Estimation of deformations between distributions by minimal Wasserstein distance

Hélène Lescornel, Jean-Michel Loubes

Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Estimation of deformations between distributions by minimal Wasserstein distance

Lescornel Hélène and Loubes Jean-Michel

Institut de Mathématiques de Toulouse
118 route de Narbonne
F-31062 Toulouse Cedex 9
e-mail: helene.lescornel@math.univ-toulouse.fr; loubes@math.univ-toulouse.fr

Abstract: We consider the issue of estimating a deformation operator acting on measures. For this we consider a parametric warping model on an empirical sample and provide a new matching criterion for cloud points based on a generalization of the registration criterion used in [12]. We study the asymptotic behaviour of the estimator of the deformation and provide some examples to some particular deformation models.

MSC 2010 subject classifications: Primary 62F12, 62F12; secondary 62E20.
Keywords and phrases: Wasserstein distance, M estimation, Density registration.

Contents

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 data samples, the usual notion of Euclidean mean is too rough 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. This problem arises naturally for a wide range of statistical research fields such as functional data analysis for instance in [14], [16], [12] and references therein, image analysis in [3] or [19], shape analysis in [13] with many applications ranging from biology in [7] to pattern recognition [17] just to name a few. To handle this issue without any assumption on the deformations, Sakoe and Chiba in [17] present a synchronization algorithm known as the Dynamic Time Warping (D.T.W.), aligning two curves by a time axis renormalization. When dealing with functional data observed in a regression scheme, this idea was generalized in [23].

For a better understanding of the deformations, another major direction has been investigated. It consists in modeling the deformations by a parametric warping operator, such as for instance, scale location parameters, rotations in [6], actions of parameters of Lie groups or in a more general way deformations parametrized by their coefficients on a given basis [5] or in an RKHS set [1]. 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. Semi-parametric technics as in [12] or [22] enable to provide sharp estimation of these parameters.

The same kind of issues arises when considering the estimation of distribution functions observed with deformations. This situation occurs often in biology, for example when considering gene expression data obtained from microarray technologies. A microarray is composed of several spots, containing copies of identical expressions of genes. From each spot, a measure is obtained but before performing any statistical analysis on such data, it is necessary to process rough data in order to remove any systematic bias inherent to the microarray technology. A natural way to handle this phenomena is to try to remove these variations in order to align the measured densities, which proves difficult since the densities are unknown. In bioinformatics and computational biology, a method to reduce this kind of variability is known as normalization.

However, when dealing with the registration of warped distributions, the literature is scarce. We mention here the method provided for biological computational issues known as quantile normalization in [7] and the related work [11]. In [10] a criterion based on Wasserstein's distance is used to match two distributions for some particular deformation framework. In this work, we consider the extension of such parametric methods to the problem of estimating a distribution of random variables, observed in a warping framework through a precise estimation of the particular deformation parameters.

Actually, assume that we observe n replications of a random variable ε of law μ, and a sample $(X_i)_{1\leqslant i\leqslant n}$ of law $\mu_\star$ which is drawn from distribution μ with some variations in the sense that there exists an unobserved warping function φ such that we have $\mu_\star = \mu\circ\varphi_\star^{-1}$. 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 is a parameter $\theta^\star$ such that $\varphi_\star = \varphi_{\theta^\star}$. This parameter represents the warping effect that undergoes the sample $(X_i)_{1\leqslant i\leqslant n}$, which must be removed by inverting the warping operator. Hence, we will estimate, in a semi-parametric framework, the parameter $\theta^\star$.

For this, inspired by the matching criterion provided in [12], we warp the observations and construct an estimator $\widehat{\theta}^n$ of $\theta^\star$ by minimizing the energy needed to *align* all the distribution $\mu_\star$ to the distribution μ . That is to say, we will minimize the cost of transport of the mass charged by $\mu_\star$ on the mass charged by μ. Hence, to quantify the alignment between the two probabilities, it seems natural to us to consider the Wasserstein distance, see for instance in [21] or [2] for the connections between this distance and mass transport. We will obtain a result of consistency under general assumptions, in particular we will not assume the compactness of the support of μ. This estimator of $\theta^\star$ will enable us to obtain an consistent estimator of the structural distribution μ. Under stronger assumptions, following the proof in [10], we will also obtain a result of convergence in law for $\widehat{\theta}^n$.

The paper is organized as follows : the description of our model and the definitions of the estimators are given in Section 2. Section 3 is devoted to the convergence results obtained for the estimators of $\theta^\star$ and μ. In Section 4, a new framework is introduced to study the asymptotic comportment of the deformation estimates with a result about their convergence in distribution. Section 5 generalize the model to the case several deformations are observed. Section 6 presents some examples of deformations which fall in the scope of our study. Finally some applications to real data are provided in Section 7. The proofs are postponed to the Appendix.

2. Statistical model for distribution deformations

In this section, we will define a model for deformations of random variables and recall some useful definitions.

First, consider the following notations. In all the paper, we denote by $\|.\|$ the euclidean norm on $\mathbb{R}^k$ for all $k \in \mathbb{N}$, $k \geqslant 2$. For a given sample $Y = (Y_1, \ldots, Y_n)$, we denote by $Y_{(1)} \leqslant \cdots \leqslant Y_{(i)} \leqslant \cdots \leqslant Y_{(n)}$ its order statistics.

For $i = 1, \ldots, n$ and $j = 1, 2$, set ε_{ij} real i.i.d. random variables with unknown distribution μ defined on an Borel set $I_a \subset \mathbb{R}$. We will consider a deformation of these real-valued observations. Hence, we consider a family of deformation functions, indexed by parameters $\theta \in \Theta$, for Θ a compact and convex subset of $\mathbb{R}^d$, which warps a point x onto another point $\varphi_\theta(x)$. The shape of the deformation is modelled by the known function φ while the amount of deformation is characterized by the parameter θ. More precisely, we consider deformation functions that verify

$$\text{For all } \theta \in \Theta, \varphi_\theta : \begin{array}{ccc} I_a & \to & I_b \\ x & \mapsto & \varphi_\theta(x) \end{array} \text{ is invertible and increasing} \tag{\textbf{A1}}$$

where I_a, I_b are subsets of $\mathbb{R}$ possibly unbounded.

We assume that we observe

$$\begin{cases} \varepsilon_{i1}, \quad 1 \leqslant i \leqslant n \\ X_i = \varphi_{\theta^\star}(\varepsilon_{i2}), \quad 1 \leqslant i \leqslant n \end{cases} \tag{2.1}$$

where $\theta^\star$ is the unknown deformation parameter in $\Theta \subset \mathbb{R}^d$. We point out that this amounts to saying that one of the signal will be taken as a reference onto which will be aligned all the other warped observations. This assumption is also necessary is most of the literature on warped data.

Our aim is to estimate the parameter $\theta^\star \in \Theta$. For this, we will study a criterion based on a registration procedure for the distributions $\mu_\star = \mu \circ \varphi_{\theta^\star}^{-1}$ of the element of the i.i.d. sample $(X_1, \ldots, X_n)$ and the distribution μ of ε_{1j}. To compute the distance between the distributions, we will need the following probabilistic tools.

If F is a distribution function , we define the quantile function associated by

$$F^{-1}(t) = \inf\{x \in \mathbb{R}, F(x) \geqslant t\}.$$

Recall that if F_n is the empirical distribution associated to a sample $(Y_1, \ldots, Y_n)$, then we have

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

A natural distance to measure the deformation cost to align two distributions is given by the Wasserstein distance. For $p \in \mathbb{N}^\star$, consider the following set

$$\mathcal{W}_2(\mathbb{R}^p) = \{P \text{ probability on } \mathbb{R}^p \text{ which admits a finite second order moment}\}.$$

Given two probabilities P and Q in $\mathcal{W}_2(\mathbb{R}^p)$ we denote by $\mathcal{P}(P,Q)$ the set of all probability measures π over the product set $\mathbb{R}^p \times \mathbb{R}^p$ 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

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

The quadratic transportation cost allows to endow the set $\mathcal{W}_2(\mathbb{R}^p)$ with a metric by setting

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

Note that we will use W_2 metrics in this work. This choice is led by the issue of optimal matching between cloud points, see for instance in [4]. 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 $\mathbb{R}^p$. In particular, the earth-mover distance which corresponds to $r = 1$ 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 founded in [15].

Hereafter, we will consider distributions on $\mathbb{R}$. In this case the Wasserstein distance can be computed directly using the inverse distribution functions, as

$$W_2^2(P,Q) = \int_0^1 \left(F^{-1}(t) - G^{-1}(t)\right)^2 dt, \tag{2.2}$$

where F (resp. G) is the distribution function associated to P (resp. Q). The registration procedure we consider is an extension to point cloud estimation of the methodology pioneered in [12] and deeply studied in [22]. Wasserstein distance is actually a powerful tool to study similarities between point distributions, see in [8] or [9].

Recall that our aim is to align the law $\mu_\star$ of the observations X_i on the law μ. Hence a natural idea is to apply the inverse deformation operator to these observations. More precisely for all candidate θ, and to each observation X_i, we can apply the inverse deformation of parameter θ. Hence we can compute the following random variables

$$Z_i(\theta) = \varphi_\theta^{-1}(X_i). \tag{2.3}$$

Now, denote by $\mu_\star(\theta)$ the common law of the elements of the i.i.d. sample $Z(\theta) = (Z_1(\theta), \ldots, Z_n(\theta))$. We have $\mu_\star(\theta) = \mu_\star \circ \varphi_\theta = \mu \circ \varphi_{\theta^\star}^{-1} \circ \varphi_\theta$.

Let

$$\mu_\star^n(\theta) = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i(\theta)} \text{ and } \mu_j^n = \frac{1}{n}\sum_{i=1}^n \delta_{\varepsilon_{ij}}$$

the empirical laws associated with the samples $(Z_i(\theta))_{1\leqslant i\leqslant n}$ and $(\varepsilon_{ij})_{1\leqslant i\leqslant n}$ for $j = 1, 2$. Then $\mu_\star^n(\theta) = \mu_2^n \circ \varphi_{\theta^\star}^{-1} \circ \varphi_\theta$.
We note $F_\star$ the distribution function associated with the law $\mu_\star$ and F the distribution function associated with the law μ, $F_\star^n$ the empirical distribution function of the random sample $(X_{12}, \ldots, X_{n2})$ and F^n the empirical distribution function of the random sample $(\varepsilon_{11}, \ldots, \varepsilon_{n1})$.

