A Parametric Registration model for warped distributions with Wasserstein's distance

Marina Agulló Antolín

Departamento de Estadística e Investigación Operativa. Facultad de Ciencias. Universidad de Valladolid. 7, paseo de Belén, 47011 Valladolid. Spain.

J.A. Cuesta Albertos

Departamento de Matemàticas. Facultad de Ciencias. Universidad de Cantabria. 39071 Santander. Spain.

Hélène Lescornel

INRIA Saclay. 1 rue Honoré d'Estienne d'Orves. Bâtiment Alan Turing, Campus de l'Ecole Polytechnique 91120 Palaiseau

Jean-Michel Loubes*

Institut de Mathématiques de Toulouse, Universite Paul Sabatier. 118, route de Narbonne F-31062 Toulouse Cedex 9

Abstract

We consider a parametric deformation model for distributions. More precisely, we assume we observe J samples of random variables which are warped from an unknown distribution template. We tackle in this paper the problem of estimating the individual deformation parameters. For this, we construct a registering criterion based on the Wasserstein distance to quantify the alignment of the distributions. We prove consistency of the empirical estimators.

Keywords: Wasserstein distance, deformation, semi-parametric model.

1. Introduction

In this paper we focus on the issue of registering measures warped by a parametric deformation operator. When confronted to large data sample, the data may present several sources of variability that include small deformations on the data such as translations, scale location models for instance or more general warping procedures. These deformations prevent the use of usual methods in data analysis and deserve a special statistical

*Corresponding author

Email addresses: marina@eio.uva.es (Marina Agulló Antolín), juan.cuesta@unican.es (J.A. Cuesta Albertos), helene.lescornel@inria.fr (Hélène Lescornel), loubes@math.univ-toulouse.fr (Jean-Michel Loubes)

The First and second authors are partially supported by the Spanish Ministerio de Ciencia y Tecnología, grants MTM2011-28657-C02-0 and MTM2011-28657-C02-02. First author is also supported by Consejería de Educación de la Junta de Castilla y León, grant VA212U13

treatment in order to align the data. We refer for instance to (Gamboa et al., 2007), (Dupuy et al., 2011), (Ramsay and Silverman, 2005), (Bercu and Fraysse, 2012) and references therein to applications to functional data analysis, to (Trouve and Younes, 2005) or (Amit et al., 1991) for applications in image analysis or to (Bolstad et al., 2003) for applications in biology.

In this paper we consider the problem of registration of distributions. Such situation occurs often in biology for example, when considering gene expression data obtained from microarray technologies, which are used to measure genome wide expression levels of genes in a given organism; the registration issue is known as normalization, see for instance in (Bolstad et al., 2003) and the related work (Gallon et al., 2013). Here we consider the extension of semi-parametric registration methods as in (Gamboa et al., 2007) or (Vimond, 2010a) to the problem of estimating a distribution of random variables, observed in a warping framework.

Actually, assume that we observe $i = 1, \dots, n$ samples of $j = 1, \dots, J$ independent random variables X_{ij} with distribution μ_j. Each sample is drawn from a *mean* distribution μ with some variations in the sense that there exist unobserved warping functions φ, such that, for all j, we have $\mu_j = \mu \circ \varphi_j^{-1}$. To deal with this issue, we assume a parametric model for the warping function. We consider that the deformations follow a known shape which depends on parameters, specific for each sample. Hence there are parameters $\theta^\star = (\theta_1^\star, \dots, \theta_J^\star)$ such that $\varphi_j = \varphi_{\theta_j^\star}, \quad$ for all $j = 1, \dots, J$. Each $\theta_j^\star$ represents the warping effect that undergoes the j^{th} sample, which must be removed to recover the unknown distribution by inverting the warping operator. So the observation model is

$$X_{ij} = \varphi_{\theta_j^\star}(\varepsilon_{ij}) \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant J, \tag{1}$$

where ε_{ij} are unobserved i.i.d. random variables with unknown distribution μ. Our objective is thus to build an estimator for the parameters $\theta_j^\star$.

When tackling registration issues, two points of view can be considered. On the one hand, a solution is given by choosing one observation as a reference and align all the data onto this chosen template. This problem has been investigated in (Lescornel and Loubes, 2012). On the other hand a more robust solution, and less sensitive to such a prior choice, is obtained by aligning the sample onto the mean of the warped distribution. This solution is discussed for the regression case in (Vimond, 2010a). In the following, we provide a generalization of this work to the case of the deformation of distribution, which moreover enables to handle the case of multidimensional warping parameters.

In this paper, to align the distribution, we use the so-called 2-Wasserstein distance to measure discrepancies (see Section 2 for details). This distance has been used previously with the same objective, the main difference being that these previous results do not use parametric models for the deformations as used in this paper. In fact, they employ a non-parametrical framework or, if any, they only handle linear transformations like those ones included in Examples 5.1 and 5.3. It is worth mentioning that, as suggested by a referee, our procedure may also handle multiple distortions which are applied in different segments of the curve under consideration (only Proposition 3.4 could fail because, in this case, the differentiability assumption **A7** could not be satisfied). However, we prefer to keep the analysis restricted to the case of single distortions in order to keep the statements and technicalities as simple as possible.

Concerning the previously known results, we would like to mention (Mémoli, 2011) where the author carries out an interesting and deep theoretical study on this problem. Apart from the references in this paper, we would like to mention (Haker et al., 2004) and (Ni et al., 2009) and, mostly, (Bonneel et al., 2014) and (Boissard et al., 2014) because their use of barycenters is quite close to the point of view we employ here. Slightly different is the problem handled in (Schmitzer and Schnorr, 2013) where the authors propose a procedure that, apart to align two images, also tries to identify the image of interest from the background.

The paper falls into the following parts. Section 2 provides a description of the parametric warping operator that acts over the deformations and the construction of the estimator. In Section 3, we prove the consistency of the estimators of the deformation parameters. In Section 4, we give a method to compute the alignment criterion. Section 5 presents examples of deformation families while some simulations are presented in Section 6. Proofs are postponed to Section 7.1.

2. A model for distribution deformation

In this section, we will precise the model considered to estimate the deformations. We start by giving some notations.

For a given sample $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, we denote by $Y_{(1)} \leqslant \cdots \leqslant Y_{(i)} \leqslant \cdots \leqslant Y_{(n)}$ its order statistics. The symbol $\rightharpoonup$ denotes convergence in distribution. In all the paper, we denote by $\| \quad \|_k$ the euclidean norm on $\mathbb{R}^k$ for all $k \in \mathbb{N}$, $k \geqslant 2$.

For $i = 1, \dots, n$ and $j = 1, \dots, J$, set ε_{ij} unobserved i.i.d. random variables with unknown distribution μ defined on an subset I_a of a complete, separable metric space (E, d). We assume that we observe only some deformations of these observations. More precisely, we consider a family of invertible deformation functions, indexed by parameters λ which warps a point x onto another point $\varphi_\lambda(x)$. The shape of the deformation is modelled by the known function φ while the amount of deformation is characterized by the parameter $\lambda \in \mathbb{R}^p$, for $p > 0$. Namely, set

$$\begin{array}{rcl} \varphi : \Lambda \times I_a & \to & I_b \\ (\lambda, x) & \mapsto & \varphi_\lambda(x) \end{array}$$

for Λ an open subset of $\mathbb{R}^p$ and I_a, I_b subsets of E which can be unbounded.

Recall that we observe the model (1)

$$X_{ij} = \varphi_{\theta_j^\star}(\varepsilon_{ij}) \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant J,$$

where $\theta_j^\star$ is the unknown deformation parameter in $\Lambda \subset \mathbb{R}^p$, associated with the j-th sample $(X_{1j}, \dots, X_{nj})$, and ε_{ij} are i.i.d random variables.

Our aim is to estimate the parameter $\theta^\star \in \Pi_{j=1}^J \Lambda$. For this, we will study a criterion based on a registration procedure for the distributions μ_j of each i.i.d. sample $(X_{1j}, \dots, X_{nj})$, for all $j = 1, \dots, J$. We denote its empirical law by $\mu_j^{(n)} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_{ij}}$, where $\mathbf{1}_x$ denotes the delta mass distribution at point x.

