Electronic Journal of Statistics
Vol. 5 (2011) 1394–1423
ISSN: 1935-7524
DOI: 10.1214/11-EJS646

Deconvolution for the Wasserstein metric and geometric inference

Claire Caillerie and Frédéric Chazal

INRIA Saclay, Parc Club Orsay Université,
2-4 rue Jacques Monod 91893 Orsay Cedex France
e-mail: claire.caillerie@inria.fr; frederic.chazal@inria.fr

Jérôme Dedecker

Laboratoire MAP5, UMR CNRS 8145 Université Paris Descartes,
45, rue des Saints Pères, 75270 Paris Cedex 06, France
e-mail: jerome.dedecker@parisdescartes.fr

Bertrand Michel

Laboratoire de Statistique Théorique et Appliquée,
Université Pierre et Marie Curie - Paris 6,
4 place Jussieu, 75252 Paris cedex 05 France
e-mail: bertrand.michel@upmc.fr

Abstract: Recently, Chazal, Cohen-Steiner and Mérigot 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.

AMS 2000 subject classifications: Primary 62H12, 60B10; secondary 28A33.
Keywords and phrases: Deconvolution, Wasserstein distance, geometric inference, computational topology.

Received July 2011.

Infering topological and geometric information from multivariate data is a problem which is attracting a lot of interest for 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

Hastie and Stuetzle (1989) 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 Genovese et al. (2009) and Genovese et al. (2010) 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 Cuevas and Fraiman (2010) for a large overview on this topic. Cluster analysis algorithms also provide geometric information. One popular approach of clustering proposed by Hartigan (1975) consists in defining clusters as connected components of the levels sets associated to a density f, see for instance Cuevas, Febrero and Fraiman (2000) and Biau, Cadre and Pelletier (2008). Another statistical work by Koltchinskii (2000) propose s 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 Cuevas, Fraiman and Rodríguez-Casal (2007) 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, Chazal and Lieutier (2008) and Chazal, Cohen-Steiner and Lieutier (2009) 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, Niyogi, Smale and Weinberger (2011) 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 Chazal, Cohen-Steiner and Mérigot have generalized the approach of Chazal, Cohen-Steiner and Lieutier (2009) 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 Rachev and Rüschendorf (1998) or Villani (2008)).

Given a probability distribution μ in $\mathbb{R}^d$ and a real parameter $0\leq m\leq 1$, Chazal, Cohen-Steiner and Mérigot 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 parameter $m_0\in[0,1]$ 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 Petrunin (2007) for more informations on the geometric properties of semiconcave functions. Using these properties Chazal, Cohen-Steiner and Mérigot 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. In other words, 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 (see Corollary 4.11 in Chazal, Cohen-Steiner and Mérigot for a precise statement). 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 Horowitz and Karandikar (1994).

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 the results of Chazal, Cohen-Steiner and Mérigot. 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 Hall and Simar (2002) and Delaigle and Gijbels (2006), 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 Meister (2006a) without assuming that the density is discontinuous at the boundary. For the multivariate case, Meister (2006b) proposes an estimator of the support based on a resampling strategy, that satisfy some consistency properties. Still in the convolution setting, Koltchinskii (2000) 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. At the end of the paper, the case of unknown error distribution is briefly addressed in a discussion section.

1. 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

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

1.1. The multivariate convolution model

Assume that one observes n i.i.d. random vectors $(Y_i = (Y_{i,1}, \ldots, Y_{i,d})^t)_{1 \leq i \leq n}$ 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 Horowitz and Karandikar (1994) 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

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

1.2. 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 1.4.

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 Carroll and Hall (1988) and Stefanski and Carroll (1990). 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 Comte and Lacour (2011) 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 Meister (2009)). Mean consistency results with respect to the $\mathbb{L}_1$ loss have also been proved by Devroye (1989). 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 $\mathbb{E}(W_2^2(\hat{\mu}_n, \mu))$.

1.3. A general decomposition

In this section, we shall always assume that

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

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:

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

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 (Villani (2008) Theorem 6.15)**.** *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 Zolotarev (1978) with the constant 4, and by Horowitz and Karandikar (1994) with the constant 3. We give here a very elementary proof which provides a better constant, that is slightly different from the proof given in Villani (2008), Theorem 6.15.

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 Rachev and Rüschendorf (1998), 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

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

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

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

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). □

1.4. 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

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

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 $h_1, h_2, \dots, h_d$. 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

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

and the estimator $\hat{\mu}_n$ of μ is deduced from $\hat{f}_n$ as in Section 1.2.

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

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

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

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

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

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

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

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

It follows that

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

In the same way, we have that

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

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 $(h_1, \dots, h_d) \in [0,1]^d$. The following upper bound holds*

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

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

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

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 $\mathbf{h} = (h_1, \dots, h_d)$ is to choose by simulation the best possible $\mathbf{h}$ in the simple case $\mu = \delta_0$. This will be done in Section 2. 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 2.2).

