Document de synthèse présenté pour l'obtention d'une

Habilitation à Diriger des Recherches

Méthodes directes pour problèmes inverses et quelques contributions en statistique non-paramétrique.

présenté par

Clément Marteau

au vu des rapports de

Cristina Butucea	Professeur, Université de Marne-La-Vallée
Iain Johnstone	Professeur, Stanford University
Axel Munk	Professeur, Georg-August Universität

soutenue le 9 décembre 2013, devant le jury composé de

Cristina Butucea	Professeur, Université de Marne-La-Vallée	(rapporteur)
Laurent Cavalier	Professeur, Université Aix-Marseille	(examinateur)
Fabrice Gamboa	Professeur, Université Paul Sabatier	(coordinateur)
Jean-Michel Marin	Professeur, Université Montpellier II	(examinateur)
Axel Munk	Professeur, Georg-August Universität	(rapporteur)

Institut Mathématiques de Toulouse
Université Paul Sabatier

[..] un plus un égal deux,
on n'a jamais trouvé mieux [..]
P.D. Burgaud.

Remerciements

Je tiens avant toute chose à remercier les trois rapporteurs de cette HDR: d'abord pour avoir accepté d'être membres de ce jury, mais également pour le temps passé à lire mon manuscrit. En particulier, un grand merci à Axel et Cristina de faire le déplacement pour cette soutenance malgré des emplois du temps très chargés. Je voudrais également remercier les autres membres de mon jury: Laurent Cavalier qui m'a mis le pied à l'étrier en 2004 et qui me fait l'honneur d'être à nouveau présent aujourd'hui, Fabrice Gamboa pour ses conseils avisés lors de la préparation de cette HDR et Jean-Michel Marin que je remercie de faire le déplacement spécialement pour cette soutenance.

Dans un second temps, je voudrais adresser un immense merci à Béatrice Laurent. La transition thésards/ maître de conférence est à mon avis un moment clé dans la vie d'un chercheur. Grâce à elle, tout s'est passé pour le mieux et mon insertion au sein de l'équipe ESP a pu se faire dans les meilleures conditions possibles. Dans le même registre, je voudrais remercier Jean-Michel Loubes pour m'avoir fait confiance dès mon arrivée à Toulouse. Plus généralement, un grand merci à tous mes co-auteurs: Jérémie Bigot, Sébastien Gadat, Yuri Ingster, Thierry Klein, Sébastien Loustau (version finale?), Peter Mathé, Cathy Maugis-Rabusseau ou plus récemment Theofanis Sapatinas. Un grand merci également aux membres de l'équipe ESP et du département GMM de l'INSA. Finalement, c'est peut-être ce que je préfère dans les mathématiques: c'est avant tout une grand aventure humaine.

En dernier lieu: un grand merci à ma femme Alexandra qui m'a suivi avec enthousiasme dans cette aventure toulousaine. La petite famille ne comptera probablement pas plus de quatre membres mais c'était pour moi la plus belle de toutes les additions...

Contents

Outline of the manuscript

This manuscript intends to summarize my research activities since I have obtained an assistant professor position in Toulouse. Since September 2008, I get the opportunity to extend my investigations that started during my PhD. Thesis. I get the chance to work with several different colleagues, working on amazing different topics.

The goal of the present document is not to provide an exhaustive list of my different contributions (I refer to my personal bibliography at the end of this document). In the following, I will rather try to explain how my different contributions are related. Then, I will provide a short overview of possible outcomes. This manuscript is decomposed in two different chapters.

Some contributions in non-parametric statistical inverse problems

In this first chapter, I will provide a brief introduction on statistical inverse problems. We will therein discuss several different models with associated problematics falling in this scope. In a first time, we briefly discuss both Gaussian white noise and error-in-variables models. Some attention will also be payed to the shifted curves model that has been widely investigated last years.

In a second time, I will briefly present my different contributions in these different topics. Few results and proofs will be given: the interested reader will be referred to the corresponding papers for more details. To this end, the list of my different contributions with complete references is provided at the end of this manuscript.

Direct methods for inverse problems

The Chapter 2 contains few formal results. In this part, we show that there exist alternatives to existing statistical procedures in an inverse problem context. For instance, in a testing framework, the inversion of the operator does not appear to be always necessary. In particular,we propose procedures that provides satisfying (minimax) behavior without inversion of an operator.

In a second time, we will see that this principle could be certainly extended to a large amount of different topics. This part is completely heuristic: no formal result are available at the present time and only informal discussions will be provided. In particular, the aim is to highlight the links between the different topics I have considered since I began to work in mathematical statistics. The different contributions that I have proposed in non-parametric estimation, testing theory or binary classification can indeed be related.

I will conclude this second chapter by some general perspectives in non-parametric statistics, that do not necessarily concern an inverse problem setting.