To recover the deformation parameter $\theta^\star$, we will minimize an energy needed to align all the distributions μ_j. For this, a natural distance to measure the deformation cost to align two distributions is given by the 2-Wasserstein distance which is at the heart of warping issues between probabilities or cloud points due to its relationship with transportation issues.

Consider the following set of probabilities

$$\mathcal{W}_2\left(E\right) = \left\{ P \text{ probability on } E \text{ s.t. } \int_E d(x_0, x)^2 dP(x) < \infty \text{ for some } x_0 \in E \right\},$$

where we recall that $d(.,.)$ is the distance on E.
Given two probabilities P and Q in $\mathcal{W}_2\left(E\right)$, we denote by $\mathcal{P}(P,Q)$ the set of all probability measures π over the product set $E \times E$ with first (resp. second) marginal P (resp. Q).

The transportation cost with quadratic cost function, or quadratic transportation cost, between these two measures P, Q is defined as

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

The quadratic transportation cost allows to endow the set $\mathcal{W}_2\left(E\right)$ with a metric by defining the 2-Wasserstein distance between P and Q as

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

We will use the W_2 metric in this work. This choice is led by the issue of optimal matching between cloud points. Yet other choices

$$W_r^r(P,Q) = \inf_{\pi \in \mathcal{P}(P,Q)} \int d(x,y)^r d\pi$$

are possible for different r and other distances d on the chosen metric space E. In particular, the earth-mover distance which corresponds to $r = 1$ and $E = \mathbb{R}^p$ could be used with more complicated calculations. However the study of this criterion falls beyond the scope of this paper. More details on Wasserstein distances and their links with optimal transport problems can be found in (Rachev and Rüschendorf, 1998) or (Villani, 2009) for instance.

It is known, see for instance (Bickel and Freedman, 1981), that the infimum defined in (2) is reached for some pair $(X^\star, Y^\star) \in \mathcal{P}(P,Q)$ which is called an optimal coupling. Hence one have $W_2(P,Q) = \mathbf{E}\left[\|X^\star - Y^\star\|^2\right]$.

An important subject of research related to the Wasserstein distance concerns the correlation structure of this optimal coupling. Especially the question of the existence of a measurable function $T : supp(P) \mapsto supp(Q)$ such that $Y^\star = T\left(X^\star\right)$ has received a lot of interest. Conditions for the existence of such a map (which is called an optimal map) are stated for instance in (Cuesta and Matran, 1989) or (Villani, 2009).

The case of Wasserstein distance over Hilbert spaces is investigated for instance in (Cuesta and Matran, 1989) and a study of more general settings can be found in (Villani, 2009). Statisticians have recently used this distance for instance in geometric inference in (Caillerie et al., 2011) or to build tests of similarity between distributions in (Alvarez-Esteban et al., 2008).

Recall that in the case of probabilities in $\mathcal{W}_2(\mathbb{R})$, a simplest expression for the Wasserstein distance is available. As shown in (Whitt, 1976), it can be written as

$$W_2^2(P,Q) = \int_0^1 \left(F^{-1}(t) - G^{-1}(t)\right)^2 dt,$$

where F (resp. G) is the distribution function associated with P (resp. Q).

In this work, we aim at aligning the law of the observations X_j. We propose the following procedure

- For a parameter θ, we compute the image of the observation by the inverse operator to *un-warp* the observations. More precisely for all candidate $\theta = (\theta_1, \dots, \theta_J)$, and to each observation $X_{ij} = \varphi_{\theta_j^\star}(\varepsilon_{ij})$, we apply the inverse deformation with parameter θ_j, which amounts to computing the following random variables

  $$Z_{ij}(\theta) = \varphi_{\theta_j}^{-1}(X_{ij}),$$

  for all $i = 1, \dots, n$ and $j = 1, \dots, J$.

- Then, we aim at picking choose the parameter which minimizes the energy needed to align the distribution of the un-warped variables with the distribution of their mean. The Wasserstein distance is chosen to measure this energy.

Actually, denote by $\mu_j(\theta)$ the common law of the elements of the i.i.d. sample $(Z_{1j}(\theta), \dots, Z_{nj}(\theta))$. We have $\mu_j(\theta) = \mu_j \circ \varphi_{\theta_j}$. Next, if $\mu_j^{(n)}(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Z_{ij}(\theta)}$ is the empirical law of the sample $(Z_{ij}(\theta))_{1 \leqslant i \leqslant n}$, then we also have $\mu_j^{(n)}(\theta) = \mu_j^{(n)} \circ \varphi_{\theta_j}$.

Now, set $\mu(\theta)$ the mean distribution of the $\mu_j(\theta)$'s, that is $\mu(\theta) = \frac{1}{J}\sum_{j=1}^J \mu_j(\theta)$.

Our procedure to recover the unknown quantity $\theta^\star$ consists in aligning using the 2-Wasserstein distance the distributions $\mu_j(\theta)$'s on $\mu(\theta)$ by varying the parameter θ.

First consider $\mu^{(n)}(\theta) := \frac{1}{J}\sum_{j=1}^J \mu_j^{(n)}(\theta)$ the mean distribution of the $\mu_j^{(n)}(\theta)$'s. Note that it corresponds to the empirical law of the sample $(Z_{ij}(\theta))$ for $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant J$ which is made of independent but not identically distributed random variables. However, it does not correspond to the empirical law associated with $\mu(\theta)$.

We quantify the alignment of the $\mu_j(\theta)$'s onto the mean distribution through the following quantity

$$M(\theta) = \sum_{j=1}^J W_2^2(\mu_j(\theta), \mu(\theta)), \text{ where } \theta \in \Pi_{j=1}^J \Lambda.$$

Remark that for all j, $\mu_j(\theta^\star) = \mu$, so $M(\theta^\star) = 0$. This function provides a characterization of our parameter of interest, $\theta^\star$, if we choose the distribution of the first sample as the reference. This is equivalent to consider that $\theta_1^\star$ is known, or to identify $\varepsilon_{i1} = X_{i1}$ for every i, and then, it happens that $\mu(\theta^*) = \mu_1$. Therefore, from now on, we will only handle the set $\Theta = \Pi_{j=2}^J \Lambda$.

However, we do not choose to align the empirical laws on the reference sample. Indeed, considering the mean law may attenuate the errors done in practice if the reference sample is not well chosen. A criterion based on the reference sample was investigated in (Lescornel and Loubes, 2012) in a slightly different setting. The same kind of problem has also been studied in (Castillo and Loubes, 2009) in the case of the estimation of shift parameters between curves. In this previous work, the matching criterion between the curves is the maximum profile likelihood maximization, which is replaced, in the case of deformations between distributions, by the Wasserstein's distance.

A natural idea is to study the empirical version of this criterion, simply obtained by considering the empirical laws instead of the real ones

$$M_n\left(\theta\right) = \sum_{j=1}^{J} W_2^2\left(\mu_j^{(n)}\left(\theta\right), \mu^{(n)}\left(\theta\right)\right). \tag{3}$$

Hence, this leads to consider as an estimator of the deformation parameters the estimate defined as any element in the set

$$\arg\min_{\theta\in\Theta} M_n\left(\theta\right). \tag{4}$$

In the following section, we investigate the asymptotic behaviour of the estimator defined in (4).

3. Estimation of the warping parameters

3.1. Assumptions

In the following we restrict the class of deformations used hereafter to the ones that satisfy the following assumptions.

$$\text{For all } \lambda \in \Lambda, x \mapsto \varphi_\lambda(x) \text{ is invertible from } I_a \text{ to } I_b. \tag{\textbf{A1}}$$

The following assumption is required to ensure that the Wasserstein distances between the samples is defined.

$$\forall\lambda \in \Lambda, \forall 1 \leqslant j \leqslant J, \int_E d(\varphi_\lambda^{-1} \circ \varphi_{\theta_j^\star}(x), x_0)^2 d\mu < \infty \text{ for one } x_0 \in E \tag{\textbf{A2}}$$

$$\text{that is } \int_E d(x, x_0)^2 d\mu_j\left(\theta\right) < \infty \quad \forall\theta \in \Theta \text{ and } x_0 \in E.$$

A regularity assumption is added on the deformation functions.

$$\varphi^{-1} \text{ is continuous on } \Lambda \times I_b. \tag{\textbf{A3}}$$

We define a ball $B = B(x_0, R)$ of the space (E, d) as $B(x_0, R) = \{x \in E, d(x, x_0) < R\}$.

