INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Deconvolution for the Wasserstein Metric and Geometric Inference

Claire Caillerie — Frédéric Chazal — Jérôme Dedecker — Bertrand Michel

N° 7678

Juillet 2011

_ Algorithms, Certification, and Cryptography _

ISSN 0249-6399 ISRN INRIA/RR--7678--FR+ENG

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche SACLAY - ÎLE-DE-FRANCE

Deconvolution for the Wasserstein Metric and Geometric Inference

Claire Caillerie *, Frédéric Chazal †, Jérôme Dedecker ‡ , Bertrand Michel §

Theme : Algorithms, Certification, and Cryptography
Équipe-Projet Geometrica

Rapport de recherche n° 7678 — Juillet 2011 — 28 pages

Abstract: Recently, [4] have defined a distance function to measures to answer geometric inference problems in a probabilistic setting. According to their result, the topological properties of a shape can be recovered by using the distance to a known measure ν, if ν is close enough to a measure μ concentrated on this shape. Here, close enough means that the Wasserstein distance W_2 between μ and ν is sufficiently small. Given a point cloud, a natural candidate for ν is the empirical measure μ_n. Nevertheless, in many situations the data points are not located on the geometric shape but in the neighborhood of it, and μ_n can be too far from μ. In a deconvolution framework, we consider a slight modification of the classical kernel deconvolution estimator, and we give a consistency result and rates of convergence for this estimator. Some simulated experiments illustrate the deconvolution method and its application to geometric inference on various shapes and with various noise distributions.

Key-words: Deconvolution, Wasserstein distance, geometric inference, computational topology

* INRIA Saclay
† INRIA Saclay
‡ Laboratoire MAP5, UMR CNRS 8145 Université Paris Descartes
§ Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie - Paris 6

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Déconvolution pour la métrique de Wasserstein et inférence géométrique

Résumé : La notion de fonction distance à une mesure récemment introduite dans [4] permet de répondre à des problèmes d'inférence géométrique dans un cadre probabiliste : les propriétés topologiques d'un compact $K \subset \mathbb{R}^d$ peuvent être estimées à l'aide de la fonction distance à une mesure de probabilité connue ν si celle-ci se trouve suffisamment proche (au sens de la distance de Wasserstein W_2) d'une mesure μ dont K est le support. En pratique lorsque les observations sont corrompues par du bruit, la mesure empirique associée aux observations n'est généralement pas assez proche de μ pour pouvoir être utilisée directement. Dans cet article, on propose une solution à ce problème en considérant un modèle de convolution pour lequel la loi du bruit est supposée connue. On considère une variante de l'estimateur par noyau de déconvolution classique dont on établit la consistence et des vitesses de convergence. On illustre la méthode proposée et ses applications en inférence géométrique sur différentes formes géométriques et différentes distributions de bruit sur les observations.

Mots-clés : Déconvolution, Distance de Wasserstein, Inférence géométrique, Topologie algorithmique

1 Introduction

Infering topological and geometric information from multivariate data is a problem which is attracting a lot of interest since a couple of decades. Many statistical methods have been developed to model and estimate geometric features from point cloud data that are usually considered as independent observations drawn according to a common distribution μ in an Euclidean space $\mathbb{R}^d$. In low dimensions, principal curves and principal surfaces have been early proposed by [17] to study simple manifolds. More elaborated structures can be also studied with density-based methods. For instance, filament estimation has been the subject of several works, see for instance [13] and [14] for recent contributions. In a more general context, set estimation deals with problems in the interplay between statistics and geometry. This field includes estimation of supports, boundaries and level sets, see [8] for a large overview on this topic. Cluster analysis algorithms also provide geometric information. One popular approach of clustering proposed by [16] consists in defining clusters as connected components of the levels sets associated to a density f, see for instance [7] and [1]. Another statistical work by [19] proposes estimators of the entropy dimension of the support of μ and of the number of clusters of the support in the case of corrupted data. The paper of [9] addresses estimation of the surface area of a d-dimensional body, as defined by the Minkowski measure. These above mentioned works propose efficient statistical methods for geometric inference but they usually do not provide topological guarantees on the estimated geometric quantities.

On the other hand many non stochastic methods have been proposed in computational geometry to infer the geometry of an unknown object from a set of data point sampled around it. In this context, distance functions to the data have shown to be efficient tools to robustly infer precise information about the geometry of the object. More precisely, [5] and [3] show that the sublevel sets of the distance function to the data can be used to recover the geometry of the unknown object. These methods offer strong geometric and topological guarantees but they rely on strong sampling assumptions that usually do not apply in a statistical framework. In particular, they fail when applied on data corrupted by outliers.

Recently, some efforts have been made to bridge the gap between the statistical and geometric approaches. For example, assuming that the observations are independently drawn from a probability measure that is the convolution of the uniform measure on a submanifold M with a Gaussian noise measure supported by the normals to M, [23] propose an algorithm to recover the Betti numbers of M. A major limitation of this method is that the noise should verify a strong variance condition.

In a different perspective [4] have generalized the approach of [3] by extending the notion of distance function from compact sets to probability measures. This new framework allows to robustly infer geometric properties of a distribution μ using independent observations drawn according to a distribution μ' "close" to μ where the closeness between probability distributions is assessed by a Wasserstein distance W_p defined by

$$W_p(\mu,\mu') = \inf_{\pi\in\Pi(\mu,\mu')} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \pi(dx,dy) \right)^{\frac{1}{p}},$$

where $\Pi(\mu,\mu')$ is the set of probability measures on $\mathbb{R}^d\times\mathbb{R}^d$ that have marginals μ and μ', $\|.\|$ is a norm and $p\geq 1$ is a real number (see [26] or [28]).

Given a probability distribution μ in $\mathbb{R}^d$ and a real parameter $0\leq m\leq 1$, [4] generalize the notion of distance to the support of μ by the function $\delta_{\mu,m}: x\in\mathbb{R}^d\mapsto \inf\{r>0\ :\ \mu(B(x,r))>m\}$ where $B(x,r)$ is the closed Euclidean ball of center x and radius r. To avoid issues due to discontinuities of the map $\mu\mapsto\delta_{\mu,m}$, the distance function to μ with paprameter $m_0\in[0,1]$ is defined by

$$\mathrm{d}_{\mu,m_0}:\mathbb{R}^d\to\mathbb{R}^+,\ x\mapsto\sqrt{\frac{1}{m_0}\int_0^{m_0}(\delta_{\mu,m}(x))^2\,dm}. \tag{1}$$

The function d_{μ,m_0} shares many properties with classical distance functions that make it well-suited for geometric inference purposes. In particular, the map $\mu\mapsto\mathrm{d}_{\mu,m_0}$ is $1/\sqrt{m_0}$-Lipschitz, i.e.

$$\sup_{x\in\mathbb{R}^d}|\mathrm{d}_{\mu,m_0}(x)-\mathrm{d}_{\mu',m_0}(x)|=\|\mathrm{d}_{\mu,m_0}-\mathrm{d}_{\mu',m_0}\|_\infty\leq\frac{1}{\sqrt{m_0}}W_2(\mu,\mu').$$

This property ensures that the distance functions associated to close measures (for the W_2 metric) have close sublevel sets. Moreover, the function d^2_{μ,m_0} is semiconcave (i.e. $x\mapsto\|x\|^2-\mathrm{d}^2_{\mu,m_0}(x)$ is convex) ensuring strong regularity properties on the geometry of its sublevel sets - see [24] for more informations on the geometric properties of semiconcave functions. Using these properties [4] (Corollary 4.11) prove, under some general assumptions, that if μ' is a probability distribution approximating μ, then the sublevel sets of d_{μ',m_0} provide a topologically correct approximation of the support of μ. The statement of such a result requires the following definitions. A probability measures is said to have dimension at most $k>0$ if there exits a constant $C(\mu)$ such that for any point x in the support of μ and any sufficiently small $\varepsilon>0$ one has $\mu(B(x,\varepsilon))\geq C(\mu)\varepsilon^k$. Given a compact set $G\subset\mathbb{R}^d$ and a real number $\alpha>0$ the α-reach of G denoted $\mathrm{reach}_\alpha(G)$ is a geometric quantity related to the critical points of the (classical) distance function d_G to G, $\mathrm{d}_G(x)=\inf\{d(x,y)\,y\in G\}$, that provides a measure of the regularity of G (see [3] for the definition).

Theorem 1. *Let μ be a measure that has dimension at most $k>0$ with compact support G such that $\mathrm{reach}_\alpha(G)\geq R>0$ for some $\alpha>0$. Let ν be another measure and ε be an upper bound on the uniform distance between d_G and d_{ν,m_0}. Then, for any $r\in[4\varepsilon/\alpha^2,R-3\varepsilon]$ and any $\eta\in]0,R[$, the r-sublevel sets of d_{μ,m_0} and the η-sublevel sets of d_G are homotopy equivalent as soon as:*

$$W_2(\mu,\nu)\leq\frac{R\sqrt{m_0}}{5+4/\alpha^2}-C(\mu)^{-1/k}m_0^{1/k+1/2}.$$

Roughly speaking, this result means that if ones knows a measure ν that is close to μ for the Wasserstein metric, the level sets of d_{ν,m_0} can be used to infer the topology of the sublevel sets of the distance function to the support G of μ. In practice if one observes a set of points independently sampled according to the distribution μ (resp. to some distribution μ' that is close to μ), a natural candidate for ν is the empirical measure of the points cloud μ_n. Indeed, $\mathbb{E}(W_2^2(\mu_n,\mu))$ (resp $\mathbb{E}(W_2^2(\mu_n,\mu'))$) converges to zero as n tends to infinity, as shown by [18].

However, in many situations the data is contaminated by noise, namely we observe some points drawn according to the convolution $\mu \star \nu$ where μ is supported by the unknown compact set G and ν is the distribution of the noise. In such a situation, $\mathbb{E}(W_2^2(\mu_n, \mu))$ does not converges to zero anymore, and μ_n may be too far from μ to apply Theorem 1. The aim of this article is to propose a deconvolution estimator $\hat{\mu}_n$ close to μ for the Wasserstein metric, and then to use the levels sets of $\mathrm{d}_{\hat{\mu}_n, m_0}$ to infer the topology of the sublevel sets of the distance function to G.