Chapter 1

Some contributions in non-parametric statistical inverse problems

In this chapter, we introduce some statistical inverse problem models. In the first section, we briefly explain and discuss the corresponding problematics and the main related outcomes. The second section is devoted to a presentation of my different contributions in these models.

1.1 Statistical inverse problems

Inverse problems arise in many different fields and are the core of many theoretical and practical investigations. Although it may be possible to provide a general description of such models, it is important to note that statistical inverse problems are related to a large amount of different situations, with different related methodologies. This is a fascinating particularity of this topic. Below, we briefly discuss the properties of statistical inverse problems (and related compact operators). Then, we will present some practical examples and associated problematics.

A statistical inverse problem can be characterized as follows. Given $\mathcal{X}$ and $\mathcal{Y}$ two Hilbert spaces, one want to provide some inference on an unknown target $f \in \mathcal{X}$. To this end, we assume that we have at our disposal noisy and indirect observations. In several cases, we assume that we observe

$$Af + \text{ ''noise''}, \tag{1.1}$$

where $A : \mathcal{X} \to \mathcal{Y}$ denotes an operator. In some sense, the object of interest f is distorted, and the operator A characterizes this distortion. The noise term can model several different situations. The properties and the behavior of this noise term will be described bellow.

The operator A and its related properties are of first importance. Indeed, in order to provide some inference on f from (1.1) , we will have to investigate the properties of Af. In many cases it will then be necessary to provide a (generalized if necessary) inverse of A.

In this manuscript, we will only deal with **compact operator** A. These operators are very interesting from a mathematical point of view. This is due in particular to the following proposition.

Proposition 1 *Let $A : \mathcal{X} \to \mathcal{Y}$ a compact operator. Then*

$$\|(A^*A)^{-1}\| = \sup_{x\in\mathcal{X}, \|x\|=1} \|(A^*A)^{-1}x\| = +\infty,$$

where A^* denotes the adjoint operator of A.

There exists several different kind of compact operators, leading to various inverse problems. We refer for instance to [46] for a review of *classical* inverse problems. Below, we provide three examples that often arise in the literature.

- **Convolution operator**. Set $\mathcal{X} = \mathcal{Y} = L^2(\mathbb{R})$, and $g \in L^1(\mathbb{R})$. The convolution operator A is then defined as

  $$Af(x) = \int_{\mathbb{R}} g(y)f(x-y)dy, \ \forall x \in \mathbb{R}, \ f \in L^2(\mathbb{R}).$$

  The function g is called *convolution kernel.* Hereafter, we will often write $Af = f * g$, where $*$ denotes the convolution product.

- **Integration operator**. Set $\mathcal{X} = \mathcal{Y} = L^2([0,1])$. For all $f \in L^2([0,1])$, we define the integration operator as

  $$Af(t) = \int_0^t f(x)dx.$$

  In such a setting, providing inference on f amounts to study the first derivative of Af. In such a case, the corresponding model (1.1) turns to be a differentiation problem.

- **Radon transform**. The Radon transform arises in medical applications. Let H be the unit disk on $\mathbb{R}^2$. The aim is to provide inference on the spatially varying density f of a cross section $\mathcal{D} \subset H$ of an human body. This inference is provided via observation obtained by non-destructive imagery, namely X-ray tomography. In such a case, given a function $f \in L^2(H)$, one measure $Rf(u,\varphi)$ which corresponds to the decay of the intensity of the X-ray in the direction φ when the receiver is at a distance u. Typically, we have

  $$Rf(u,\varphi) = \frac{\pi}{2\sqrt{1-u^2}} \int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} f(u\cos(\varphi) - t\sin(\varphi), u\sin(\varphi) + t\cos(\varphi))dt,$$

  for all $(u,\varphi) \in [0,1] \times [0,2\pi)$.

The main consequence of Proposition 1 is that A^*A is not continuously invertible. Hence, nothing guarantees that a small amount of noise will provide a solution close to the target with rough (for instance least-square) methods. In order to overcome this problem, several approaches have been proposed in the literature. These methods are called regularization algorithms and will be briefly discussed in the following sections.

Below, we introduce and present three different inverse problem models. In each case, we explicit conditions on the noise term in (1.1) and discuss some related challenging problems.

1.1.1 Gaussian white noise model

The Gaussian white noise model has received a lot of attention over last years. Although it covers several practical situations, this model is often considered as a toy-model. In particular, it allows a sharp study and a deep understanding of all the difficulties and outcomes related to inverse problems. In such a setting, we assume that we observe

$$Y = Af + \epsilon\xi, \tag{1.2}$$

where $A : \mathcal{X} \to \mathcal{Y}$ is a compact operator, ϵ a positive noise level. In the numerical inverse problem literature, the noise ξ is assumed to be deterministic and bounded. In a statistical framework, we assume that this noise is Gaussian and white. In this case, the Gaussian white noise model (1.2) is a slight abuse of notation. In fact, one assume that we observe

$$\langle Y, g\rangle = \langle Af, g\rangle + \epsilon\langle \xi, g\rangle, \ \forall g \in \mathcal{Y}, \tag{1.3}$$

and that for all $g, g_1, g_2 \in \mathcal{Y}$, one has

$$\langle \xi, g\rangle \sim \mathcal{N}(0, \|g\|^2), \text{ and } \mathbb{E}\left[\langle \xi, g_1\rangle\langle \xi, g_2\rangle\right] = \langle g_1, g_2\rangle.$$

Now, the following question arises: having at hand Y following the model (1.2), how can we provide some inference on f?