The following tightness assumption ensures that the mass charged by the law $\mu_j\left(\theta\right)$ goes to sets with quite small μ_j probability if $\|\theta_j\|$ is large.

$$\begin{gathered}\text{For any ball } B \text{ and any numbers } \nu > 0, \text{ there are a closed set } S \text{ and a constant } H > 0 \\ \text{such that } \|\lambda\| > H \text{ implies } \varphi_\lambda\left(B\right) \subset S \text{ with } \mu_j\left[S\right] < \nu \quad \forall j = 1, \ldots, J.\end{gathered} \tag{\textbf{A4}}$$

We point out that the set S is not necessarily bounded, as shown in the examples considered later.

Finally, the following assumption provides the identifiability of the model.

$$M \text{ has an unique minimizer.} \tag{\textbf{A5}}$$

Let Id denotes the identity function. Remark that **A5** is verified as soon as there exists j with $2 \leqslant j \leqslant J$ such that $\varphi_{\theta_j}^{-1} \circ \varphi_{\theta_j^\star} \neq Id$ for $\theta \neq \theta^\star$ on a set of positive μ-measure. We provide some examples of deformations that undergo these assumptions in Section 5.

3.2. Main Result

The assumptions stated in previous subsection enable to obtain a result of almost sure convergence for the estimator $\widehat{\theta}^{(n)}$ defined in (4).

Theorem 3.1. *Under Assumptions* **A1** *to* **A5**, $\widehat{\theta}^{(n)} \to \theta^\star$ *almost surely (a.s.) when* $n \to \infty$.

The proof of this result follows is inspired by (Cuesta and Matran, 1988). It is carried out in the Appendix and it is divided in two steps:

- We first establish that a sequence defined by (4) is a.s. bounded through the two following lemmata.

  Lemma 3.2. *For all* $\theta \in \Theta$, $M_n(\theta)$ *converges a.s. to* $M(\theta)$.

  Lemma 3.3. *We have* $\mathbf{P}\left[\left\{\widehat{\theta}^{(n)}\right\}_{n\in\mathbb{N}} \text{ is bounded }\right] = 1$.

  This part mainly relies on the assumption **A4** and some classical tools of Probability theory.

- Next, we prove that on a set of probability one, the sequence $\left\{\widehat{\theta}^{(n)}\right\}_{n\in\mathbb{N}}$ has an unique accumulation point equal to $\theta^\star$.

  This part requires Assumption **A5**.

Another argument to choose to align the distributions on the mean law instead of the reference sample is the following. If θ is well chosen, then $\mu^{(n)}(\theta)$ can be viewed as an approximation of the empirical law associated with an i.i.d. sample of law μ of size nJ. On the other hand, the reference sample $\mu_1^{(n)}$ is the empirical law associated with a sample which is really of i.i.d. elements of law μ, but only of size n.

Then, the aligned empirical mean distribution $\mu^{(n)}\left(\widehat{\theta}^{(n)}\right) = \frac{1}{J}\sum_{j=1}^{J} \mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right)$ will be a proper estimate of the true distribution μ. The following result justifies this intuition.

Consider the case where E is a vectorial space endowed with a norm $\|\cdot\|$. Assume in addition that

$$\text{For all } x \in I_b, \varphi_\lambda^{-1} : \begin{array}{ccc} \Lambda & \to & I_a \\ \lambda & \mapsto & \varphi_\lambda^{-1}(x) \end{array} \text{ is continuously differentiable,} \tag{\textbf{A6}}$$

and

$$\forall 1 \leqslant j \leqslant J, \text{ the family } \left(\partial\varphi_\lambda^{-1}(\cdot)\right)_{\lambda\in\Lambda} \text{ has an envelope in } L^2(\mu_j), \qquad \textbf{(A7)}$$
$$\text{that is } \sup_{\lambda\in\Lambda} \left\|\partial\varphi_\lambda^{-1}(x)\right\| \leqslant H(x), H \in L^2(\mu_j).$$

Then, we have the following proposition

Proposition 3.4. *Under Assumptions* **A1** *to* **A7***,*

$$W_2\left(\mu^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu\right) \xrightarrow{n\to+\infty} 0 \ a.s.$$

This result is inspired by the paper (Lescornel and Loubes, 2012). The proof is postponed to the appendix.

4. Computation of the criterion

When the template measure is defined on some subset of $\mathbb{R}$, the value of M_n is easily available using the expression of the Wasserstein distance giving by (2) and the order statistics of the deformed observations. Indeed, recall that if F_n is the empirical distribution associated to a sample $(Y_1, \ldots, Y_n)$, then we have

$$F_n^{-1}(t) = Y_{(i)}, \text{ for } \frac{i-1}{n} < t \leqslant \frac{i}{n}.$$

Hence, setting $Z_{(k)}(\theta)$ the k-th order statistic of the sample $(Z_{ij}(\theta))_{\substack{1\leqslant i\leqslant n \\ 1\leqslant j\leqslant J}}$, and $Z_{(k)j}$ to the corresponding one in the sample $(Z_{ij}(\theta))_{1\leqslant i\leqslant n}$, standard computations lead to

$$M_n(\theta) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{Jn}\sum_{i=1}^{n}\sum_{k=1}^{J}\left[Z_{(i)j}(\theta) - Z_{(J(i-1)+k)}(\theta)\right]^2.$$

If the measure is defined on a more general space, we can use the ideas in (Alvarez-Esteban, 2009) to give a procedure to compute the Wasserstein distance between these laws denoted by $\mu^{(n)}(\theta)$ and $\mu_j^{(n)}(\theta)$. First note that those laws are uniform on $A = \{x_1, \ldots, x_{nJ}\}$ and $B = \{y_1, \ldots, y_n\}$. We denote these laws by $\mathcal{U}_A$ and $\mathcal{U}_B$, however note that A (resp. B) is not necessarily a set of nJ (resp. n) different elements.

Set $X \sim \mathcal{U}_A$ and $Y \sim \mathcal{U}_B$. Then, we have that

$$\mathbf{E}\left[d(X,Y)^2\right] = \sum_{(i,k)=(1,1)}^{(nJ,n)} d(x_i, y_k)^2 c_{ik} \qquad (5)$$

where $c = \{c_{ik}\}_{\substack{1\leqslant i\leqslant nJ \\ 1\leqslant k\leqslant n}} \in \mathcal{C}$, is such that

$$c_{ik} \geqslant 0 \quad \forall i, k,$$

$$\sum_{i=1}^{nJ} c_{ik} = \frac{1}{n} \quad \forall k,$$

and

$$\sum_{k=1}^{n} c_{ik} = \frac{1}{nJ} \quad \forall i.$$

Indeed, setting

$$c_{ik} = \mathbf{P}\left[(X,Y) = (x_i, y_k)\right], \tag{6}$$

we obtain equality (5) for some $X \sim \mathcal{U}_A$ and $Y \sim \mathcal{U}_B$ with joint law defined by (6).

Hence, computing the Wasserstein distance between $\mathcal{U}_A$ and $\mathcal{U}_B$ turns to solve a linear optimization problem

$$W_2^2\left(X,Y\right) = \inf_{c \in \mathcal{C}} \sum_{(i,k)=(1,1)}^{(nJ,n)} d\left(x_i, y_k\right)^2 c_{ik} := \inf_{c \in \mathcal{C}} L(A,B,c).$$

So for $\theta \in \Theta$ we follow this procedure to compute for all j a sequence $c^j\left(\theta\right) \in \mathcal{C}$ such that

$$W_2^2\left(\mu_j^{(n)}\left(\theta\right), \mu^{(n)}\left(\theta\right)\right) = L\left(\left(Z_{ij}\left(\theta\right)\right)_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant J}}, \left(Z_{ij}\left(\theta\right)\right)_{1 \leqslant i \leqslant n}, c^j\left(\theta\right)\right),$$

and we obtain $M_n\left(\theta\right) = \frac{1}{J}\sum_{j=1}^{J} L\left(\left(Z_{ij}\left(\theta\right)\right)_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant J}}, \left(Z_{ij}\left(\theta\right)\right)_{1 \leqslant i \leqslant n}, c^j\left(\theta\right)\right)$.

5. Applications

Now we provide some examples of admissible deformations, which undergo previous set of assumptions.