Many papers deal with the convolution model from a statistical point of view. We focus here on works related to support estimation or geometric inference. Support estimation in the convolution setting has been the subject of recent works mostly in the univariate case. In [15] and [10], the boundary of the support is detected *via* the large values of the derivate of the density estimator under the assumption that the density of μ has a discontinuity at the boundary. An alternative method based on moment estimation is proposed by [22] without assuming that the density is discontinuous at the boundary. For the multivariate case, [21] proposes an estimator of the support based on a resampling strategy, that satisfy some consistency properties. Still in the convolution setting, [19] gives estimates of the entropy dimension of the support G and of the number of clusters of G.

In this paper, we study the behavior of a deconvolution estimator with respect to the Wassertein metric W_2. In the applications we have in mind, μ is typically supported by a submanifold of $\mathbb{R}^d$ with dimension strictly less than d. Consequently, we shall not assume that μ has a density with respect to the Lebesgue measure on $\mathbb{R}^d$. In fact, except that it is compactly supported, we shall make no further assumptions on μ.

Besides the geometric applications we have in mind, studying the properties of probability estimators for the W_2 metric is also interesting in itself. Firstly, contrary to the $\mathbb{L}_p$-distances between probability densities (except for $p = 1$, which conincides with the total variation distance), the distances W_p are true distances between probability distributions. Secondly, many natural estimators $\hat{\mu}_n$ of μ are singular with respect to μ (think of the empirical measure in most cases), and consequently the total variation distance between $\hat{\mu}_n$ and μ is equal to 2 for any n. This is the case of our deconvolution estimator, if the support G is a submanifold in $\mathbb{R}^d$ with dimension srictly less than d. Wasserstein metrics appear as natural distances to evaluate the performance of such estimators.

The first section of this paper is devoted to the theoretical aspects of the paper. We first define the deconvolution estimator $\hat{\mu}_n$ and then we give rates of convergence for $\mathbb{E}(W_2^2(\hat{\mu}_n, \mu))$. The second section presents some numerical experiments with applications to geometric inference.

2 Deconvolution for the Wasserstein metric

We start with some notation. The inner product $< \cdot, \cdot >$ from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}$ is defined as follows: for $x = (x_1, \ldots, x_d)^t$ and $y = (y_1, \ldots, y_d)^t$, $< x, y >= x_1 y_1 + \cdots + x_d y_d$. The euclidean norm of x is denoted by $\|x\| = \sqrt{< x, x >}$.

In the following, we denote by μ^* (respectively f^*) the Fourier transform of the probability measure μ (respectively of the integrable function f), that is:

$$\mu^*(x) = \int_{\mathbb{R}^d} e^{i<t,x>}\mu(dt) \quad \text{and} \quad f^*(x) = \int_{\mathbb{R}^d} e^{i<t,x>} f(t)dt\,.$$

For two probability measures μ, ν on $\mathbb{R}^d$, we denote by $\mu \star \nu$ the convolution product of μ and ν, that is the image measure of $\mu \otimes \nu$ by the application $(x, y) \to x + y$ from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}^d$. If ν has a density g on $\mathbb{R}^d$, we denote by $\mu \star g$ the density of $\mu \star \nu$, that is

$$\mu \star g(x) = \int_{\mathbb{R}^d} g(x - z)\mu(dz)\,.$$

2.1 The multivariate convolution model

Assume that one observes n i.i.d. random vectors $(Y_i = (Y_{i,1}, \ldots, Y_{i,d})^t)_{1 \le i \le n}$ with values in $\mathbb{R}^d$ in the model

$$Y_i = X_i + \varepsilon_i, \tag{2}$$

where the random vectors $X_i = (X_{i,1}, \ldots, X_{i,d})^t$ are i.i.d and distributed according to an unknown probability measure μ supported on an unknown compact subset G of $\mathbb{R}^d$. The random vectors $\varepsilon_i = (\varepsilon_{i,1}, \ldots \varepsilon_{i,d})^t$'s are also i.i.d. random and distributed according to a probability measure μ_ε which is supposed to be known and symmetric (that is $-\varepsilon_1$ has the same distribution μ_ε). Hence, the distribution of the Y_i's is given by $\nu = \mu \star \mu_\varepsilon$.

Since μ_ε is symmetric, its Fourier transform μ_ε^* is a real-valued function. We also assume that

$$\int_{\mathbb{R}^d} \|x\|^6 \mu_\varepsilon(dx) < \infty\,, \tag{3}$$

which implies in particular that μ_ε^* is six times continuously differentiable. Finally, we assume that μ_ε^* is positive on $\mathbb{R}^d$.

Let μ_n be the empirical measure of the observations, that is

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}\,. \tag{4}$$

Under suitable assumptions, it follows from [18] that

$$\lim_{n\to\infty} \mathbb{E}(W_2^2(\mu_n, \mu)) = W_2^2(\mu \star \mu_\varepsilon, \mu)\,,$$

and the term on right hand is nonzero if μ_ε is not the Dirac measure at 0. Our aim is to provide an estimator $\hat{\mu}_n$ of the unknown distribution μ such that

$$\lim_{n\to\infty} \mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) = 0.$$

2.2 Deconvolution estimators

Let K be a symmetric density probability on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} \|x\|^2 K(x)dx < \infty\,.$$

Assume moreover that its Fourier transform K^* is compactly supported and two times differentiable with Lipschitz second derivatives. We shall give an example of such a kernel in Section 2.4.

Let H be an invertible matrix from $\mathbb{R}^d$ to $\mathbb{R}^d$, H^t be the transpose of H, and $|H|$ be the absolute value of the determinant of H. Define the preliminary estimator

$$\hat{f}_n(x) = \frac{1}{n|H|} \sum_{i=1}^{n} \tilde{K}_H(H^{-1}(x - Y_i)) \,, \tag{5}$$

where

$$\tilde{K}_H(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i<u,x>} \frac{K^*(u)}{\mu_\varepsilon^*((H^{-1})^t u)} du \,. \tag{6}$$

The kernel $\tilde{K}_H$ is called the deconvolution kernel. It is well defined since K^* is compactly supported and μ_ε^* is continuous and positive. Moreover $\tilde{K}_H$ belongs to $\mathbb{L}^1(\mathbb{R}^d)$: this follows from the fact that the function $u \to K^*(u)/\mu_\varepsilon^*((H^{-1})^t u)$ is compactly supported and two times differentiable.

The estimator (5) is the multivariate version of the standard deconvolution kernel density estimator which was first introduced in [2] and [27]. This estimator has been the subject of many works, in particular in the non-parametric univariate setting. Only few papers study the multidimensional deconvolution problem, see [6] for a recent work on this subject.

Note that $\hat{f}_n$ is not necessarily a density, since it has no reason to be non negative. Since our estimator has to be a probability measure, we define

$$\hat{g}_n(x) = \alpha_n \hat{f}_n^+(x), \quad \text{where} \quad \alpha_n = \frac{1}{\int_{\mathbb{R}^d} \hat{f}_n^+(x) dx} \quad \text{and} \quad \hat{f}_n^+ = \max\{0, \hat{f}_n\} \,.$$

The estimator $\hat{\mu}_n$ of μ is then the probability measure with density $\hat{g}_n$.

The first step is to prove a consistency result for this estimator, and to do this, we need to specify a loss function. The pointwise (semi-) metric and the $\mathbb{L}_2$ metric between probability densities are the most currently used (see for instance the monograph of [20]). Mean consistency results with respect to the $\mathbb{L}_1$ loss have also been proved by [12]. However, these loss functions are not adapted to our context, since we do not assume that μ has a density.

In this paper we take W_2^2 as our loss function, and we give rates of convergence for the quantity $\mathbb{E}(W_2^2(\hat{\mu}_n, \mu))$.

2.3 A general decomposition

In this section, we shall always assume that

$$\int_{\mathbb{R}^d} (1 + \|x\|^2) \sqrt{\mathrm{Var}(f_n(x))} dx < \infty \,,$$

which implies that $\mathbb{E}(W_2^2(\hat{\mu}_n, \mu))$ is finite. More precisely, we shall prove the following "bias-variance" decomposition:

Proposition 1. *Let*

$$B(H) = \int_{\mathbb{R}^d} \|Hu\|^2 K(u) du \quad \textit{and} \quad C(H) = B(H) + \int_{\mathbb{R}^d} \|x\|^2 \mu(dx) \,.$$

The following upper bound holds:

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq 2B(H) + 4\int_{\mathbb{R}^d} (2C(H) + \|x\|^2)\sqrt{\mathrm{Var}(f_n(x))}dx\,.$$

Proof of Proposition 1. We first define the kernel K_H by

$$K_H(x) = \frac{1}{|H|}K(H^{-1}x)\,.$$

As usual in deconvolution problems, the estimator $\hat{f}_n$ is build in such a way that $\mathbb{E}(\hat{f}_n(x)) = \mu \star K_H(x)$. Indeed, by Plancherel's identity

$$\begin{aligned}\mathbb{E}(\hat{f}_n(x)) &= \frac{1}{|H|}\int_{\mathbb{R}^d} \tilde{K}_H(H^{-1}(x-z))\mu \star \mu_\varepsilon(z)dz \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \Big(\frac{1}{|H|}\tilde{K}_H(H^{-1}(x-\cdot)\Big)^*(u)\mu^*(u)\mu_\varepsilon^*(u)du\,,\end{aligned}$$

Since K_H is symmetric, we have that

$$\Big(\frac{1}{|H|}\tilde{K}_H(H^{-1}(x-\cdot)\Big)^*(u) = e^{i<u,x>}\tilde{K}_H^*(-H^t u) = e^{i<u,x>}\tilde{K}_H^*(H^t u)\,,$$

and by definition of $\tilde{K}_H$,

$$\begin{aligned}\mathbb{E}(\hat{f}_n(x)) &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{i<u,x>}\frac{K^*(H^t u)}{\mu_\varepsilon^*(u)}\mu^*(u)\mu_\varepsilon^*(u)du \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \Big(\frac{1}{|H|}K(H^{-1}(x-\cdot)\Big)^*(u)\mu^*(u)du = \mu \star K_H(x)\,.\end{aligned}$$

Now, by the triangle inequality

$$W_2^2(\hat{\mu}_n, \mu) \leq 2W_2^2(\mu \star K_H, \mu) + 2W_2^2(\hat{\mu}_n, \mu \star K_H)\,. \tag{7}$$