Suppose that $Y \in \mathcal{Y}$. In a such a situation, it would be natural to estimate f via the least square methods, namely

$$\hat{f}_{LS} = \arg\min_{\nu\in\mathcal{X}} \|Y - A\nu\|^2 = (A^*A)^{-1}A^*Y,$$

where A^* denotes the adjoint of A. Nevertheless, as discussed in Proposition 1, the operator A^*A is not continuously invertible. Hence, it is not possible to control the associated quality of reconstruction. A possible outcome is to impose a constraint on the estimator in order to ensure the stability. As an example, we can mention for instance the Tikhonov estimator defined as

$$\hat{f}_\tau := \arg\min_{\nu\in H} \left[\|Y - A\nu\| + \tau\|\nu\|^2\right],$$

where τ denotes a so-called *regularization parameter*. This method can be seen as a generalization of the least-square algorithm where one impose some constraint on the recovered function f, here a control on the norm. In particular, it is possible to prove that

$$\hat{f}_\tau = (A^*A + \tau I)^{-1}A^*Y := g_\tau(A^*A)A^*Y.$$

Roughly speaking, the parameter τ hence control the level of the constraint on the norm of f. Small values of τ will provide solution close to the least-square estimator, while large values of τ will induce bias in the estimation.

Several alternative approaches have been proposed over years. In each cases, the proposed algorithms replace the operator $(A^*A)^{-1}$ by $g_\tau(A^*A)$ for some function g_τ. The so-called regularization parameter controls the *proximity* between $g_\tau(A^*A)$ and $(A^*A)^{-1}$. More precisely, we will require the following properties for the function g_τ.

Definition 1 *A family of functions*

$$g_\tau : (0, \|A^\star A\|] \to \mathbb{R}, \ \tau \in (0, \|A^\star A\|],$$

is called a regularization family if the functions g_τ are piece-wise continuous in τ and if the following properties holds:

- *For all $0 < t \leq \|A^\star A\|$, we have $|r_\tau(t)| := |1 - tg_\tau(t)| \to 0$ as $\tau \to 0$,*
- *There exists a constant γ_1 such that*

  $$\sup_{0<t\leq\|A^\star A\|} |r_\tau(t)| \leq \gamma_1,$$

  for all $\tau \in (0, \|A^\star A\|]$,
- *There exists a constant $\gamma_\star$ such that*

  $$\sup_{0<t\leq\|A^\star A\|} \tau|g_\tau(t)| \leq \gamma_\star,$$

  for all $0 < \tau < +\infty$.

For more details, we refer for instance to [46], [54], [78].

Concerning the estimation issue, the choice of the regularization parameter and the study of the related (quadratic) risk in this setting has concentrated a lot of attention in the two last decades. We will mention [14], [15], [79] or [78] among others.

As a particular case of the previous approach, the spectral regularization provides interesting properties. The main underlying idea is to project the observation in a particular basis of $\mathcal{X}$, for which the representation matrix of A^*A will be diagonal. This basis is associated to the singular value decomposition $(b_k^2, \phi_k, \psi_k)_{k\in\mathbb{N}}$ of A^*A, where for all $k \in \mathbb{N}$,

$$\begin{cases} A\phi_k = b_k\psi_k, \\ A^*\psi_k = b_k\phi_k. \end{cases}$$

Then, for all $k \in \mathbb{N}$, replacing g by ψ_k in (1.3), we obtain the sequence space model

$$y_k := \langle Y, \psi_k \rangle = b_k\theta_k + \epsilon\xi_k, \ \forall k \in \mathbb{N}, \tag{1.4}$$

where $\theta_k := \langle f, \phi_k \rangle$ for all $k \in \mathbb{N}$ and the ξ_k are i.i.d. standard Gaussian random variables. Such a model appears to be an interesting generalization of classical non-parametric *direct* model for which $b_k = 1$ for all $k \in \mathbb{N}$ (see for instance [88] or [3]).

In this setting, a regularization method g_τ can be identified to a linear (spectral) estimator defined as

$$\hat{f}_\lambda = \sum_{k\in\mathbb{N}} \lambda_k(\tau) b_k^{-1} y_k, \tag{1.5}$$

where $\lambda(\tau) = (\lambda_k(\tau))_{k\in\mathbb{N}}$ denotes a filter, i.e. a real sequence having values in $[0, 1]$. In particular, the spectral cut-off (projection) filter $\lambda_k(\tau) = \mathbf{1}_{\{k\leq\tau^{-1}\}}$ for some $\tau^{-1} \in \mathbb{N}$ has attracted a lot of attention over years. For more details on such a model and related issues, we mention for instance [31].