5.1. Example 1 : Location/scale model

$$\varphi_\lambda\left(x\right) = \frac{x}{\lambda_2} + \lambda_1.$$

Here $E = \mathbb{R}^p$, and φ_λ is invertible on $\mathbb{R}^p$ if $\lambda_2 \neq 0$ so $\Lambda \subset \mathbb{R}^p \times \mathbb{R} - \{0\}$.

We have $\varphi_\lambda^{-1}\left(x\right) = \lambda_2 x - \lambda_1\lambda_2 = \varphi_{(-\lambda_1\lambda_2, \frac{1}{\lambda_2})}(x)$, and $\varphi_\lambda^{-1}\left(\varphi_\beta\left(x\right)\right) = \frac{\lambda_2}{\beta_2}x + \beta_1 - \lambda_1\lambda_2$ which is in $L^2\left(\mu\right)$ if $\mu \in \mathcal{W}_2\left(\mathbb{R}^p\right)$. Hence Assumptions **A1** to **A3** and **A5** take place if μ is in $\mathcal{W}_2\left(\mathbb{R}^p\right)$ and atom-less.

Now consider the assumption **A4**. Assume that μ has a bounded density with respect to the Lebesgue measure. Then for all $\nu > 0$, we can find η such that for all x, $\mu_j\left(\bar{B}(x,\eta)\right) < \nu$ and M such that $\mu_j\left(B(0,M)^c\right) < \nu$.

We have $\varphi_\lambda\left(B(y_0,r)\right) = B(\frac{1}{\lambda_2}y_0 + \lambda_1, \frac{1}{|\lambda_2|}r)$. Hence if only $|\lambda_2| \to \infty$, then $\varphi_\lambda\left(B(y_0,r)\right) \subset \bar{B}(x_\lambda,\eta)$ if $\|\lambda\|$ is sufficiently large. Now, if $\|\lambda_1\| \to \infty$, $\varphi_\lambda\left(B(y_0,r)\right) \subset B(0,M)^c$ if $|\lambda_2| \geqslant \alpha > 0$. Then, if in addition $\Lambda \subset \mathbb{R}^p \times]-\infty; -\alpha] \cup [\alpha; +\infty[$ with $\alpha > 0$, **A4** is valid.

This example can also be connected to a more general model of ANOVA, where the variables ε_{ij} have different variances. More precisely it corresponds to the case where we observe

$$X_{ij} = \mu_j^\star + \sigma_j^\star \tilde{\varepsilon}_{ij} \quad 1 \leqslant i \leqslant n \quad 1 \leqslant j \leqslant J$$

where $\tilde{\varepsilon}_{ij}$ are random independent variables of same law with mean 0.

One particular case is the translation model

$$\varphi_\lambda(x) = x + \lambda.$$

With the same arguments, it can be shown that Assumptions **A1** to **A4** are satisfied if μ is in $\mathcal{W}_2\left(\mathbb{R}^d\right)$.

We can also consider the scale model

$$\varphi_\lambda(x) = \frac{1}{\lambda}x.$$

In this case, all the assumptions are valid if $\Lambda = \mathbb{R} - \{0\}$ and if μ is in $\mathcal{W}_2(\mathbb{R}^p)$ with $\mu(0) = 0$. Note that such deformations constitute the extension to the density warping case of the deformations studied in Vimond (2010a) in the case of functional warping.

5.2. *Example 2 : Logarithmic transformation*

$$\varphi_\lambda(x) = \frac{1}{\lambda}\log(x).$$

φ_λ is invertible from $(0, +\infty)$ to $\mathbb{R}$ for all $\lambda \neq 0$, so here Λ must be contained in $(0, +\infty)$ and the support of μ in $(0, +\infty)$. We have $\varphi_\lambda^{-1}(x) = \exp(x\lambda)$, and $\varphi_\lambda^{-1}(\varphi_\beta(x)) = \exp\left(\frac{\lambda\log(x)}{\beta}\right) = x^{\frac{\lambda}{\beta}}$. Hence $\varphi_\lambda^{-1} \in L^2(X_{1j})$ if $\mathbf{E}\left[\varepsilon^{\frac{2\lambda}{\theta^\star_j}}\right] < \infty$ for all $\lambda \in \Lambda$. Assumptions **A1** to **A3** and **A5** take place if the support of μ is contained in $(0, +\infty)$ and if $\mathbf{E}\left[\varepsilon^{\frac{2\lambda}{\theta^\star_j}}\right] < \infty$ for all $\lambda \in \Lambda$.

In this case the conditions are more restrictive on the law μ, but notice that, for instance, the exponential distribution satisfies them.

Moreover, for y_0 and r such that $B(y_0, r) \subset (0, +\infty)$, we have $\log(B(y_0, r)) \subset \log\left(\bar{B}(y_0, r)\right) \subset \bar{B}(z_0, R)$ because the image of a compact set through a continuous function still is a compact set. Hence $\varphi_\lambda(B(y_0, r)) \subset \bar{B}\left(\frac{z_0}{\lambda}, \frac{R}{|\lambda|}\right)$ and **A4** is verified if $\mu(0) = 0$.

5.3. *Example 3 : Affine transformation*

$$\varphi_\lambda(x) = A^{-1}x + b.$$

Here $x \in \mathbb{R}^p$, and $\lambda = (A, b) \in GL(\mathbb{R}^p) \times \mathbb{R}^p$. We have $\varphi_\lambda^{-1}(x) = A(x - b)$. Hence $\varphi_{\lambda_1}^{-1} \circ \varphi_{\lambda_2}(x) = A_1\left(A_2^{-1}x + b_2 - b_1\right)$, so Assumptions **A1** to **A3** and **A5** are valid if μ is in $\mathcal{W}_2(\mathbb{R}^p)$ and atom-less.

Now, for $y_0 \in \mathbb{R}^p$ and $r > 0$ we have $\varphi_\lambda(B(y_0, r)) \subset B\left(A^{-1}y_0 + b, r\left\|A^{-1}\right\|\right)$. Hence, as in Example 1, Assumption **A4** is verified if we assume that μ has a bounded density with respect to the Lebesgue measure and if one chooses A in a subset of $GL(\mathbb{R}^p)$ with $\|A\| \geqslant \alpha > 0$.

5.4. Example 4 : Composition

$$\varphi_\lambda(x) = f \circ \tilde{\varphi}_\lambda(x)$$

Consider a function $\tilde{\varphi}_\lambda(x)$, a law μ and a parameter $\theta^\star$ which satisfies all the assumptions **A1** to **A5**. Then, if f is an homeomorphism from I_b to I_c the deformation function $\varphi_\lambda(x) = f \circ \tilde{\varphi}_\lambda(x)$ with the same law μ and parameter $\theta^\star$ verifies also these assumptions replacing I_b by I_c. Indeed, Assumptions **A1** and **A3** are easily checked, and we have

$$\varphi_\lambda^{-1} \circ \varphi_\beta = \tilde{\varphi}_\lambda^{-1} \circ f^{-1} \circ f \circ \tilde{\varphi}_\beta = \tilde{\varphi}_\lambda^{-1} \circ \tilde{\varphi}_\beta .$$

Hence $\tilde{\mu}_j(\lambda) = \mu \circ \tilde{\varphi}^{-1}_{\theta^\star_j} \circ \tilde{\varphi}_\lambda = \mu \circ \varphi^{-1}_{\theta^\star_j} \circ \varphi_\lambda = \mu_j(\lambda)$ so Assumptions **A2**, **A5** are also satisfied. In particular the criterion $\tilde{M}(\theta)$ corresponding to $\tilde{\varphi}$ is exactly the same than the criterion $M(\theta)$ corresponding to φ.

Choose $\nu > 0$, a ball B and S a closed set such that $\tilde{\varphi}_\lambda(B) \subset S$ with $\tilde{\mu}_j[S] < \nu$ for all j if λ is sufficiently large. We have $\varphi_\lambda(B) \subset f(S)$ which is closed because f is an homeomorphism. Moreover $\mu_j[f(S)] = \tilde{\mu}_j[S] < \nu$ so **A4** is also verified.

This allows us to consider a lot of new deformations. For instance, in the same setting of the scale model we can consider the logit model, with the deformation function $\varphi_\lambda(x) = (1 + \exp(x/\lambda))^{-1}$.

