_2(\mathbb{R}^d)$, we have*

$$W_2(\mu_B, \nu) \leq \|\frac{1}{J}\sum_{j=1}^{J} T_j - T_\nu\|_{L^2(\mu)}$$

where T_ν is the Brenier map from μ to ν.

Proof. With the explicit expression of the barycenter, we know that the Brenier map from μ to μ_B is $1/J \sum_{j=1}^J T_j$, which implies that

$$\pi = (\frac{1}{J}\sum_{j=1}^J T_j, T_\nu)_{\#}\mu$$

is a coupling of μ_B and ν. Consequently,

$$W_2^2(\mu_B, \nu) \leq \int |\frac{1}{J}\sum_{j=1}^J T_j(x) - T_\nu(x)|^2 \mu(dx).$$

□

5 Statistical Applications

5.1 Distribution Template estimation from empirical observations

In many situations, the issue of estimating the mean behaviour of random observations plays a crucial role to analyze the data, in image analysis, kinetics in biology for instance. For this, we propose to use the iterative barycenter of the empirical distribution as a good estimate of the *mean* information conveyed by the data. Moreover, this estimate has the advantage that if the different distribution are warped from an unknown distribution, the empirical iterative barycenter converges to this pattern when the number of replications grows large.

Assume we observe $j = 1, \ldots, J$ samples of $i = 1, \ldots, n$ points $X_{i,j} \in \mathbb{R}^d$ which are i.i.d realizations of measures μ_j. Hence we observe cloud points or in an equivalent way $\mu_j^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_{i,j}}$ empirical versions of the measures μ_j. It is well-known that considering the mean with respect to the number of samples J of all observation points does not provide a good model of the *mean* behaviour. Instead we here consider the iterative barycenter $\mu_B^J = IB(\mu_j, \frac{1}{J})$ defined in Definition 3.1. The following proposition shows that the barycenter of the empirical distributions provides a good estimate for this *mean shape*. We point out that this estimator corresponds to the so-called Fréchet mean of the empirical measures.

Proposition 5.1. *Assume that the observations $X_{.,j} \sim \mu_j$ are warped by a centered admissible deformation process from an unknown template distribution μ continuous with respect to Lebesgue measure. Set $\mu_B^{n,J} \in Bar(\mu_j^n, \frac{1}{J})$, an empirical mean of the empirical distribution. As $n \to +\infty$, we have*

$$\mu_B^{n,J} \longrightarrow \mu_B^J.$$

Moreover, when $n \to +\infty$ and $J \to +\infty$, $\mu_B^{n,J}$ is a consistent estimate of μ, in the sense that

$$\mu_B^{n,J} \longrightarrow \mu \quad \text{in } W_2 \text{ distance.}$$

We point out that $\mu_B^{n,J}$ exits but is not unique. Actually, to ensure uniqueness, one may consider a regularized version of the empirical measures. For instance let γ_ε denotes a $\mathcal{N}(0, \varepsilon \mathrm{I}_d)$ measure. Set $\widehat{\mu_j^n} = \mu_j^n * \gamma_{1/n}$. In this case $\widehat{\mu_B^{n,J}} = \mathrm{Bar}(\mu_j^n, \frac{1}{J})$ is uniquely defined and as $n \to +\infty$, we have $\widehat{\mu_B^{n,J}} \longrightarrow \mu_B$ in Wasserstein distance. Note that any other regularization scheme may be used as soon as the corresponding measures converge to the true measures in Wasserstein distance when n goes to infinity.

An important application is given by the issue of ensuring equality between the candidates in an exam with several different referees. This constitutes a natural extension of the work in [14] to higher dimensions.