Consider the following assumption necessary to compute the Wasserstein's distance between the warped samples

$$\text{For all } \theta \in \Theta, \varphi_\theta^{-1}(\cdot) \text{ is in } L^2(\mu_\star), \text{ that is } \varphi_\theta^{-1} \circ \varphi_{\theta^\star}(\cdot) \in L^2(\mu). \tag{\textbf{A2}}$$

Then introduce the following criterion

$$M: \quad \theta \mapsto M(\theta) = W_2^2(\mu_\star(\theta), \mu). \tag{2.4}$$

For $\theta = \theta^\star$, we get $\mu_\star(\theta^\star) = \mu$. Hence the distributions are the same for the true parameter $\theta^\star$, and the criterion M reaches its minimum at this point.

The estimation of this criterion is given by its corresponding empirical version, which is

$$M_n(\theta) = W_2^2(\mu_\star^n(\theta), \mu_1^n). \tag{2.5}$$

It can be computed using (2.2) and the order statistics associated with the sample $(Z_i(\theta))_{1 \leqslant i \leqslant n}$ and $(\varepsilon_{i1})_{1 \leqslant i \leqslant n}$

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left[Z_{(i)}(\theta) - \varepsilon_{(i)1} \right]^2.$$

The estimator of $\theta^\star$ is finally defined as

$$\widehat{\theta}^n \in \arg\min_{\theta \in \Theta} M_n(\theta). \tag{2.6}$$

Our aim is thus twofold.

- First, study the asymptotic comportment of this M-estimator.
- Then, using this estimator, estimate the template measure μ with a plug-in procedure

$$\widehat{\mu}^n = \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n \delta_{\varphi_{\widehat{\theta}^n}^{-1}(X_i)} + \mu_1^n \right) = \frac{1}{2} \left(\mu_\star^n \left(\widehat{\theta}^n \right) + \mu_1^n \right) \tag{2.7}$$

We point out that we restrict ourselves to distributions on $\mathbb{R}$ and not $\mathbb{R}^p$. As a matter of fact, the statistical analysis of the estimates and their asymptotic behaviour in distribution require a particular study of the asymptotic expansion of M_n that can not be achieved using the general expression of Wasserstein metrics. Indeed, we will need to express W_2 with quantile functions, estimated by the corresponding order statistics, which can only be done in the one dimensional case. The extension of this work to the case where distributions are multidimensional deserves a specific method which will be the subject of a future work.

3. Consistent estimation of the deformation parameters and the distribution template

The main objective of this section is to study the consistency of the estimator defined in (2.6) as

$$\widehat{\theta}^n \in \arg\min_{\theta\in\Theta} M_n(\theta).$$

In addition of assumptions **A1** and **A2**, we consider the following regularity assumptions on the deformation functions.

$$\text{For all } x \in I_b, \varphi_\theta^{-1}: \begin{array}{ccc} \Lambda & \to & I_a \\ \lambda & \mapsto & \varphi_\theta^{-1}(x) \end{array} \text{ is continuously differentiable.} \qquad (\mathbf{A3})$$

We denote its partial differential with respect to the variable θ on θ_0 by $\partial\varphi_{\theta_0}^{-1}(x) \in \mathbb{R}^d$.

$$\begin{gathered}\text{The family } \left(\partial\varphi_\theta^{-1}(\cdot)\right)_{\theta\in\Theta} \text{ has an envelope in } L^2(\mu_\star), \\ \text{that is } \sup_{\theta\in\Theta}\left\|\partial\varphi_\theta^{-1}(x)\right\| \leqslant H(x), H \in L^2(\mu_\star).\end{gathered} \qquad (\mathbf{A4})$$

It remains to have the following inequality

$$\sup_{\theta\in\Theta}\left\|\partial\varphi_\theta^{-1}\circ\varphi_{\theta^\star}(x)\right\| \leqslant G(x)$$

with $G \in L^2(\mu)$.

These two assumptions are required in order to get bounds for the empirical processes.

The last assumption is related to the identifiability of the model. More precisely it ensures that M admits an unique minimum on Θ at the parameter of interest, $\theta^\star$.

$$\begin{gathered}\text{For all } \theta \neq \theta^\star \in \Theta, \text{ there exists a set } A \\ \text{such that } \mu(A) > 0 \text{ and } \varphi_\theta^{-1}\circ\varphi_{\theta^\star} \neq Id \text{ on } A.\end{gathered} \qquad (\mathbf{A5})$$

Finally, recall that Θ is a compact and convex subset of $\mathbb{R}^d$.

3.1. Estimation of $\theta^\star$

Assume we observe $X_i, i = 1, \ldots, n$ and $\varepsilon_{i1}, i = 1, \ldots, n$, defined in (2.1). The following theorem proves the consistency of the estimator of the deformation parameter.

Theorem 3.1. *Under assumptions* **A1** *to* **A5**, $\widehat{\theta}^n \in argmin_{\theta\in\Theta} M_n(\theta)$ *converges in probability to* $\theta^\star$ *when* n *tends to infinity .*

The estimate of $\theta^\star$ is defined as an M-estimator. Hence its study follows the classical guidelines stated for instance in [20]. More precisely, its consistency can be obtained by establishing the uniform convergence of the criterion, that is

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{n\to\infty} 0 \text{ in probability}$$

under the following condition of identifiability

$$\text{for all } \varepsilon > 0, \inf_{\Theta\cap B(\theta^\star,\varepsilon)^c} M(\theta) > 0.$$

So according to Theorem 5.7 p.45 in [20], these two results enable to obtain Theorem 3.1.

The uniform convergence is obtained through the followings steps

- We first prove the pointwise convergence of M_n to M in probability. It involves classical properties of the Wasserstein distance about the convergence of empirical measures.
- Next we obtain the following property of "uniform continuity"

$$\text{for all } \varepsilon > 0, \limsup_{n\to\infty} \mathbb{P}\left(\sup_{\|\theta^1-\theta^2\|\leqslant\nu} \left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| > \varepsilon\right) \xrightarrow{\nu\to 0} 0.$$

  This part is the most important, and especially requires assumption **A4**.

We conclude by using arguments of compactness and continuity. The latter, in addition to assumption **A5**, are used to obtain the condition of identifiability. The details of the proof are given in the Appendix.

3.2. Reconstruction of the measure μ

Theorem 3.1 enables to get a sharp approximation of the true parameters of deformations with the estimator $\widehat{\theta}^n$. This entails that the observations can be aligned by computing the inverse transformation applied to the observations. Actually when n is sufficiently large, $\varphi^{-1}_{\widehat{\theta}^n}(X_i) = \varphi^{-1}_{\widehat{\theta}^n} \circ \varphi_{\theta^\star}(\varepsilon_{i2})$ is very close to ε_{i2}. So a natural estimator of the measure μ is given by

$$\mu^n_\star\left(\widehat{\theta}^n\right) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\varphi^{-1}_{\widehat{\theta}^n}(X_i)}.$$

The following theorem proves the consistency of $\mu^n_\star\left(\widehat{\theta}^n\right)$.

Theorem 3.2. *Under assumptions* **A1** *to* **A5**, $\mu^n_\star\left(\widehat{\theta}^n\right)$ *converges in the Wasserstein distance sense to the measure* μ *in probability :*

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

Using the warped observations to estimate the template distribution has some advantages. First it can be viewed as a mean to increase the size n of the sample $(\varepsilon_{i1})_{1\leqslant i\leqslant n}$. Of course, $\mu_\star^n$ will not perform as well as μ_1^n, but since $\widehat{\theta}_n$ is close to θ, we can expect that the plugged-in distribution estimate behaves roughly as the empirical measure. However, to quantify this conjecture, results on the exact rate of convergence of the Wasserstein distances are needed, which are difficult Theorems out of the scope of this paper.

4. Asymptotic analysis of the deformation parameters

4.1. Assumptions

Now we add to assumptions **A1** to **A5** the following regularity conditions on the deformation functions.

$$\varphi^{-1} \text{ is } C^2 \text{ with respect to } (\theta, x) \text{ on } \theta \times I_b. \tag{\textbf{AL1}}$$

We denote by $\partial\varphi_\theta^{-1}(x)$ its partial derivative with respect to the first variable at the point (θ, x) and by $d\varphi_\theta^{-1}(x)$ its partial derivative with respect to the second variable at the point (θ, x).

Consider the following restrictions on the distribution $\mu_\star$, which is the distribution of observations $(X_i)_{1\leqslant i\leqslant n}$.

$$\mu_\star \text{ is a law with compact support } [\alpha;\beta] \subset I_b. \tag{\textbf{AL2}}$$

$$F_\star \text{ is } C^1 \text{ and } F_\star' = f_\star > 0 \text{ on its support.} \tag{\textbf{AL3}}$$

Actually these assumptions are required to prove the convergence in distribution of the empirical quantile functions.

Note that using the relation $F_\star = F \circ \varphi_{\theta^\star}$ which is due to **A1**, we obtain that **AL2** and **AL3** imply that F is continuously differentiable with strictly positive derivative denoted by f.

4.2. Asymptotic distribution of the deformation estimates

Theorem 4.1.

Set $\Phi = \int_0^1 \partial\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right) \partial\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)^T dt \in \mathbb{R}^{d\times d}$.

Under assumptions **A1** *to* **A5** *and* **AL1** *to* **AL3**, *and if* Φ *is invertible, then*

$$\sqrt{n}\left(\widehat{\theta}^n - \theta^\star\right) \rightharpoonup (\Phi)^{-1} \int_0^1 \frac{\partial\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)}{f(F^{-1}(t))} \left[\mathbb{G}_2(t) - \mathbb{G}_1(t)\right] dt \tag{4.1}$$

where $\mathbb{G}_1$ *and* $\mathbb{G}_2$ *are independent standard Brownian bridges.*

The proof, given in the Appendix, is done for $d = 1$ that is $\theta^\star \in \Theta \subset \mathbb{R}^d = \mathbb{R}$. The generalization in higher dimension comes straightforward.