The first term on the right hand side of (7) is deterministic, and can be easily bounded as follows: let Y_H be a random variable with distribution K_H and independent of X_1, in such a way that the distribution of $X_1 + Y_H$ is $\mu \star K_H$. By definition of W_2, one has

$$W_2^2(\mu \star K_H, \mu) \leq \mathbb{E}(\|X_1 + Y_H - X_1\|^2) = \mathbb{E}(\|Y_H\|_2^2) = B(H)\,. \tag{8}$$

To control the second term of (7), we shall use the following lemma

Lemma 1. *Let μ and ν be two probability measures on $\mathbb{R}^d$, and let $|\mu - \nu|$ be the total variation measure of $\mu - \nu$. Then*

$$W_2^2(\mu, \nu) \leq 2\min_{a\in\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-a\|^2|\mu-\nu|(dx)\,.$$

In particular, if μ and ν have respective densities f and g with respect to the Lebesgue measure

$$W_2^2(\mu, \nu) \leq 2\min_{a\in\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-a\|^2|f(x)-g(x)|dx\,. \tag{9}$$

Remark 1. *The inequality (9) has been proved by [29] with the constant 4, and by [18] with the constant 3. We give here a very elementary proof which provides a better constant.*

Remark 2. *If μ has a density f_μ with respect to the Lebesgue measure on $\mathbb{R}^d$, we can use the inequality (9) to obtain the following upper bound for the first term of (7)*

$$W_2^2(\mu \star K_H, \mu) \leq 2 \int_{\mathbb{R}^d} \|x\|^2 |f_\mu(x) - \mu \star K_H(x)| dx \,. \tag{10}$$

Now, as in density deconvolution, if f_μ is smooth enough, this upper bound may be more precise than the simple upper bound $W_2^2(\mu \star K_H, \mu) \leq B(H)$. However, the fact that f_μ exists and is smooth on $\mathbb{R}^d$ is a very restrictive assumption in our context. Indeed, the basic case that we want to recover is that where μ is uniformly distributed on the compact set G. In that case, the density f_μ may not exist at all, and if it exists, it is not regular at the boundary of G, so that the upper bound $W_2^2(\mu \star K_H, \mu) \leq B(H)$ is always better than (10). Note also that the upper bound $W_2^2(\mu \star K_H, \mu) \leq B(H)$ is in fact an equality if μ is a Dirac measure, and hence it cannot be improved without additional assumptions on μ.

Proof of Lemma 1. Let $\mu - \nu = \pi_+ - \pi_-$ be the Hahn-Jordan decomposition of $\mu - \nu$. From the proof of Theorem 2.6.1 of [26], we know that

$$W_2^2(\mu, \nu) = W_2^2(\pi_+, \pi_-) \,.$$

By the triangle inequality, for any $a \in \mathbb{R}^d$,

$$W_2^2(\pi_+, \pi_-) \leq 2W_2^2(\pi_+(\mathbb{R}^d)\delta_a, \pi_+) + 2W_2^2(\pi_-(\mathbb{R}^d)\delta_a, \pi_-) \,.$$

Now

$$W_2^2(\pi_+(\mathbb{R}^d)\delta_a, \pi_+) = \int_{\mathbb{R}^d} \|x - a\|^2 \pi_+(dx)$$

and

$$W_2^2(\pi_-(\mathbb{R}^d)\delta_a, \pi_-) = \int_{\mathbb{R}^d} \|x - a\|^2 \pi_-(dx) \,.$$

Finally,

$$W_2^2(\mu, \nu) = W_2^2(\pi_+, \pi_-) \leq 2 \int \|x - a\|^2 (\pi_+ + \pi_-)(dx) \,,$$

and the result follows.□

We continue the proof of Proposition 1. Applying Lemma 1, we have successively

$$\begin{aligned} & W_2^2(\hat{\mu}_n, \mu \star K_H) \leq 2 \int_{\mathbb{R}^d} \|x\|^2 |\alpha_n \hat{f}_n^+(x) - \mathbb{E}(\hat{f}_n(x))| dx \\ \leq \; & 2\alpha_n \int_{\mathbb{R}^d} \|x\|^2 |\hat{f}_n^+(x) - \mathbb{E}(\hat{f}_n(x))| dx + 2(1 - \alpha_n) \int_{\mathbb{R}^d} \|x\|^2 \mathbb{E}(\hat{f}_n(x)) dx \\ \leq \; & 2 \int_{\mathbb{R}^d} \|x\|^2 |f_n(x) - \mathbb{E}(\hat{f}_n(x))| + 2(1 - \alpha_n) \int_{\mathbb{R}^d} \|x\|^2 \mathbb{E}(\hat{f}_n(x)) dx \,. \end{aligned} \tag{11}$$

Note that

$$\int_{\mathbb{R}^d} \|x\|^2 \mathbb{E}(\hat{f}_n(x))dx \leq 2C(H) \quad \text{and} \quad (1-\alpha_n) \leq \int_{\mathbb{R}^d} (\hat{f}_n^+(x) - \hat{f}_n(x))dx\,,$$

and consequently

$$\begin{aligned}\mathbb{E}\Big((1-\alpha_n)\int_{\mathbb{R}^d} \|x\|^2 \mathbb{E}(\hat{f}_n(x))dx\Big) &\leq 2C(H)\mathbb{E}\Big(\int_{\mathbb{R}^d} (\hat{f}_n^+(x) - \mathbb{E}(\hat{f}_n(x)))dx\Big) \\ &\leq 2C(H)\mathbb{E}\Big(\int_{\mathbb{R}^d} |\hat{f}_n(x) - \mathbb{E}(\hat{f}_n(x))|dx\Big)\,. \quad (12)\end{aligned}$$

Since $\mathbb{E}(|\hat{f}_n(x) - \mathbb{E}(\hat{f}_n(x))|) \leq (\mathrm{Var}(\hat{f}_n(x)))^{1/2}$, Proposition 1 follows from (7), (8), (11) and (12). □

2.4 Errors with independent coordinates

In this section, we assume that the random variables $(\varepsilon_{1,j})_{1\leq j\leq d}$ are independent, which means that ε_1 has the distribution $\mu_\varepsilon = \mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_d$.

In this context, we shall use the kernel

$$K = k^{\otimes n}, \quad \text{where} \quad k(x) = \frac{3}{8\pi}\left(\frac{4\sin(x/4)}{x}\right)^4. \quad (13)$$

Note that $k^*(x) = 3g(4|t|)/16$, with

$$g(t) = \left(\frac{t^3}{2} - 2t^2 + \frac{16}{3}\right)\mathbf{1}_{[0,2[}(t) + \left(\frac{-t^3}{6} + 2t^2 - 8t + \frac{32}{3}\right)\mathbf{1}_{[2,4[}(t)\,.$$

The kernel K is a symmetric density, and K^* is supported over $[-1,1]^d$. Moreover, since $t \to g(|t|)$ is two times differentiable with Lipschitz second derivative, the kernel K satisfies all the required conditions.

We choose a diagonal matrix H with positive diagonal terms $h_1, h_2, \ldots, h_d$. The kernel $\tilde{K}_H$ defined in (6) is given by

$$\tilde{K}_H = \tilde{k}_{1,h_1} \otimes \tilde{k}_{2,h_2} \otimes \cdots \otimes \tilde{k}_{d,h_d} \quad \text{where} \quad \tilde{k}_{j,h_j}(x) = \frac{1}{2\pi}\int e^{iux}\frac{k^*(u)}{\mu_j^*(u/h_j)}du\,.$$

The preliminary estimator $\hat{f}_n$ defined in (5) is then

$$\hat{f}_n(x_1,\ldots,x_d) = \frac{1}{n}\sum_{i=1}^n \prod_{j=1\ldots d} \frac{1}{h_j}\tilde{k}_{j,h_j}\Big(\frac{x_j - Y_{i,j}}{h_j}\Big)\,, \quad (14)$$

and the estimator $\hat{\mu}_n$ of μ is deduced from $\hat{f}_n$ as in Section 2.2.

Note that $B(H) = \beta(h_1^2 + \cdots + h_d^2)$, with $\beta = \int u^2 k(u)du$. To ensure the consistency of the estimator, the bias term $B(H)$ has to tend to zero as n tends to infinity. Without loss of generality, we assume in the following that H is such that $B(H) \leq 1$. Hence, the variance term

$$V_n = 4\int_{\mathbb{R}^d} (2C(H) + \|x\|^2)\sqrt{\mathrm{Var}(\hat{f}_n(x))}dx$$

in Proposition 1 is such that

$$V_n \leq C \int_{\mathbb{R}^d} \Big(1 + \sum_{i=1}^d x_i^2\Big) \sqrt{\mathrm{Var}(\hat{f}_n(x_1, \ldots, x_n))}\, dx_1 \ldots dx_d$$

for some positive constant C that only depends on μ via the quantity $M = \sup_{1 \leq i \leq d} |X_{1,i}|_\infty$ where $|\cdot|_\infty$ is the essential-supremum norm. Now

$$\sqrt{\mathrm{Var}(\hat{f}_n(x_1, \ldots, x_n))} \leq \frac{1}{\sqrt{n}} \sqrt{\mathbb{E}\Big(\Big(\prod_{i=1}^d \frac{1}{h_i} \tilde{k}_{i,h_i}\Big(\frac{x_i - Y_{1,i}}{h_i}\Big)\Big)^2\Big)}\,.$$