Consider an examination with a large number of candidates, such that it is impossible to evaluate the candidates one after another. The students are divided into J groups, and J boards of examiners are charged to grade these groups: each board grades one group of candidates. The evaluation is performed by assigning p scores. The J different boards of examiners are supposed to behave the same way, so as to respect the equality among the candidates. Moreover it is assumed that the sampling of the candidates is perfect in the sense that it is done in such a way that each board of examiners evaluates candidates with the same global level. Hence, if all the examiners had the same requirement levels, the distribution of the ranks would be the same for all the boards of examiners. Here, we aim at balancing the effects of the differences between the examiners, gaining equity for the candidates. The situation can be modeled as follows. For each group j among J groups of candidates, let $\mathbf{X}^j = \left\{X_i^j \in \mathbb{R}^p,\ i = 1, \ldots, n\right\}$ denote the scores of the students within this group. Let μ_j and μ_j^n be respectively the measure and the empirical measure of the scores in the j-th group.

We aim at finding the average way of ranking, with respect to the ranks that were given within the p bunches of candidates. For this, assume that there is such an average measure, and that each group-specific measure is warped from this reference measure by a random process. A good choice is given by the barycenter measure In order to obtain a global common ranking for the N candidates, one can now replace the p group-specific rankings by the sole ranking based on barycenter measure. Indeed each measure can be pushed towards the barycenter. As a result, we obtain a new set of scores for the N candidates, which can be interpreted as the scores that would have been obtained, had the candidates been judged by an average board of examiners.

5.2 Principal Component Analysis with Wasserstein distance

Once we have succeeded in defining a mean of a collection of distributions, then the second step consists in understanding the variability of the the different experiments with respect to this *average* distribution, which is, in statistics, the aim of the so-called PCA analysis. In a Euclidean space, a natural way to define principal components is through the minimization of the variance of the residuals. This concept has been extended to non Euclidean situations such as manifolds, Kendall's shape spaces in [24]. The principal component directions are replaced by principal component *curves* from a suitable family of curves, e.g. geodesics. In our framework, we generalize this idea to the Wasserstein distance.

As previously, let $\mu_1, \ldots, \mu_J$ be measures in $\mathcal{P}_2(\mathbb{R}^d)$ and let μ_B be the mean defined as the Barycenter $\mu_B = Bar(\mu_j, \frac{1}{J})$. Let $S_j,\ j = 1, \ldots, J$ be the transport plan between the μ_j's and μ_B and write $\mu_j = S_{j\#}\mu_B$. Assume the (Id, S_j) are an admissible family of transformations. Clustering the experiments in order to build coherent groups is usually achieved by comparing a distance between these distributions. Here by choosing the Wasserstein distance we get that

$$W_2^2(\mu_B, \mu_j) = \int |S_j(x) - x|^2 d\mu_B = \|S_j - \mathrm{id}\|^2_{L^2(\mu_B)}.$$

Hence statistical analysis of the distributions μ_j's amounts to clustering their Wasserstein square distance $\|S_j - \mathrm{id}\|^2_{L^2(\mu_B)} \in \mathbb{R}^+$.

It is known that the Wasserstein metric endows the space of probability measures with a formal Riemannian structure, in which it is possible to define geodesics, tangent spaces, etc., see [19], [4]. We propose here a method of principal component analysis using Wasserstein distance based on geodesics of the intrinsic metric, which follows the ideas developed in [19]. For this, consider a geodesic segment γ at point μ with direction T, which can be written as

$$\forall t \in [0, 1],\ \gamma(t) = ((1 - t)Id + tT)_{\#}\mu.$$

We extend the definition of γ to every $t \in \mathbb{R}$, with the important provision that γ is in general *not* a geodesic curve for the whole range of $t \in \mathbb{R}$. We perform PCA with respect to this family of curves which we somewhat abusively refer to as "geodesic curves" on their extended range. We will come back to this discussion at the end of our analysis.