Remark that

$$\int_0^1 \frac{\partial \varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)}{f(F^{-1}(t))}\left[\mathbb{G}_2(t) - \mathbb{G}_1(t)\right] dt \sim \mathcal{N}\left(0; 2\int_{[0;1]\times[0;1]} \frac{\partial \varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)}{f(F^{-1}(t))}\frac{\partial \varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(s)\right)}{f(F^{-1}(s))}\left(\min(s,t) - st\right) dsdt\right).$$

The matrix Φ is invertible for instance in the case where the vector space generated by the family $\left\{\partial \varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)\right\}_{t\in(0;1)}$ has an orthogonal complementary reduced to zero. In the classical deformation families studied later, this matrix is always invertible.

5. Extensions to multiple deformations

Our model can be easily extended to the case where we observe several deformations of a single signal. In this case, the observation model can be written as

$$\begin{cases} \varepsilon_{i1}, & 1 \leqslant i \leqslant n \\ X_{i1} = \varphi_{\theta_1^\star}\left(\varepsilon_{i2}\right), & 1 \leqslant i \leqslant n \\ \dots & \\ X_{iJ} = \varphi_{\theta_J^\star}\left(\varepsilon_{iJ+1}\right), & 1 \leqslant i \leqslant n. \end{cases} \tag{5.1}$$

In this case, the aim is to estimate the vector $\theta^\star = (\theta_1^\star, \dots \theta_J^\star)$ by a quantity $\widehat{\theta}^n = \left(\widehat{\theta}_1^n, \dots, \widehat{\theta}_J^n\right)$.

We call $\mu_{\star,j}$ the law of X_{1j} and its distribution function is denoted by $F_{\star,j}$.

Following our method, we consider for all j

$$M_n\left(\theta_j\right) = \frac{1}{n}\sum_{i=1}^n \left[Z_{(i)j}\left(\theta\right) - \varepsilon_{(i)1}\right]^2,$$

where $Z_{ij}\left(\theta\right) = \varphi_{\theta_j}\left(X_{ij}\right)$, and choose

$$\widehat{\theta}_j^n \in \arg\min_{\theta_j \in \Theta} M_n\left(\theta_j\right).$$

Then, assume assumption **A1** to **A3**. Instead of assumption **A4**, one has to assume that for all j

$$\sup_{\theta\in\Theta}\left\|\partial \varphi_\theta^{-1} \circ \varphi_{\theta_j^\star}\left(x\right)\right\| \leqslant G_j(x)$$

with $G_j \in L^2\left(\mu\right)$.

The last assumption related to the identifiability of the model should also be reformulated as "for all $\theta \neq \theta_j^\star \in \Theta$, there exists a set A such that $\mu\left(A\right) > 0$ and $\varphi_\theta^{-1} \circ \varphi_{\theta_j^\star} \neq Id$ on A."

Then the convergence in probability of the whole vector $\widehat{\theta}^n$ comes straightforward from Theorem 3.1.

The convergence in Wasserstein sense of the measures $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varphi^{-1}_{\widehat{\theta}^n_j}(X_{ij})}$ is also a simple consequence of Theorem 3.2.

Concerning the convergence in law, assume $d=1$, **AL1**, **AL2** for all $\mu_{\star,j}$ and **AL3** for all $F_{\star,j}$, $1 \leqslant j \leqslant J$.

Then, slight modifications of the proof of Theorem 4.1 lead to the following result of convergence in law

$$\sqrt{n}\left(\widehat{\theta}^n - \theta^\star\right) \rightharpoonup Z$$

where

$$Z_j = \left(\int_0^1 \partial\varphi^{-1}_{\theta^\star_j}\left(F^{-1}_{\star,j}(t)\right)^2 dt\right)^{-1}\int_0^1 \frac{\partial\varphi^{-1}_{\theta^\star_j}\left(F^{-1}_{\star,j}(t)\right)}{f(F^{-1}(t))}\left[\mathbb{G}_j(t)-\mathbb{G}_0(t)\right]dt$$

with $(\mathbb{G}_0,\mathbb{G}_1,\ldots,\mathbb{G}_J)$ are independent standard brownian bridges.

We point out that we only consider the asymptotic with respect to n, the number of points per individuals. Another interesting but yet different point of view would be to tackle the case where J is large with respect to n.

6. Examples of deformation families

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

6.1. Example 1 : Location/scale model

$$\varphi_\theta(x) = \theta_2 x + \theta_1$$

This choice of deformation is related to observations

$$X_{ij} = \mu^\star_j + \sigma^\star_j\varepsilon_{ij} \quad 1 \leqslant i \leqslant n \quad 1 \leqslant j \leqslant J$$

where ε_{ij} are random independent variables drawn from an unknown distribution μ. It corresponds to an ANOVA model with heterogenous variances.

Here $\theta = (\theta_1,\theta_2) \in \Theta \subset \mathbb{R}^2$. The deformation function φ_θ is invertible on $\mathbb{R}$ if $\theta_2 \neq 0$. φ_θ is non decreasing if $\theta_2 > 0$, then we must choose Θ as a compact convex subset of $\mathbb{R}\times(0;+\infty)$.

We have $\varphi^{-1}_\theta(x) = \frac{x-\theta_1}{\theta_2} = \varphi_{\left(\frac{-\theta_1}{\theta_2},\frac{1}{\theta_2}\right)}(x)$, and $\varphi^{-1}_\theta(\varphi_\beta(x)) = \frac{x\beta_2+\beta_1-\theta_1}{\theta_2}$ which is in $L^2(\mu)$ if $\mu \in \mathcal{W}_2(\mathbb{R})$.

Moreover $\partial\varphi^{-1}_\theta(x) = \left(\frac{-1}{\theta_2},\frac{\theta_1-x}{\theta_2^2}\right)$ and $\left\|\partial\varphi^{-1}_\theta(x)\right\| = \sqrt{\left(\frac{-1}{\theta_2}\right)^2+\left(\frac{\theta_1-x}{\theta_2^2}\right)^2}$. Hence $\sup_{\theta\in\Theta}\left\|\partial\varphi^{-1}_\theta(\cdot)\right\| \in L^2(\mu_\star)$ if $\mu \in \mathcal{W}_2(\mathbb{R})$.

In conclusion, assumptions **A1** to **A5** are verified as soon as Θ is a compact convex of $\mathbb{R} \times (0;+\infty)$ and $\mu \in \mathcal{W}_2(\mathbb{R})$ is different from the Dirac mass at zero.

Now, remark that $\partial\varphi_\theta^{-1}\left(F_\star^{-1}(t)\right) = \frac{-1}{\theta_2^\star}\left(1, F^{-1}(t)\right)$. Hence, the matrix Φ defined in Theorem 4.1 is

$$\Phi = \frac{1}{\left(\theta_2^\star\right)^2}\begin{pmatrix} 1 & \mathbb{E}[\varepsilon] \\ \mathbb{E}[\varepsilon] & \mathbb{E}\left[\varepsilon^2\right] \end{pmatrix}.$$

Then it is invertible if ε is not constant a.e. which is necessary to get the invertibility of $F_\star$.

In the particular case of a translation model, $\theta_2 = 0$, i.e $\varphi_\theta(x) = x + \theta$, the assumptions are also easily tractable : if Θ is compact and convex in $\mathbb{R}$ and $\mu \in \mathcal{W}_2(\mathbb{R})$, **A1** to **A5** are verified.

The case of the scale model, that is $\varphi_\theta(x) = \theta x$, can also be considered as a particular case. Here, assumptions **A1** to **A4** are verified if $\mu \in \mathcal{W}_2(\mathbb{R})$, and Θ is a compact interval included in $(0;+\infty)$. **A5** holds if μ is different from the Dirac mass at zero.

6.2. Example 2 : Logarithmic transform

$$\varphi_\theta(x) = \theta\log(x)$$

φ_θ is invertible from $(0;+\infty)$ to $\mathbb{R}$ for all $\theta \neq 0$, and φ_θ is non decreasing if θ is positive: here Θ must be contained in $(0;+\infty)$ and ε take its values in $(0;+\infty)$. We have $\varphi_\theta^{-1}(x) = \exp\left(\frac{x}{\theta}\right)$, and $\varphi_\theta^{-1}(\varphi_\beta(x)) = \exp\left(\frac{\beta\log(x)}{\theta}\right) = x^{\frac{\beta}{\theta}}$. Hence $\varphi_\theta^{-1} \in L^2(\mu_\star)$ if $\mathbb{E}\left[\varepsilon^{\frac{2\theta^\star}{\theta}}\right] < \infty$ for all $\theta \in \Theta$.

Moreover $\partial\varphi_\theta^{-1}(x) = \frac{-x}{\theta^2}\exp\left(\frac{x}{\theta}\right)$, so $\partial\varphi_\theta^{-1}(\varphi_\beta(x)) = \frac{-\beta}{\theta^2}x^{\frac{\beta}{\theta}}\log(x)$, and $\sup_{\theta\in\Theta}\left|\partial\varphi_\theta^{-1}(\cdot)\right| \in L^2(\mu_\star)$ if $\mathbb{E}\left[\varepsilon^{\frac{2\theta^\star}{\theta_{min}}}\log^2(\varepsilon)\right] < \infty$ and $\mathbb{E}\left[\varepsilon^{\frac{2\theta^\star}{\theta_{Max}}}\log^2(\varepsilon)\right] < \infty$ where $\theta_{Max} = \max\{\theta \in \Theta\}$ and $\theta_{min} = \min\{\theta \in \Theta\}$. In this case the conditions are more restrictive on the law μ, but remark that the exponential distribution verifies these conditions. Assumption of identifiability holds if μ is different from the Dirac mass at point 1.