Applying Cauchy-Schwarz's inequality d-times, we obtain that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \sqrt{\mathrm{Var}(\hat{f}_n(x_1, \ldots, x_n))}\, dx_1 \ldots dx_d \\
\leq\ & \frac{D_1}{\sqrt{n}} \sqrt{\mathbb{E}\Big(\prod_{i=1}^d \int (1 \vee x_i^2)\Big(\frac{1}{h_i} \tilde{k}_{i,h_i}\Big(\frac{x_i - Y_{1,i}}{h_i}\Big)\Big)^2 dx_i\Big)} \\
\leq\ & \frac{D_2}{\sqrt{n}} \sqrt{\mathbb{E}\Big(\prod_{i=1}^d (1 \vee Y_{1,i}^2) \int (1 + u_i^2 h_i^2) \frac{1}{h_i} (\tilde{k}_{i,h_i}(u))^2 du_i\Big)}
\end{aligned}$$

where D_1 and D_2 are positive constants depending on d. Now, $Y_{1,i}^2 \leq 2(M^2 + \varepsilon_{1,i}^2)$ and using the independence of the coordinates of ε_1, we obtain that

$$\int_{\mathbb{R}^d} \sqrt{\mathrm{Var}(\hat{f}_n(x))}\, dx \leq \frac{D_3}{\sqrt{n}} \sqrt{\Big(\prod_{i=1}^d (M^2 + \mathbb{E}(\varepsilon_{1,i}^2)) \int (1 + u_i^2 h_i^2) \frac{1}{h_i} (\tilde{k}_{i,h_i}(u))^2 du_i\Big)}\,, \tag{15}$$

It follows that

$$\int_{\mathbb{R}^d} \sqrt{\mathrm{Var}(\hat{f}_n(x))}\, dx \leq \frac{A_0}{\sqrt{n}} \sqrt{\prod_{i=1}^d \int (1 + u_i^2 h_i^2) \frac{1}{h_i} (\tilde{k}_{i,h_i}(u))^2 du_i}\,. \tag{16}$$

In the same way, we have that

$$\begin{aligned}
\int_{\mathbb{R}^d} x_k^2 \sqrt{\mathrm{Var}(\hat{f}_n(x))}\, dx \ \leq\ & \frac{A_k}{\sqrt{n}} \sqrt{(M^6 + \mathbb{E}(\varepsilon_{1,i}^6)) \int (1 + u_k^6 h_k^6) \frac{1}{h_k} (\tilde{k}_{k,h_k}(u))^2 du_k} \\
& \times \sqrt{\prod_{i \neq k} (M + \mathbb{E}(\varepsilon_{1,i}^2)) \int (1 + u_i^2 h_i^2) \frac{1}{h_i} (\tilde{k}_{i,h_i}(u))^2 du_i}. \quad (17)
\end{aligned}$$

Note that $\mathbb{E}(\varepsilon_{1,i}^2)$ and $\mathbb{E}(\varepsilon_{1,i}^6)$ are finished according to (3). Starting from these computations, one can prove the following Proposition.

Proposition 2. *Let $r_i(x) = 1/\mu_i^*(x)$, and let $(h_1, \ldots, h_d) \in [0,1]^d$. The following upper bound holds*

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq 2\beta(h_1^2 + \cdots + h_d^2) + \frac{L}{\sqrt{n}} \left(\prod_{i=1}^d I_i(h_i) + \sum_{k=1}^d J_k(h_k)\Big(\prod_{i=1, i \neq k}^d I_i(h_i)\Big)\right)$$

where L is some positive constant depending on d, M and $(\mathbb{E}(\varepsilon_{1,i}^2), \mathbb{E}(\varepsilon_{1,i}^6))_{1\leq i\leq d}$, and

$$\begin{aligned} I_i(h) &\leq \sqrt{\int_{-1/h}^{1/h} (r_i(u))^2 + (r_i'(u))^2 du}\,, \\ J_i(h) &\leq \sqrt{\int_{-1/h}^{1/h} (r_i(u))^2 + (r_i'''(u))^2 du} \\ &\quad + h\sqrt{\int_{-1/h}^{1/h} (r_i''(u))^2 du} + h^2\sqrt{\int_{-1/h}^{1/h} (r_i'(u))^2 du}\,. \end{aligned}$$

Remark 3. *Note that the upper bound in Proposition 2 depends on the unknown distribution μ only through the constant M (which appears in (15) and (17)). Hence the rate of convergence of $\hat{\mu}$ obtained from Proposition 2 does not depend on μ. Note that in the classical context of $\mathbb{L}_2$-density deconvolution, the variance term is exactly*

$$\int_{\mathbb{R}^d} \mathrm{Var}(\hat{f}_n(x))dx\,,$$

which can be bounded independently of μ. This leads to the idea that $\mathbb{E}(W_2^2(\hat{\mu}_n, \mu))$ depends very poorly of the unknown distribution μ. If this intuition is correct, a possible way to select the bandwidth parameter $\mathbf{h} = (h_1, \ldots, h_d)$ is to choose by simulation the best possible $\mathbf{h}$ in the simple case $\mu = \delta_0$. This will be done in Section 3. We shall see that this selected $\mathbf{h}$ leads to very good results for different choices of μ, even when μ has a density (see Section 3.2).

Proof of Proposition 2. By Plancherel's identity,

$$\begin{aligned} \int \frac{1}{h}(\tilde{k}_{i,h}(u))^2 du = \frac{1}{2\pi}\int \frac{1}{h}\frac{(k^*(u))^2}{(\mu_i^*(u/h))^2}du &= \frac{1}{2\pi}\int \frac{(k^*(hu))^2}{(\mu_i^*(u))^2}du \\ &\leq \frac{1}{2\pi}\int_{-1/h}^{1/h} r_i^2(u)du\,. \end{aligned}$$

the last upper bound being true because k^* is supported over $[-1,1]$ and bounded by 1.

Let C be a positive constant, which may vary from line to line. Let $q_{i,h}(u) = r_i(u/h)k^*(u)$. Since $q_{i,h}$ is differentiable with compactly supported derivative, we have that

$$-iu\tilde{k}_{i,h}(u) = (q_{i,h}')^*(u)\,.$$

Applying Plancherel's identity again,

$$\begin{aligned} \int hu^2(\tilde{k}_{i,h}(u))^2 du &= \frac{1}{2\pi}\int h(q_{i,h}'(u))^2 du \\ &\leq C\ \Big(\int_{-1/h}^{1/h}(r_i'(u))^2 du + h^2\int_{-1/h}^{1/h} r_i^2(u)du\Big)\,, \end{aligned}$$

the last inequality being true because k^* and $(k^*)'$ are compactly supported over $[-1,1]$. Consequently

$$\sqrt{\int (1+u_i^2h_i^2)\frac{1}{h_i}(\tilde{k}_{i,h_i}(u))^2 du_i} \leq CI_i(h_i)\,.$$

In the same way

$$-iu^3\tilde{k}_{i,h}(u) = (q'''_{i,h})^*(u) \quad \text{and} \quad \int h^5 u^6 (\tilde{k}_{i,h}(u))^2 du = \frac{1}{2\pi}\int h^5 (q'''_{i,h}(u))^2 du\,,$$

Now, since $k^*, (k^*)', (k^*)''$ and $(k^*)'''$ are compactly supported over $[-1,1]$,

$$\begin{aligned}\int h^5 (q'''_{i,h}(u))^2 du &\leq C\Big(\int_{-1/h}^{1/h} (r_i'''(u))^2 du + h^2\int_{-1/h}^{1/h}(r_i''(u))^2 du \\ &\quad + h^4\int_{-1/h}^{1/h}(r_i'(u))^2 du + h^6\int_{-1/h}^{1/h}(r_i(u))^2 du\Big).\end{aligned}$$

Consequently

$$\sqrt{\int (1+u_k^6 h_k^6)\frac{1}{h_k}(\tilde{k}_{k,h_k}(u))^2 du_k} \leq C J_k(h_k)\,.$$

The results follows. □

2.5 Linear transform of errors with independent coordinates

In this section, we assume that $\varepsilon = A\eta$, where the distribution μ_η of η is such that $\mu_\eta = \mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_d$, and A is some known invertible matrix. Applying A^{-1} to the random variables Y_i in (2), we obtain the new model

$$A^{-1}Y_i = A^{-1}X_i + \eta_i\,,$$

that is: a convolution model in which the error has independent coordinates.

To estimate the image measure $\mu^{A^{-1}}$ of μ by A^{-1}, we use the preliminary estimator of Section 2.4, that is

$$\hat{f}_{n,A^{-1}}(x_1,\ldots,x_d) = \frac{1}{n}\sum_{i=1}^n \prod_{j=1\ldots d}\frac{1}{h_j}\tilde{k}_{j,h_j}\Big(\frac{x_j - (A^{-1}Y_i)_j}{h_j}\Big)\,,$$

and the estimator $\hat{\mu}_{n,A^{-1}}$ of $\mu^{A^{-1}}$ is deduced from $\hat{f}_{n,A^{-1}}$ as in Section 2.2. This estimator $\hat{\mu}_{n,A^{-1}}$ has the density $\hat{g}_{n,A^{-1}}$ with respect to the Lebesgue measure.

To estimate μ, we define $\hat{\mu}_n = \hat{\mu}^A_{n,A^{-1}}$ as the image measure of $\hat{\mu}_{n,A^{-1}}$ by A. This estimator has the density $\hat{g}_n = |A|^{-1}\hat{g}_{n,A^{-1}} \circ A^{-1}$ with respect to the Lebesgue measure. It can be deduced from the preliminary estimator $\hat{f}_n = |A|^{-1}\hat{f}_{n,A^{-1}} \circ A^{-1}$ as in Section 2.2. Now

$$\begin{aligned}W_2^2(\hat{\mu}_n,\mu) &= \min_{\lambda\in\Pi(\hat{\mu}_n,\mu)}\int \|x-y\|^2\lambda(dx,dy) \\ &= \min_{\pi\in\Pi(\hat{\mu}_{n,A^{-1}},\mu^{A^{-1}})}\int \|A(x-y)\|^2\pi(dx,dy)\,.\end{aligned}$$

Consequently, if $\|A\| = \sup_{\|x\|=1}\|Ax\|$, we obtain that

$$W_2^2(\hat{\mu}_n,\mu) \leq \|A\| W_2^2(\hat{\mu}_{n,A^{-1}},\mu^{A^{-1}})\,,$$

which is an equality if A is an unitary matrix. Hence the upper bound given in Proposition 2 for the quantity $\mathbb{E}(W_2^2(\hat{\mu}_{n,A^{-1}},\mu^{A^{-1}}))$ is also valid for $\mathbb{E}(W_2^2(\hat{\mu}_n,\mu))$.

Note that $\hat{f}_n$ can be written as in (5), with the kernel $K = |A|^{-1}k^{\otimes n}\circ A^{-1}$ and the diagonal matrix H with diagonal terms $h_1,\ldots,h_d$.