For every μ_j, the natural distance to the geodesic curve γ is given by

$$d^2(\mu_j, \gamma) = \inf_{t \in \mathbb{R}} W_2^2(\mu_j, \gamma(t)).$$

Definition 5.1. A geodesic γ_1 is called a *first generalized principal component geodesic (GPCG)* to the μ_j's if it minimizes the following quantity

$$\gamma \mapsto \frac{1}{J} \sum_{j=1}^{J} d^2(\mu_j, \gamma) \tag{4}$$

Then we define the *second GPCG*, a geodesic γ_2 which minimizes (4) over all geodesics that have at least one point in common with γ_1 and that are orthogonal to γ_1 at all points in common.
Every point $\mu^\star$ that minimizes $\mu \mapsto \frac{1}{J}\sum_{j=1}^{J} W_2^2(\mu_j, \mu)$ over all common points of γ_1 and γ_2 will be called a *principal component geodesic mean*. Given the first and the second principal component geodesics γ_1 and γ_2 with principal component geodesic mean $\mu^\star$ we say that a geodesic γ_3 is a *third principal component geodesic* if it minimizes (4) over all geodesics that meet previous principal components orthogonally at $\mu^\star$. Analogously, principal component geodesics of higher order are defined.

Here we will focus on the computation of the first geodesic component γ_1. We will only consider geodesic curves from μ_B, in that case note that $\mu_B = \mu^\star$. Hence we root our

analysis at the central point given by μ_B which plays the role of the mean of the sample of distributions. In this setting, we first define a geodesic starting at a measure μ_B directed by a map T as $\gamma(t) = ((1-t)Id + tT)_{\#}\mu_B$. We consider a family of maps T such that (Id, S_j, T) is an admissible family of deformations. Hence, in that case, the distance of any measure μ_j with respect to such a geodesic can be written as

$$\begin{aligned} d^2(\mu_j, \gamma) &= \inf_{t\in\mathbb{R}} \int \left[((1-t)Id + tT) \circ S_j^{-1} - Id\right]^2 d\mu_j(x) \\ &= \inf_{t\in\mathbb{R}} \int \left[((1-t)Id + tT) - S_j\right]^2 d\mu_B(x) \\ &= -\frac{< S_j - Id, T - Id >^2_{L^2(\mu_B)}}{\|T - Id\|^2_{L^2(\mu_B)}} \end{aligned}$$

where $\|.\|_{L^2(\mu_B)}$ denotes the quadratic norm with respect to the measure μ_B with corresponding scalar product $< ., . >_{L^2(\mu_B)}$. Finally PCA with respect to Wasserstein distance amounts to minimizing with respect to T the quantity $\sum_{j=1}^J d^2(\mu_j, \gamma)$, which can be written as

$$T \mapsto -\sum_{j=1}^J | < S_j - Id, \frac{(T - Id)}{\|T - Id\|_{L^2(\mu_B)}} > |^2_{L^2(\mu_B)}.$$

If we set $v = T - Id$, this maximization can be written as finding the solution to

$$\arg \max_{v,\, \|v\|_{L^2(\mu_B)}=1} \sum_{j=1}^J | < S_j - Id, v > |^2_{L^2(\mu_B)},$$

which corresponds to the functional principal component analysis of the maps S_j, $j = 1, \ldots, J$ in the space $L^2(\mu_B)$. This analysis can be achieved using tools defined for instance in [24]. Finally, if we get $T^{(1)}$ the map corresponding to the first functional principal component, the corresponding principal geodesic if obtained by setting $\gamma^{(1)}(t) = ((1-t)Id + tT^{(1)})_{\#}\mu_B$. The other principal components can be computed using the same procedure.
In the one dimensional case, the situation is simpler since, the distance between μ_j with distribution function F_j and a geodesic γ from μ_B with distribution function F_B to $T_{\#}\mu_B$ is given by

$$d^2(\mu_j, \gamma) = \inf_{t\in\mathbb{R}} \int \left[((1-t)Id + tT) \circ F_B^{-1} - F_j^{-1}\right]^2 dt.$$

Hence PCA analysis amounts to maximizing for all functions T

$$T \mapsto \sum_{j=1}^J | < S_j \circ F_B^{-1} - F_B^{-1}, \frac{(T - Id) \circ F_B^{-1}}{\|(T - Id) \circ F_B^{-1}\|} > |^2,$$

which corresponds to the functional PCA of the maps S_j, $j = 1, \ldots, J$ in the space $L^2(\mu_B)$ without any restriction.