6.3. Example 3 : Composition $\varphi_\theta(x) = f \circ \tilde{\varphi}_\theta(x)$

Consider a function $\tilde{\varphi}_\theta(x)$ which verifies all the assumptions **A1** to **A5**. Then, if f is an increasing function invertible from I_b to I_c, the deformation function $\varphi_\theta(x) = f \circ \tilde{\varphi}_\theta(x)$ verifies also these assumptions replacing I_b by I_c. Indeed, assumptions **A1**, and **A3** are immediately verified, and we have

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

and

$$\partial\varphi_\theta^{-1} = \partial\left(\tilde{\varphi}_\theta^{-1} \circ f^{-1}\right) = \partial\tilde{\varphi}_\theta^{-1} \circ f^{-1}.$$

So

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

Hence assumptions of integrability (**A2**, **A4**) and identifiability (**A5**) are also verified for the function $\varphi_\theta(x)$.

The composition action allows a large number of new admissible deformations. For instance, the logit model $\varphi_\theta(x) = \frac{1}{1+\exp(-\theta x)}$ can be obtained by the composition of the scale model with the function $f(x) = \frac{1}{1+\exp(-x)}$.

The study of the example 2 gives also the conditions under which the deformation $\varphi_\theta(x) = x^\theta$ can be handled by our method.

7. Application to real data

In this section, we asses our methodology with application to different real datasets coming from genetics and meteorology. In all cases, our aim is to align the distribution of the datas, and to control the quality of our estimator through their alignment. For that, we consider that data can be modelled through (5.1). More precisely, we consider that one of the sample is a reference, that is it corresponds to $(\varepsilon_{i1})_{1\leqslant i\leqslant n}$ and that the others are different deformations $X_{ij} = \varphi_{\theta_j^\star}(\varepsilon_{ij+1})$, for $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant J$. We compute the estimators of the deformation parameters $\widehat{\theta}_j^n$ and the quantities $Z_{ij}\left(\widehat{\theta}_j^n\right) = \varphi^{-1}_{\widehat{\theta}_j^n}(X_{ij})$ for $1 \leqslant j \leqslant J$, $1 \leqslant i \leqslant n$. Finally we plot the densities of the samples $(\varepsilon_{i1})_{1\leqslant i\leqslant n}$, and $(X_{ij})_{1\leqslant i\leqslant n}$ for all $j \geqslant 1$ and the densities of the aligned variables $\left(Z_{ij}\left(\widehat{\theta}^n\right)\right)_{1\leqslant i\leqslant n}$.

For its versatility, we consider the scale/location model and choose for reference the first sample.

7.1. Genes of Zebrafish

Gene expression data obtained from microarray technologies are used to measure genome wide expression levels of genes in a given organism. A microarray may contain thousands of spots, each one containing a few million copies of identical DNA molecules that uniquely correspond to a gene. From each spot, a measure is obtained but observed with a systematic deformation inhering to the microarray technology: differential efficiency of the two fluorescent dyes, different amounts of starting mRNA material, background noise, hybridization reactions and conditions. A natural way to handle this phenomena is to try remove these variations. We apply our method to align two-channel (two-color) spotted microarrays, studying the underlying molecular mechanisms of differentially expressed genes of the popular Swirl data set, which can be downloaded from http://bioinf.wehi.edu.au/limmaGUI/DataSets.html. This experiment was conducted using zebrafish as a model organism to study early development in vertebrates. Swirl is a point mutation in the BMP2 gene that affects the dorsal-ventral body axis. Ventral fates such as blood are reduced, whereas dorsal structures such as somites and the notochord are expanded. One of the goals of the experiment is to identify genes with altered expression in the swirl mutant compared to wild-type zebrafish. A total of four arrays were performed in two dye-swap pairs with 8448 probes. The Fig. 1 and Fig. 2 show the estimated densities of unnormalized individual-channel intensities for two-color microarrays and its corresponding print-tip loess normalization within arrays, respectively. We can see that the normalization on the observations provide a good normalization of the densities, which enable a better comparison of the gene effects. Note that this dataset has been fully studied in [11].

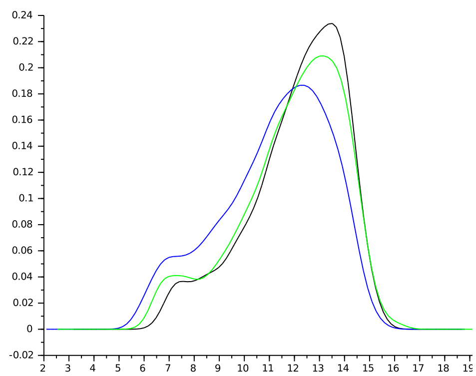

Figure 1: Densities for individual-channel intensities for zebrafish microarray data.

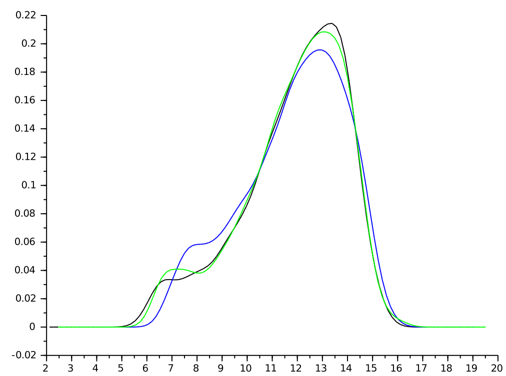

Figure 2: Densities for individual-channel intensities for zebrafish microarray data after normalization.

7.2. Temperature probes

We observe temperatures measured at 5 different probes located around the same area but with different conditions (different heights, different sun exposure conditions, different wind conditions ...). The mean temperature is recorded daily during more than 5 years. The outcome of the experiment is then 5 cloud points, consisting of 19918 random sample. To provide a template of the distribution of the temperatures, the 5 different distributions have to be normalized in order to remove the variability due to the location of the probes and not of the variability of the weather.

We present in Figure 3 and 4 the initial densities and the aligned densities of the data. Estimating the deformation parameters enable here to learn a correction rule. Then, this modification can be applied directly to the data in order to get an online rescaled information of the observed temperature at this point without the need for registrating the data.

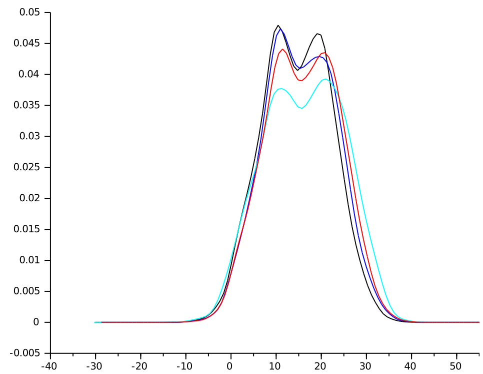

Figure 3: Density of temperature data

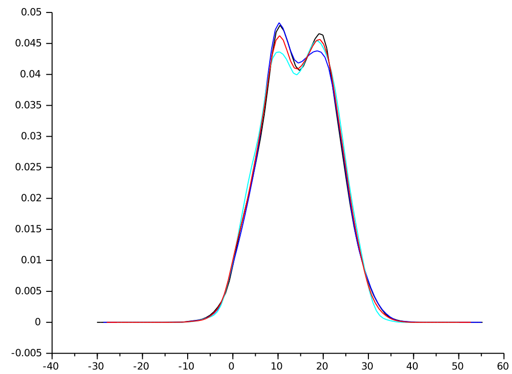

Figure 4: Density of normalized temperature data

7.3. Conclusion

In all the cases, we can see that our methodology performs well. Working directly on the data enables a better comparison by reducing the variability due to extra effects. To our knowledge, there are few methodologies aligning random variables by matching their distribution even if many authors have focused on registration methods for functional data. Yet, our aim is no to compete with normalization procedures but rather to show that the semi-parametric model we consider behaves well in practice with the advantage of providing particular estimates of the deformation parameters, which can be used for statistical inference on the data. Moreover, aligning the data enables to learn an automatic correction that can be applied to new data in order to align the data automatically.

8. Appendix section

8.1. Proof of Theorem 3.1

We start by proving the uniqueness of the minimum of the criterion $M(\theta)$.

STEP 0 : Identifiability

We have already remarked that $M(\theta^\star) = 0 = \min_{\theta \in \Theta} M(\theta)$.

Set $\theta \in \Theta$. We have $M(\theta) = 0$ if and only if $W_2^2(\mu_\star(\theta), \mu) = 0$, that is

$$\varphi_\theta^{-1} \circ \varphi_{\theta^\star} = Id \quad \mu \text{ a.s.}$$

Hence under assumption **A5** $\theta^\star$ is the only minimizer of M.

Now we aim to show that the empirical criterion M_n converges uniformly to M in probability. The proof follows three steps, beginning with the study of the pointwise convergence.

STEP 1

For all θ in Θ,

$$\left|M_n\left(\theta\right)-M\left(\theta\right)\right| \xrightarrow{n\to\infty} 0 \text{ in probability .}$$

Proof.

We have to prove that $W_2\left(\mu_\star^n\left(\theta\right),\mu_1^n\right) \xrightarrow{n\to\infty} W_2\left(\mu_\star\left(\theta\right),\mu\right)$

It comes almost directly from the following result about the convergence in the Wasserstein sense of the empirical measures which is stated in [18] p. 63.