2.6 Examples of rates of convergence

In this section we shall always assume that $\mu_\varepsilon = \mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_d$. According to the comments of Section 2.5, the rates of convergence are also valid for any linear invertible transform of such noises.

Case 1: no noise. In that case $\mu_1^* = \mu_2^* = \ldots = \mu_d^* = 1$. Taking $h_1 = h_2 = \cdots = h_d = h$, Proposition 2 gives the upper bound

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq C\Big(h^2 + \frac{1}{\sqrt{nh^d}}\Big).$$

Taking $h = n^{-1/(d+4)}$, we obtain the rate of convergence

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{n^{2/(d+4)}}\,.$$

Note that this is the same rate as that obtained by [18] for the empirical measure μ_n defined in (4). To our knowledge, this rate of convergence has not been improved without making additional assumptions on μ. [25] proved some upper and lower bounds (Theorem 11.1.6) for $\mathbb{E}(W_2^2(\mu_n, \mu))$ under entropy conditions on μ. It follows from his estimates that if μ has a density and d is even, then the rate $\mathbb{E}(W_2^2(\mu_n, \mu)) \leq Cn^{-2/d}$ is optimal.

Case 2: convolution of Laplace noise. We consider the case where

$$\mu_i^*(u) = \frac{1}{(1+u^2)^{k_i}}\,,$$

k_i being nonnegative integers. This corresponds to the case where the density of $\varepsilon_{1,i}$ is the k_i-times convolution of the Laplace density.

Proposition 2 gives the upper bound

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq C\Big(h_1^2 + \cdots + h_d^2 + \frac{1}{\sqrt{n}} \prod_{i=1}^{d} h_i^{-(4k_i+1)/2}\Big).$$

This bound is similar to the $\mathbb{L}_2$-risk bound obtained by [6] in the context of multivariate density deconvolution with an ordinary smooth noise (see Section 3.2.1 in their paper). Their computations show that one can take $h_i = n^{-1/(d+4(1+k_1+\cdots+k_d))}$ and then obtain the rate of convergence

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{n^{2/(d+4(1+k_1+\cdots+k_d))}}\,.$$

In particular:

- If $k_i = 0$ for all i (no noise), we obtain the same rate as previously.
- If $k_i = 1$ for all i (isotropic Laplace noise), we obtain the rate

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{n^{2/(5d+4)}}\,.$$

- If $k_\ell = 1$ and $k_i = 0$ for $i \neq \ell$ (Laplace noise in one direction), we obtain the rate

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{n^{2/(d+8)}}\,.$$

The index $I(d) = k_1 + \cdots + k_d$ can be seen as an index of global regularity of the error distribution, summing all the regularities k_i of the marginal distributions. As usual in deconvolution problems, the worst rates of convergence are obtained for very regular error distributions: more precisely, the rate of convergence becomes slower as $I(d)$ increases. Note also that a single marginal distribution with regularity N gives the same rates as N marginal distributions with regularity 1.

Case 3: isotropic Gaussian noise. We consider the case where

$$\mu_1^*(u) = \mu_2^*(u) = \ldots = \mu_d^*(u) = \exp(-u^2/2).$$

Proposition 2 gives the upper bound

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq C\Big(h_1^2 + \cdots + h_d^2 + \frac{1}{\sqrt{n}} \prod_{i=1}^{d} h_i^{-5/2} \exp(h_i^{-2}/2)\Big).$$

Following again [6], one can take $h_i = \sqrt{2/\log(n)}$, and we obtain the rate of convergence

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{\log(n)}\,.$$

Case 4: Gaussian noise in one direction. We consider the case where $\mu_1^*(u) = \exp(-u^2/2)$, and $\mu_2^* = \cdots = \mu_d^* = 1$. Taking $h_2 = h_3 = \cdots = h_d = h$, Proposition 2 gives the upper bound

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq C\Big(h_1^2 + h^2 + \frac{1}{\sqrt{n h^{d-1} h_1^5}} \exp(h_1^{-2}/2)\Big).$$

Taking $h_1 = \sqrt{2/\log(n)}$ and $h = n^{-1/(5d-5)}$, we obtain the rate of convergence

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{\log(n)}\,.$$

Hence, a Gaussian noise in one single direction gives the same rate of convergence as an isotropic Gaussian noise. This is coherent with the discussion in Section 3.2.2 of [6] about density deconvolution in $\mathbb{R}^d$.

3 Experiments

In this section, we take $d = 2$ and we consider the case where $\mu_\varepsilon = \mu_1 \otimes \mu_2$. For all the following experiments the preliminary estimator $\hat{f}_n$ is defined as in (14) with the bandwidth parameter $\mathbf{h} = (h_1, h_2)$ and the kernel

$$K = k^{\otimes n}, \quad \text{where} \quad k(x) = \frac{3}{16\pi} \left(\frac{8\sin(x/8)}{x}\right)^4 . \tag{18}$$

The only difference with the kernel given in (13) is that K is now supported over $[-1/2, 1/2]^d$. The estimator $\hat{\mu}_n$ of μ is then deduced from $\hat{f}_n$ as in Section 2.2.

In practice, the deconvolution estimator $\hat{\mu}_n = \hat{\mu}_{n,\mathbf{h}}$ is only computed on a finite set of locations. Let $\mathcal{P} = \{p\}$ be a finite regular grid of points in $\mathbb{R}^2$, a discrete version $\tilde{\mu}_n = \tilde{\mu}_{n,\mathbf{h}}$ of $\hat{\mu}_{n,\mathbf{h}}$ is defined by

$$\tilde{\mu}_{n,\mathbf{h}} = \sum_{p\in\mathcal{P}} \hat{\alpha}_p(\mathbf{h})\delta_p$$

where

$$\hat{\alpha}_p(\mathbf{h}) = \frac{\hat{f}_n^+(p)}{\sum_{p\in\mathcal{P}} \hat{f}_n^+(p)}.$$

Note that the W_2 distance between $\tilde{\mu}_{n,\mathbf{h}}$ and $\hat{\mu}_{n,\mathbf{h}}$ tends to zero as the grid resolution tends to zero. In the following, it is assumed that the grid resolution is chosen small enough, namely it is assumed that

$$W_2^2(\tilde{\mu}_n, \hat{\mu}_n) \ll W_2^2(\mu, \hat{\mu}_n).$$

3.1 Dirac experiment and bandwidth selection.

One situation for which the Wasserstein distance W_2 is computable is the case where μ is a Dirac measure. Obviously this framework has no interest in practice but it allows us to validate the results proved in Section 2. As we shall see, it is also a way to select a bandwidth which will be a reasonable candidate in a general context.

Let $\mu = \delta_0$, which corresponds to the case where $Y_i = \varepsilon_i$ in the convolution model (2). Assume that $\mu_1^*(u) = \mu_2^*(u) = (1+u^2)^{-1}$, which means that $\varepsilon_{1,1}$ and $\varepsilon_{1,2}$ have a standard Laplace distribution with variance 2. For this Laplace isotropic noise, we choose $\mathbf{h} = (h, h)$.

For the empirical measure μ_n defined in (4), one has

$$\mathbb{E}(W_2^2(\mu_n, \delta_0)) = \mathbb{E}\Big(\int_{\mathbb{R}^2} \|x\|^2 \mu_n(x)dx\Big) = \mathrm{Var}(\varepsilon_{1,1}) + \mathrm{Var}(\varepsilon_{1,2}) = 4\,.$$

For $\tilde{\mu}_n$, one has

$$\mathbb{E}(W_2^2(\tilde{\mu}_n, \delta_0)) = \mathbb{E}\Big(\sum_{p\in\mathcal{P}} \|p\|^2 \hat{\alpha}_p(\mathbf{h})\Big)\,.$$

Let $I_n(h) = W_2^2(\tilde{\mu}_n, \delta_0)$ be the Wasserstein distance between δ_0 and $\tilde{\mu}_n$. For a given h in a grid $\mathcal{H}$ of possible bandwidths, $\mathbb{E}(I_n(h))$ can be approximated with an elementary Monte Carlo method by repeating the simulation N_s times. Figure 1 shows the boxplot of the distribution of $I_n(h)$ on a rough grid of bandwidth with $n = 20000$. For such a sample size, the deconvolution estimator $\tilde{\mu}_n$ performs better than the empirical measure on a large scale of bandwidth values.

For each n, an approximation of $h_* = \mathrm{argmin}\mathbb{E}(W^2(\hat{\mu}_{n,h}, \delta_0))$ can be computed as follows

$$\hat{h}_*(n) = \mathrm{argmin}_{h\in\mathcal{H}} \bar{I}_n(h) \quad \text{where} \quad \bar{I}_n(h) = \frac{1}{N_s}\sum_{s=1}^{N_s} I_{n,s}(h)\,.$$

and $I_{n,s}(h)$ is the computation of $I_n(h)$ corresponding to the s-th simulation. Table 1 gives the value of $\hat{h}_*(n)$ computed for different sample sizes and the

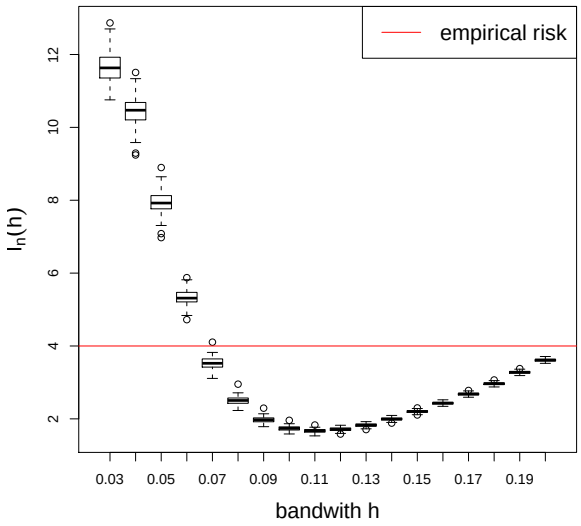

Figure 1: Boxplots of $I_n(h)$ for different bandwidth h. These results correspond to $N_s = 100$ computations of the deconvolution estimator based on samples of size $n = 20000$.

n	30	100	500	1000
$\hat{h}_*$	0.172	0.159	0.143	0.137
$\bar{I}_n(\hat{h}_*)$	4.7 ∓ 0.2	3.8 ∓ 0.2	3.14 ∓ 0.05	2.78 ∓ 0.05
n	5000	7500	10000	20000
$\hat{h}_*$	0.123	0.119	0.117	0.111
$\bar{I}_n(\hat{h}_*)$	2.12 ∓ 0.02	1.98 ∓ 0.02	1.87 ∓ 0.02	1.67 ∓ 0.01

Table 1: Estimations of $\hat{h}_*$ and estimated risks for several values of the sample size n. These results have been computed thanks to $N_s = 100$ computations of the deconvolution estimator.

corresponding estimation $\bar{I}_n(\hat{h}_*)$ of $\mathbb{E}(I_n(h_*))$. For $n = 7500$, $\bar{I}_n(\hat{h}_*)$ is about one half of $\mathbb{E}(W^2(\mu_n, \delta_0))$.