Proof of Proposition 2. By Plancherel's identity,

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

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,

$$\int hu^2(\tilde{k}_{i,h}(u))^2du \quad = \quad \frac{1}{2\pi}\int h(q'_{i,h}(u))^2du$$
$$\leq \quad C\Big(\int_{-1/h}^{1/h}(r'_i(u))^2du + h^2\int_{-1/h}^{1/h} r_i^2(u)du\Big)\,,$$

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

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

In the same way

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

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

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

Consequently

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

The results follows. □

1.5. 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 1.4, 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 1.2. 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 1.2. 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

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

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 $h_1,\ldots,h_d$.

1.6. 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 1.5, 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

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

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 Horowitz and Karandikar (1994) 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. Rachev (1991) 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: Rates of convergence for a family of noise distributions. For $\alpha \in (0, 2)$, consider the density on $\mathbb{R}$:

$$f_\alpha(x) = \frac{\alpha}{2\Gamma(1/\alpha)} \exp(-|x|^\alpha) .$$

The Fourier transform of this density is proportional to the symetric α-stable distribution. It has an explicity form for $\alpha = 1$ and $\alpha = 2$ only. The case $\alpha = 2$ corresponds to the Gaussian density, and will be treated in the next paragraph. The case $\alpha = 1$ corresponds to the Laplace density, with Fourier transform

$$f_1^*(u) = \frac{1}{(1 + u^2)} .$$

Clearly, f_α^* is infinitely differentiable. From Lemma 2.2.7 page 44 in Koldobsky (2005), we know that f_α^* is positive and symmetric on $\mathbb{R}$. Bergström (1952) has given a precise asymptotic expansion of f_α^*. In particular, it follows from Bergström's result that there exists a positive constant $C(\alpha)$ such that, for $x > 0$,

$$f_\alpha^*(x) = \frac{C(\alpha)}{x^{\alpha+1}} + R(x)$$

where $|R(x)| = O(x^{-2\alpha-1})$ as $x \to +\infty$. The residual part $R(x)$ is expressed as an integral depending on x, which can be differentiated to obtain the following asymptotic expansions, as $x \to +\infty$:

$$\begin{aligned}
(f_\alpha^*)'(x) &= \frac{-(\alpha+1)C(\alpha)}{x^{\alpha+2}} + O(x^{-2\alpha-2}) \\
(f_\alpha^*)''(x) &= \frac{(\alpha+1)(\alpha+2)C(\alpha)}{x^{\alpha+3}} + O(x^{-2\alpha-3}) \\
(f_\alpha^*)'''(x) &= \frac{-(\alpha+1)(\alpha+2)(\alpha+3)C(\alpha)}{x^{\alpha+4}} + O(x^{-2\alpha-4}) . \qquad (18)
\end{aligned}$$

Let k be some nonnegative integer, and let $r_{\alpha,k}(x) = 1/(f_\alpha^*(x))^k$. It follows from (18) that, for different constants $C_1(\alpha, k)$, $C_2(\alpha, k)$, $C_3(\alpha, k)$, $C_4(\alpha, k)$, as $x \to +\infty$,

$$\begin{array}{ll}
r_{\alpha,k}(x) \sim C_1(\alpha,k)x^{k\alpha+k}, & r'_{\alpha,k}(x) \sim C_2(\alpha,k)x^{k\alpha+k-1}, \\
r''_{\alpha,k}(x) \sim C_3(\alpha,k)x^{k\alpha+k-2}, & r'''_{\alpha,k}(x) \sim C_4(\alpha,k)x^{k\alpha+k-3}.
\end{array} \qquad (19)$$

In this paragraph, we consider the case where

$$\mu_i^*(u) = (f_{\alpha_i}^*(u))^{k_i} \, ,$$

with $\alpha_i \in (0,2)$, and $k_i \in \mathbb{N}$. This corresponds to the case where the density of $\varepsilon_{1,i}$ is the k_i-times convolution of the density f_{α_i}.