In an estimation purpose, it is then possible to measure the error associated to such a linear estimator via its quadratic risk $R(\theta, \lambda)$ defined as

$$R(\theta, \lambda) := \mathbb{E}_f \|f_\lambda - f\|^2 = \sum_{k=1}^{+\infty} (1 - \lambda_k(\tau))^2 \theta_k^2 + \epsilon^2 \sum_{k=1}^{+\infty} \lambda_k^2(\tau) b_k^{-2}.$$

Such a term can be controlled uniformly over wide smoothness classes of functions with an appropriate choice for τ (minimax point of view), or compared with the smallest possible one when we consider data-driven choices for τ (this is the oracle point of view). Different kind of inference are also available: signal detection, estimation of quadratic functional, and so on... Some of these topics will be detailed in Section 1.2 bellow.

Remark: In order to get the sequence space model (1.4), we have projected the observations in the basis associated to the singular value decomposition of the operator A. In some sense, this basis is optimal w.r.t. the operator. Nevertheless, this basis may not be convenient for the signal of interest. For instance, the eigenvectors of the convolution operator correspond to the real trigonometric basis. This basis does not describe non-continuous signals in a convenient way.

In order to overcome this drawback, alternative strategies have been proposed in the last twenty years: wavelet-vaguelette decomposition in [40], projection on Meyer wavelet basis in [66] (or more recently [10]) for convolution problems, or construction of alternative bases as in [69].

1.1.2 Error-in-variables model

Assume that we have at our disposal a sample $\mathcal{S} = (X_1, \dots, X_n)$ of i.i.d. random variables, having a common density f with respect to the Lebesgue measure. Classical non-parametric problems in this setting are related to the estimation of alternatively f, a functional of f or by the construction of goodness-of-fit testing procedures. In each case, one want to provide a precise study of minimax rates of convergence associated to given smoothness constraints (see [88]), or a precise description of minimax separation rates when working in a testing framework (see [4], [17] or Section 1.2 for a formal definition).

In order to provide some inference on the function f (in an estimation purpose for instance), several methods have been proposed. For instance, one can use a kernel estimator defined as

$$\tilde{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} \mathcal{K}\left(\frac{X_i - x}{\lambda} \right), \ \forall x \in \mathbb{R}, \tag{1.6}$$

where $\mathcal{K}$ denotes a kernel ($\mathcal{K} \in L^2(\mathbb{R})$, $\int_{\mathbb{R}} \mathcal{K}(x)dx = 1$) and $\lambda \in \mathbb{R}^+$ is a bandwidth. In particular, for all $x \in \mathbb{R}$,

$$\mathbb{E}[\tilde{f}_\lambda(x)] = \frac{1}{\lambda} \mathbb{E}\left[\mathcal{K}\left(\frac{X_1 - x}{\lambda} \right) \right] = \frac{1}{\lambda} \int_{\mathbb{R}} \mathcal{K}\left(\frac{y - x}{\lambda} \right) f(y) dy := \mathcal{K}_\lambda * f(x),$$

$\mathcal{K}_\lambda = \lambda^{-1}\mathcal{K}(./\lambda)$. It is then possible to control the proximity between $\mathcal{K}_\lambda * f$ and f assuming some smoothness properties (often characterized by the decay of the Fourier transform of f). Following the purpose (estimation, test, ...), this allows to establish upper bounds for the minimax rate of convergence. For more details, we refer for instance to [88].

When dealing with the sample $\mathcal{S}$, one implicitly assume that each observation is perfectly gathered, i.e. that we observe the X_i with no error. Nevertheless, in several practical situations (see [17], [81], or [57]), it is rather natural to assume that these observations are associated to some error measurements: the X_i are not observable, but only noisy approximations of the form

$$Z_i = X_i + \epsilon_i, \ \forall i \in \{1, \dots, n\}. \tag{1.7}$$

where the ϵ_i denote observations errors. These random variables are assumed to be i.i.d., independent of the X_i, with common density η w.r.t. the Lebesgue measure. This density will assumed to be **known** in the following.

In the model (1.7), it is not possible to replace the X_i by observations Z_i in the kernel estimator (1.6). Indeed, since the variables of interest X_i are independent of the measurement errors, the density of the Z_i corresponds to the convolution product $f * \eta$. Hence,

$$\mathbb{E}[\tilde{f}_\lambda(x)] = \frac{1}{h}\mathbb{E}\left[K\left(\frac{Z_i - x}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{y-x}{h}\right) f * \eta(y)dy := \mathcal{K}_\lambda * \{f * \eta\}(x),$$

where $\tilde{f}_\lambda$ is the kernel estimator (1.6) (replacing the X_i by the Z_i). In general, the term $\mathcal{K}_\lambda * \{f * \eta\}$ does not provide a good approximation of the unknown target f, even under strong smoothness assumptions. We are in fact faced to an inverse (convolution) problem. As in the Gaussian white noise model, it is necessary to introduce a regularization (deconvolution) step if one want to recover the function f.

