Support and distribution inference from noisy data

Jérémie Capitao-Miniconi[1] & Élisabeth Gassiat [2] & Luc Lehéricy [3]

[1] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France, Jeremie.Capitao-Miniconi@universite-paris-saclay.fr*
[2] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France, Elisabeth.Gassiat@universite-paris-saclay.fr*
[3] *Université Côte d'Azur, CNRS, LJAD, France, Luc.Lehericy@univ-cotedazur.fr*

Résumé. Nous considérons des observations bruitées d'une distribution G dont le support est inconnu. Dans le modèle de déconvolution, il a récemment été démontré [6] que, sous des hypothèses très faibles, il est possible de résoudre le problème de déconvolution sans connaître la distribution du bruit et sans aucun échantillon du bruit. Dans ce même modèle, Nous nous intéréssons à estimer la distribution G et son support. Nous commençons par définir les paramètres généraux dans lesquels la théorie s'applique et présentons des classes de supports pouvant être récupérées dans ce contexte. Nous exposons ensuite des classes de distributions pour lesquelles nous prouvons des vitesses minimax adaptatifs (jusqu'à un facteur log log) pour l'estimation du support en distance de Hausdorff. De plus, pour la classe de distributions à support compact, nous fournissons des estimateurs de la distribution inconnue (généralement singulière) et prouvons des vitesses maximums en distance de Wasserstein. Nous démontrons également une borne inférieure presque correspondante sur le risque minimax associé.

Mots-clés. Déconvolution, Apprentissage de variété, statistiques non paramétriques, distance de Hausdorff, distance de Wasserstein, bruit inconnu.

Abstract. We consider noisy observations of a distribution G with unknown support. In the deconvolution model, it has been proved recently [6] that, under very mild assumptions, it is possible to solve the deconvolution problem without knowing the noise distribution and with no sample of the noise. In this same model, we are interested in estimating the distribution G and its support. We first give general settings where the theory applies and provide classes of supports that can be recovered in this context. We then exhibit classes of distributions over which we prove adaptive minimax rates (up to a log log factor) for the estimation of the support in Hausdorff distance. Moreover, for the class of distributions with compact support, we provide estimators of the unknown (in general singular) distribution and prove maximum rates in Wasserstein distance. We also prove an almost matching lower bound on the associated minimax risk.

Keywords. Déconvolution, Manifold learning, Non-paramétric statistics, Hausdorff distance, Wasserstein distance, unknown noise.

1 Introduction

It is a common observation that high dimensional data has a low intrinsic dimension. The computational geometry point of view gave rise to a number of interesting algorithms (see [2] and references therein) for the reconstruction of a non linear shape from a point cloud, and in the statistical community, past years have seen increasing interest for manifold estimation. The case of non noisy data, that is when the observations are sampled on the unknown manifold, is by now relatively well understood. When the loss is measured using the Hausdorff distance, minimax rates for manifold estimation are known and have been proved recently. The rates depend on the intrinsic dimension of the manifold and differ when the manifold has a boundary or does not have a boundary, due to the particular way points accumulate near boundaries (see [1] for the most recent results, together with an overview of the subject and references).

When considering the estimation of a distribution with unknown non linear low dimensional support, one has to choose a loss function. The Wasserstein distance allows to compare distributions that can be mutually singular, and is thus useful to compare distributions having possibly different supports. Moreover, approximating an unknown probability distribution μ by a good estimator $\hat{\mu}$ with respect to the Wasserstein metric allows to infer the topology of the support of μ, see [4]. When using non noisy data, one can look at [5] and [7] for the most recent results and for an overview of the references. However, despite these fruitful developments, geometric inference from noisy data remains a theoretical and practical widely open problem.

In this work, we are interested in the estimation of possibly low dimensional supports, and of distributions supported on such supports, when the observations are corrupted with *unknown* noise. We aim at giving a new contribution on the type of noise which can affect the data without preventing to build consistent estimators of the support and of the law of the noisy signal.

2 Setting

We consider independent and identically distributed observations Y_i, $i = 1, \dots, n$ coming from the model

$$Y = X + \varepsilon, \tag{1}$$

in which the signal X and the noise ε are independent random variables. We assume that the observation has dimension at least two, and that its coordinates can be partitioned in such a way that the corresponding blocks of noise variables are independently distributed, that is

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} + \begin{pmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{pmatrix} = X + \varepsilon \tag{2}$$

in which $Y^{(1)}, X^{(1)}, \varepsilon^{(1)} \in \mathbb{R}^{d_1}$ and $Y^{(2)}, X^{(2)}, \varepsilon^{(2)} \in \mathbb{R}^{d_2}$, for $d_1, d_2 \geq 1$ with $d_1 + d_2 = D$, and we assume that the noise components $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ are independent random variables.