Let us come back to the caveat that the curves chosen are not Wasserstein geodesics on the entire parameter range. It is easy to check in the one dimensional case (see [4]) that

a curve $\gamma(t) = ((1-t)Id + tT)_{\#}\mu$ is a geodesic curve for all $t \in \mathbb{R}$ such that $(1-t)\mathrm{Id} + tT$ is an increasing function. Assuming T' takes values in the interval $[a, b]$, $0 < a < 1 < b$, this means that γ is a geodesic curve for all $t \in [1/(a-1), 1/(b-1)]$. Once the analysis above yields the expression of $T^{(1)}$ and the t_j^* minimizing $d^2(\mu_j, \gamma)$, it is possible to check whether they fall in this range. Actually,

$$t_j^* = \frac{< S_j - Id, T^{(1)} - Id >_{L^2(\mu_B)}}{\|T^{(1)} - Id\|^2_{L^2(\mu_B)}}.$$

Hence, when the measures μ_j are not too far from their barycenter (i.e. when the $\|S_j - \mathrm{Id}\|_\infty$ are small) these conditions are met.

Within this framework, we can analyze the toy example of translation effect, studied in [16] or in [15]. Here consider i.i.d random variables $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ who are translated by parameters $\theta_j = (\theta_j^1, \ldots, \theta_j^p)$, $j = 1, \ldots, J$. Hence the observation model is $X_{ij} = X_i + \theta_j$. Let μ be the distribution of the X_i's and assume that this distribution admits a density f with respect to the Lebesgue measure. Hence μ_j's, the distributions of the X_{ij} are given by

$$\mu_j = T_{j\#}\mu$$

with $T_j(x) = x + \theta_j$. They admit densities with respect to Lebesgue measure, f_j's which are such that $\forall x \in \mathbb{R}^p$, $f_j(x) = f(x - \theta_j)$. In this case, μ_B the Barycenter of the μ_j's exists and is characterized by its density $f_B(x) = f(x - \bar{\theta})$, with $\bar{\theta} = \frac{1}{J}\sum_{j=1}^J \theta_j$.

In this context, each distribution μ_j can be expressed as

$$\mu_j = S_j \# \mu_B, \quad S_j(x) = \frac{1}{J}\sum_{k=1}^J T_k^{-1} \circ T_j(x) = x + \bar{\theta} - \theta_j.$$

Now finding the first geodesic component rooted in μ_B amounts to maximize with respect to T the quantity

$$\begin{aligned} T \mapsto & \sum_{j=1}^J | < S_j - Id, \frac{(T - Id)}{\|T - Id\|_{L^2(\mu_B)}} >_{L^2(\mu_B)} |^2 \\ & = \sum_{j=1}^J \|\theta_j - \bar{\theta}\|^2 \frac{(\int (T - Id) d\mu_B)^2}{\|T - Id\|^2_{L^2_{(\mu_B)}}} \end{aligned}$$

which is achieved by choosing $T = Id + c$ for all constant c, where $\|.\|$ denotes the norm in $\mathbb{R}^p$. Hence the first principal geodesic component is given by μ_t^1 with density $f(x - \bar{\theta} - t)$ while the variance explained is given by $\sum_{j=1}^J \|\theta_j - \bar{\theta}\|^2$. This corresponds actually to the variance of the deformations.

6 Appendix

Proof of Lemma 3.2

Proof. The existence of a solution of the multimarginal problem (1) follows from a classical compactness argument.