In this context, the construction of a deconvolution estimator is generally based on Fourier calculus. In the following, given a function $g \in L^2(\mathbb{R})$, we denote by $\mathcal{F}[g]$ its corresponding Fourier transform. It appears that

$$\mathcal{F}[f * \eta] = \mathcal{F}[f] \times \mathcal{F}[\eta].$$

In the frequency domain, the operator has a multiplicative effect on the unknown signal f. In order to guarantee some identifiability properties, we will assume that $\mathcal{F}[\eta](t) > 0$ for all $t \in \mathbb{R}$. This property provides a first (naive) approach. One can for instance compute the empirical Fourier transform of $f * \eta$ from the Z_i, and then divide this term by $\mathcal{F}[\eta]$ (which is allowed since the density η is assumed to have a non-null Fourier transform). Nevertheless, since $\eta \in L^2(\mathbb{R})$, its Fourier transform is close to 0 for large frequencies, hence leading to an unstable estimator.

In order to get round of this effect, a classical approach consists in introducing a regularization step. As before, let $\mathcal{K}$ be a kernel, with associated Fourier transform $\mathcal{F}[\mathcal{K}]$ and let $\lambda > 0$ be a given bandwidth. The related deconvolution kernel $\mathcal{K}_\eta$ is then defined as

$$\mathcal{F}[\mathcal{K}_\eta](.) = \frac{\mathcal{F}[\mathcal{K}](.)}{\mathcal{F}[\eta](./\lambda)} \ \Leftrightarrow \mathcal{K}_\eta(.) = \int_{\mathbb{R}} e^{it.} \frac{\mathcal{F}[\mathcal{K}](t)}{\mathcal{F}[\eta](t/\lambda)} dt. \tag{1.8}$$

A particular interesting example is the sinc kernel, defined as $\mathcal{K}(x) = \sin(\pi x)/(\pi x)$ for all $x \in \mathbb{R}$. Indeed, with such a kernel, we get

$$\mathcal{F}[\mathcal{K}](t) = \mathbf{1}_{\{t\in[-1;1]\}}, \text{ and } \mathcal{K}_\eta(.) = \int_{-1/\lambda}^{1/\lambda} e^{ihs.} \mathcal{F}^{-1}[\eta](s) ds. \tag{1.9}$$

The deconvolution kernel $\mathcal{K}_\eta$ hence corresponds in this case to a projection (spectral cut-off) estimator of f. Remark that there a strong analogy with the spectral cut-off estimator defined in (1.5) in the Gaussian white noise model.

Now, we have all the requirements in order to construct a (deconvolution) estimator of the unknown density of f from the noisy data (1.7). For all $x \in \mathbb{R}$, define $\hat{f}_\lambda(x)$ as

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^n \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right).$$

Then, we can investigate basic properties of this estimator. Concerning the expectation of $\hat{f}_\lambda$, for all $x \in \mathbb{R}$, we get

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_\lambda(x)\right] &= \frac{1}{\lambda}\mathbb{E}\left[\mathcal{K}_\eta\left(\frac{Z_1 - x}{\lambda}\right)\right], \\
&= \frac{1}{\lambda}\mathbb{E}\left[\int_{\mathbb{R}} e^{it\left(\frac{Z_1 - x}{\lambda}\right)} \frac{\mathcal{F}[\mathcal{K}](t)}{\mathcal{F}[\eta](t/\lambda)} dt\right], \\
&= \frac{1}{\lambda}\mathbb{E}\left[\int_{\mathbb{R}} e^{it\left(\frac{X_1 + \epsilon_1 - x}{\lambda}\right)} \frac{\mathcal{F}[\mathcal{K}](t)}{\mathcal{F}[\eta](t/\lambda)} dt\right], \\
&= \frac{1}{\lambda}\mathbb{E}\left[\int_{\mathbb{R}} e^{it\left(\frac{X_1 - x}{\lambda}\right)} \frac{\mathcal{F}[\mathcal{K}](t)}{\mathcal{F}[\eta](t/\lambda)} \mathbb{E}\left[e^{it\epsilon_1/\lambda}/X_1\right] dt\right], \\
&= \frac{1}{\lambda}\mathbb{E}\left[\int_{\mathbb{R}} e^{it\left(\frac{X_1 - x}{\lambda}\right)} \mathcal{F}[\mathcal{K}](t) dt\right], \\
&= \frac{1}{\lambda}\mathbb{E}\left[\mathcal{K}\left(\frac{X_1 - x}{\lambda}\right)\right].
\end{aligned} \tag{1.10}$$