Figure 2 shows a linear relation between $\log \hat{h}_*$ and $\log n$. A linear regression leads to an estimation of the slope of $-0.067 = -1/14.9$, which is close (a little larger) to the theoretical slope: $-1/14$ (see Section 2.6, Case 2: isotropic Laplace noise with $d = 2$).

As pointed out in Remark 3 of Section 2.4, it seems that h_* does not strongly depend on the geometric shape G. Hence, the bandwidth $\hat{h}_*$ computed for the Dirac measure should be a reasonable bandwidth for estimating other distributions μ when the error distribution is an isotropic Laplace noise. This intuition is confirmed *via* the simulations presented in the next section.

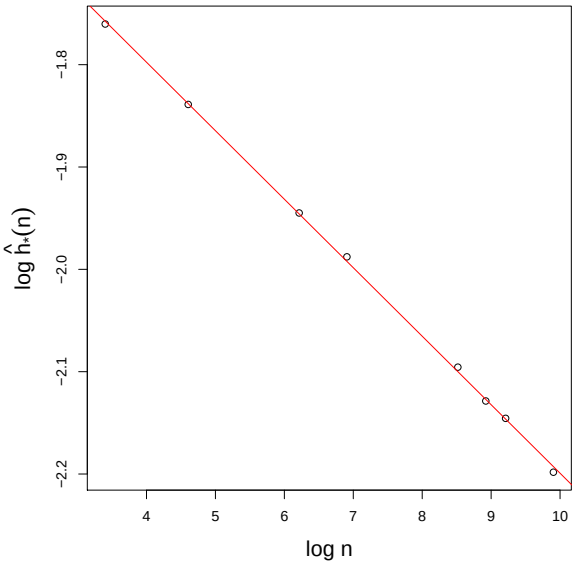

Figure 2: Estimation of the bandwidths $\hat{h}_*(n)$ (left) against the sample size in logarithm scales. The estimated slope for the regression of $\log \hat{h}_*(n)$ by $\log n$ is $-0.067 \approx -1/14.9$.

3.2 Geometric inference

This section illustrates with some simulations how to take advantage from the estimator $\hat{\mu}_n$ and its consistency properties for geometric inference purposes. As already explained in the introduction, the geometry of the unknown object G can be inferred thanks to the levels of the distance function to a measure d_{ν,m_0} defined by (1) if ν is close enough to μ for the Wasserstein metric. The following simulations compare the geometry recovered from the distance d_{μ_n,m_0} to the empirical measure as in [4], and the distance $\mathrm{d}_{\hat{\mu}_n,m_0}$ to the deconvolution estimator $\hat{\mu}_n$. The scale parameter m_0 is fixed to $m_0 = 0.01$ for all the computations of the section. Hence we shall note d_ν for d_{ν,m_0} in the sequel.

Three disks and Laplace noise

For this first example, we consider the geometric shape in $\mathbb{R}^2$ composed of three disks of radius one whose centers are at a distance $\frac{5}{2}$ of each other. A total set of 20000 points is sampled uniformly on these disks and observed with an isotropic Laplace noise, as in Section 3.1. Figure 3 allows us to compare the distance function to the empirical measure μ_n and the distance function to the estimator $\tilde{\mu}_n$ deduced from the deconvolution estimator. For the bandwidth, we take $h = \hat{h}_* = 0.11$, where $\hat{h}_*$ has been computed in Section 3.1 for the Dirac measure (see Table 1).

The deconvolution allows us to enlarge the numbers of levels which recovers the three disks : only the levels of d_{μ_n} between 0.29 and 0.5 have the correct topology whereas the levels between 0.16 and 0.57 are valid for $\mathrm{d}_{\tilde{\mu}_n}$. Further-

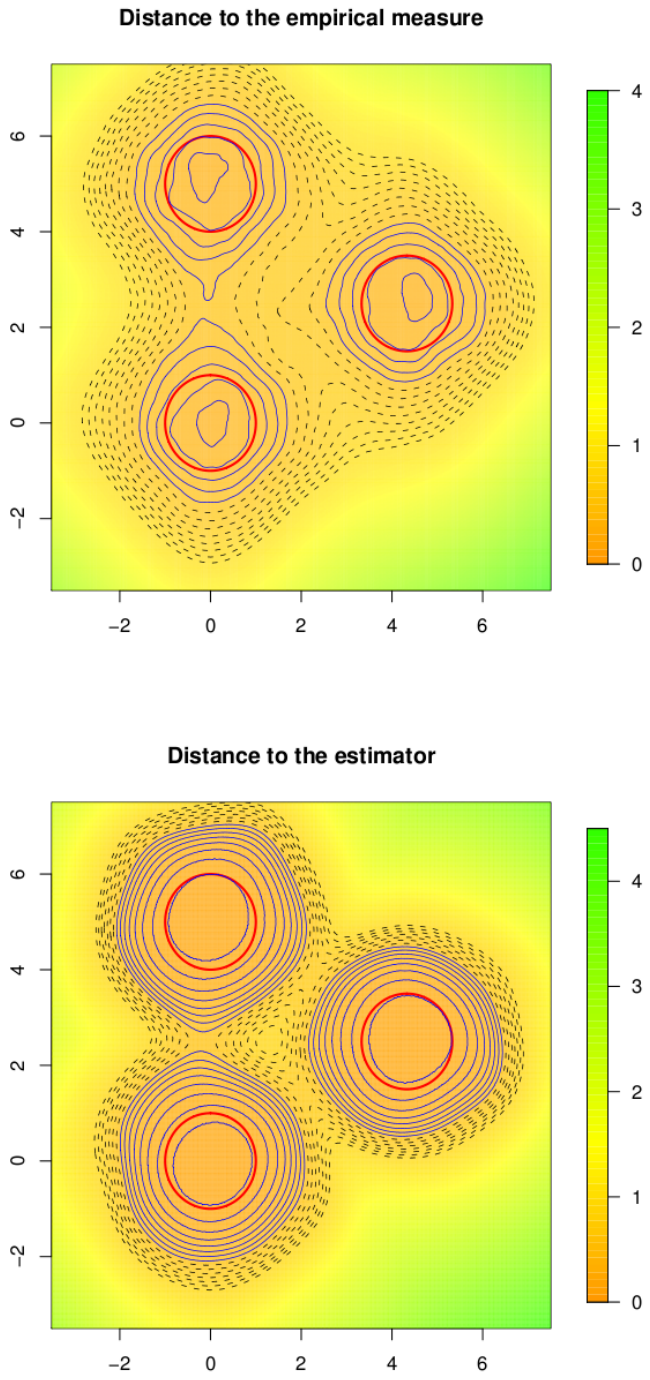

Figure 3: Distance d_{μ_n} to the empirical measure and distance $\mathrm{d}_{\tilde{\mu}_n(0.11)}$ to the estimator for the three disks experiment with Laplace noise. The three circles delimiting the disks are drawn in red and the levels of the distance function which have the correct topology are drawn in blue. The other levels are the black dashed lines. The same grid of levels is used on the two pictures.

more, by drawing and comparing the levels of $\mathrm{d}_{\tilde{\mu}_n(h)}$ for different bandwidth h, it can be checked that $h = 0.11$ is around the optimal topological bandwidth, namely it corresponds to the larger scale of levels of correct topology.

Two circles and Laplace noise

The geometric shape of this second experiment is composed of two circles of radius 4 and 7. A total set of 20000 points is sampled uniformly on these two circles and the sample is observed with an isotropic Laplace noise, as in Section 3.1. The benefit of using a deconvolution estimator is obvious in this context, since no levels of d_{μ_n} can reach the correct topology, whereas the levels of $\mathrm{d}_{\tilde{\mu}_n}$ between 0.56 and 0.63 give the correct topology, see Figure 4. The bandwidth used here is again $h = \hat{h}_* = 0.11$, as calibrated in Section 3.1.

One Gaussian example

As explained in Section 2.6 a Gaussian noise will give a logarithmic rate of convergence of the deconvolution estimator $\hat{\mu}_n$: this makes the application more difficult in this framework. Anyway, a Gaussian example is proposed here, but we use a large sample to be able to observe the topological effects.

The geometric shape to be recovered is composed of two embedded closed filaments. One set of $n = 100000$ points are uniformly sampled on these two filaments and this sample is observed with a standard isotropic Gaussian noise, which means that $\varepsilon_{1,1}$ and $\varepsilon_{1,2}$ have a standard normal distribution. The two filaments are drawn on the two pictures of Figure 5. No one of the drawn levels of d_{μ_n} recovers the correct topology. In fact it can be checked by drawing the contour plot with a thiner resolution that only the levels between 1.04 and 1.06 have the correct topology. We use a bandwidth $h = 0.12$ for our deconvolution estimator. A larger scale of levels of $\mathrm{d}_{\tilde{\mu}_n}$ between 0.72 and 0.91 allows us to recover the correct topology.

One directional measurement error

For this example, we take $\mu_1 = \delta_0$ and $\mu_2^*(u) = (1+u^2)^{-1}$, which means that $\varepsilon_{1,1} = 0$ and that $\varepsilon_{1,2}$ has a standard Laplace distribution.