Let $\gamma \in \Gamma(\mu_1, ..., \mu_J)$ and set $\nu = T_{\#}\gamma$. For all $1 \leq j \leq J$,

$$W_2^2(\nu, \mu_j) \leq \int \|T(x_1, ..., x_J) - x_j\|^2 d\gamma(x_1, ..., x_J),$$

and thus,

$$\sum_{1 \leq j \leq J} \lambda_j W_2^2(\nu, \mu_j) \leq \int \sum_{1 \leq j \leq J} \lambda_j \|T(x_1, ..., x_J) - x_j\|^2 d\gamma(x_1, ..., x_J). \tag{5}$$

For $1 \leq j \leq J$ and a probability measure $\hat{\nu}$, denote $\hat{\pi}_j$ a minimiser of

$$\int \|x - x_j\|^2 d\pi(x, x_j)$$

over all $\pi \in \Gamma(\nu, \mu_j)$ and define Π by

$$\Pi(A \times B_1 \times ... \times B_J) = \hat{\nu}(A) \frac{\pi_1(A \times B_1)}{\hat{\nu}(A)} ... \frac{\pi_J(A \times B_J)}{\hat{\nu}(A)} \tag{6}$$

Suppose now that γ is moreover a minimizer of (1), we want to show that $\nu = T_{\#}\gamma$ is a barycenter. Indeed,

$$\begin{aligned}
\sum_{1 \leq j \leq J} \lambda_j W_2^2(\hat{\nu}, \mu_j) &= \sum_{1 \leq j \leq J} \lambda_j \int \|x - x_j\|^2 d\Pi(x, x_1, ..., x_J) && (7)\\
&= \int \sum_{1 \leq j \leq J} \lambda_j \|x - x_j\|^2 d\Pi(x, x_1, ..., x_J) \\
&\geq \int \inf_{z \in E} \sum_{1 \leq j \leq J} \lambda_j \|z - x_j\|^2 d\Pi(x, x_1, ..., x_J) && (8)\\
&= \int \sum_{1 \leq j \leq J} \lambda_j \|T(x_1, ..., x_J) - x_j\|^2 d\Pi(x, x_1, ..., x_J) && (9)\\
&\geq \int \sum_{1 \leq j \leq J} \lambda_j \|T(x_1, ..., x_J) - x_j\|^2 d\gamma(x_1, ..., x_J) && (10)\\
&\geq \sum_{1 \leq j \leq J} \lambda_j W_2^2(\nu, \mu_j), && (11)
\end{aligned}$$

where (7) holds by definition (6) and (9) holds since for fixed $x_1, ..., x_J$, the sum $\sum_{1 \leq j \leq J} \lambda_j \|x - x_j\|^2$ attains its minimum at $x = T(x_1, ..., x_J) = \sum_{1 \leq j \leq J} \lambda_j x_j$. The inequality (10) holds since γ is optimal and (11) holds by (5).

Since $\hat{\nu}$ was arbitrary, this shows that ν is a barycenter.

On the other hand, taking $\hat{\nu}$ a barycenter, inequality (8) becomes an equality, so that, for Π-almost all $(x, x_1, ..., x_J) \in \mathbb{R}^{d \times (J+1)}$,

$$\sum_{1 \leq j \leq J} \lambda_j \|x - x_j\|^2 = \inf_{z \in E} \sum_{1 \leq j \leq J} \lambda_j \|z - x_j\|^2$$

which shows that $x = T(x_1, ..., x_J)$ Π-almost surely, and thus that $\hat{\nu} = T_{\#}\Pi_p$, where Π_p is the projection of Π over the last J marginals. The fact that Π_p is a solution of (1) is a consequence of (5) and equality (10).