If P_n is the empirical law of an i.i.d. sample $Y_1,\ldots,Y_n$ with law $P\in\mathcal{W}_2\left(\mathbb{R}\right)$, then

$$W_2\left(P_n,P\right) \xrightarrow{n\to\infty} 0 \text{ a.s.}$$

Indeed, using the triangular inequality we can write

$$\begin{aligned} W_2\left(\mu_\star^n\left(\theta\right),\mu_1^n\right) \leqslant W_2\left(\mu_\star^n\left(\theta\right),\mu_\star\left(\theta\right)\right)+W_2\left(\mu_\star\left(\theta\right),\mu\right)\\ +W_2\left(\mu,\mu_1^n\right)\end{aligned}$$

and

$$\begin{aligned} W_2\left(\mu_\star\left(\theta\right),\mu\right) \leqslant W_2\left(\mu_\star\left(\theta\right),\mu_\star^n\left(\theta\right)\right)+W_2\left(\mu_\star^n\left(\theta\right),\mu_1^n\right)\\ +W_2\left(\mu_1^n,\mu\right).\end{aligned}$$

Hence

$$\begin{gathered} W_2\left(\mu_\star\left(\theta\right),\mu\right)-W_2\left(\mu_\star^n\left(\theta\right),\mu_\star\left(\theta\right)\right)-W_2\left(\mu,\mu_1^n\right)\\ \leqslant W_2\left(\mu_\star^n\left(\theta\right),\mu_1^n\right)\\ \leqslant W_2\left(\mu_\star^n\left(\theta\right),\mu_\star\left(\theta\right)\right)+W_2\left(\mu_\star\left(\theta\right),\mu\right)+W_2\left(\mu,\mu_1^n\right).\end{gathered}$$

So because $\mu_\star^n\left(\theta\right)$ (resp. μ_1^n) is the empirical law associated with an i.i.d. sample of law $\mu_\star\left(\theta\right)\in\mathcal{W}_2\left(\mathbb{R}\right)$ (resp. $\mu\in\mathcal{W}_2\left(\mathbb{R}\right)$) we conclude that for all θ

$$M_n\left(\theta\right)=W_2\left(\mu_\star^n\left(\theta\right),\mu_1^n\right) \xrightarrow{n\to\infty} W_2\left(\mu_\star\left(\theta\right),\mu\right)=M\left(\theta\right) \text{ a.s.}$$

and consequently the convergence in probability holds, implied by the a.s. convergence.

STEP 2

For all $\varepsilon>0$

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta^1-\theta^2\|\leqslant\nu}\left|M_n\left(\theta^1\right)-M_n\left(\theta^2\right)\right|>\varepsilon\right) \xrightarrow{\nu\to 0} 0$$

Proof.
Recall that
$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[Z_{(i)}(\theta) - \varepsilon_{(i)1}\right]^2.$$
For θ^1 and θ^2 in Θ, we have

$$\left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| \leqslant \frac{1}{n}\sum_{i=1}^{n}\left|\left[Z_{(i)}\left(\theta^1\right) - \varepsilon_{(i)1}\right]^2 - \left[Z_{(i)}\left(\theta^2\right) - \varepsilon_{(i)1}\right]^2\right|.$$

It can be bounded using the equality $a^2 - b^2 = (a-b)(a+b)$ by
$$\left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| \leqslant \frac{1}{n}\sum_{i=1}^{n} A_i\left(\theta^1,\theta^2\right) B_i\left(\theta^1,\theta^2\right),$$

where we have set
$$A_i\left(\theta^1,\theta^2\right) = \left|Z_{(i)}\left(\theta^1\right) - Z_{(i)}\left(\theta^2\right)\right|$$
and
$$B_i\left(\theta^1,\theta^2\right) = \left|Z_{(i)}\left(\theta^1\right) + Z_{(i)}\left(\theta^2\right) - 2\varepsilon_{(i)1}\right|.$$
Using Cauchy-Schwarz's inequality we get
$$\left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n} B_i\left(\theta^1,\theta^2\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} A_i\left(\theta^1,\theta^2\right)^2}.$$

We first consider
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} B_i\left(\theta^1,\theta^2\right)^2}.$$
Using **A1** and the triangular inequality
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} B_i\left(\theta^1,\theta^2\right)^2} \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}\varphi_{\theta^1}^{-1}\left(X_{(i)}\right)^2} + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{(i)1}^2}$$
$$+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varphi_{\theta^2}^{-1}\left(X_{(i)}\right)^2}.$$

Hence
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} B_i\left(\theta^1,\theta^2\right)^2} \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}\varphi_{\theta^1}^{-1}\left(X_i\right)^2} + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i1}^2}$$
$$+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varphi_{\theta^2}^{-1}\left(X_i\right)^2}.$$

So

$$\sup_{\|\theta^1-\theta^2\|\leqslant\nu}\sqrt{\frac{1}{n}\sum_{i=1}^{n}B_i\left(\theta^1,\theta^2\right)^2}\leqslant 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{\lambda\in\Theta}\varphi_{\lambda}^{-1}\left(X_i\right)^2}+2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i1}^2}.$$

Now we will show that

$$\sup_{\|\theta^1-\theta^2\|\leqslant\nu}\sqrt{\frac{1}{n}\sum_{i=1}^{n}B_i\left(\theta^1,\theta^2\right)^2}=O_{\mathbb{P}}(1).$$

Using assumption **A3**, for λ^1, $\lambda^2\in\Theta$ we can write

$$\varphi_{\lambda^1}^{-1}(x)-\varphi_{\lambda^2}^{-1}(x)=\partial\varphi_{\lambda^{1,2}}^{-1}(x)\left(\lambda^1-\lambda^2\right)$$

for $\lambda^{1,2}$ on the segment between λ^1 and λ^2. Then

$$\left|\varphi_{\lambda^1}^{-1}(x)-\varphi_{\lambda^2}^{-1}(x)\right|\leqslant\sup_{\lambda\in\Theta}\left\|\partial\varphi_{\lambda}^{-1}(x)\right\|\left\|\lambda^1-\lambda^2\right\|.$$

So for all $\lambda\in\Theta$, using **A4**

$$\left|\varphi_{\lambda}^{-1}(x)\right|\leqslant H(x)\Delta+\left|\varphi_{\lambda^0}^{-1}(x)\right|$$

where $\lambda^0\in\Theta$ and Δ is the diameter of Θ. Hence **A2** implies that

$$\sup_{\lambda\in\Theta}\left|\varphi_{\lambda}^{-1}(\cdot)\right|^2\in L^1\left(\mu_\star\right) \tag{8.1}$$

and so we can use the Law of Large Numbers to obtain that $\frac{1}{n}\sum_{i=1}^{n}\sup_{\lambda\in\Theta}\varphi_{\lambda}^{-1}\left(X_i\right)^2$ and $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i1}^2$ converge in probability, hence we get

$$\sup_{\theta^1,\theta^2\in\Theta^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}B_i\left(\theta^1,\theta^2\right)^2}=O_{\mathbb{P}}(1).$$

Now we focus on $\sqrt{\frac{1}{n}\sum_{i=1}^{n}A_i\left(\theta^1,\theta^2\right)^2}$.
Using again assumption **A1**, we can write

$$\begin{aligned}A_i\left(\theta^1,\theta^2\right)&=\left|Z_{(i)}\left(\theta^1\right)-Z_{(i)}\left(\theta^2\right)\right|\\&=\left|\varphi_{\theta^1}^{-1}\left(X_{(i)}\right)-\varphi_{\theta^2}^{-1}\left(X_{(i)}\right)\right|.\end{aligned}$$

Now using again a Taylor-Lagrange expansion

$$\begin{aligned}\left|\varphi_{\theta^1}^{-1}\left(X_i\right)-\varphi_{\theta^2}^{-1}\left(X_i\right)\right|&=\left|\partial\varphi_{\tilde{\theta}_i^{1,2}}^{-1}\left(X_i\right)\left(\theta^1-\theta^2\right)\right|\\&\leqslant\sup_{\lambda\in\Theta}\left\|\partial\varphi_{\lambda}^{-1}\left(X_i\right)\right\|\left\|\theta^1-\theta^2\right\|\end{aligned}$$

so

$$\sup_{\|\theta^1-\theta^2\|\leqslant\nu}\frac{1}{n}\sum_{i=1}^{n}A_i\left(\theta^1,\theta^2\right)^2\leqslant\frac{1}{n}\sum_{i=1}^{n}\sup_{\lambda\in\Theta}\left\|\partial\varphi_\lambda^{-1}\left(X_i\right)\right\|^2\nu^2.$$

But under assumption **A3** we can apply the Law of Large Numbers to get that $\frac{1}{n}\sum_{i=1}^{n}\sup_{\lambda\in\Theta}\left\|\partial\varphi_\lambda^{-1}\left(X_i\right)\right\|^2$ converges in probability, and so

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\lambda\in\Theta}\left\|\partial\varphi_\lambda^{-1}\left(X_i\right)\right\|^2=O_{\mathbb{P}}(1).$$

In conclusion

$$\sup_{\|\theta^1-\theta^2\|\leqslant\nu}\left|M_n\left(\theta^1\right)-M_n\left(\theta^2\right)\right|\leqslant V_n\nu^2$$

where $V_n=O_{\mathbb{P}}\left(1\right)$ is independent of ν and we obtain

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta^1-\theta^2\|\leqslant\nu}\left|M_n\left(\theta^1\right)-M_n\left(\theta^2\right)\right|>\varepsilon\right)\xrightarrow{\nu\to 0}0.$$

STEP 3
The function $\theta\mapsto M\left(\theta\right)$ is continuous on Θ.

Proof.
Let $\left(\theta^n\right)_{n\in\mathbb{N}}$ be a sequence of Θ such that $\theta^n\xrightarrow{n\to\infty}\theta^0$. We will show that $W_2^2\left(\mu_\star\left(\theta^n\right),\mu_\star\left(\theta^0\right)\right)\xrightarrow{n\to\infty}0$, that is $M\left(\theta^n\right)\xrightarrow{n\to\infty}M\left(\theta^0\right)$. For this we will use the following equivalence.

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

$$W_2\left(P_n,P\right)\xrightarrow{n\to\infty}0$$

if and only if

$$P_n\rightharpoonup P\text{ and }\mathbb{E}\left[Y_n^2\right]\to\mathbb{E}\left[Y^2\right]$$

where Y_n follows the law P_n and Y the law P.

This characterization of the convergence in the Wasserstein's sense is proved for instance in [18]. Recall that $Z_1\left(\theta\right)=\varphi_\theta^{-1}\circ\varphi_{\theta^\star}\left(\varepsilon_{12}\right)$.

We first show that $\mathbb{E}\left[Z_1^2\left(\theta^n\right)\right]\xrightarrow{n\to\infty}\mathbb{E}\left[Z_1^2\left(\theta^0\right)\right]$. Thanks to (8.1) , we have for all $\theta\in\Theta$,

$$\left|Z_1\left(\theta\right)\right|=\left|\varphi_\theta^{-1}\left(X_1\right)\right|\leqslant\tilde{H}\left(X_1\right)$$

with $\tilde{H}\left(X_1\right)\in L^2$.