A set of 10000 points is sampled uniformly along an incomplete circle, and then observed with this one directional Laplace noise. The top picture of Figure 6 shows the sample and the incomplete circle in red, and the bottom picture shows a sample drawn according to $\tilde{\mu}_{n,\mathbf{h}}$: the hole is impossible to see on these contaminated data whereas it is whith the deconvolved measure. However, due to the oscillations of the deconvolution estimator (which is a well known drawback of these kind of estimators) a small amount of the mass appears at a large distance of the circle. Using the distance function $\mathrm{d}_{\tilde{\mu}_n,m_0}$ is then particularly appropriate there, because this distance will ignore these outliers provided m_0 is not too small. Figure 7 compares the distance to the empirical measure with the distance to the estimator: the hole can be recovered only by the levels of $\mathrm{d}_{\tilde{\mu}_n}$. The correct levels of $\mathrm{d}_{\tilde{\mu}_n}$ are between 0.31 and 0.57. The estimator $\tilde{\mu}_n$ has been computed here with the bandwidths $h_1 = 0.07$ and $h_2 = 0.25$, this choice leading to a correct inference of the geometric shape.

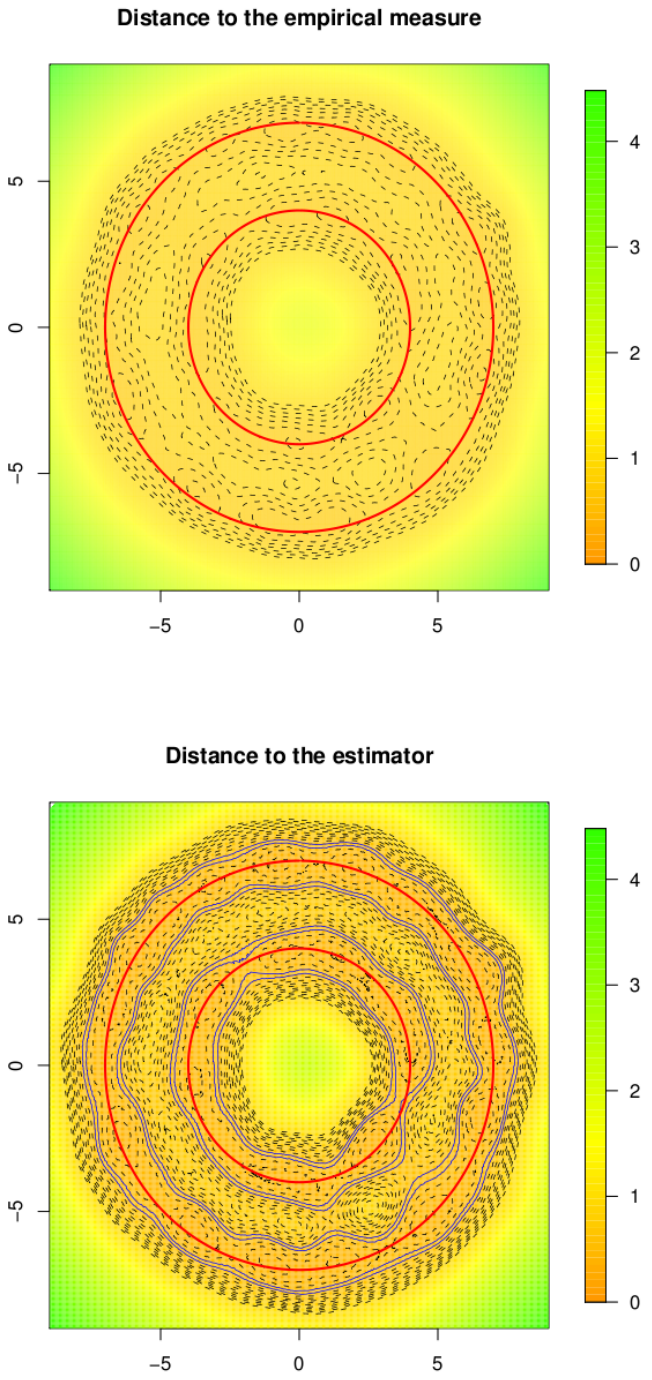

Figure 4: Distance d_{μ_n} and distance $d_{\tilde{\mu}_n(0.11)}$ for the two circles experiment with Laplace noise. See Figure 3 for more details about the legend.

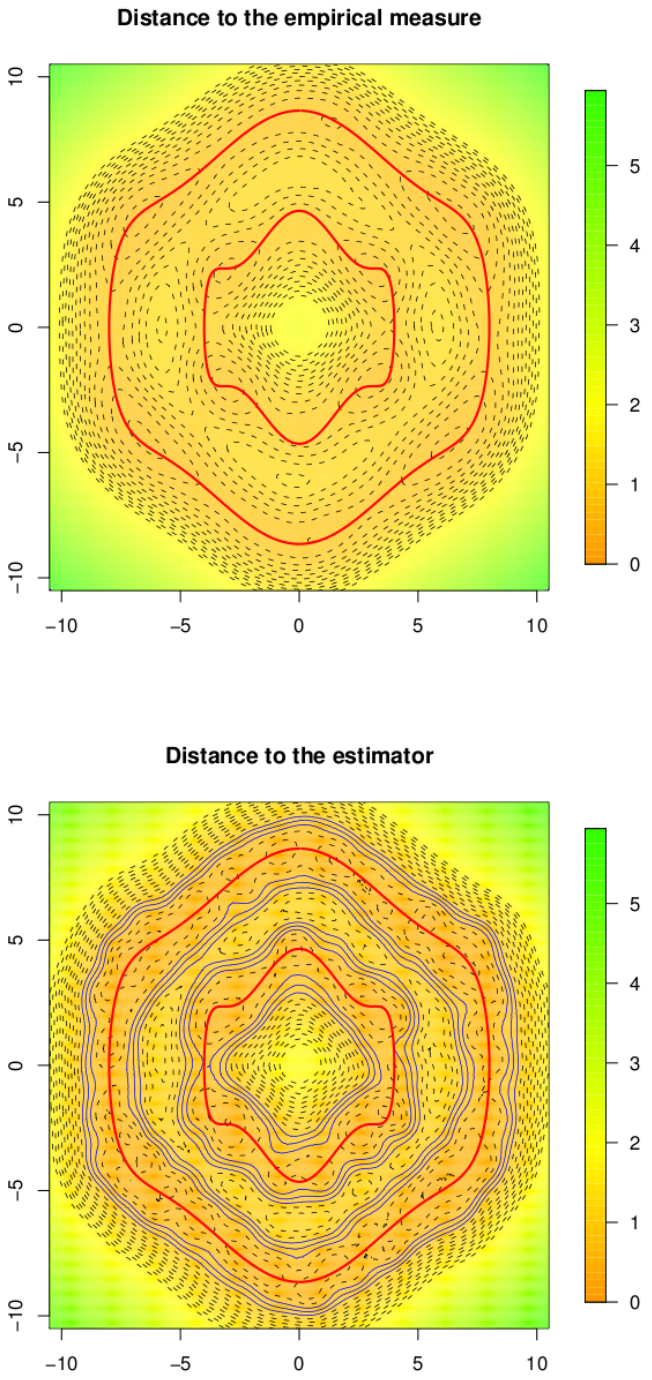

Figure 5: Distance d_{μ_n} and distance $d_{\tilde{\mu}_n(0.12)}$ for the two filaments experiment with Gaussian noise. See Figure 3 for more details about the legend.

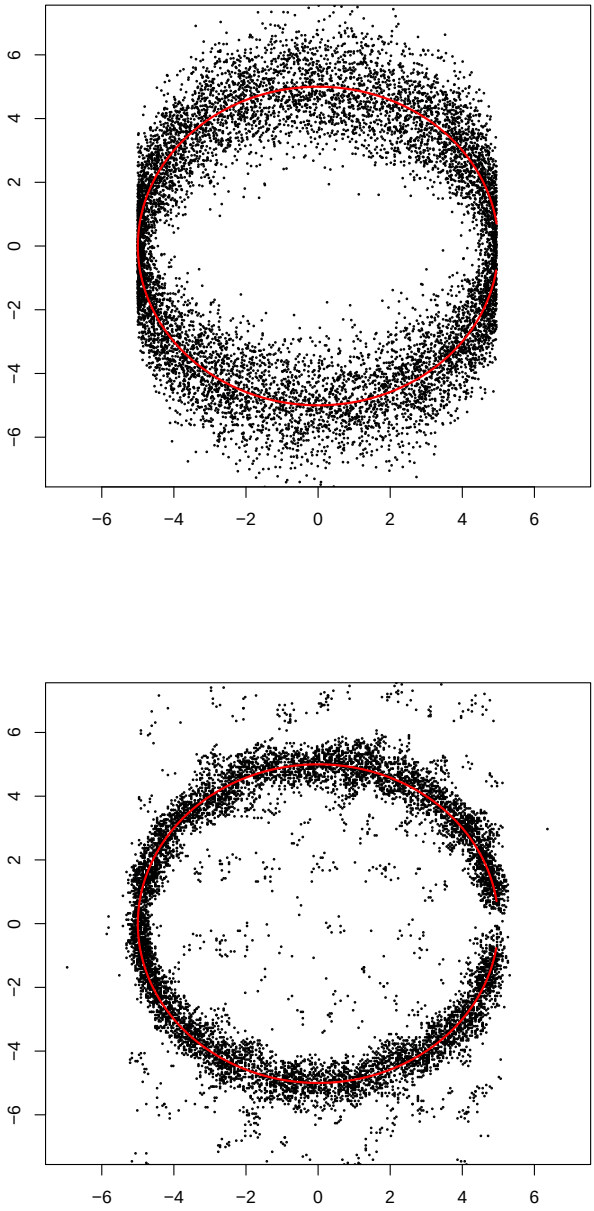

Figure 6: Circle with hole in red and 10000 points sampled on it with an unidirectional Laplace measurement error (top) and simulation of 10000 points according to $\tilde{\mu}_{n,\mathbf{h}}$ with $h_1 = 0.07$ and $h2 = 0.25$ (bottom).