Indeed, for all $t \in \mathbb{R}$,

$$\mathbb{E}\left[e^{it\epsilon_1/\lambda}/X_1\right] = \int_{\mathbb{R}} e^{\frac{its}{\lambda}} \eta(s) ds = \mathcal{F}[\eta](t/\lambda).$$

Hence, the expectation of is exactly the same than in the free-noise case ($\epsilon_i = 0$):

$$\mathbb{E}\left[\hat{f}_\lambda(x)\right] = \mathbb{E}\left[\tilde{f}_\lambda(x)\right] = \frac{1}{\lambda}\mathbb{E}\left[\mathcal{K}\left(\frac{X_1 - x}{\lambda}\right)\right] = \mathcal{K}_\lambda * f(x), \ \forall x \in \mathbb{R}.$$

We can now describe the behavior of the quadratic risk associated to such kind of estimator. For the sake of readability , we will deal with the *sinc* kernel introduced above. Using similar computations, we get

$$\begin{aligned}
\mathbb{E}\|\hat{f}_h - f\|^2 &= \|\mathcal{K}_\lambda * f - f\|^2 + \mathbb{E}\|\hat{f}_h - \mathcal{K}_\lambda * f\|^2, \\
&\leq \|\mathcal{K}_\lambda * f - f\|^2 + \frac{1}{\lambda^2}\int_{\mathbb{R}} \left|K_\eta^2(z/\lambda)\right|^2 dz, \\
&\leq \underbrace{\|\mathcal{K}_\lambda * f - f\|^2}_{\text{Bias}} + \underbrace{\frac{C}{\lambda} \sup_{t\in[-1/\lambda;1/\lambda]} \mathcal{F}^{-2}[\eta](t)}_{\text{Variance}},
\end{aligned}$$

for some constant $C > 0$. In the free noise case (i.e. when $\epsilon = 0$), the variance term can be bounded by $1/\lambda$. When controlling the quadratic risk in an estimation purpose, one has to find a trade-off between bias and variance under some smoothness assumptions in order to derive a minimax rate of convergence. In the error in variable model, the variance is larger since it involves $\mathcal{F}^{-1}[\eta](t)$, the inverse of Fourier transform of η. This term explodes as $t \to +\infty$. In such a case, it is necessary to introduce assumptions on the behavior of this term (and conditions on the kernel K) in order to get rates of convergence.

A large amount of statistical issues can be investigated when dealing with the model (1.7). We refer for instance to [49] for seminal investigations in this context, [20], [18], [37] for recent contributions, [81] for a review of existing method in an estimation purpose or [17], [57] in a goodness-of-fit testing setting. Some contributions in binary supervised classification and testing theory are provided in Section 1.2.

1.1.3 Shifted curves model

The shifted curves model appears to be an interesting example of inverse problems with random operator. It provides surprising and interesting outcomes. In this setting, we assume that we can observe n noisy curves, which may be seen as realisations of the following processes

$$dY_j(t) = f(t - \tau_j)dt + \epsilon dW_j(t), \ j = 1 \ldots n, \tag{1.11}$$

where $f \in L^2([0,1])$ denotes the unknown 1-periodic function of interest (common to all observations), ϵ a positive noise level and the W_j independent Brownian motions. The sequence $(\tau_j)_j$ represents i.i.d. random variables. These variables are supposed to admit a common **known** density g w.r.t. the Lebesgue measure, but are not observable. For such a model, one can alternatively be interested in the estimation of the law of the shifts as in [24], [35] [11], [91] or in the estimation of the common shape f as in [10]. We will focus on this last task in the sequel.

In this context, in an asymptotic purpose, we will assume that the noise level is fixed, but that the number n of observed curves tends to infinity. Following [10], it is possible to prove that the related estimation problem can be characterized as an inverse problem. We will not dwell into details concerning this assertion. We will just remark that for all $j \in \{1, \ldots, n\}$ and $t \in \mathbb{R}$, we have $\mathbb{E}[Y_j(t)] = f * g(t)$ where $f * g$ denotes the convolution product between f and g. Hence, computing for instance the empirical Fourier coefficient of the observations will not suffice. We have to introduce a regularization (deconvolution) step.

In the following, for all $k \in \mathbb{Z}$, we denote by θ_k the k^{th} Fourier coefficient of the function f, defined as

$$\theta_k = \int_0^1 e^{-2ik\pi x} f(x)dx.$$

Hence, using the model (1.11) and the properties of the Gaussian white noise, we can observe, for all $k \in \mathbb{Z}$, $j \in \{1, \ldots, n\}$,

$$c_{j,k} := \int_0^1 e^{-2ik\pi x} dY_j(x) = \theta_k e^{-i2\pi k\tau_j} + \epsilon z_{k,j} \tag{1.12}$$

where $z_{k,j}$ are i.i.d. $\mathcal{N}_{\mathbb{C}}(0,1)$ variables, i.e. complex Gaussian random variables with zero mean and such that $\mathbb{E}|z_{k,j}|^2 = 1$. Then, for all $k \in \mathbb{Z}$, computing the empirical mean of the $c_{j,k}$, we get