From (19) and Proposition 2, we obtain the upper bound

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

This bound is similar to the $\mathbb{L}_2$-risk bound obtained by Comte and Lacour (2011) 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

$$h_i = n^{-1/(d+4+2k_1(1+\alpha_1)+\cdots+2k_d(1+\alpha_d)))}$$

and then obtain the rate of convergence

$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{n^{2/(d+4+2k_1(1+\alpha_1)+\cdots+2k_d(1+\alpha_d)))}} \, .$$

In particular:

- If $k_i = 0$ for all i (no noise), we obtain the same rate as previously.
- If $k_i = 1$ and $\alpha_i = \alpha$ for all i (isotropic noise with marginal density f_α), we obtain the rate
$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{n^{2/((3+2\alpha)d+4)}} \, . \tag{20}$$
- If $k_\ell = 1$ and $k_i = 0$ for $i \neq \ell$ (noise in the first direction only, with density f_{α_1}), we obtain the rate
$$\mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq \frac{C}{n^{2/(d+4+2(1+\alpha_1))}} \, .$$

The index $I(d) = k_1(1+\alpha_1) + \cdots + k_d(1+\alpha_d)$ can be seen as an index of global regularity of the error distribution, summing all the regularities $k_i(1+\alpha_i)$ 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 $k(1+\alpha)$ gives the same rates as k marginal distributions with regularity $(1+\alpha)$.

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

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

Following again Comte and Lacour (2011), one can take $h_i=\sqrt{2/\log(n)}$, and we obtain the rate of convergence

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

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

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

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

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

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 Comte and Lacour (2011) about density deconvolution in $\mathbb{R}^d$.

2. 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

$$K=k^{\otimes n}, \quad \text{where} \quad k(x)=\frac{3}{16\pi}\left(\frac{8\sin(x/8)}{x}\right)^4\,. \tag{21}$$

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 1.2.

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

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

2.1. 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 1. 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 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 (20), with $\alpha = 1$ and $d = 2$).

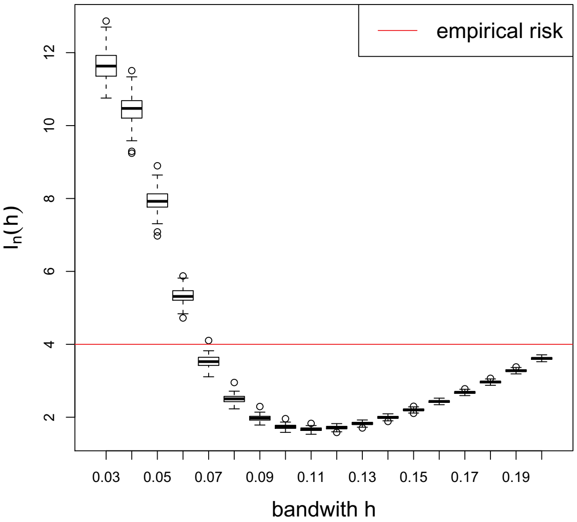


FIG 1. *Boxplots of $I_n(h)$ for different bandwidth h. These results correspond to $N_s = 100$ computations of the deconvolution estimator based on samples of size $n = 20000$.*

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

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

As pointed out in Remark 3 of Section 1.4, 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.

2.2. 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

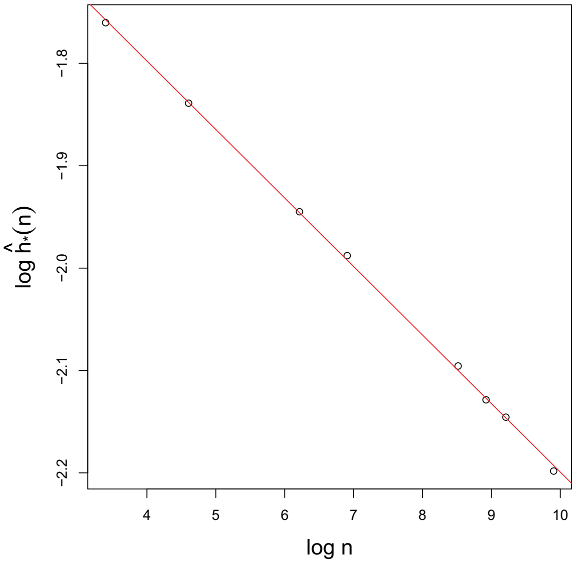


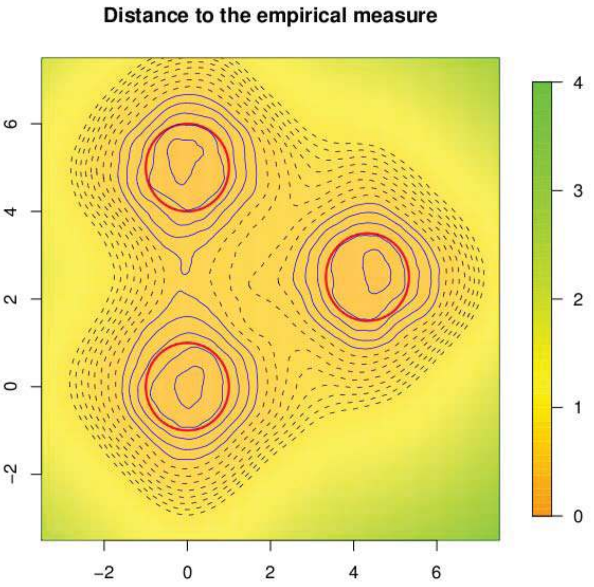
FIG 2. *Estimation of the bandwidths $\hat{h}_*(n)$ (left) against the sample size in logarithm scales. The estimated slope for the regression of $\log \hat{h}_*(n)$ by $\log n$ is $-0.067 \approx -1/14.9$.*

simulations compare the geometry recovered from the distance d_{μ_n,m_0} to the empirical measure as in Chazal, Cohen-Steiner and Mérigot, 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 2.1. 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 2.1 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 $\mathrm{d}_{\tilde{\mu}_n}$. Furthermore, 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.



