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; Classification.

e-mail: boissard,le-gouic,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 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 [12], [21] and references therein, image analysis in [25] [24] or [4], shape analysis in [17] or [14] with many applications ranging from biology in [8] to pattern recognition [22] 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 instances scale location parameters, rotations, actions of parameters of Lie groups as in [7] 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 statistics or

semi-parametric 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 [5] or [20] 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 as pointed out in [6]. 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. 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 applications for real data case are pointed out in Section 5.

2 Barycenters in the Wasserstein space: Notations and general results

Let (E, d, Ω) denotes 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 $x_1,\dots,x_J$ with weights $\lambda_1,\dots,\lambda_J$, $\lambda_j \geq 0$, $\sum_{j=1}^J = 1$, 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 $\mu_1,\dots,\mu_J \in \mathcal{P}_2(E)$ endowed with weights $\lambda_1,\dots,\lambda_J$, 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 $E = \mathbb{R}^d$ and the measures μ_j, $1 \leq j \leq J$ satisfy suitable assumptions, the barycenter exists and 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 [9]). *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be compactly supported 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.

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.

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 in Wasserstein topology.

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 measures absolutely continuous with respect to the Lebesgue measure. Let $\lambda_1, \ldots, \lambda_J$ be positive weights. Let*

$$\hat{\mu}^n(\lambda) = Bar((\mu_j^n, \lambda_j)_{1 \leq j \leq J}).$$

Let $\mu_1, \ldots, \mu_J \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous w.r.t. Lebesgue measure, and let

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

then when $n \to +\infty$

$$W_2(\hat{\mu}^n(\lambda), \mu_B(\lambda)) \longrightarrow 0.$$

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 :

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

Remark. Iterated barycenters are well-suited to computations, since there exist efficient numerical methods to compute McCann's interpolant, see e.g. [15], [23]. 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.

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

Theorem 3.2. *The iterated barycenter is consistent : if $\mu_j^n \to \mu_j$ in W_2 distance for $j = 1, \ldots, J$, 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.

4 Deformations of a template measure

We now would like to use Wasserstein barycenters or iterated barycenters in the following framework : assume that we observe probability measures $\mu_1, \ldots, \mu_J$ 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 [10]. 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 $F \circ h_1^{-1}, \ldots, F \circ h_n^{-1}$

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, $j = 1, \ldots, J$ 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 $S : x \mapsto (F_1(x_1), \dots, F_n(x_n))$ and $T : x \mapsto (G_1(x_1), \dots, G_n(x_n))$. The map $S \circ T^{-1}$ is given by

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

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.

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 :*

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

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.

Proposition 4.3. *Let $\mu_j = (T_j)_{\#}\mu$, where $(T_i)_{i\in I}$ is an admissible family of deformations and $\mu << \lambda$. Then*

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

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 [10] or [11]: 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 $\gamma_1, \ldots, \gamma_J$ on $\mathbb{R}^n$, defined by their covariances matrices $S_1, \ldots S_J \in \mathcal{S}_n^{++}$.

   According e.g. to [19], 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 4.4. *Assume that μ is compactly supported. As soon as $\varphi = \mathrm{id}$, the barycenter μ_B 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 bounds :

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

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 4.5. *Let $(T_i)_{i\in I}$ be an admissible family of deformations on a domain $\Omega \subset \mathbb{R}^n$, and let $\mu \in \mathcal{P}_2(\Omega)$, $\mu << \lambda$. 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 play a crucial to analyze the data, in image analysis, kinetics in biology for instance. For this, we propose to use the iterative barycenter of a smooth approximation 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

Assume we observe $j = 1, \dots, J$ samples of $i = 1, \dots, n$ 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 $\widehat{\mu}_j^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_{i,j}}$ 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 $\mu_B = IB(\mu_j, \frac{1}{J})$ defined in Definition 3.1. The following proposition shows that the regularization of the empirical distributions provides a consistent estimator of the true barycenter of the corresponding distributions.

Proposition 5.1. *Let γ_ε denote a $\mathcal{N}(0, \varepsilon \mathrm{I}_d)$ measure. Set*

$$\mu_j^n = \widehat{\mu}_j^n * \gamma_{1/n}.$$

Set $\mu_B^{n,J} = Bar(\mu_j^n, \frac{1}{J})$. As $n \to +\infty$, we have

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

Moreover, if the observations $X_{.,j} \sim \mu_j$ are warped from an unknown template distribution μ by a centered admissible deformation process, hence $\mu_B^{n,J}$ is a consistent estimate of μ_B, in the sense that when $n \to +\infty$ and $J \to +\infty$, we get

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

We point out that we have used here a Gaussian kernel regularization of the empirical measures $\widehat{\mu}_j$. Actually, this regularization is needed in order to obtain the existence of the barycenter of the data. 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.