We write G the distribution of X and $\mathcal{M}_G$ its support. For $i \in \{1,2\}$, we write $\mathbb{Q}^{(i)}$ the distribution of $\varepsilon^{(i)}$, so that $\mathbb{Q} = \mathbb{Q}^{(1)} \otimes \mathbb{Q}^{(2)}$ is the distribution of ε.

We shall not make any more assumption on the distribution of the noise ε, and we shall not assume that its distribution is known. Indeed in [6], it is proved that under very mild conditions on the distribution of the signal X, model (2) is fully identifiable, that is one can recover G, and thus its support, and $\mathbb{Q}$ from $G * \mathbb{Q}$.

We introduce the assumptions on the distribution of the signal we shall use. The first one is about the tail of G. Let ρ be a positive real number.

A(ρ) There exist $a, b > 0$ such that for all $\lambda \in \mathbb{R}^D$, $\mathbb{E}\left[\exp\left(\lambda^\top X\right)\right] \leq a \exp\left(b\|\lambda\|_2^\rho\right)$.

Under A(ρ), the characteristic function of the signal can be extended into the multivariate analytic function

$$\begin{aligned} \Phi_X : \mathbb{C}^{d_1} \times \mathbb{C}^{d_2} &\longrightarrow \mathbb{C} \\ (z_1, z_2) &\longmapsto \mathbb{E}\left[\exp\left(iz_1^\top X^{(1)} + iz_2^\top X^{(2)}\right)\right]. \end{aligned}$$

The second assumption is a mild dependence assumption.

(Adep) For any $z_0 \in \mathbb{C}^{d_1}$, $z \mapsto \Phi_X(z_0, z)$ is not the null function and for any $z_0 \in \mathbb{C}^{d_2}$, $z \mapsto \Phi_X(z, z_0)$ is not the null function.

The authors of [6] proved the following identifiability Theorem.

Theorem 1 *if the distribution of the signal satisfies A(ρ) for some $\rho < 2$ and (Adep), then the distribution of the signal and the distribution of the noise can be recovered from the distribution of the observations up to translation.*

The identifiability result above is the base upon which our estimators are built. A key part of its proof is the following result: if a multi-analytic function ϕ satisfies $\phi(0) = 1$, $\phi(t) = \overline{\phi(-t)}$ for all t, as well as assumptions A(ρ) for some $\rho < 2$, (Adep) and

$$\phi(t_1, t_2)\Phi_X(t_1, 0)\Phi_X(0, t_2) = \Phi_X(t_1, t_2)\phi(t_1, 0)\phi(0, t_2)$$

for all (t_1, t_2) in a neighborhood of zero in $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, then $\phi = \Phi_X$. In particular, fixing some $\nu_{\text{est}} > 0$, the only function ϕ such that

$$\int_{B_{\nu_{\text{est}}}^{d_1} \times B_{\nu_{\text{est}}}^{d_2}} |\phi(t_1, t_2)\Phi_X(t_1, 0)\Phi_X(0, t_2) - \Phi_X(t_1, t_2)\phi(t_1, 0)\phi(0, t_2)|^2 |\Phi_{\varepsilon^{(1)}}(t_1)\Phi_{\varepsilon^{(2)}}(t_2)|^2 dt_1 dt_2 = 0$$

is Φ_X; the addition of the characteristic functions of the noises $\Phi_{\varepsilon^{(i)}}$, $i \in \{1,2\}$, does not result in any loss of generality since they are continuous and equal to 1 in zero. From there, we construct an empirical criterion M_n that estimates the above integral as, for any multi-analytic function ϕ,

$$M_n(\phi) = \int_{B^{d_1}_{\nu_{\text{est}}} \times B^{d_2}_{\nu_{\text{est}}}} |\phi(t_1,t_2)\tilde{\phi}_n(t_1,0)\tilde{\phi}_n(0,t_2) - \tilde{\phi}_n(t_1,t_2)\phi(t_1,0)\phi(0,t_2)|^2 dt_1 dt_2, \qquad (3)$$

where for all $(t_1,t_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$,

$$\tilde{\phi}_n(t_1,t_2) = \frac{1}{n}\sum_{\ell=1}^{n} \exp\left\{ it_1^\top Y_\ell^{(1)} + it_2^\top Y_\ell^{(2)} \right\}.$$

We minimize M_n to construct an estimator of Φ_X: let

- Υ_ρ be a set of multivariate analytic functions on $\mathbb{C}^D$, which is in essence the set of functions satisfying A(ρ) and such that $\phi(0) = 1$ and $\phi(t) = \overline{\phi(-t)}$ for all t,
- $\mathcal{H}$ be any closed set of $L^2([-\nu_{\text{est}}, \nu_{\text{est}}]^D)$ of functions satisfying (Adep)
- $\mathbb{C}_m[X]$ be the set of polynomial functions of degree m in D indeterminates. The degree m must satisfy $m \in [2\rho \log n/\log\log n, C\log n/\log\log n]$ for some $C > 2\rho$ arbitrarily large. In the following, we will take $m = \lceil 4\frac{\log n}{\log\log n}\rceil$.