Moreover using the regularity of φ^{-1} with respect to the deformation parameter we have the a.s. convergence

$$Z_1\left(\theta^n\right)^2=\varphi_{\theta^n}^{-1}\left(X_1\right)^2\xrightarrow{n\to\infty}\varphi_{\theta^0}^{-1}\left(X_1\right)^2=Z_1\left(\theta^0\right)^2.$$

Hence we obtain $\mathbb{E}\left[Z_1^2\left(\theta^n\right)\right] \xrightarrow{n\to\infty} \mathbb{E}\left[Z_1^2\left(\theta^0\right)\right]$.

In addition, we proved the a.s. convergence of $Z_{12}^2\left(\theta^n\right)$ to $Z_{12}^2\left(\theta^0\right)$, which implies the weak convergence $\mu_\star\left(\theta^n\right) \rightharpoonup \mu_\star\left(\theta^0\right)$.

From this we deduce that $W_2^2\left(\mu_\star\left(\theta^n\right), \mu_\star\left(\theta^0\right)\right) \xrightarrow{n\to\infty} 0$ and consequently $M\left(\theta^n\right) \xrightarrow{n\to\infty} M\left(\theta^0\right)$ if $\theta^n \xrightarrow{n\to\infty} \theta^0$: M is continuous on Θ.

CONSEQUENCE

If Θ is compact, then

$$\sup_{\theta\in\Theta}\left|M_n(\theta) - M(\theta)\right| \xrightarrow{n\to\infty} 0 \text{ in probability.}$$

Proof.

Set ε and δ two real positive numbers. Thanks to the steps 2 and 3, we can choose ν_0 such that

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta^1-\theta^2\|\leqslant\nu_0}\left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| > \varepsilon\right) \leqslant \delta$$

and

$$\sup_{\|\theta^1-\theta^2\|\leqslant\nu_0}\left|M\left(\theta^1\right) - M\left(\theta^2\right)\right| \leqslant \varepsilon.$$

With the compactness of Θ, we can find a sequence $\left(\theta^k\right)_{1\leqslant k\leqslant m}$ in Θ such that $\Theta \subset \cup_{k=1}^m B\left(\theta^k, \nu_0\right)$. Now for $\theta \in \Theta \cap B\left(\theta^p, \nu_0\right)$

$$\left|M_n(\theta) - M(\theta)\right| \leqslant \left|M_n(\theta) - M_n\left(\theta^p\right)\right| + \left|M_n\left(\theta^p\right) - M\left(\theta^p\right)\right| + \left|M\left(\theta^p\right) - M(\theta)\right|$$

$$\begin{aligned}\sup_{\theta\in\Theta}\left|M_n(\theta) - M(\theta)\right| \leqslant &\sup_{\|\theta^1-\theta^2\|\leqslant\nu_0}\left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| + \\ &\max_{1\leqslant k\leqslant m}\left|M_n\left(\theta^k\right) - M\left(\theta^k\right)\right| + \sup_{\|\theta^1-\theta^2\|\leqslant\nu_0}\left|M\left(\theta^1\right) - M\left(\theta^2\right)\right|\end{aligned}$$

Hence

$$\left(\sup_{\theta\in\Theta}\left|M_n(\theta) - M(\theta)\right| > 3\varepsilon\right) \subset$$
$$\left(\sup_{\|\theta^1-\theta^2\|\leqslant\nu_0}\left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| > \varepsilon\right) \cup \left(\max_{1\leqslant k\leqslant m}\left|M_n\left(\theta^k\right) - M\left(\theta^k\right)\right| > \varepsilon\right).$$

So

$$\begin{aligned}\mathbb{P}\left(\sup_{\theta\in\Theta}\left|M_n(\theta) - M(\theta)\right| > 3\varepsilon\right) \leqslant &\mathbb{P}\left(\sup_{\|\theta^1-\theta^2|\leqslant\nu_0}\left|M_n\left(\theta^1\right) - M_n\left(\theta^2\right)\right| > \varepsilon\right) \\ &+ \sum_{k=1}^m \mathbb{P}\left(\left|M_n\left(\theta^k\right) - M\left(\theta^k\right)\right| > \varepsilon\right)\end{aligned}$$

And with the step 1, we deduce that for all δ and $\varepsilon > 0$

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| > 3\varepsilon\right) \leqslant \delta.$$

Hence,

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{n\to\infty} 0 \text{ in probability.}$$

Finally we complete the proof as follows.

Using the result of identifiability together with the continuity of M and the compactness of Θ, we deduce that for all $\varepsilon > 0$

$$\inf_{\Theta\cap B(\theta^\star,\varepsilon)^c} M > 0.$$

Following the M-estimation theorem of [20] (th 5.7 p.45), this result combining with the uniform convergence in probability of M_n to M leads to the consistency of the estimator.

8.2. Proof of Theorem 2.7

Recall that μ_2^n is the empirical law of the sample $(\varepsilon_{12}, \dots \varepsilon_{n2})$. Then we have

$$W_2\left(\mu_\star^n\left(\widehat{\theta}^n\right), \mu\right) \leqslant W_2\left(\mu_\star^n\left(\widehat{\theta}^n\right), \mu_2^n\right) + W_2\left(\mu_2^n, \mu\right).$$

With the convergence of empirical measures in the Wasserstein sense used in the step 1 in the proof of Theorem 3.1 we get the a.s. convergence of $W_2(\mu_2^n, \mu)$ to 0 when n tends to infinity.

Second, φ_θ is non decreasing for all θ, so we have

$$\begin{aligned} W_2^2\left(\mu_\star^n\left(\widehat{\theta}^n\right), \mu_2^n\right) &= \frac{1}{n}\sum_{i=1}^n \left(\varphi_{\widehat{\theta}^n}^{-1}\left(\varphi_{\theta^\star}\left(\varepsilon_{(i)2}\right)\right) - \varepsilon_{(i)2}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left(\varphi_{\widehat{\theta}^n}^{-1}\left(\varphi_{\theta^\star}\left(\varepsilon_{i2}\right)\right) - \varepsilon_{i2}\right)^2, \end{aligned}$$

and with a Taylor expansion of $\theta \mapsto \varphi_\theta^{-1}(X_i)$ between $\widehat{\theta}^n$ and $\theta^\star$, we obtain

$$\varphi_{\widehat{\theta}^n}^{-1}\left(\varphi_{\theta^\star}\left(\varepsilon_{i2}\right)\right) = \varepsilon_{i2} + \partial\varphi_{\widetilde{\theta}_i^n}^{-1}(X_i)\left(\widehat{\theta}^n - \theta^\star\right)$$

for $\widetilde{\theta}_i^n$ in the segment between $\widehat{\theta}^n$ and $\theta^\star$. So

$$W_2^2\left(\mu_\star^n\left(\widehat{\theta}^n\right), \mu_2^n\right) \leqslant \left[\frac{1}{n}\sum_{i=1}^n \sup_{\lambda\in\Theta} \left\|\partial\varphi_\lambda^{-1}(X_i)\right\|^2\right] \left\|\widehat{\theta}^n - \theta^\star\right\|^2.$$

But we showed in the step 2 of the proof of Theorem 3.1 that $\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{\lambda\in\Theta}\left\|\partial\varphi_{\lambda}^{-1}\left(X_i\right)\right\|^2\right] = O_{\mathbb{P}}(1)$, and the consistency of the estimator $\widehat{\theta}^n$ implies that $\left\|\widehat{\theta}^n - \theta^{\star}\right\| \xrightarrow{n\to\infty} 0$ in probability. Hence we deduce that $W_2^2\left(\mu_{\star}^n\left(\widehat{\theta}^n\right), \mu_2^n\right) \xrightarrow{n\to\infty} 0$ in probability.

In conclusion,

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

8.3. *Proof of Theorem 4.1*

For sake of simplicity, we prove the theorem in the case where $d = 1$.

Here we introduce new notations.

We note $\mathbb{D}]\alpha;\beta]$ the set of distribution functions of measures that concentrate on $]\alpha;\beta]$ and $\mathbb{S}$ the Skorohod space, that is the space of cadlag functions on $\bar{\mathbb{R}}$ endowed with the supremum norm $\|\quad\|_{\infty}$. Recall that the cadlag functions are defined as the right continuous functions which admit a limit from the left.

$\ell_{\infty}(0;1)$ is the set of functions bounded on $(0;1)$, and for $I = I_b$ or $I = [\alpha;\beta]$, $\ell_{\infty}\left((0;1);I\right)$ is the set of functions bounded on $(0;1)$ with values in I. $\ell_{\infty,m}(0;1)$ denotes the set of bounded and measurable functions on $(0;1)$. Recall that $[\alpha;\beta] \subset I_b$.

On the spaces $\ell_{\infty}^2(0;1)$ and $\ell_{\infty,m}^2(0;1)$ we consider the norm $\|h\|_{\infty,2} = max\left(\|h_1\|_{\infty}, \|h_2\|_{\infty}\right)$ where $h = (h_1, h_2)$. Finally we denote by $Q_{\star}^n$ the empirical quantile function $(F_{\star}^n)^{-1}$ and $Q^n = (F^n)^{-1}$.
We start by the computation of the first and second derivatives of M_n.