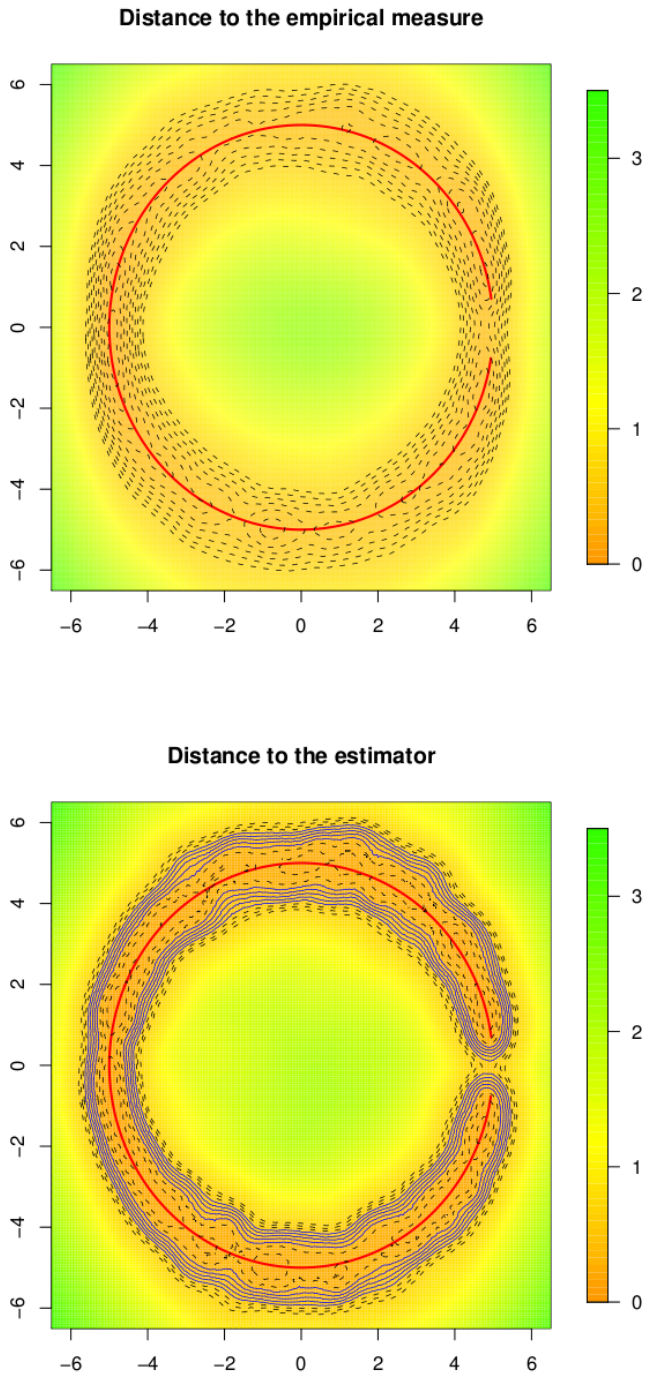

Figure 7: Distance d_{μ_n} and distance $d_{\tilde{\mu}_n}$ for the circle with hole with unidirectionnal noise. See Figure 3 for more details about the legend.

$k_1(x) := \sin(x)/x$	$k_1^*(t) = 1_{[-1,1]}$
$k_2(x) := 48(\cos x)(1-15x^{-2})(\pi x^4)^{-1}$	$k_2^*(t) = (1-t^2)^3 1_{[-1,1]}$
$k_3(x) := k(x)$, see (13)	$k_3^* = k^*$, see (13)
$k_4(x) := (2\pi)^{-1/2}\exp(-x^2/2)$	$k_4^*(t) = \exp(-x^2/2)$

Table 2: Four kernels and their Fourier transform.

A Kernel performances

The section shortly discusses the kernel choice by comparing the performances of the deconvolution estimator for the most used kernels. In the density estimation framework with $\mathbb{L}_2$ risk in dimension one, [11] compare the performances of four kernels given in Table 2.

Only k_3 fulfills all the required conditions (see Section 2.2) to prove the consistency result. Nevertheless, note that k_1^*, k_2^* have a compact support $[-1,1]$. k_2^* is also C^2 and its second derivate is Lipschitz, so this second kernel nearly fulfills the required assumptions. On the other hand k_1^* is the less regular, and concerning the Gaussian kernel, k_4^* has the required regularity but it has not a compact support.

The four kernels are compared in the simple situation in $\mathbb{R}$ for which the Wasserstein distance can be computed, namely a Dirac mass at 0. For each simulation, we consider a set of $n = 500$ independent Laplace variables of variance 2. An accurate grid $\mathcal{P}$ of points p over $\mathbb{R}$ is fixed, and for each simulation and a given bandwidth h, let $I_n(h)$ be the Wasserstein distance between δ_0 and $\tilde{\mu}_n(h)$: $I_n(h) := \sum_{p\in\mathcal{P}} x_i^2 \hat{\alpha}_i(h)$. The Wasserstein risk is then estimated by computing $\bar{I}_n(h)$ over 100 simulations of this experience.

It appears that the two kernels k_1 and k_4 have very bad performances. Even for their estimated minimal bandwidth $\hat{h}_*$, the mean risk of their estimator is over 6. This first observation tends to confirm that the kernel assumptions of Section 2.2 are not too restrictive. On the other hand, Figure 8 shows that k_2 and k_3 lead to estimators whose performance are quite similar in this context. This is not surprising since these two kernels have similar regularity properties. In spite of this observation, note that the consistency result is not proved for k_2 since it is not positive, and thus the control of the bias proposed in this paper is not valid for this kernel.

References

[1] Gérard Biau, Benoît Cadre, and Bruno Pelletier. Exact rates in density support estimation. *J. Multivariate Anal.*, 99(10):2185–2207, 2008.

[2] Raymond J. Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404):1184–1186, 1988.

[3] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean spaces. *Discrete Comput Geom*, 41:461–479, 2009.

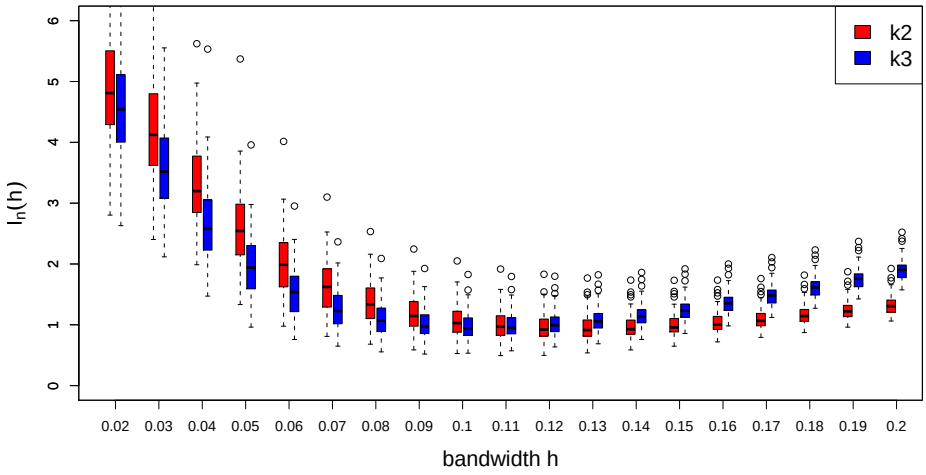

Figure 8: Comparing the performances of the deconvolution estimators defined by the two kernels k_2 and k_3.

[4] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *J. Foundations of Computational Mathematics*. to appear.

[5] F. Chazal and A. Lieutier. Smooth manifold reconstruction from noisy and non uniform approximation with guarantees. *Comp. Geom: Theory and Applications*, 40:156–170, 2008.

[6] Fabienne Comte and Claire Lacour. Anisotropic adaptive kernel deconvolution. hal-00579608, 2011.

[7] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Estimating the number of clusters. *Canad. J. Statist.*, 28(2):367–382, 2000.

[8] Antonio Cuevas and Ricardo Fraiman. Set estimation. In *New perspectives in stochastic geometry*, pages 374–397. Oxford Univ. Press, Oxford, 2010.

[9] Antonio Cuevas, Ricardo Fraiman, and Alberto Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.*, 35(3):1031–1051, 2007.

[10] A. Delaigle and I. Gijbels. Estimation of boundary and discontinuity points in deconvolution problems. *Statist. Sinica*, 16(3):773–788, 2006.

[11] Aurore Delaigle and Peter Hall. On optimal kernel choice for deconvolution. *Statist. Probab. Lett.*, 76(15):1594–1602, 2006.

[12] Luc Devroye. Consistent deconvolution in density estimation. *Canad. J. Statist.*, 17(2):235–239, 1989.

[13] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. On the path density of a gradient field. *Ann. Statist.*, 37(6A):3236–3271, 2009.

[14] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. The geometry of nonparametric filament estimation. arXiv:1003.5536v2, 2010.

[15] Peter Hall and Léopold Simar. Estimating a changepoint, boundary, or frontier in the presence of observation error. *J. Amer. Statist. Assoc.*, 97(458):523–534, 2002.

[16] John A. Hartigan. *Clustering algorithms*. John Wiley & Sons, New York-London-Sydney, 1975. Wiley Series in Probability and Mathematical Statistics.

[17] Trevor Hastie and Werner Stuetzle. Principal curves. *J. Amer. Statist. Assoc.*, 84(406):502–516, 1989.

[18] Joseph Horowitz and Rajeeva L. Karandikar. Mean rates of convergence of empirical measures in the Wasserstein metric. *J. Comput. Appl. Math.*, 55(3):261–273, 1994.

[19] V. I. Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *Ann. Statist.*, 28(2):591–629, 2000.

[20] A. Meister. *Deconvolution problems in nonparametric statistics*. Lecture Notes in Statistics 193. Springer-Verlag, 2009.

[21] Alexander Meister. Estimating the support of multivariate densities under measurement error. *J. Multivariate Anal.*, 97(8):1702–1717, 2006.

[22] Alexander Meister. Support estimation via moment estimation in presence of noise. *Statistics*, 40(3):259–275, 2006.

[23] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646–663, 2011.

[24] A. Petrunin. Semiconcave functions in Alexandrov's geometry. In *Surveys in differential geometry. Vol. XI*, pages 137–201. Int. Press, Somerville, MA, 2007.

[25] Svetlozar T. Rachev. *Probability metrics and the stability of stochastic models*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1991.

[26] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass transportation problems. Vol. II*. Probability and its Applications. Springer-Verlag, 1998.

[27] Leonard Stefanski and Raymond J. Carroll. Deconvoluting kernel density estimators. *Statistics*, 21(2):169–184, 1990.

[28] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, 2003.

[29] V. M. Zolotarev. Pseudomoments. *Teor. Verojatnost. i Primenen.*, 23(2):284–294, 1978.

Contents

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399