Our estimator of Φ_X is a $1/n$-minimizer of M_n over the intersection of these three sets:

For any integer m and any $\rho > 1$, let $\widehat{\Phi}_{n,m,\rho}$ be a (up to $1/n$) measurable minimizer of the functional $\phi \mapsto M_n(\phi)$ over $\Upsilon_{\rho,S} \cap \mathcal{H} \cap \mathbb{C}_m[X]$.

Provided Φ_X itself is in $\Upsilon_\rho \cap \mathcal{H}$, this estimator converges toward Φ_X. As for its rate of convergence, by carefully analyzing the behaviour of M_n near Φ_X. We show the following Proposition.

Proposition 1 (Variant of Proposition 1 in [3]) *For all $\rho_0 < 2$, $\nu \in (0, \nu_{est}]$, $S, c(\nu), E, C > 0$ and $\delta, \delta', \delta'' \in (0,1)$ with $\delta' > \delta$, there exist positive constants c and n_0 such that the following holds: let $\rho \in [1, \rho_0]$, for all $\Phi_X \in \Upsilon_{\rho,S} \cap \mathcal{H}$ and $\mathbb{Q} \in \mathcal{Q}^{(D)}(\nu, c(\nu), E)$, for all $n \geq n_0$ and $x \in [1, n^{1-\delta'}]$, with probability at least $1 - 2e^{-x}$,*

$$\sup_{\rho' \in [\rho,\rho_0],\ m \in [2\rho' \frac{\log n}{\log\log n}, C\frac{\log n}{\log\log n}]} \int_{B^{d_1}_\nu \times B^{d_2}_\nu} |\widehat{\Phi}_{n,m,\rho'}(t) - \Phi_X(t)|^2 dt \leq c\left(\frac{x}{n^{1-\delta}}\right)^{1-\delta''}.$$

3 Estimation

From there, we first consider the estimation of its support $\mathcal{M}_G$. While it may not be possible to directly recover the distribution of X from Φ_X by inverse Fourier transform, since G could be singular, it is possible to recover the function $G * \Psi$ for any properly chosen kernel Ψ.

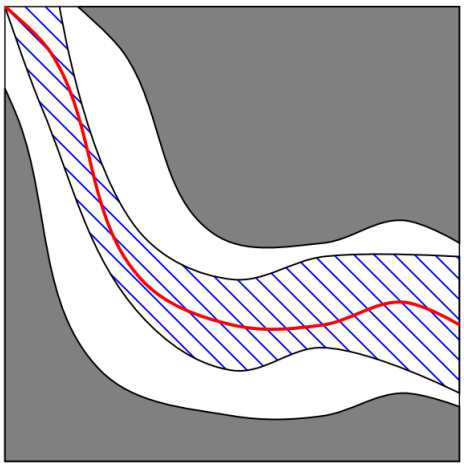

Figure 1: In red, the support of the signal distribution $\mathcal{M}_G$, the blue hatched area represents the set $\{\bar{g} \geq \lambda_n + \|\bar{g} - \widehat{g}_n\|_\infty\}$ and the gray area represents the set $\{\bar{g} \leq \lambda_n - \|\bar{g} - \widehat{g}_n\|_\infty\}$, so that the estimator of the support lies in between the gray and the blue areas.

Moreover, when Ψ is an approximation of the Dirac, the support of G can be approximated by an upper level set of $G * \Psi$. Intuitively, if Ψ were the density of a random variable Z, $G * \Psi$ would be the density of $X + Z$. When Ψ is an approximation of the Dirac, the distribution of Z will be close to the Dirac in zero, thus $X + Z$ will concentrate close to the support of X. Figure 1 illustrates this idea.

The only geometric assumption we need on G is that it is (a, d)-standard, *i.e.* that the weight of any ball of radius r centered on a point of its support is at least ar^d. In particular, we do not require that the reach of the support $\mathcal{M}_G$ is lower bounded, or even that $\mathcal{M}_G$ is a manifold.