This example is also related to the usual logistic regression. The aim is to explain Bernoulli variables $(X_j)_{j=1}^J$ as functions of a random variable ε, and the model can be understood as follows. For all j there exists $\theta^\star_j$ such that $\mathbf{P}(X_j = 1/\varepsilon) = \left(1 + \exp\left(\varepsilon/\theta^\star_j\right)\right)^{-1}$. Hence if the data collected permits to approach the quantity $\mathbf{P}(X_j = 1/\varepsilon)$, our method gives also an alternative to estimate the parameters $\theta^\star_j$.

The study of Example 2 also gives the conditions under which the deformation $\varphi_\lambda(x) = x^{\frac{1}{\lambda}}$ can be studied by our method.

5.5. Deformations coming from differential equations

In Younes (2004) is given a general class of deformations used in image registration as follows. Set $v : \Gamma \times \mathbb{R}^p \times [0;T] \to \mathbb{R}$ a function C^2 on $\Gamma \times \mathbb{R}^p \times [0;T]$, depending on parameters in Γ is an open subset of $\mathbb{R}^d$. Assume that for all $(\lambda, t) \in \Gamma \times [0;T]$, the support of $v(\lambda, \cdot, t)$ is contained in Ω, an open bounded subset of $\mathbb{R}^p$. For $t \in [0;T]$ and $x \in \Omega$, consider the equations

$$(E) \left\{ \begin{array}{l} \partial_4 \Phi(\lambda, x; t, s) = v(\lambda, \Phi(\lambda, x; t, s), s) \\ \Phi(\lambda, x; t, t) = x \end{array} \right.$$

where $\partial_4 \Phi(\lambda, x; t, s)$ denotes the partial derivative of Φ with respect to the variable s. Then let $\Phi(\lambda, x; t, \cdot)$ be the unique solution to the system of equations (E) defined on $[0;T]$. Hence, if (s,t) is fixed, the solution to the system (E) at point s, $\Phi(\lambda, x; t, s) = \varphi_\lambda(x)$ is a deformation function based on a vector field indexed by the parameters λ. Such diffeomorphisms used in image deformations (see for instance (Bigot et al., 2009) and references therein) fall into the scope of this paper and the parameters λ can be estimated with our methodology. Yet the computation costs are heavy since the dimension of the set of parameters play an important role when finding the minimum of the criterion (3) and for such examples d is large. Hence the application of our methodology to such class of deformations deserve a specific treatment that falls out of the scope of this paper.

6. Simulations

In this section, we challenge our method with different simulations. For this we have generated n-sized samples of independent random variables ε_{ij} with a Gaussian standard distribution where $j = 1, .., J$ and $i = 1, \ldots, n$. We can simulate the following observations $X_{ij} = \varphi_{\theta_j^*}(\varepsilon_{ij})$ for given parameters θ_j^*, which will be estimated. To obtain an estimation of the deformation parameters, we have to minimize the criterion defined in (4). The main difficulty comes from the minimization of the Wasserstein distance which requires a special treatment. For this, using the software R, we have used the function *optim* with the option L-BFGS-B which minimizes a given function with a quasi-Newton method and the function Partial.Transport (a new package for transport cost optimization that can be obtained upon request) which computes the Wasserstein distance between 2 probability distributions. The simulations are fast but the complexity grows with the dimension of the parameters θ_j^*. We have considered two different deformation functions

- A scale location-model
$$\varphi_{\theta_j^*}(x) = \frac{x}{\theta_1^*} + \theta_2^*$$
We have made simulations for 5 pairs of the parameters $(\theta_1^\star, \theta_2^\star)$: (2,1), (7,2), (2.5,2), (5,5) and (10,8). Recall that $\theta_1^\star$ is the scale parameter and $\theta_2^\star$ is the location parameter.

- A matrix as the scale parameter
$$\varphi_{\theta_j^*}(x) = x * (\theta_1^*)^{-1} + \theta_2^*$$
where x is a 1×2 vector (an observation from the bivariate normal distribution) and θ_1^* is a 2×2 matrix of the form $\begin{pmatrix} \theta_{11}^* & 0 \\ 0 & \theta_{12}^* \end{pmatrix}$. The simulations have been made for 4 sets of parameters which are (2,2,1), (4,3,2), (2.5,3,1.5) and (5,2,2). The first two parameters correspond to θ_{11}^* and θ_{12}^* respectively and the last one to the location parameter.

All simulations have been done for n = 20, 50, 100, 200, 300 and $J = 2$. To analyze the estimation error, we provide in the figure 6 the boxplots of the errors in the estimations for each pair of parameters. The results of the other simulations are available upon request.

As expected, the error decreases as the sample size increases leading to good estimators for large sample sizes. We point out that the location parameter estimator behaves better than the scale parameter estimator, as shown in the tables 1 and 2, presenting the mean and the standard deviation for these parameters and for different sample sizes.

Note that in the case of the location parameter, the mean of the estimator in the simulations is very close to the real value of the deformation parameter and the standard deviation is small, even for small sample sizes. For the scale parameter the results are not as good for small sample size but they significantly improve as the sample size grows.

Last we made a simulation for $J = 3$, with the scale location deformation function. In the boxplot 6 and the table 3 with the mean and standard deviation we can see that although the results for small sample sizes are worse than the ones for $J = 2$, in the biggest sample sizes the approximation is becoming accurate.

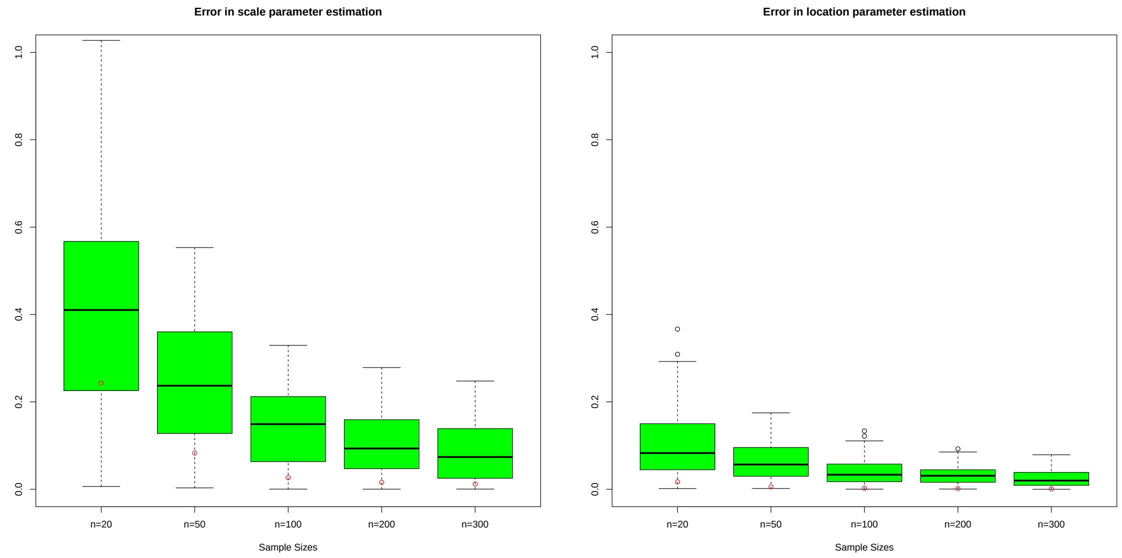


Figure 1: Boxplot of the errors for parameters 2,1

Table 1: Mean and Standard Deviation for Scale Parameter estimation

Mean

Sample Size\Parameters	**2**,1	**2.5**,7	**5**,5	**7**,2	**10**,8
20	1,627001701	2,0839746	4,167520282	5,9679944	8,389333779
50	1,778303097	2,2620969	4,470069471	6,4297052	9,211445528
100	1,891638304	2,3781902	4,70858878	6,5357411	9,379041831
200	1,922227623	2,3981952	4,791570246	6,7249666	9,668745199
300	1,950083809	2,4351132	4,875349531	6,8101462	9,71924409

Standard Deviation

Sample Size\Parameters	**2**,1	**2.5**,7	**5**,5	**7**,2	**10**,8
20	0,323957502	0,3636535	0,771711943	1,181713	1,549372177
50	0,184464246	0,2189685	0,4213675	0,7855247	1,007653165
100	0,122020757	0,1480172	0,315504577	0,4910598	0,69810673
200	0,098394056	0,1054195	0,25755501	0,3594169	0,525405374
300	0,098315417	0,0958916	0,215009943	0,2329655	0,425525537