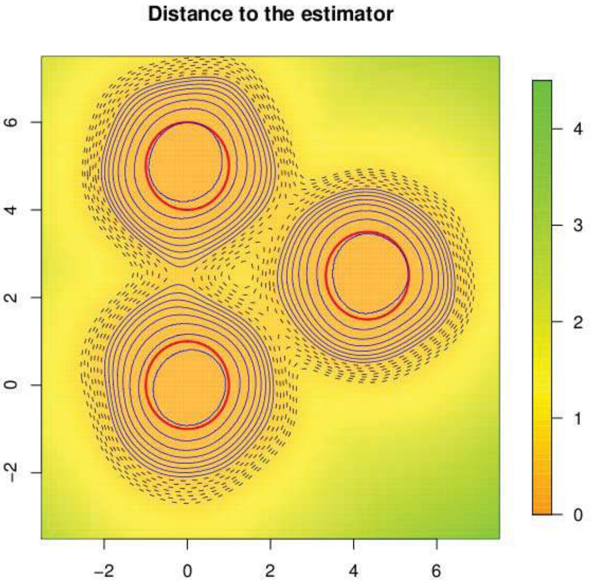


FIG 3. *Distance* d_{μ_n} *to the empirical measure and distance* $\mathrm{d}_{\tilde{\mu}_n(0.11)}$ *to the estimator for the three disks experiment with Laplace noise. The three circles delimiting the disks are drawn in red and the levels of the distance function which have the correct topology are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

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 2.1. 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 2.1.

One Gaussian example

As explained in Section 1.6 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.