Differentiability of $\mathbf{M_n}$

We have

$$M_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\varphi_{\theta}^{-1}\left(X_{(i)}\right) - \varepsilon_{(i)1}\right]^2.$$

Hence M_n is C^2 on Θ under **AL1**, and

$$\partial M_n\left(\theta\right) = \frac{2}{n}\sum_{i=1}^{n}\partial\varphi_{\theta}^{-1}\left(X_{(i)}\right)\left[\varphi_{\theta}^{-1}\left(X_{(i)}\right) - \varepsilon_{(i)1}\right].$$

We can also write,

$$\partial M_n\left(\theta\right) = 2\int_0^1 \partial\varphi_{\theta}^{-1}\left(Q_{\star}^n(t)\right)\left[\varphi_{\theta}^{-1}\left(Q_{\star}^n(t)\right) - Q^n(t)\right]dt. \tag{8.2}$$

Moreover

$$\partial^2 M_n\left(\theta\right) = \frac{2}{n}\sum_{i=1}^{n}\partial\varphi_{\theta}^{-1}\left(X_{(i)}\right)\left[\partial\varphi_{\theta}^{-1}\left(X_{(i)}\right)\right] + \frac{2}{n}\sum_{i=1}^{n}\partial^2\varphi_{\theta}^{-1}\left(X_{(i)}\right)\left[\varphi_{\theta}^{-1}\left(X_{(i)}\right) - \varepsilon_{(i)1}\right].$$

So

$$\partial^2 M_n(\theta) = 2\int_0^1 \partial\varphi_\theta^{-1}\left(Q_\star^n(t)\right)^2 + \partial^2\varphi_\theta^{-1}\left(Q_\star^n(t)\right)\left[\varphi_\theta^{-1}\left(Q_\star^n(t)\right) - Q^n(t)\right] dt. \tag{8.3}$$

The regularity of M_n allows a Taylor expansion

$$\partial M_n\left(\widehat{\theta}^n\right) = \partial M_n(\theta^\star) + \partial^2 M_n\left(\widetilde{\theta}^n\right)\left(\widehat{\theta}^n - \theta^\star\right)$$

for $\widetilde{\theta}^n$ between $\widehat{\theta}^n$ and $\theta^\star$. Using that M_n admits a minimum on $\widehat{\theta}^n$ we have

$$-\partial M_n(\theta^\star) = \partial^2 M_n\left(\widetilde{\theta}^n\right)\left(\widehat{\theta}^n - \theta^\star\right).$$

We set $\partial M_n(\theta^\star) = \Psi(F^n, F_\star^n)$, $\partial^2 M_n\left(\widetilde{\theta}^n\right) = \Phi\left(F^n, F_\star^n, \widetilde{\theta}^n\right)$.

The aim of the following is to show that Ψ is Hadamard differentiable in order to apply a Delta method to get $\sqrt{n}(-\partial M_n(\theta^\star)) \rightharpoonup Z$ for some random variable Z.

Convergence in law of $\partial\mathbf{M_n}(\theta^\star)$.
We have $\Psi = \chi \circ \Psi_0$ where

$$\Psi_0(F, F_\star) = \left(F^{-1}, F_\star^{-1}\right)$$

is defined on $\mathbb{D}]\alpha;\beta]^J$ with values in $\ell^2_{\infty,m}((0;1),[\alpha;\beta])$.
χ is defined from $\ell^2_{\infty,m}((0;1),[\alpha;\beta])$ to $\mathbb{R}$ with

$$\chi(g_1, g_2) = 2\int_0^1 \partial\varphi_{\theta^\star}^{-1}(g_2(t))\left[\varphi_{\theta^\star}^{-1}(g_2(t)) - g_1(t)\right] dt.$$

Now consider the following lemma.

Lemma 8.1. *Let $G : I_b^J \to \mathbb{R}$ a continuous function. Then, if $[\alpha;\beta] \subset I_b$,*

$$\tilde{G} : \left(\ell^J_{\infty,m}((0;1);[\alpha;\beta]), \|\quad\|_{\infty,J}\right) \to \mathbb{R}$$

$$(g_1, \ldots, g_J) \mapsto \int_0^1 G(g_1(u), \ldots, g_J(u))\, du$$

is continuous. If G is continuously differentiable, then $\tilde{G}$ is Hadamard differentiable tangentially to $\ell^J_{\infty,m}((0;1))$ with

$$D\tilde{G}(g_1, \ldots, g_J)[h_1, \ldots, h_J] = \int_0^1 DG(g_1(u), \ldots, g_J(u))[h_1(u), \ldots, h_J(u)]\, du.$$

Using **AL1**, we apply this lemma for $J = 2$ to

$$G(x_1, x_2) = \partial\varphi_{\theta^\star}^{-1}(x_2)\left[\varphi_{\theta^\star}^{-1}(x_2) - x_1\right]$$

and we deduce that χ is Hadamard differentiable tangentially to $\ell^2_{\infty,m}(0;1)$. Moreover, for $k_1, k_2 \in \ell^2_{\infty,m}(0;1)$

$$\begin{aligned} D\chi(g_1, g_2)[k_1, k_2] =& 2\int_0^1 d\partial\varphi_{\theta^\star}^{-1}(g_2(t))[k_2(t)]\left[\varphi_{\theta^\star}^{-1}(g_2(t)) - g_1(t)\right] dt \\ &+ 2\int_0^1 \partial\varphi_{\theta^\star}^{-1}(g_2(t))\left[d\varphi_{\theta^\star}^{-1}(g_2(t))[k_2(t)] - k_1(t)\right] dt \end{aligned}$$

Under **AL2** and **AL3** we can apply Theorem 8.2 in Section 8.5 which ensures that Ψ_0 is Hadamard differentiable at $(F, F_\star)$ tangentially to $C^2[\alpha;\beta]$, with

$$D\Psi_0(F, F_\star)[h_1, h_2] = -\left(\frac{h_1 \circ F^{-1}}{f \circ F^{-1}}, \frac{h_2 \circ F_\star^{-1}}{f_\star \circ F_\star^{-1}}\right)$$

for $(h_1, h_2) \in C^2[\alpha;\beta]$. Hence, with the regularity of the functions $F_\star$ and F, we obtain that $D\Psi_0(F, F_\star)\left(C^2[\alpha;\beta]\right) \subset \ell^2_{\infty,m}(0;1)$. Thus we can apply the chain rule to the composed function $\Psi = \chi \circ \Psi_0$ to get that Ψ is Hadamard differentiable at $(F, F_\star)$ tangentially to $C^2[\alpha;\beta]$ with

$$D\Psi(F, F_\star)[h] = D\chi(\Psi_0(F, F_\star))[D\Psi_0(F, F_\star)[h]]$$

for $h = (h_1, h_2) \in C^2[\alpha;\beta]$.

Under **A1**, we have $F_\star = F \circ \varphi_{\theta^\star}^{-1}$. Hence, $F_\star^{-1} = \left(F \circ \varphi_{\theta^\star}^{-1}\right)^{-1} = \varphi_{\theta^\star} \circ F^{-1}$ and we obtain $\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right) = F^{-1}(t)$. This leads to

$$\begin{aligned} &D\Psi(F, F_\star)[h_1, h_2] = \\ &2\int_0^1 \partial\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)\left[d\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)\left[\frac{-h_2\left(F_\star^{-1}(t)\right)}{f_\star\left(F_\star^{-1}(t)\right)}\right] - \frac{-h_1\left(F^{-1}(t)\right)}{f\left(F^{-1}(t)\right)}\right] dt, \end{aligned}$$

Moreover, if we differentiate the equality $F_\star(x) = F \circ \varphi_{\theta^\star}^{-1}(x)$ we obtain that $f_\star(x) = d\varphi_{\theta^\star}^{-1}(x)\, f \circ \varphi_{\theta^\star}^{-1}(x)$.

Hence $f_\star(F_\star^{-1}(t)) = d\varphi_{\theta^\star}^{-1}(F_\star^{-1}(t)) f \circ \varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right) = d\varphi_{\theta^\star}^{-1}(F_\star^{-1}(t)) f \circ F^{-1}(t)$, and we can simplify

$$D\Psi(F, F_\star)[h_1, h_2] = 2\int_0^1 \frac{\partial\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)}{f(F^{-1}(t))}\left[h_1\left(F^{-1}(t)\right) - h_2\left(F_\star^{-1}(t)\right)\right] dt$$

With the independence of the different samples and the convergence in law of the empirical distribution functions which is stated in Theorem 8.4 in Section 8.5, we know that

$$\sqrt{n}\begin{pmatrix} F^n - F \\ F_\star^n - F_\star \end{pmatrix} \rightharpoonup \begin{pmatrix} \mathbb{G}_1 \circ F \\ \mathbb{G}_2 \circ F_\star \end{pmatrix}$$

in the product space $(\mathbb{S}^2, \|\quad\|_{\infty,2})$ where $\mathbb{G}_1$ and $\mathbb{G}_2$ are independent standard Brownian bridges .

Hence we can apply Theorem 8.3, the functional Delta method which is stated in section 8.5 with the following correspondences : A is the Skohorod space, $A_\phi = \mathbb{D}]\alpha;\beta]^2$, $A_0 = C^2[\alpha;\beta]$ (we have $(\mathbb{G}_1 \circ F, \mathbb{G}_2 \circ F_\star) \in A_0$). Hence, computing $\Psi\left(F, F_\star\right) = 0$ we obtain

$$\sqrt{n}(-\partial M_n\left(\theta^\star\right)) \rightharpoonup -D\Psi\left(F, F_\star\right)\left[\mathbb{G}_1 \circ F, \mathbb{G}_2 \circ F_\star\right]$$

in $\mathbb{R}$.

Next we will show that $\Phi\left(F^n, F_2^n, \widetilde{\theta}^n\right) \to \Phi\left(F, F_\star, \theta^\star\right)$ in probability.

Convergence of $\partial^2 \mathbf{M_n}$.

We can write $\Phi\left(F^n, F_\star^n, \widetilde{\theta}^n\right) = \phi\left(\Psi_0(F^n, F_\star^n), \widetilde{\theta}^n\right) = \partial^2 M_n\left(\widetilde{\theta}^n\right)$where

$$\phi\left(g_1, g_2, \theta\right) = 2\int_0^1 \partial\varphi_\theta^{-1}\left(g_2(t)\right)^2 + \partial^2\varphi_\theta^{-1}\left(g_2(t)\right)\left[\varphi_\theta^{-1}\left(g_2(t)\right) - g_1(t)\right] dt.$$

Using **AL1** and a slight modification of Lemma 8.1, we get that the function ϕ is continuous on $\left(\ell^2_{\infty,m}\left((0;1)\,;[\alpha;\beta]\right) \times \mathbb{R}, \max\left(\|\quad\|_{\infty,2}, \|\quad\|\right)\right)$. Moreover,

$$\widetilde{\theta}^n \xrightarrow{n\to\infty} \theta^\star \text{ in probability}$$

and

$$\Psi_0\left(F^n, F_\star^n\right) \xrightarrow{n\to\infty} \Psi_0\left(F, F_\star\right) = \left(F^{-1}, F_\star^{-1}\right) \text{ in probability}$$

in the space $\left(\ell^2_{\infty,m}\left((0;1)\,;[\alpha;\beta]\right), \|\quad\|_{\infty,2}\right)$. Hence

$$\Phi\left(F^n, F_\star^n, \widetilde{\theta}^n\right) \xrightarrow{n\to\infty} \Phi\left(F, F_\star, \theta^\star\right) \text{ in probability,}$$

with

$$\Phi\left(F, F_\star, \theta^\star\right) = 2\int_0^1 \partial\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)^2 + \partial^2\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)\left[\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right) - F^{-1}(t)\right] dt$$

that is

$$\Phi\left(F, F_\star, \theta^\star\right) = 2\int_0^1 \partial\varphi_{\theta^\star}^{-1}\left(F_\star^{-1}(t)\right)^2 dt. \tag{8.4}$$

Hence one get the result using Slutsky's lemma.