In particular, kernel estimates can be used in this framework. For instance, if there exist for all $j = 1, \dots, J$, a density with respect to λ, the Lebesgue measure on $\mathbb{R}^d$, f_j such that, $\frac{d\mu_j}{d\lambda} = f_j$, one may use a kernel estimation of the density of the data. Let K be a multidimensional kernel in $\mathbb{R}^d$. Let $f_{j,n} = \frac{1}{nh^d}\sum_{i=1}^n K_h(., X_{i,j})$ be an estimator of the density f_j. In this case, set $\mu_{j,n}$ the distribution such that $\frac{d\mu_{j,n}}{d\lambda} = f_{j,n}$. Let $h = h_n$ goes to zero. In this case, we clearly have $W_2(\mu_{j,n}, \mu) \to 0$ when n goes to infinity. Hence previous Proposition entails the consistency of the iterate barycenter of the $\mu_{j,n}$'s.

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 [10] 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 m 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 $\mu_{j,n}$ 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 Discriminant analysis with Wasserstein distance

Once we have succeeded in defining a mean of a collection of distributions, then the second step consists in trying to differentiate the different experiments with respect to this *average* distribution. For this consider $S_j,\ j = 1, \ldots, J$ the transport plan between the μ_j's and μ_B and write $\mu_j = {S_j}_{\#}\mu_B$. 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 $\|S_j - \mathrm{id}\|^2_{L^2(\mu_B)} \in \mathbb{R}^+$, which can be easily achieved by any clustering methodology.

Moreover PCA analysis can be conducted by generalizing the ideas of the usual PCA on Euclidean space. To extend the framework, one replaces the principal component directions with principal component *curves* from a suitable family of curves, e.g. geodesics. Following this idea, classical principal component analysis has been extended to situations such as manifolds, Kendall's shape spaces, and functional settings, see [21].

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, tangents spaces, etc., see [16], [3]. We propose here a method of principal component analysis using Wasserstein distance based on geodesics of the intrinsic metric for the one dimensional case, which follows the ideas developed in [16]. 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, the natural distance to the geodesic curve γ is given by

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

Hence for any μ_j, $j = 1, \ldots, J$ let F_B and respectively F_j be the distribution functions of the barycenter measure μ_B and respectively the measures μ_J, then define

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

A geodesic γ_1 is called a first component geodesic to the μ_j's if it minimizes the following quantity

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

Then we call a geodesic γ_2 minimizes (1) 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 (1) 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 first define a geodesic starting at a measure μ with distribution function F and directed by their increasing function T as $\gamma(t) = ((1-t)Id + tT)_{\#}\mu$. Hence, in that case, the distance of any measure μ_j with respect to such a geodesic can be written as

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

where $\|.\|$ denotes the usual L^2 norm with corresponding scalar product $< ., . >$. Finally PCA with respect to Wasserstein distance amounts to minimizing with respect to T and F the following quantity

$$\begin{aligned} T &\mapsto \sum_{j=1}^{J} d^2(\mu_j, \gamma) \\ &= \sum_{j=1}^{J} \|S_j \circ F_B^{-1} - F^{-1}\|^2 - \sum_{j=1}^{J} | < S_j \circ F_B^{-1} - F^{-1}, \frac{(T - Id) \circ F^{-1}}{\|(T - Id) \circ F^{-1}\|} > |^2. \end{aligned}$$

This optimization program can not be solved easily. Note that it requires to maximize in T the quantity

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

If we set $v = (T - Id) \circ F^{-1}$, this maximization can be written as finding the solution to

$$\arg \max_{v,\, \|v\|=1} \sum_{j=1}^{J} | < S_j \circ F_B^{-1} - F^{-1}, v > |^2,$$

which corresponds to the functional principal component analysis of the maps $S_j \circ F_B^{-1}, j = 1, \ldots, J$ where F plays the role of the *mean* of the data. Hence, in this framework, choosing F equal to F_B is a natural approximation of the solution of the initial optimization program. Then the final PCA is conducted by considering geodesics starting at μ_B with direction found by maximizing

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

which corresponds to a functional PCA in $L^2_{(\mu_B)}$. This analysis can be achieved using tools defined for instance in [21]. Finally, if we get $T^{(1)}$ the map corresponding to the first functional principal component, the corresponding principal geodesic if obtained by setting $\gamma^{(1)} = ((1-t)Id + tT^{(1)})_{\#}\mu_B$. The other principal components can be computed using the same procedure.