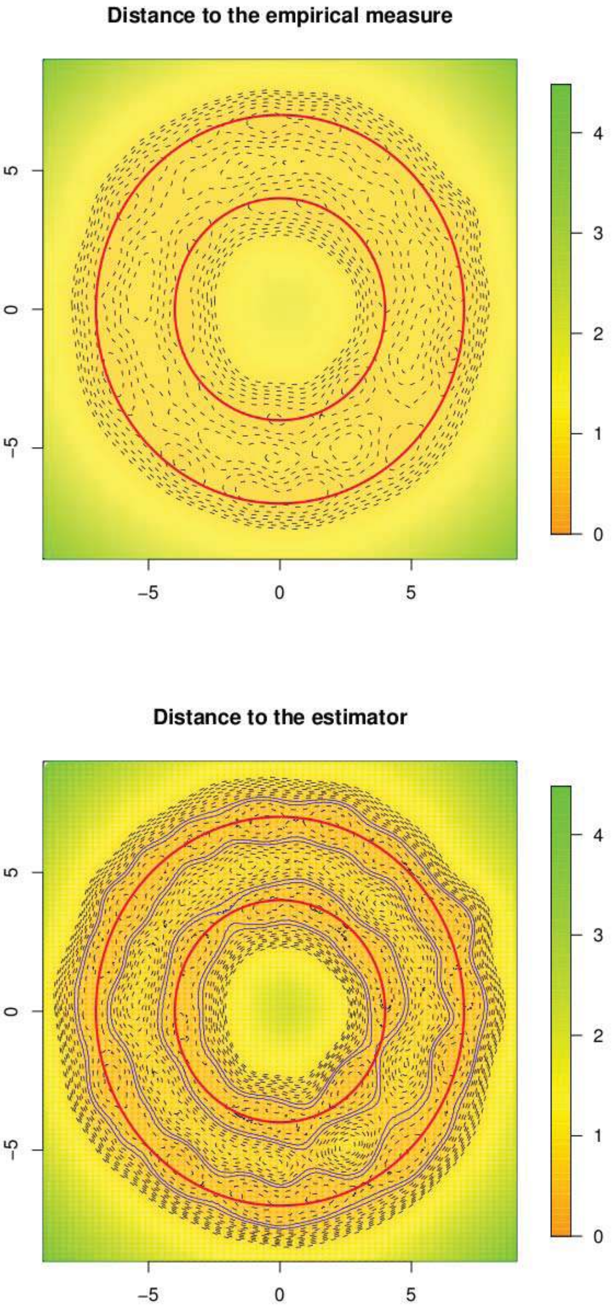
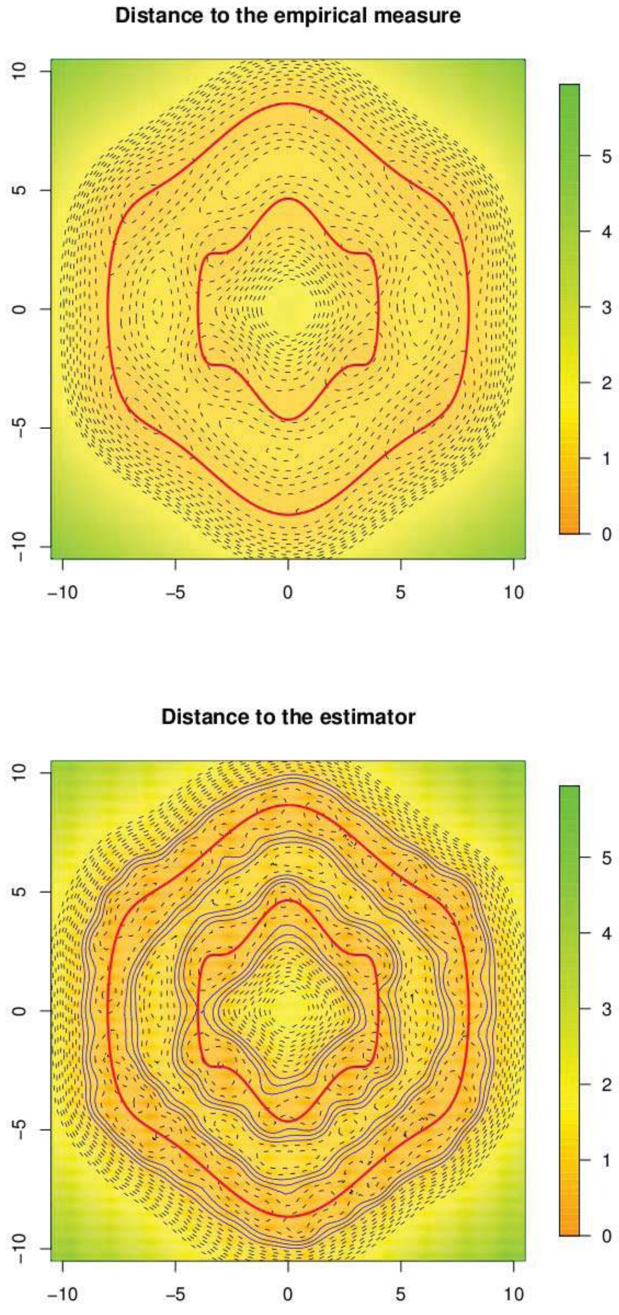


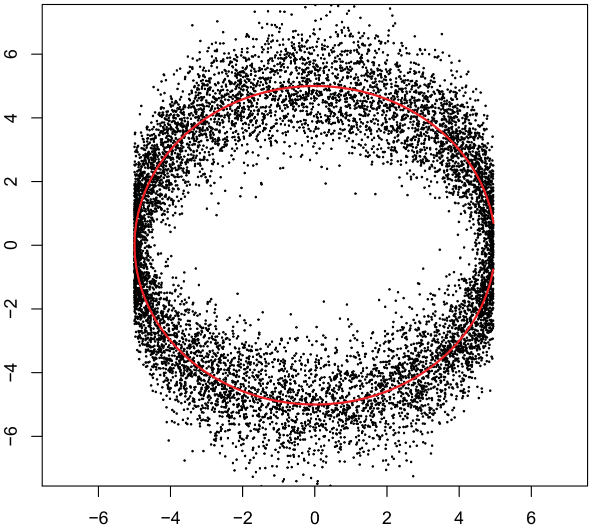
FIG 4. *Distance* d_{μ_n} *and distance* $\mathrm{d}_{\tilde{\mu}_n(0.11)}$ *for the two circles experiment with Laplace noise. See Figure 3 for more details about the legend.*