8.4. Proof of Lemma 8.1

We first prove the continuity of $\tilde{G}$.

Choose $g = (g_1, \dots, g_J) \in \ell^J_{\infty,m}((0;1), [\alpha;\beta])$. G is uniformly continuous on the compact $[\alpha;\beta]^J \subset I_b^J$.

For all ε, set $\nu(\varepsilon)$ such that $|x-y|_\infty = max(|x_1-y_1|, \dots, |x_J-y_J|) \leqslant \nu(\varepsilon)$ implies $|G(x_1,\dots,x_J) - G(y_1,\dots,y_J)| \leqslant \varepsilon$ if $x, y \in [\alpha;\beta]^J$.

Set $h = (h_1,\dots,h_J) \in \ell^J_{\infty,m}((0;1),[\alpha;\beta])$ such that $\|h-g\|_{\infty,J} \leqslant \nu(\varepsilon)$. Then

$$\left|\tilde{G}(h) - \tilde{G}(g)\right| \leqslant \int_0^1 |G(g(u)) - G(h(u))|\, du \leqslant \int_0^1 \varepsilon du = \varepsilon :$$

$\tilde{G}$ is continuous.

Now we consider the Hadamard differentiability.

Let $g = (g_1,\dots,g_J) \in \ell^J_{\infty,m}((0;1),[\alpha;\beta])$, $h = (h_1,\dots,h_J) \in \ell^J_{\infty,m}(0;1)$ and $h^t = (h_1^t,\dots,h_J^t)$ such that $h^t \xrightarrow{t\to 0} h \in \ell^J_{\infty,m}((0;1))$ and $g + th^t \in \ell^J_{\infty,m}((0;1),[\alpha;\beta])$ for t sufficiently small. For v and w in $\mathbb{R}^J$, we denote by $[v;w]$ the segment between these two vectors, that is

$$[v;w] = \{sv + (1-s)w, s \in [0;1]\}.$$

Recall that we have set

$$D\tilde{G}(g)[h] = \int_0^1 DG(g_1(u),\dots,g_J(u))[h_1(u),\dots,h_J(u)]\, du.$$

First remark that $D\tilde{G}(g_1,\dots,g_J)$ is well definite, linear and continuous on $\ell^J_{\infty,m}(0;1)$.

Next, write

$$\left|\tilde{G}\left(g+th^t\right) - \tilde{G}(g) - t\int_0^1 DG(g(u))[h(u)]\, du\right|$$

$$\leqslant \int_0^1 \left|G\left((g(u)) + t\left(h^t(u)\right)\right) - G(g(u)) - tDG(g(u))\left[h^t(u)\right]\right| du$$
$$+ \int_0^1 \left|tDG(g(u))\left[h^t(u)\right] - tDG(g(u))[h(u)]\right| du$$

$$\leqslant \int_0^1 \sup_{k(u)\in[g(u);g(u)+th^t(u)]} \|DG(k(u)) - DG(g(u))\| \left\|t\left[h^t(u)\right]\right\| du$$
$$+ \int_0^1 \|DG(g(u))\| \left\|t\left[h^t(u)\right] - t[h(u)]\right\| du$$

with the Mean theorem applied to the function $F(x) = G(g(u) + tx) - tDG((g(u))\,x$ between $x = h^t(u)$ and $x = (0, \ldots, 0)$.

Hence for $t \neq 0$

$$\frac{1}{|t|}\left|\tilde{G}\,(g + th) - \tilde{G}\,(g) - t\int_0^1 DG\,(g_1(u), \ldots, g_J(u))\,[h_1(u), \ldots, h_J(u)]\,du\right|$$

$$\begin{aligned}&\leqslant \int_0^1 \sup_{k(u)\in[g(u);g(u)+th^t(u)]} \|DG\,(k(u)) - DG(g(u))\|\,du\,\left\|h^t\right\|_{\infty,J}\\ &\qquad + \int_0^1 \|DG\,(g_1(u), \ldots, g_J(u))\|\,du\,\left\|h - h^t\right\|_{\infty,J}.\end{aligned}$$

But for all u, $th^t(u)$ tends to 0 while t tends to 0 , and by continuity of DG we deduce that

$$\sup_{k(u)\in[g(u);g(u)+th^t(u)]} \|DG\,(k(u)) - DG(g(u))\| \xrightarrow{t\to 0} 0$$

for all u.

Moreover $u \mapsto DG\,(g_1(u), \ldots, g_J(u))$ is bounded thanks to the continuity of DG and the fact that $g \in \ell^J_{\infty,m}\,((0;1), [\alpha;\beta])$. Same arguments leads to the fact that $u \mapsto DG\,(k(u))$ is bounded for k between g and $g + th^t$ if t is sufficiently small. Hence we can apply the dominated convergence theorem to obtain that

$$\int_0^1 \sup_{k(u)\in[g(u);g(u)+th^t(u)]} \|DG\,(k(u)) - DG(g(u))\|\,du \xrightarrow{t\to 0} 0.$$

So with the convergence of h^t we conclude that

$$\frac{1}{|t|}\left|\tilde{G}\left(g + th^t\right) - \tilde{G}\,(g) - \int_0^1 tDG\,(g(u))\,[h(u)]\,du\right| \xrightarrow{t\to 0} 0$$

that is, $\tilde{G}$ is Hadamard differentiable tangentially to $\ell^J_{\infty,m}(0;1)$ with

$$D\tilde{G}\,(g_1, \ldots, g_J)\,[h_1, \ldots, h_J] = \int_0^1 DG\,(g_1(u), \ldots, g_J(u))\,[h_1(u), \ldots, h_J(u)]\,du.$$

8.5. Auxiliary theorems

The following theorems are taken from [20]. The first one is Lemma 21.4 p. 307.

Theorem 8.2. *Set*

$$\Psi_0\,(F_1, \ldots F_J) = \left(F_1^{-1}, \ldots, F_J^{-1}\right)$$

defined on $\mathbb{D}]\alpha;\beta]^J$ *with values in* $\ell^J_\infty(0;1)$

Assume that for all j, F_j has a compact support $[\alpha;\beta]$ and is continuously differentiable on its support with strictly positive derivative f_j. Then Ψ_0 is Hadamard differentiable on $(F_1, \dots F_J)$ tangentially to $C[\alpha;\beta]^J$. The derivative is the map defined on $C[\alpha;\beta]^J$:

$$(h_1, \dots, h_J) \mapsto \ -\left(\frac{h_1 \circ F_1^{-1}}{f_1 \circ F_1^{-1}}, \dots, \frac{h_J \circ F_J^{-1}}{f_J \circ F_J^{-1}} \right)$$

This one is the statement of the functional Delta method labelled as Theorem 20.8 p. 297.

Theorem 8.3. *Let A and B normed linear spaces, and $\phi : A_\phi \subset A \to B$ Hadamard differentiable at a tangentially to A_0. Let X_n random variables with values in A_ϕ such that $r_n (X_n - a) \rightharpoonup X$, where X takes its values in A_0 and $r_n \to \infty$.*

Then $r_n (\phi(X_n) - \phi(a)) \rightharpoonup D\phi(a)X$.

And finally Donsker's Theorem corresponds to Theorem 19.3 p. 266.

Theorem 8.4. *If $X_1, \dots, X_n$ are i.i.d. random variables with distribution function F and empirical distribution function F_n, the sequence $\sqrt{n}\,(F_n - F)$ converges in distribution in $(\mathbb{S}, \|\cdot\|_\infty)$ to $\mathbb{G} \circ F$ where $\mathbb{G}$ is a standard Brownian bridge.*

References

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

[2] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Trimmed comparison of distributions. *J. Amer. Statist. Assoc.*, 103(482):697–704, 2008.

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

[4] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

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

[6] J. Bigot, F. Gamboa, and M. Vimond. Estimation of translation, rotation, and scaling between noisy images using the Fourier-Mellin transform. *SIAM J. Imaging Sci.*, 2(2):614–645, 2009.

[7] 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.

[8] J. A. Cuesta and C. Matrán. Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.*, 17(3):1264–1276, 1989.

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

[10] G. Freitag and A. Munk. On Hadamard differentiability in k-sample semi-parametric models—with applications to the assessment of structural relationships. *Journal of Multivariate Analysis*, 94(1):123–158, 2005.

[11] S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129 – 142, 2013.

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

[13] 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.

[14] A. Kneip and K. J. Utikal. Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.*, 96(454):519–542, 2001. With comments and a rejoinder by the authors.

[15] S. T. Rachev. The Monge-Kantorovich problem on mass transfer and its applications in stochastics. *Teor. Veroyatnost. i Primenen.*, 29(4):625–653, 1984.

[16] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2nd edition, 2005.

[17] 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.

[18] G. Shorack and J. Wellner. *Empirical processes with applications to statistics*, volume 59. Society for Industrial Mathematics, 2009.

[19] A. Trouvé and L. Younes. Metamorphoses Through Lie Group Action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005.

[20] A. Van der Vaart. *Asymptotic statistics*. Number 3. Cambridge Univ Pr, 2000.

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

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

[23] K. Wang and T. Gasser. Synchronizing sample curves nonparametrically. *Ann. Statist.*, 27(2):439–460, 1999.