□

Proof of Theorem 3.1 We know by Lemma 3.2 that for all $n \geq 1$, there exists $\gamma^n \in \Gamma(\mu_1, ..., \mu_J)$ such that $\mu^n = T_{\#}\gamma^n$. We first show that the sequence $(\gamma^n)_{n\geq 1}$ is tight. Let $B_1, \ldots, B_J$ be large balls in $\mathbb{R}^d$, we have

$$\begin{aligned}\gamma^n((B_1 \times \ldots B_J)^c) &= \gamma^n(\cup_{j=1}^J E \times \ldots \times E \times B_j^c \times E \ldots \times E) \\ &\leq \sum_{j=1}^J \gamma_n(E \times \ldots \times E \times B_j^c \times E \ldots \times E) \\ &= \sum_{j=1}^J \mu_j^n(B_j^c).\end{aligned}$$

Thus, tightness of the sequences $(\mu_j^n)_{n\geq 1}$ guarantees tightness of $(\gamma^n)_{n\geq 1}$. Note that the under the assumption of the convergence of μ_j^n, $n \geq 1$ in Wasserstein distance, we recover the compactness of $(\gamma^n)_{n\geq 1}$ in Wasserstein topology. Indeed, denote γ any weak limit of the tight sequence $(\gamma^n)_{n\geq 1}$, the second moments are converging:

$$\begin{aligned}\int |x|^2 d\gamma^n &= \sum_{j=1}^J \int |x_j|^2 d\mu_j^n \\ &\to \int |x_j|^2 d\mu_j = \int |x|^2 d\gamma.\end{aligned}$$

Here we used the fact that Wasserstein's convergence coincides with weak convergence together with the convergence of second order moments. The above implies tightness of the sequence of barycenters $\mu_B^n, n \geq 1$: indeed, it is the push-forward of the tight sequence $(\gamma^n)_{n\geq 1}$ by the application $T : \mathbb{R}^{d\times J} \to \mathbb{R}^d$, which is Lipschitz continuous (with Lipschitz constant bounded by 1). It is readily checked that this operation preserves tightness, as it preserves convergence (in weak and Wasserstein topologies).

We conclude by showing that any limiting point μ^∞ is a minimizer for the barycenter problem associated with $\mu_1, \ldots, \mu_J$. Denote by μ_B a barycenter of $\mu_1, \ldots, \mu_J$. Since μ_B^n is a barycenter for $\mu_1^n, \ldots, \mu_J^n$, we have

$$\sum_{j=1}^J \lambda_j W_2^2(\mu_B^n, \mu_j^n) \leq \sum_{j=1}^J \lambda_j W_2^2(\mu_B, \mu_j^n).$$

Since, up to a subsequence, $\mu_B^n \to \mu^\infty$ in Wasserstein distance, letting $n \to +\infty$ shows

$$\sum_{j=1}^J \lambda_j W_2^2(\mu^\infty, \mu_j) \leq \lim \sum_{j=1}^J \lambda_j W_2^2(\mu_B, \mu_j^n). \tag{12}$$

Proof of Theorem 4.2

Proof. For the first part, (2), we use induction on J. For $J = 1$, the result is obvious. Suppose then that it is established for $J \geq 1$. Choose $T_1, \ldots, T_{J+1}$ from a family of admissible deformations, and fix $\lambda_1, \ldots, \lambda_{J+1}$ with $\sum_{j=1}^{J+1} \lambda_j = 1$. Using the definition of the iterated barycenter, we have

$$\begin{aligned} IB((\mu_j, \lambda_j)_{1\leq j\leq J+1}) &= \mathrm{Bar}\left(IB\left((\mu_j, \lambda_j)_{1\leq j\leq J}), \sum_{j=1}^{J} \lambda_{,j} \right), (\mu_{J+1}, \lambda_{J+1}) \right) \\ &= \mathrm{Bar}\left(\left((\frac{1}{\Lambda_j} \sum_{j=1}^{J} \lambda_j T_j)_{\#}\mu, \Lambda_J \right), (\mu_{J+1}, \lambda_{J+1}) \right) \end{aligned}$$