Distance to the empirical measure

Distance to the estimator

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



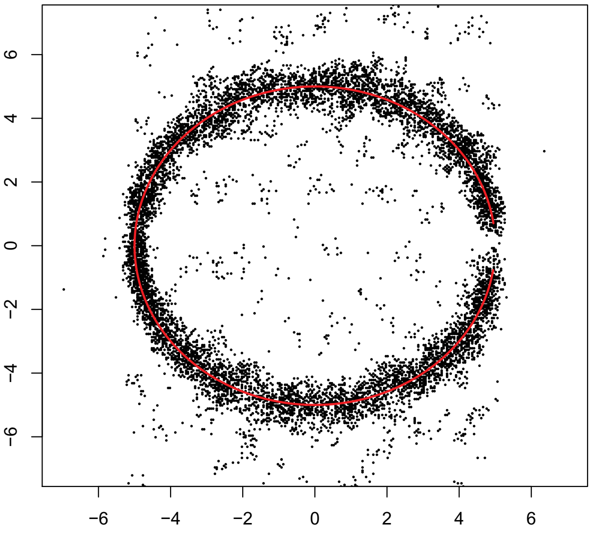
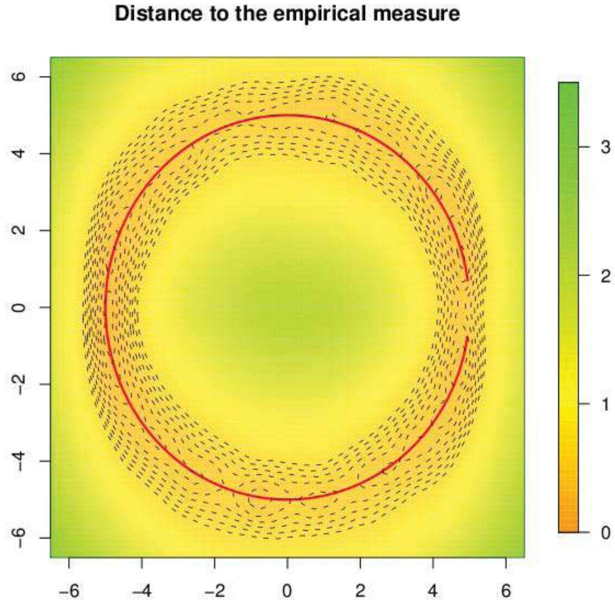


Fig 6. *Circle with hole in red and 10000 points sampled on it with an unidirectional Laplace measurement error (top) and simulation of 10000 points according to $\tilde{\mu}_{n,\mathbf{h}}$ with $h_1 = 0.07$ and $h_2 = 0.25$ (bottom).*



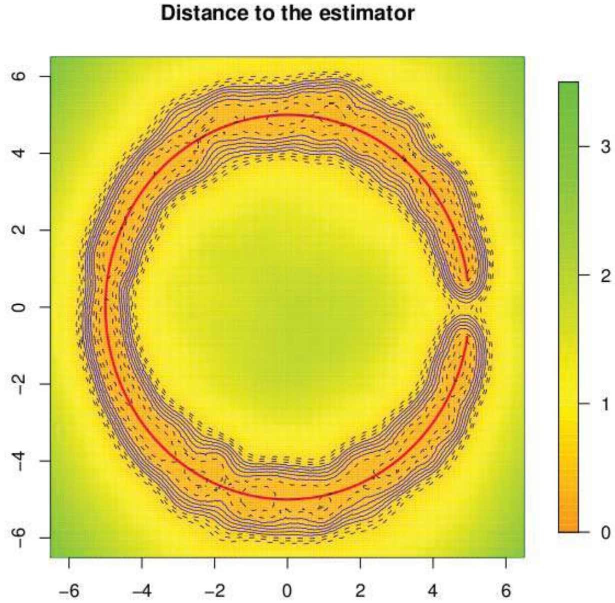


Fig 7. *Distance* d_{μ_n} *and distance* $\mathrm{d}_{\tilde{\mu}_n}$ *for the circle with hole with unidirectionnal noise. See Figure 3 for more details about the legend.*

3. Unknown error distribution

In order to define the deconvolution estimator $\hat{\mu}_n$, the distribution of the error needs to be completely known. We briefly discuss this issue in this section. It may be addressed in more details in future works.

The problem of unknown error has been often advanced against the estimator of Carroll and Hall (1988) and Stefanski and Carroll (1990) in the context of density deconvolution. This question has been the subject of many papers, most of them dealing with the $\mathbb{L}^2$ metric. Whatever the metric considered, two main questions arise: what is the effect of a misspecified error distribution, and what can be done if the error distribution is unknown.

The impact of a misspecified error distribution on the convergence of the devolution estimator (5) has been studied by Meister (2004) for the $\mathbb{L}^2$ metric. Under suitable conditions, the mean integrated squared error converges to a particular functional that depends of the true error distribution and of the wrong distribution used in the estimator. With such a result in mind, it seems likely that the convergence of $\hat{\mu}_n$ is also lost for the Wasserstein metric if the error distribution is not well specified. However, it may not be too catastrophic if the wrong distribution is not too far from the true one, as illustrated in the following example. Assume that $d = 1$, that the true error distribution is $\mathcal{N}(0, \sigma^2)$, and that we use instead the distribution $\mathcal{N}(0, (1 - \eta^2)\sigma^2)$, for $\eta < 1$. Following the lines of the proof of Proposition 1, it can be easily shown that $\mathbb{E}(\hat{f}_n(x)) = \mu \star K_H \star G_{\eta\sigma}$ where $G_{\eta\sigma}$ is a centered Gaussian distribution with variance $\eta^2\sigma^2$. By the triangle inequality, for any $\alpha > 0$,

$$W_2^2(\hat{\mu}_n, \mu) \leq (1 + \alpha) W_2^2(\mu \star G_{\eta\sigma} \star K_H, \mu) + \frac{\alpha + 1}{\alpha} W_2^2(\hat{\mu}_n, \mu \star K_H \star G_{\eta\sigma}) \,. \quad (22)$$