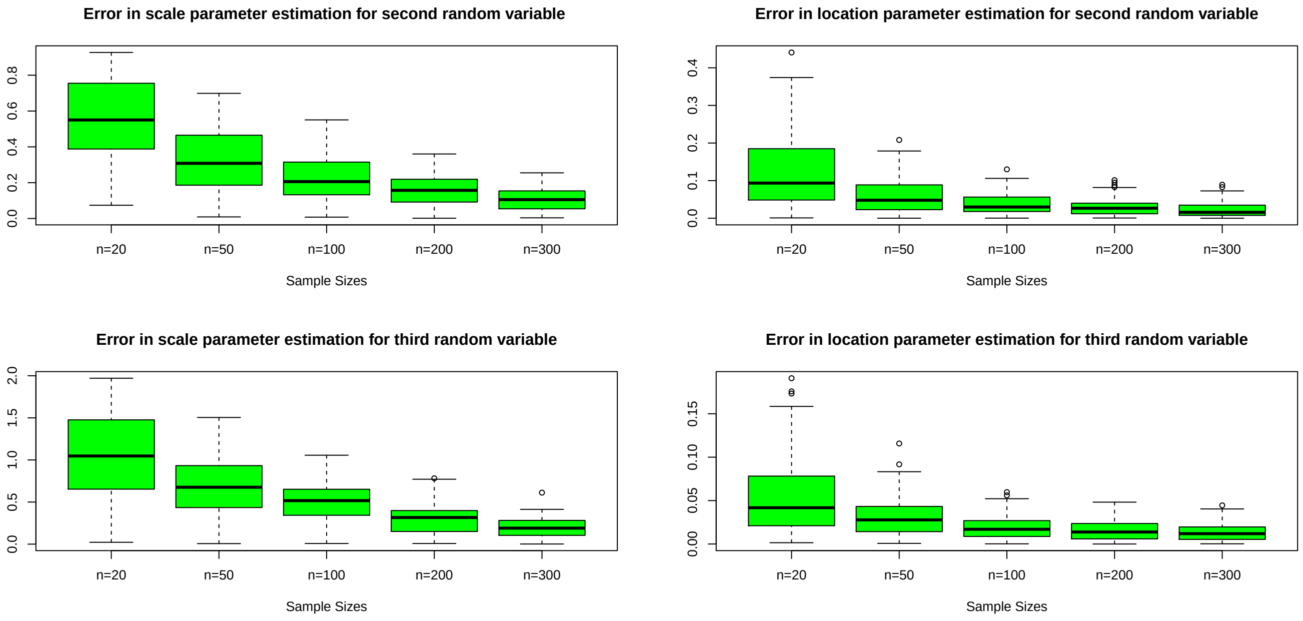
Table 2: Mean and Standard Deviation for Location Parameter estimation

Mean

Sample Size\Parameters	2,**1**	2.5,**7**	5,**5**	7,**2**	10,**8**
20	1,0023873	6,9945413	5,0027399	1,9982004	7,9984572
50	0,9936314	7,0010124	5,0039185	2,0037834	7,9990901
100	0,9963651	7,0030782	4,9968519	2,0000012	7,9993247
200	1,0008455	6,9993688	4,9993167	2,0007894	8,0000355
300	1,0015348	6,9970673	5,0006839	1,9993071	8,0007074

Standard Deviation

Sample Size\Parameters	2,**1**	2.5,**7**	5,**5**	7,**2**	10,**8**
20	0,1304445	0,1007973	0,0491171	0,040827	0,0265132
50	0,074524	0,0574477	0,032775	0,0228102	0,0144892
100	0,0483674	0,040702	0,0200511	0,0144279	0,0093959
200	0,0386074	0,0283701	0,0154285	0,0099725	0,0079499
300	0,0306963	0,0220878	0,0124193	0,008336	0,0061003



Error in scale parameter estimation for second random variable

Error in location parameter estimation for second random variable

Error in scale parameter estimation for third random variable

Error in location parameter estimation for third random variable

Figure 2: Boxplot of the errors for parameters when $J = 3$

Table 3: Mean and Standard Deviation for the 4 parameters when $J = 3$

Mean

Sample Sizes\Parameter	**Scale j=2**	**Location j=2**	**Scale j=3**	**Location j=3**
20	1,464537092	1,008042501	2,967974141	2,997307357
50	1,683731845	0,98915318	3,347790181	2,997213004
100	1,784915087	1,000058216	3,51445098	3,00086421
200	1,849162384	0,991362845	3,712380117	2,997154113
300	1.903418458	0.999896672	3.813571177	2.998477961

Standard Deviation

Sample Sizes\Parameter	**Scale j=2**	**Location j=2**	**Scale j=3**	**Location j=3**
20	0,257727866	0,152380211	0,518586349	0,070961424
50	0,180780109	0,077025035	0,375416689	0,038774036
100	0,132750853	0,047759921	0,24000684	0,024065151
200	0,095342177	0,036894564	0,192043519	0,019273895
300	0.075536603	0.028990515	0.13926473	0.016831891

7. Appendix.

7.1. Proof of Theorem 3.1

The proof of Theorem 3.1 is split in two steps. We begin with the proof of Lemmas 3.2 and 3.3.

Proof of Lemma 3.2

Proof. We will use the following equivalence:

If $\{P_n\}_{n\in\mathbb{N}}$ is a sequence in $\mathcal{W}_2(E)$ and $P \in \mathcal{W}_2(E)$, then

$$W_2(P_n, P) \xrightarrow{n\to\infty} 0 \text{ if and only if } \left\{ \begin{array}{l} P_n \rightharpoonup P \\ \int_E d(x,x_0)^2 dP_n(x) \xrightarrow{n\to\infty} \int_E d(x,x_0)^2 dP(x), \end{array} \right. \tag{7}$$

where x_0 is any point of E.

This characterization of the convergence in the Wasserstein's sense is proved for instance in (Shorack and Wellner, 2009) p.63 in the particular case of probabilities on $\mathbb{R}$ and in (Bickel and Freedman, 1981) or (Rachev, 1982) in abstract spaces.

From the properties of the empirical distribution we have that for all j and θ, a.s.

$$\mu_j^{(n)}(\theta) \quad \rightharpoonup \quad \mu_j(\theta) \tag{8}$$

$$\int_E d(x,x_0)^2 d\mu_j^{(n)}(\theta) \quad \xrightarrow{n\to\infty} \quad \int_E d(x,x_0)^2 d\mu_j(\theta). \tag{9}$$

So -using for instance the characterization of convergence in distribution with the bounded and continuous functions- from (8), we have that a.s.

$$\mu^{(n)}(\theta) = \frac{1}{J}\sum_{j=1}^{J} \mu_j^{(n)}(\theta) \rightharpoonup \frac{1}{J}\sum_{j=1}^{J} \mu_j(\theta) = \mu(\theta), \tag{10}$$

and from (9), a.s.

$$\int_E d(x,x_0)^2 d\mu^{(n)}(\theta) \xrightarrow{n\to\infty} \int_E d(x,x_0)^2 d\mu(\theta)\,. \tag{11}$$

Using equivalence (7) we get that (10) and (11) imply

$$W_2\left(\mu^{(n)}(\theta), \mu(\theta)\right) \xrightarrow{n\to\infty} 0 \text{ a.s.}$$

Equivalence (7), now joined to (8) and (9), also gives that for all j and θ fixed

$$W_2^2\left(\mu_j^{(n)}(\theta), \mu_j(\theta)\right) \xrightarrow{n\to\infty} 0, \text{a.s.}$$

The last two relations finally give that

$$M_n(\theta) = \sum_{j=1}^{J} W_2^2\left(\mu_j^{(n)}(\theta), \mu^{(n)}(\theta)\right) \xrightarrow{n\to\infty} M(\theta) = \sum_{j=1}^{J} W_2^2\left(\mu_j(\theta), \mu(\theta)\right) \text{ a.s.}$$

□

Proof of Lemma 3.3

Proof. Remember that we are assuming that $\mu_1 = \mu$, and that $\mu_1^{(n)}\left(\widehat{\theta}^{(n)}\right) = \mu_1^{(n)}$ because the first sample remains unchanged. Thus, Lemma 3.2 and the properties of the empirical distributions imply that there exists a set Ω_0 of probability 1 on which:

$$\begin{aligned} W_2\left(\mu_1^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu\right) &\to 0, && (12)\\ M_n(\theta^\star) &\to M(\theta^\star)\,, && (13)\\ \mu_j^{(n)} &\rightharpoonup \mu_j, \quad \text{for every } j = 1,\dots J. && (14) \end{aligned}$$

From now on, we fix an element in the set Ω_0, and we will show that every sequence $\widehat{\theta}^{(n)} \in \arg\min_{\theta\in\Theta} M_n(\theta)$ is bounded. By contradiction, assume that $\limsup_{n\to\infty}\left\|\widehat{\theta}^{(n)}\right\|_{p(J-1)} = \infty$ and choose $j \in \{2\dots J\}$ such that $\limsup_{n\to\infty}\left\|\widehat{\theta}_j^{(n)}\right\|_p = \infty$.