where we set $\Lambda_J = \sum_{j=1}^{J} \lambda_j$.
Set $\nu = (\frac{1}{\Lambda_j} \sum_{j=1}^{J} \lambda_j T_j)_{\#}\mu$. As $\mu_{J+1} = T_{J+1\#}\mu$, we have also $\mu = (T_{J+1})_{\#}^{-1}\mu_{J+1}$, and

$$\begin{aligned} \nu &= (\frac{1}{\Lambda_j} \sum_{j=1}^{J} \lambda_j T_j) \circ (T_{J+1})_{\#}^{-1}\mu \\ &= (\frac{1}{\Lambda_j} \sum_{j=1}^{J} \lambda_j T_j \circ (T_{J+1})^{-1})_{\#}\mu_{J+1}. \end{aligned}$$

Now, observe that by assumption all the maps $T_j \circ (T_{J+1})^{-1}$ are gradients of convex functions, so that their convex combination also is. By Brenier's theorem, the map

$$\mathcal{T} = \frac{1}{\Lambda_j} \sum_{j=1}^{J} \lambda_j T_j \circ (T_{J+1})^{-1}$$

is the Brenier map from μ_{J+1} to ν. We deduce that the barycenter of ν and μ_{J+1} is

$$\begin{aligned} &(\lambda_{J+1}\mathrm{Id} + \Lambda_J \mathcal{T})_{\#}\, \mu_{J+1} \\ &= (\lambda_{J+1} T_{J+1} + \Lambda_J \mathcal{T} \circ T_{J+1})_{\#}\, \mu \\ &= (\sum_{j=1}^{J+1} \lambda_j T_j)_{\#}\mu. \end{aligned}$$

This finishes the first part of the proof.

For the identification of the barycenter and the iterative barycenter given in (3), we proceed as follows. Set $T(x_1, \ldots, x_J) = \sum_{j=1}^{J} \lambda_j x_j$ for $x_1, \ldots, x_J \in \mathbb{R}^d$. Proposition 4.2 of [1] claims that the barycenter of $(\mu_j, \lambda_j)_{1\leq j\leq J}$, denoted by μ_B, satisfies $\mu_B = T_{\#}\gamma$ where $\gamma \in \mathcal{P}((R^d)^J)$ is the unique solution of the optimization problem

$$\inf \left\{ \int \sum_{j=1}^{J} \lambda_j |T(x) - x_j|^2 d\gamma(x_1, \ldots, x_J), \quad \gamma \in \Pi(\mu_1, \ldots, \mu_J) \right\}$$

where $\Pi(\mu_1, \ldots, \mu_J)$ is the set of probability measures on $\mathbb{R}^{dJ}$ with j-th marginal μ_j, $1 \leq j \leq J$. This can be rewritten as

$$\frac{1}{2} \inf \left\{ \int \sum_{i,j=1}^{J} \lambda_i \lambda_j |x_i - x_j|^2 d\gamma(x_1, \ldots, x_J), \quad \gamma \in \Pi(\mu_1, \ldots, \mu_J) \right\}.$$

The integral is bounded below by $\sum_{i,j=1}^{J} \lambda_i \lambda_j W_2^2(\mu_i, \mu_j)$ (because each term of the sum is bounded by $W_2^2(\mu_i, \mu_j)$). On the other hand, choosing

$$\gamma = (T_1, \ldots, T_j)_{\#}\mu,$$

we see that $\gamma \in \Pi(\mu_1, \ldots, \mu_J)$, and that

$$\int |x_j - x_i|^2 d\gamma = \int |T_j(x) - T_i(x)|^2 d\mu(x) = \int [T_j \circ {T_i}^{-1}(x) - x|^2 \mu_i(dx) = W_2^2(\mu_i, \mu_j).$$