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 (see [3]) 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,

$$t_j^* = \frac{< F_j^{-1} - F_B^{-1}, (T^{(1)} - Id) \circ F_B^{-1} >}{\|(T^{(1)} - Id) \circ F_B^{-1}\|^2}.$$

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.

6 Appendix

Proof of Theorem 3.1

Proof. Let

$$T : (x_1, \ldots, x_J) \mapsto \sum_{j=1}^{J} \lambda_j x_j.$$

Following [1], we call $\gamma \in \mathcal{P}(\mathbb{R}^{d\times J})$ a solution of the multi-marginal problem associated with $\mu_1, \dots, \mu_J$ if it is a minimizer for the functional

$$F(\tilde{\gamma}) = \int (\sum_{j=1}^{J} \lambda_j |x_i - T(x)|^2) \tilde{\gamma}(dx_1, \dots, dx_J)$$

among all measures $\tilde{\gamma} \in \mathcal{P}(\mathbb{R}^{d\times J})$ with marginals $\mu_1, \dots, \mu_J$. Theorem 4.1 in [1], quoting from W.Gangbo and A.Święch [13], shows that when the μ_j are absolutely continuous w.r.t. Lebesgue measure the multi-marginal problem has a unique solution γ (actually absolute continuity of the measures is more than is required in the theorem). Moreover (Proposition 4.2 in [1]) the barycenter of the μ_j is obtained as $\mu_B = T_{\#}\gamma$.

For every $n \geq 1$, we associate to μ_j^n, $1 \leq j \leq J$, the solution γ^n of the multi-marginal problem. We also denote by γ^* the solution of the multi-marginal problem w.r.t. $\mu_1, \dots, \mu_J$.

We show that the sequence γ^n is weakly tight. Let $B_1, \dots, B_J$ be large balls in $\mathbb{R}^d$, we have

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

Tightness of the sequences μ_j^n guarantees tightness of γ^n. If convergence of μ_j^n, $n \geq 1$, holds in Waserstein distance, we also recover tightness of γ^n in Wasserstein topology. Indeed, the second moments are bounded as they form a convergent sequence :

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

The above implies tightness of the sequence of barycenters : 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 $\hat{\mu}_\infty$ is a minimizer for the barycenter problem associated with $\mu_1, \dots, \mu_J$, and by invoking the uniqueness of the barycenter. Since $\hat{\mu}^n$ is the barycenter for $\mu_1^n, \dots, \mu_J^n$, we have

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

Assume now that up to a subsequence, $\hat{\mu}^n \to \hat{\mu}_\infty$ in Wasserstein distance, and let $n \to +\infty$. Since W_2 is weakly lower semi-continuous, we get

$$\sum_{j=1}^{J} \lambda_j W_2^2(\hat{\mu}_\infty, \mu_j) \leq \liminf \sum_{j=1}^{J} \lambda_j W_2^2(\hat{\mu}^n, \mu_j^n) \leq \liminf \sum_{j=1}^{J} \lambda_j W_2^2(\hat{\mu}^*, \mu_j^n).$$

The right-hand side converges to the value $\sum_{j=1}^{J} \lambda_j W_2^2(\hat{\mu}^*, \mu_j)$, which is minimal by definition. This shows that the inequalities are equalities and it concludes the proof.

□

Proof of Theorem 3.2

Proof. One sees from the definition that it is sufficient to prove the result for two measures, because then the result may be obtained by recurrence. Consider then $\mu_n \to \mu$ and $\nu_n \to \nu$ (convergence is understood in Wasserstein topology), and fix $t \in (0,1)$. The Brenier transport map between μ and ν (resp. μ_n and ν_n) will be denoted by T (resp. T_n). Also denote by μ^t, resp. μ_n^t the point in the Wasserstein geodesic between μ and ν (resp. μ_n and ν_n) at time t, that is to say

$$\mu^t = ((1-t)\mathrm{Id} + tT)_\# \mu \tag{2}$$
$$\mu_n^t = ((1-t)\mathrm{Id} + tT_n)_\# \mu_n. \tag{3}$$

As noted earlier, μ^t is the barycenter of μ and ν with weights $1-t$ and t, so that we need only prove weak continuity of $(\mu, \nu) \mapsto \mu^t$. We first take care of weak convergence, i.e. we assume that $\mu_n \rightharpoonup \mu$, $\nu_n \rightharpoonup \nu$ and we show that $\mu_n^t \to \mu^t$. The measure $\pi \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ defined by $\pi = (\mathrm{Id} \times T)_\# \mu$ is the unique optimal transport plan between μ and ν. Likewise, $\pi_n = (\mathrm{Id} \times T_n)_\# \mu_n$ is the optimal transport plan between μ_n and ν_n. Now, Theorem 5.20 in [26] (stability of transport plans) ensures that π_n weakly converges to π. Let $f_t : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ be defined by $f_t(x,y) = (1-t)x + ty$. The map f_t is continuous, so that pushing forward by f_t is a weakly continuous mapping. Therefore, $f_{t\#}\pi_n \rightharpoonup f_{t\#}\pi$. It only remains to check that in fact, $f_{t\#}\pi_n = \mu_n^t$ and $f_{t\#}\pi = \mu^t$.