Consider an extraction $\{n_h\}_{h\geqslant 1}$ such that $\lim_{h\to\infty}\left\|\widehat{\theta}_j^{(n_h)}\right\|_p = \infty$.

With Assumption **A4**, for all ball B and all $\nu > 0$ there exist a closed set S and an integer h_0 such as, for all $h \geqslant h_0$, $\varphi_{\widehat{\theta}_j^{(n_h)}}(B) \subset S$ and so $\mu_j\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \leqslant \mu_j[S] \leqslant \nu$. Hence

$$\limsup_{h\to\infty} \mu_j^{(n_h)}\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \leqslant \limsup_{h\to\infty} \mu_j^{(n_h)}[S]\,.$$

Now recall the assertion of Portmanteau's theorem (stated in (Dudley, 2002), Theorem 11.1.1 p.386 for instance), which claims that a sequence of probabilities P_n defined on a metric space converges weakly to a probability P if and only if for all closed sets F,

$\limsup_{n\to\infty} P_n(F) \leqslant P(F)$. Hence using the convergence in law of the measure $\mu_j^{(n_h)}$ stated in (14), we can write

$$\limsup_{h\to\infty} \mu_j^{(n_h)}[S] \leqslant \mu_j[S],$$

so

$$\limsup_{h\to\infty} \mu_j^{(n_h)}\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \leqslant \nu.$$

This inequality holds for all $\nu > 0$, so we can conclude that for every ball B

$$\mu_j^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)[B] = \mu_j^{(n_h)}\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \xrightarrow{h\to\infty} 0.$$

Moreover, by definition of $\widehat{\theta}^{(n)}$

$$M_{n_h}(\theta^\star) \geqslant M_{n_h}\left(\widehat{\theta}^{(n_h)}\right)$$

so, from (13),

$$0 = M(\theta^\star) \geqslant \liminf_{h\to\infty} M_{n_h}\left(\widehat{\theta}^{(n_h)}\right). \tag{15}$$

From here, we have that

$$0 = \liminf_{h\to\infty} W_2^2\left(\mu_1^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right), \mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)\right).$$

First, recall that on Ω_0, (12) holds. Hence, considering the subsequence such that

$$0 = \lim_{q\to\infty} W_2^2\left(\mu_1^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right), \mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)\right)$$

we necessarily have $\mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right) \xrightarrow{q\to\infty} \mu$ for the Wasserstein distance. Hence we have obtained a sub-sequence such that for one j and all ball B

$$\mu_j^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)[B] \xrightarrow{q\to\infty} 0 \tag{16}$$

and the "mean" law converges in distribution

$$\mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right) \rightharpoonup \mu.$$

Set $\delta > 0$ and $x_0 \in E$. Using (16), we know that for all $H \in \mathbb{N}$, one can find q_H such that for all $q \geqslant q_H$

$$\mu_j^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)[B(x_0; H)] \leqslant \delta. \tag{17}$$

Hence, we can construct a new subsequence (which we still denote by $(n_h)_q$ for sake of simplicity) such that for all $H \geqslant 0$ (17) holds and

$$\mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right) \rightharpoonup \mu.$$

Since $\mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)$ is a sequence of measures in a complete separable metric space which converges in distribution, it is a tight sequence. In consequence, for all $\nu > 0$, there exists a compact K such that for all $q \in \mathbb{N}$, $\mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)[K] \geqslant 1-\nu$. However K is bounded, then there exists M such that $K \subset B\left(x_0; M\right)$. Then

$$\begin{aligned}\nu \geqslant \mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)[K^c] \geqslant \mu^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)&[B\left(x_0; M\right)^c] \\ \geqslant &\mu_j^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)[B\left(x_0; M\right)^c] \geqslant 1-\delta.\end{aligned}$$

If we take $\nu = \frac{1-\delta}{2}$, we have a contradiction with the tightness of the sequence $\mu_j^{\left((n_h)_q\right)}\left(\widehat{\theta}^{\left((n_h)_q\right)}\right)$. Hence $\limsup_{n\to\infty}\left\|\widehat{\theta}^{(n)}\right\|_{p(J-1)} < \infty$ on Ω_0.

□

After Lemma 3.3, Theorem 3.1 will be proved if we show the following lemma:

Lemma 7.1. *The sequence* $\left(\widehat{\theta}^{(n)}\right)$, $n \in \mathbb{N}$ *satisfies that*

$$\mathbf{P}\left[\left(\widehat{\theta}^{(n)}\right) \text{ has an unique accumulation point equal to } \theta^\star\right] = 1.$$

In the proof of this lemma, we will employ Skohorod's Theorem, stated in (Dudley, 2002) (Theorem 11.7.2 p.415), which gives an almost sure representation for the convergence in distribution.

Theorem 7.2. *Let S be a separable metric space, and $\{P_n\}_{n\in\mathbb{N}}$ be probability laws on S converging in distribution to a probability P. Then, on some probability space there exist random variables $\{X_n\}_{n\in\mathbb{N}}$ and X with values in S such that $X_n \sim P_n$ for all n and $X \sim P$, and $X_n \to X$ (a.s).*

Proof of Lemma 7.1.

Proof. To prove this lemma, let us consider the set Ω_0 of probability one, introduced in Lemma 3.3, whose points satisfy (12), (13) and (14).

From now on, we reason in a deterministic setting by fixing one element of this set. According to Lemma 3.3, the sequence $\left(\widehat{\theta}^{(n)}\right)_{n\geqslant 1}$ is bounded on Ω_0 and we only need to show that if θ^0 is the limit of any subsequence of $\left(\widehat{\theta}^{(n)}\right)_{n\geqslant 1}$, then, it happens that $\theta^\star = \theta^0$. For sake of simplicity we keep the same notation for the subsequence, so we assume that $\widehat{\theta}^{(n)} \to \theta^0$.

As shown in Lemma 3.3 inequality (15) holds. From this inequality, we have that, for every $j = 1, \ldots, J$,

$$\liminf_{n\to\infty} W_2\left(\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu^{(n)}\left(\widehat{\theta}^{(n)}\right)\right) = 0.$$

Thus, perhaps taking a new subsequence (which we still denote with the same notation), we have that

$$W_2\left(\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu^{(n)}\left(\widehat{\theta}^{(n)}\right)\right) \xrightarrow{n\to\infty} 0, \text{ for every } j = 1, \ldots, J. \tag{18}$$

This, when applied to $j = 1$, and (12), gives that

$$W_2\left(\mu^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu\right) \xrightarrow{n\to\infty} 0.$$

Now, if $j = 1, \ldots, J$, from here and (18), we have that

$$\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right) \rightharpoonup \mu. \tag{19}$$

On the other hand, (14) is satisfied. Thus, we can apply Theorem 7.2 to obtain random vectors $Z_n, n = 0, 1, \ldots,$ such that, the sequence $\{Z_n\}$ converges a.s. to Z_0 and the distribution of Z_0 is μ_j and that of Z_n is $\mu_j^{(n)}$. Since $\widehat{\theta}^{(n)} \to \theta^0$, Assumption **A3** gives that

$$\varphi^{-1}_{\hat{\theta}_j^{(n)}}(Z_n) \xrightarrow{n\to\infty} \varphi^{-1}_{\theta_j^0}(Z_0), \text{ a.s.}$$

Therefore, we have that

$$\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right) \rightharpoonup \mu_j\left(\theta^0\right). \tag{20}$$

This, joined to (19), gives that $\mu_j\left(\theta^0\right) = \mu$ for all j. Finally, it happens that $M\left(\theta^0\right) = 0 = M\left(\theta^\star\right)$. Thus, from Assumption **A5** it follows that $\theta_j^0 = \theta_j^*$ and the proof ends. □

7.2. Proof of Proposition 3.4

Proof. Proposition 3.4 is easily tractable using the caracterization of convergence in the Wasserstein sense given by (7).