$$y_k := \frac{1}{n}\sum_{j=1}^{n} c_{j,k} = \theta_k \times \frac{1}{n}\sum_{j=1}^{n} e^{-i2\pi k\tau_j} + \epsilon\frac{1}{n}\sum_{i=1}^{n} z_{k,j} := \tilde{\gamma}_k\theta_k + \frac{\epsilon}{\sqrt{n}}\xi_j, \tag{1.13}$$

with

$$\tilde{\gamma}_k = \frac{1}{n}\sum_{j=1}^{n} e^{-i2\pi k\tau_j} \ \forall k \in \mathbb{Z},$$

and where the ξ_k are i.i.d. standard Gaussian (complex) random variables. The model (1.13) more or less corresponds to the sequence space model associated to (1.11). Remark that using this formulation, we can find strong analogies with the Gaussian white noise model described in (1.4). The main difference is contained in the expression of the *eigenvalues* $\tilde{\gamma}_k$. Indeed, the $\tilde{\gamma}_k$ corresponds to the empirical Fourier coefficients of the density g which describes the law of the shifts. Since the realization of the shifts are not observable, the $\tilde{\gamma}_k$ are unknown. Hence, we can not use the sequence $(\tilde{\gamma}_k^{-1}y_k)_{k\in\mathbb{N}}$ in order to estimate the corresponding coefficient θ_k. In some sense, we are face to a random inverse problem, in the sense that the involved operator is random and unobservable.

In order to get round of this problem, we can remark that due to the law of large numbers

$$\tilde{\gamma}_k \xrightarrow{P} \gamma_k := \int_{\mathbb{R}} e^{itx} g(t)dt, \ \forall k \in \mathbb{Z}, \ \text{as } n \to +\infty,$$

where the convergence holds in probability, and the γ_k corresponds to the Fourier coefficients of the density g. Since this density is assumed to be known, we are able to compute these coefficients. In this context, each coefficient θ_k can be estimated by $\gamma_k^{-1}y_k$. Using (1.13), it appears that

$$\gamma_k^{-1}y_k = \theta_k + \frac{\epsilon}{\sqrt{n}}\gamma_k^{-1}\xi_k + \left(1 - \frac{\tilde{\gamma}_k}{\gamma_k}\right)\theta_k, \ k \in \mathbb{Z}. \tag{1.14}$$

Using this formulation we can remark that we deal with two random sources: the first one depends on both the noise level ϵ and the number of observed curves n. It is rather classical in the sense that it corresponds to the expression of the noise in standard non-parametric model (except that the noise level depends on two different parameters). The second random source is related to the fact that we approximate a random operator by a deterministic one. The corresponding approximation term only depends on n and has to be considered carefully.

As for the Gaussian white noise model, we can construct linear estimators from the sequence $\gamma_k^{-1}y_k$. Many statistical issues are then of first interest (estimation, model selection, test, etc...). Some of them will be briefly described below.

1.1.4 Econometric models

The econometric theory provides to the statistician several different models that are of first interest from a mathematical point of view. I will not try to provide an exhaustive list, but rather focus on one of them: the instrumental variable regression model.

We assume that we have at hand a sample $(Y_i, X_i, W_i)_{i=1\dots n}$ satisfying

$$Y_i = \varphi(X_i) + U_i \text{ and } \mathbb{E}[U_i/W_i] = 0 \ \forall i \in \{1, \dots, n\}, \tag{1.15}$$

where φ denotes the function of interest and U_i noise measurements. The sequence $(W_i)_{i=1\dots n}$ is called an instrument. When $W_i = X_i$, we obtain the classical model of regression with a random design. In an estimation purpose, one can use for instance a model selection approach: see for instance [3] or [1] among others. In the model (1.15), the noise $(U_i)_{i=1\dots n}$ is not assumed to be centered conditionally to the design. As proved in [32], we are in fact faced to an inverse problem.

In order to shed light on this assertion, first introduce the space L^2_X and L^2_W respectively defined as

$$L^2_X = \left\{ h : \mathbb{R} \to \mathbb{R}, \ \|h\|^2_X := \mathbb{E}[h^2(X)] < +\infty, \right\},$$

and

$$L^2_W = \left\{ g : \mathbb{R} \to \mathbb{R}, \ \|g\|^2_W := \mathbb{E}[g^2(W)] < +\infty, \right\}.$$

We denote by $\langle ., . \rangle_X$ and $\langle ., . \rangle_W$ the corresponding scalar products. Then, the model (1.15) can be rewritten as follows. For all $i \in \{1, \dots n\}$, we write

$$\begin{aligned} Y_i &= \varphi(X_i) + U_i, \\ &= \mathbb{E}[\varphi(X_i)/W_i] + \varphi(X_i) - \mathbb{E}[\varphi(X_i)/W_i] + U_i, \\ &= \mathbb{E}[\varphi(X_i)/W_i] + V_i, \end{aligned}$$

where

$$V_i = \varphi(X_i) - \mathbb{E}[\varphi(X_i)/W_i] + U_i \ \forall i \in \{1, \dots n\}.$$