We now look at the convergence in W_2 distance. Observe that the transport plans converge for the W_2 metric over $\mathcal{P}_2(\mathbb{R}^n \times \mathbb{R}^n)$: to see this, we use the fact that convergence in W_2 topology is equivalent to weak convergence plus convergence of second moments. And indeed, as we noted, $\pi_n \rightharpoonup \pi$, and on the other hand,

$$\begin{aligned}
\int |(x,y)|^2 d\pi_n(x,y) &= \int \left(|x|^2 + |y|^2\right) d\pi_n(x,y) \\
&= \int |x|^2 d\mu_n(x) + \int |y|^2 d\nu_n(y) \\
&\overset{n\to+\infty}{\longrightarrow} \int |x|^2 d\mu(x) + \int |y|^2 d\nu(y) \\
&= \int |(x,y)|^2 d\pi.
\end{aligned}$$

Now we observe that $\mu_n^t = f_{t\#}\pi_n$ and $\mu^t = f_{t\#}\pi$. But f_t is Lipschitz, and it suffices to use the fact that $W_2(f_{t\#}\pi_n, f_{t\#}\pi) \leq \|f_t\|_{\mathrm{Lip}} W_2(\pi_n, \pi)$.

□

Proof of Theorem 4.2

Proof. 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 proof.

□

Proof of Proposition 4.3

Proof. Set $T(x_1, \dots, x_J) = \sum_{j=1}^J \lambda_j x_j$ for $x_1, \dots, x_J \in \mathbb{R}^d$. 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

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

where $\Pi(\mu_1, \dots, \mu_J)$ 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

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

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

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

we see that $\gamma \in \Pi(\mu_1, \dots, \mu_J)$, 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 4.4

Proof. Using the results of Corollary 4.5, 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 [18], 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. □

Proof of Proposition 5.1

Proof. By the empirical law of large numbers, $\mu_j^n \rightharpoonup \mu_j$ weakly. Moreover $W_2(\widehat{\mu_j^n}, \mu_j^n) \leq \frac{1}{n}$, so that $\widehat{\mu_j^n} \to \mu_j$ in W_2 metric, see for instance in [26]. Hence using Theorem 3.2, we obtain the consistency of the barycenter. Moreover, if $\mu_j = T_{j\#}\mu$ for all $j = 1, \ldots, J$, where the T_j's are an admissible family of bounded deformations, hence Theorem 4.4 enables to get that

$$W_2(\mu_B^{n,J}, \mu_B) \longrightarrow 0$$

when both n and J goes to infinity. □

References

[1] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *Preprint*, 2010.

[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] 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.

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

[5] Mukund Balasubramanian and Eric L. Schwartz. The isomap algorithm and topological stability. *Science*, 295, 2002.

[6] Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.*, 31(1):1–29, 2003.

[7] J. Bigot, J-M. Loubes, and M. Vimond. Semiparametric estimation of shifts on compact lie groups for image registration. *Probability Theory and Related Fields*, 1, 2011.

[8] 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.

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

[10] 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.

[11] Santiago Gallón, Jean-Michel Loubes, and Elie Maza. Statistical Properties of the Quantile Normalization Method for Density Curve Alignment, hal-00593476. Technical report, 2011.

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

[13] W. Gangbo and A. Święch. Optimal maps for the multidimensional monge-kantorovich problem. *Communications on pure and applied mathematics*, 51(1):23–45, 1998.

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

[15] S. Haker and A. Tannenbaum. Optimal mass transport and image registration. In *Variational and Level Set Methods in Computer Vision, 2001. Proceedings. IEEE Workshop on*, pages 29–36. IEEE, 2001.

[16] 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.

[17] 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.

[18] 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.

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

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

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

[22] 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.

[23] S. Srivastava, M.A. Peletier, et al. Numerical algorithm for computation of the 2-wasserstein distance, and applications to the foundations of diffusion. 2009.

[24] Alain Trouvé. Probability measures and statistics on function spaces and nonlinear infinite dimensional spaces. *SAMSI, Special Issue on Program on the Geometry and Statistics of Shape Spaces*, Summer 2007.

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

[26] C. Villani. *Optimal transport: old and new*, volume 338. Springer Verlag, 2009.

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