Thus γ is optimal, and we have

$$\mu_B = T_{\#}\gamma = (\sum_{j=1}^{J} \lambda_j T_j)_{\#}\mu.$$

□

Proof of Theorem 4.3

Proof. Using the results of Proposition 4.4, we get that

$$W_2^2(\mu_B, \mu) \leq \int |\frac{1}{J} \sum_{j=1}^{J} T_j(x) - x|^2 \mu(dx).$$

Almost sure convergence towards 0 of $\frac{1}{J} \sum_{j=1}^{J} (T_j - \mathrm{id})$ is directly deduced from Corollary 7.10 (p. 189) in [21], which is an extension of the Strong Law of Large Numbers to Banach spaces. Then the result follows from dominated convergence.

Likewise, obtaining error bounds is straightforward. Assuming that $\|T - id\|_{L^2} \leq M$ a.s., we can use Yurinskii's version of Bernstein's inequality in Hilbert spaces ([27], p. 491) to get the result announced. □

Acknowledgements: We thank an anonymous referee for his/her comments and suggestions which contribute to numerous improvements in the paper.

References

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

[2] Stéphanie Allassonière, Yali Amit, and Alain Trouvé. Toward a coherent statistical framework for dense deformable template estimation. *Journal of the Statistical Royal Society (B)*, 69:3–29, 2007.

[3] Pedro César Álvarez-Esteban, Eustasio del Barrio, Juan Antonio Cuesta-Albertos, and Carlos Matrán. Trimmed comparison of distributions. *J. Amer. Statist. Assoc.*, 103(482):697–704, 2008.

[4] 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), 2004.

[5] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable template. *Journal of the American Statistical Association*, 86:376–387, 1991.

[6] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[7] B. Bercu and P. Fraysse. A Robbins-Monro procedure for estimation in semiparametric regression models. *ArXiv e-prints*, January 2011.

[8] Jérémie Bigot, Jean-Michel Loubes, and Myriam Vimond. Semiparametric estimation of shifts on compact Lie groups for image registration. *Probab. Theory Related Fields*, 152(3-4):425–473, 2012.

[9] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

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

[11] Juan Antonio Cuesta and Carlos Matrán. Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.*, 17(3):1264–1276, 1989.

[12] Eustasio del Barrio, Juan A. Cuesta-Albertos, Carlos Matrán, and Jesús M. Rodríguez-Rodríguez. Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.*, 27(4):1230–1239, 1999.

[13] Eustasio del Barrio, Evarist Giné, and Carlos Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.*, 27(2):1009–1071, 1999.

[14] J.F. Dupuy, J.M. Loubes, and E. Maza. Non parametric estimation of the structural expectation of a stochastic increasing function. *Statistics and Computing*, pages 1–16, 2011.

[15] Santiago Gallón, Jean-Michel Loubes, and Elie Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129 – 142, 2013.

[16] F. Gamboa, J-M. Loubes, and E. Maza. Semi-parametric estimation of shits. *Electronic Journal of Statistics*, 1:616–640, 2007.

[17] U. Grenander. General pattern theory|a mathematical study of regular structures, oxford university press. *New*, York:1994.

[18] Eldad Haber, Tauseef Rehman, and Allen Tannenbaum. An efficient numerical method for the solution of the L_2 optimal mass transfer problem. *SIAM J. Sci. Comput.*, 32(1):197–211, 2010.

[19] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica*, 20(1):1–58, 2010.

[20] D. G. Kendall, D. Barden, T. K. Carne, and H. Le. *Shape and shape theory*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1999.

[21] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[22] R.J. McCann. A convexity principle for interacting gases* 1. *advances in mathematics*, 128(1):153–179, 1997.

[23] Xavier Pennec. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *J. Math. Imaging Vision*, 25(1):127–154, 2006.

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

[25] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[26] Alain Trouvé and Laurent Younes. Metamorphoses through lie group action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005.

[27] VV Yurinski. Exponential inequalities for sums of random vectors. *Journal of multivariate analysis*, 6(4):473–499, 1976.