Taking $h = \sqrt{2/\log(n)}$, the variance term $\mathbb{E}(W_2^2(\hat{\mu}_n, \mu \star K_H \star G_{\eta\sigma}))$ tends to zero, and the bias term $W_2^2(\mu \star G_{\eta\sigma} \star K_H, \mu)$ converges to $W_2^2(\mu \star G_{\eta\sigma}, \mu)$. The upper bound (22) being true for any $\alpha > 0$, we obtain that

$$\limsup_{n\to\infty} \mathbb{E}(W_2^2(\hat{\mu}_n, \mu)) \leq W_2^2(\mu \star G_{\eta\sigma}, \mu) \,,$$

while, for the empirical measure μ_n defined in (4),

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

Hence, as n tends to infinity, the deconvolution estimator becomes closer to μ than the empirical measure: it is still relevant to use this deconvolution estimator.

For most applications, the error distribution is partially or completely unknown. If one wants to exhibit a consistent estimator in this context, the first step is to check that the distributions are identifiable. Many examples of identifiable or non identifiable problems can be found in Section 2.6 of Meister (2009). For instance, the distributions are identifiable if μ is an ordinary smooth density and if the error is a centered Gaussian with unknown variance.

To estimate the error distribution and then use it to recover the distribution of the signal, some authors assume that an independent sample from the error distribution is observed (see Neumann (1997) and Comte and Lacour (2011)), while other authors consider longitudinal data (see Li and Vuong (1998)). Another way is to make more hypotheses on the distributions of the signal and of the error. In Butucea and Matias (2005), the distribution of the error is a s-stable distribution with s belonging to $(0, 2)$, and the distribution of the signal is known up to a scale parameter σ. Meister (2007) supposes that the density of the signal is compactly supported, and that the Fourier transform of the error distribution is known on a line segment. Schwarz and Van Bellegem (2010) assume that the error is Gaussian with an unknown variance and that the distribution of the signal puts no mass on a set of Lebesgue measure zero.

We think that some of the approaches described above in the context of density deconvolution could be adapted to the W_2-metric. However, as we have seen in Section 1, one needs to be careful when dealing with this particular metric, and none of these possible extensions seem to be easy.

Appendix 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, Delaigle and Hall (2006) compare the performances of four kernels given in Table 2.

Only k_3 fulfills all the required conditions (see Section 1.2) 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

Table 2
Four kernels and their Fourier transform

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

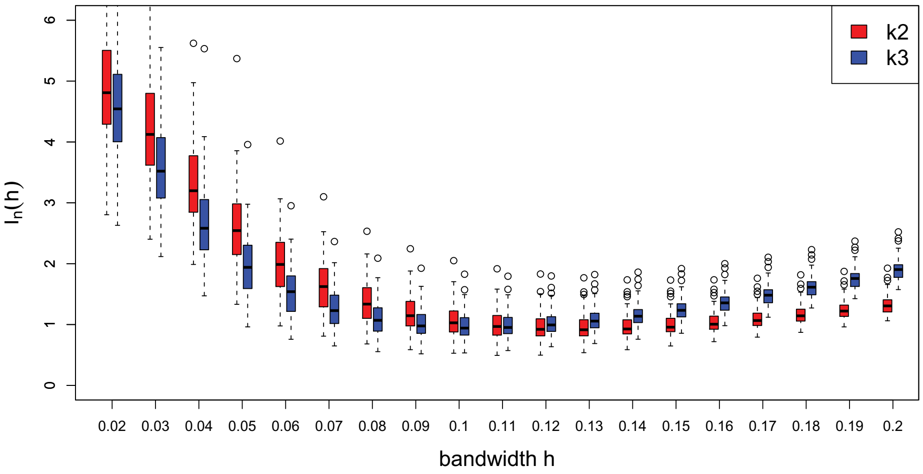


FIG 8. *Comparing the performances of the deconvolution estimators defined by the two kernels k_2 and k_3.*

over 6. This first observation tends to confirm that the kernel assumptions of Section 1.2 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.

Acknowledgements

The authors acknowledge the support of the GIGA grant ANR-09-BLAN-0331-01 and the European project CG-Learning No. 255827.