Step 1 *It happens that* $\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right) \rightharpoonup \mu$ *a.s.*

All the arguments are contained in the proof of Lemma 7.1.
Indeed, we have already shown in (20) that if $\widehat{\theta}^{(n)} \to \theta^0$ a.s, then

$$\mu_j^{(n)}\left(\hat{\theta}^{(n)}\right) \rightharpoonup \mu_j\left(\theta^0\right).$$

Hence, using Theorem 3.1, we get $\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right) \rightharpoonup \mu$ a.s.

Step 2 *Let ε be a r.v. with the same distribution as the r.v.'s ε_{ij}'s introduced in (1). Then,* $\frac{1}{n}\sum_{i=1}^n \left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|^2 \xrightarrow{n\to\infty} \mathbf{E}\left[\|\varepsilon\|^2\right]$ *a.s.*

Using the Strong Law of Large Numbers, we know that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \left\|\varphi^{-1}_{\theta_j^\star}\left(X_{ij}\right)\right\|^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \|\varepsilon_{ij}\|^2 = \mathbf{E}\left[\|\varepsilon\|^2\right] \text{ a.s.}$$

Hence it is sufficient to show that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \left\|\varphi^{-1}_{\theta_j^\star}\left(X_{ij}\right)\right\|^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|^2$.

We have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|^2-\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|^2\right| \\
&\quad= \left|\frac{1}{n}\sum_{i=1}^{n}\left(\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|-\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|\right)\left(\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|+\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|\right)\right| \\
&\quad\leqslant \frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)-\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|\left(\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|+\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|\right) \\
&\quad\leqslant \frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)-\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|\left(\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)-\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|+2\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|\right) \\
&\quad\leqslant \frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)-\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|^2 \\
&\qquad+2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)-\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|^2}.
\end{aligned}$$

From here, Assumption **A6**, and the Mean Inequality with Assumption **A7** give

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|^2-\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|^2\right| \\
&\leqslant \left(\frac{1}{n}\sum_{i=1}^{n}H\left(X_{ij}\right)^2+2\sqrt{\frac{1}{n}\sum_{i=1}^{n}H\left(X_{ij}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\theta_j^{\star}}\left(X_{ij}\right)\right\|^2}\right)\left\|\widehat{\theta}_j^{(n)}-\theta_j^{\star}\right\|,
\end{aligned}$$

which converges (a.s) to 0 using the Strong Law of Large Numbers and Theorem 3.1. Hence, (a.s)

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}\left(X_{ij}\right)\right\|^2=\mathbf{E}\left[\|\varepsilon\|^2\right].$$

Finally, using the characterization of the convergence in the Wasserstein sense given in (7), Proposition 3.4 is proved. □

References

Alvarez-Esteban, P., 2009. Aplicaciones de los recortes imparciales en la comparación de distribuciones. PhD Thesis. Universidad de Valladolid.

Alvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A., Matran, C., 2008. Trimmed comparison of distributions. J. Amer. Statist. Assoc. 103 (482), 697–704.

Amit, Y., Grenander, U., Piccioni, M., 1991. Structural Image Restoration through deformable template. Journal of the American Statistical Association 86, 376–387.

Bercu, B., Fraysse, P., 2012. A Robbins-Monro procedure for estimation in semiparametric regression models. Ann. Statist. 40 (2), 666–693.

Bickel, P. J., Freedman, D. A., 1981. Some asymptotic theory for the bootstrap. Ann. Statist. 9 (6), 1196–1217.

Bigot, J., Gadat, S., Loubes, J.-M., 2009. Statistical M-estimation and consistency in large deformable models for image warping. J. Math. Imaging Vision 34 (3), 270–290.

Boissard, E., Gouic, T. L., Loubes, J.-M., 2014. Distribution's template estimate with Wasserstein metrics. Bernoulli.
URL http://arxiv.org/pdf/1111.5927.pdf

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

Bonneel, N., Rabin, J., Peyré, G., Pfister, H., 2014. Sliced and radon Wasserstein barycenters of measures. J. Math. Imaging Vis.
URL http://link.springer.com/article/10.1007%2Fs10851-014-0506-3

Caillerie, C., Chazal, F., Dedecker, J., Michel, B., 2011. Deconvolution for the Wasserstein metric and geometric inference. Electron. J. Stat. 5, 1394–1423.

Castillo, I., Loubes, J.-M., 2009. Estimation of the distribution of random shifts deformation. Math. Methods Statist. 18 (1), 21–42.

Cuesta, J., Matran, C., 1988. The strong law of large numbers for k-means and best possible nets of Banach valued random variables. Probability Theory and Related Fields 78 (4), 523–534.

Cuesta, J. A., Matran, C., 1989. Notes on the Wasserstein metric in Hilbert spaces. Ann. Probab. 17 (3), 1264–1276.

Dudley, R. M., 2002. Real Analysis and Probability. Vol. 74 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, revised reprint of the 1989 original.

Dupuy, J.-F., Loubes, J.-M., Maza, E., 2011. Non parametric estimation of the structural expectation of a stochastic increasing function. Stat. Comput. 21 (1), 121–136.

Gallon, S., Loubes, J.-M., Maza, E., 2013. Statistical properties of the quantile normalization method for density curve alignment. Mathematical Biosciences 242 (2), 129–142.

Gamboa, F., Loubes, J.-M., Maza, E., 2007. Semi-parametric estimation of shifts. Electron. J. Stat. 1, 616–640.

Haker, S., Zhu, L., Tannenbaum, A., Angement, S., 2004. Optimal mass transport for registration and warping. Int. J. Comput. Vis. 60 (3), 225–240.

Lescornel, H., Loubes, J.-M., 2012. Estimation of deformations between distributions by minimal Wasserstein distance.
URL http://hal.archives-ouvertes.fr/hal-00749519

Mémoli, F., 2011. Gromov-Wasserstein distances and the metric approach to object matching. Found. Comput. Math. 11 (4), 417–487.

Ni, K., Bresson, X., Chan, T., Esedoglu, S., 2009. Local histogram based segmentation using the Wasserstein distance. Int. J. Comput. Vis. 84 (1), 97–111.

Rachev, S. T., 1982. Minimal metrics in the random variables space. In: Probability and statistical inference (Bad Tatzmannsdorf, 1981). Reidel, Dordrecht, pp. 319–327.

Rachev, S. T., Rüschendorf, L., 1998. Mass transportation problems. Vols. I and II. Probability and its Applications (New York). Springer-Verlag, New York.

Ramsay, J. O., Silverman, B. W., 2005. Functional data analysis, 2nd Edition. Springer Series in Statistics. Springer, New York.

Schmitzer, B., Schnorr, C., 2013. Object Segmentation by Shape Matching with Wasserstein Modes. In: Energy Minimization Methods in Computer Vision and Pattern Recognition, 9th International Conference, EMMCVPR 2013, Lund, Sweden, August 19-21, 2013. Proceedings. Springer-Verlag, Berlin Heidelberg, pp. 123–136.

Shorack, G., Wellner, J., 2009. Empirical Processes with Applications to Statistics. Vol. 59. Society for Industrial Mathematics, Philadelphia.

Trouve, A., Younes, L., 2005. Metamorphoses through Lie group action. Foundations of Computational Mathematics 5 (2), 173–198.

Villani, C., 2009. Optimal transport. Vol. 338 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin.

Vimond, M., 2010a. Efficient estimation for a subclass of shape invariant models. Ann. Statist. 38 (3), 1885–1912.

Whitt, W., 1976. Bivariate distributions with given marginals. Ann. Statist. 4 (6), 1280–1289.

Younes, L., 2004. Invariance, déformations et reconnaissance de formes. Vol. 44 of Mathématiques & Applications [Mathematics & Applications]. Springer-Verlag, Berlin.