The precise choice of Ψ is non trivial and will not be discussed here. For f an integrable function from $\mathbb{R}^D$ to $\mathbb{R}$, we denote by $\mathcal{F}[f]$ (resp. $\mathcal{F}^{-1}[f]$) the (resp. inverse) Fourier transform of f defined, for all $y \in \mathbb{R}^D$, by

$$\mathcal{F}[f](y) = \int e^{it^\top y} f(t) dt \quad \text{and} \quad \mathcal{F}^{-1}[f](y) = \left(\frac{1}{2\pi}\right)^D \int e^{-it^\top y} f(t) dt.$$

Among the properties of Ψ, one is that its Fourier transform $\mathcal{F}[\Psi]$ is compactly supported. We show that the estimator

$$\widehat{g} := \mathcal{F}^{-1}\left[\widehat{\phi} \cdot \mathcal{F}[\Psi]\right]$$

is close to $\bar{g} := \mathcal{F}^{-1}\left[\Phi_X \cdot \mathcal{F}[\Psi]\right] = G * \Psi$, whose upper level sets are close to $\mathcal{M}_G$. Our estimator is thus taken as

$$\widehat{\mathcal{M}} = \{y : \widehat{g}(y) > \lambda\},$$

where λ has to be choosen. We also show that, when H is the Hausdorff distance and $\mathcal{K}$ is a known compact in $\mathbb{R}^D$,

$$\mathbb{E}[H(\mathcal{M}_G \cap \mathcal{K}, \widehat{\mathcal{M}} \cap \mathcal{K})] = O\left(\frac{(\log\log n)^{1/\rho+1+\delta}}{(\log n)^{1/\rho}}\right)$$

for some $\delta > 0$ arbitrarily small (that depends on Ψ). A caveat of the above result is that we can only control what happens inside a compact set $\mathcal{K}$. On the other hand, $\mathcal{K}$ can be taken arbitrarily large.

We use the two-point method to prove that this result is quasi-optimal, in the sense that for any estimator $\widehat{\mathcal{M}}$, there exists G such that the error $\mathbb{E}[H(\mathcal{M}_G \cap \mathcal{K}, \widehat{\mathcal{M}} \cap \mathcal{K})]$ is at least $\frac{1}{(\log n)^{1/\rho}}$ when $\rho > 1$, and $\frac{1}{(\log n)^{1+\delta}}$ when $\rho = 1$ (i.e. when the support of X is compact), for $\delta > 0$ arbitrarily small.

All these results can be made adaptive in the tail parameter ρ; we provide a method based on Goldenshluger and Lepski's method.

We consider the estimation of G in Wasserstein distance, when G is compactly supported. We do so by restricting $\max(\widehat{g}, 0)$ to the estimated support $\widehat{\mathcal{M}}$, then renormalizing it to obtain the probability density of a measure $\widehat{P}$, which satisfies

$$\mathbb{E}[W_2(G, \widehat{P})] = O\left(\frac{\log\log n}{\log n}\right),$$

Again, we use the two-point method to prove that this result is quasi-optimal, showing that for any estimator $\widehat{P}$, there exists G such that the error $\mathbb{E}[W_2(G, \widehat{P})]$ is at least $\frac{1}{(\log n)^{1+\delta}}$, for $\delta > 0$ arbitrarily small

References

[1] Eddie Aamari, Catherine Aaron, and Clément Levrard. *Minimax Boundary Estimation and Estimation with Boundary*. 2021. DOI: 10.48550/ARXIV.2108.03135. URL: https://arxiv.org/abs/2108.03135.

[2] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and topological inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2018, pp. xii+233. ISBN: 978-1-108-41089-2; 978-1-108-41939-0. DOI: 10.1017/9781108297806. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1017/9781108297806.

[3] Jérémie Capitao-Miniconi and Élisabeth Gassiat. "Deconvolution of spherical data corrupted with unknown noise". In: *Electron. J. Stat.* (to appear).

[4] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. "Geometric inference for probability measures". In: *Found. Comput. Math.* 11.6 (2011), pp. 733–751. ISSN: 1615-3375. DOI: 10.1007/s10208-011-9098-0. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1007/s10208-011-9098-0.

[5] Vincent Divol. "Measure estimation on manifolds: an optimal transport approach". In: *Probab. Theory Related Fields* 183.1-2 (2022), pp. 581–647. ISSN: 0178-8051. DOI: 10.1007/s00440-022-01118-z. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1007/s00440-022-01118-z.

[6] Élisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. "Deconvolution with unknown noise distribution is possible for multivariate signals". In: *Ann. Statist.* 50.1 (2022), pp. 303–323.

[7] Jonathan Niles-Weed and Quentin Berthet. "Minimax estimation of smooth densities in Wasserstein distance". In: *Ann. Statist.* 50.3 (2022), pp. 1519–1540. ISSN: 0090-5364. DOI: 10.1214/21-aos2161. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1214/21-aos2161.