References

Bergström, H. (1952). On some expansions of stable distribution functions. *Ark. Mat.* **2** 375–378. MR0065053

Biau, G., Cadre, B. and Pelletier, B. (2008). Exact rates in density support estimation. *J. Multivariate Anal.* **99** 2185–2207. MR2463383

Butucea, C. and Matias, C. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli* **11** 309–340. MR2132729

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186. MR0997599

Chazal, F., Cohen-Steiner, D. and Lieutier, A. (2009). A Sampling Theory for Compact Sets in Euclidean Spaces. *Discrete Comput Geom* **41** 461-479. MR2486371

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

Chazal, F. and Lieutier, A. (2008). Smooth Manifold Reconstruction from Noisy and Non Uniform Approximation with Guarantees. *Comp. Geom: Theory and Applications* **40** 156-170. MR2400541

Comte, F. and Lacour, C. (2011). Data driven density estimation in presence of unknown convolution operator. *J. Royal Stat. Soc., Ser B* **73** 601–627.

Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters. *Canad. J. Statist.* **28** 367–382. MR1792055

Cuevas, A., Fraiman, R. and Rodríguez-Casal, A. (2007). A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.* **35** 1031–1051. MR2341697

Cuevas, A. and Fraiman, R. (2010). Set estimation. In *New perspectives in stochastic geometry* 374–397. Oxford Univ. Press, Oxford. MR2654684

Delaigle, A. and Gijbels, I. (2006). Estimation of boundary and discontinuity points in deconvolution problems. *Statist. Sinica* **16** 773–788. MR2281301

Delaigle, A. and Hall, P. (2006). On optimal kernel choice for deconvolution. *Statist. Probab. Lett.* **76** 1594–1602. MR2248846

Devroye, L. (1989). Consistent deconvolution in density estimation. *Canad. J. Statist.* **17** 235–239. MR1033106

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2009). On the path density of a gradient field. *Ann. Statist.* **37** 3236–3271. MR2549559

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2010). The Geometry of Nonparametric Filament Estimation. arXiv:1003.5536v2.

Hall, P. and Simar, L. (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *J. Amer. Statist. Assoc.* **97** 523–534. MR1941469

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, New York-London-Sydney Wiley Series in Probability and Mathematical Statistics. MR0405726

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84** 502–516. MR1010339

Horowitz, J. and Karandikar, R. L. (1994). Mean rates of convergence of empirical measures in the Wasserstein metric. *J. Comput. Appl. Math.* **55** 261–273. MR1329874

Koldobsky, A. (2005). *Fourier analysis in convex geometry. Mathematical Surveys and Monographs* **116**. American Mathematical Society, Providence, RI. MR2132704

Koltchinskii, V. I. (2000). Empirical geometry of multivariate data: a deconvolution approach. *Ann. Statist.* **28** 591–629. MR1790011

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivariate Anal.* **65** 139–165. MR1625869

Meister, A. (2004). On the effect of misspecifying the error density in a deconvolution problem. *Canad. J. Statist.* **32** 439–449. MR2125855

Meister, A. (2006a). Support estimation via moment estimation in presence of noise. *Statistics* **40** 259–275. MR2236207

Meister, A. (2006b). Estimating the support of multivariate densities under measurement error. *J. Multivariate Anal.* **97** 1702–1717. MR2298884

Meister, A. (2007). Deconvolving compactly supported densities. *Math. Methods Statist.* **16** 63–76. MR2319471

Meister, A. (2009). *Deconvolution problems in nonparametric statistics. Lecture Notes in Statistics 193*. Springer-Verlag. MR2768576

Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametr. Statist.* **7** 307–330. MR1460203

Niyogi, P., Smale, S. and Weinberger, S. (2011). A Topological View of Unsupervised Learning from Noisy Data. *SIAM Journal on Computing* **40** 646-663.

Petrunin, A. (2007). Semiconcave functions in Alexandrov's geometry. In *Surveys in differential geometry. Vol. XI* 137–201. Int. Press, Somerville, MA. MR2408266 MR2408266

Rachev, S. T. (1991). *Probability metrics and the stability of stochastic models. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons Ltd., Chichester. MR1105086

Rachev, S. T. and Rüschendorf, L. (1998). *Mass transportation problems. Vol. II. Probability and its Applications*. Springer-Verlag. MR1619171

Schwarz, M. and Van Bellegem, S. (2010). Consistent density deconvolution under partially known error distribution. *Statist. Probab. Lett.* **80** 236–241. MR2575451

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. MR1054861

Villani, C. (2008). *Optimal Transport: Old and New. Grundlehren Der Mathematischen Wissenschaften*. Springer-Verlag. MR2459454

Zolotarev, V. M. (1978). Pseudomoments. *Teor. Verojatnost. i Primenen.* **23** 284–294. MR0517340