In particular, we can remark that $\mathbb{E}[V_i/W_i] = 0$ for all $i \in \{1, \dots n\}$. Hence, for all φ, defining the operator T as

$$\begin{aligned} T: \ & L^2_X \to L^2_W \\ & \varphi \mapsto T\varphi(.) = \mathbb{E}[\varphi(X)/W = .], \end{aligned} \tag{1.16}$$

we can rewrite the model (1.15) as

$$Y_i = T\varphi(W_i) + V_i \ \forall i \in \{1, \dots, n\},$$

where the noise $(V_i)_{i=1\dots n}$ is centered conditionally to the design $(W_i)_{i=1\dots n}$.

We are in fact faced to an inverse regression model, where the operator T is unknown. Indeed, we can see from (1.16) that T explicitly depends on the joint distribution of the couple (X, W), which is not assumed to be known in practice. Hence, we have to provide an estimator for both the function φ of interest and the operator T from the same sample $(Y_i, X_i, W_i)_{i=1\dots n}$.

Concerning the operator, the corresponding estimation can be performed in an empirical way. Indeed, let $(\phi_k)_k$ and $(\psi_k)_k$ be two bases of respectively L^2_X and L^2_W. Then, we can express T via its corresponding representation matrix $T = (T_{ij})_{i,j\in\mathbb{N}}$, where $\forall i, j \in \mathbb{N}$,

$$\begin{aligned} T_{ij} = \langle T\phi_j, \psi_j \rangle_W &= \mathbb{E}[T\phi_i(W)\psi_j(W)], \\ &= \mathbb{E}[\mathbb{E}[\phi_i(X)/W]\psi_j(W)], \\ &= \mathbb{E}[\phi_i(X)\psi_j(W)]. \end{aligned}$$

Each coefficient of the matrix T can then be empirically estimated from the sample $(Y_i, X_i, W_i)_{i=1\dots n}$ by

$$\hat{T}_{ij} = \frac{1}{n} \sum_{i=1}^{n} \phi_i(X_i)\psi_j(W_i).$$

Then, one can address the following questions

- How can we estimate the whole operator from the pointwise estimation of the different coefficients?
- What kind of estimation algorithms can be proposed in such a setting?
- What is the influence of the empirical estimation of the operator on the quality of estimation?

Different kind of strategies have been proposed in the literature in an estimation purpose. We mention for instance [83], [23] or [53] among others. We will also present a small contribution at the end of this chapter.

1.2 Different contributions in these models

In this section, I present a brief description of my different contributions proposed over last years in statistical inverse problems. In particular, I have been interested in signal detection, classification and estimation with noise in the operator. These different frameworks cover all the models presented in Section 1.1.

1.2.1 Signal detection for inverse problems

While estimation is a quantitative problem (one want to estimate the whole signal), signal detection corresponds to a qualitative task. Given noisy observations, the aim is to determine whether these observations contain signal or not. In a non-parametric setting, this statistical problem has been mainly popularized by Y. Ingster in the 90's. We refer in particular to its seminal series of papers [60],[61] and [62] for a complete description.

The formal framework of this topic is the following. Given a sample of observations Y, involving a function of interest f, our aim is to test

$$H_0 : f = f_0, \text{ against } H_1 : f \neq f_0, \tag{1.17}$$

where f_0 denotes some given benchmark function. If we deal with the Gaussian white noise model (1.2), the particular case where $f_0 = 0$ corresponds to a signal detection problem: one want to assess whether we are observing signal or not. In the error-in-variable model (1.7), the function f_0 can correspond to a benchmark density. We know that under some particular condition, the density of the variable of interest should be f_0. In such a case, one want to verify this assertion.

As in classical parametric testing theory, we first have to construct a decision rule (a test), i.e. a measurable function of the data. By convention, a test Φ will take values in $\{0, 1\}$: we reject H_0 if $\Phi = 1$ and do not reject this hypothesis in the other case. Given a prescribed level $\alpha \in]0,1[$, we will only deal with level-α tests Φ_α satisfying

$$P_{H_0}(\Phi_\alpha = 1) \leq \alpha.$$

Then, given such a decision rule, one might want to investigate its power or its second kind error. If one use an alternative as expressed in (1.17), the control of this error will not be possible: the condition $f \neq f_0$ is indeed to rich (a similar phenomenon occurs even in the parametric setting). Hence, the alternative is typically expressed via two conditions on the signal

- a smoothness constraint: $f \in \mathcal{F}$ for some functional space $\mathcal{F} \subset \mathcal{X}$,
- a *signal-to-noise ratio* constraint that measures the amount of available signal in the observations.

In the following, we will test

$$H_0 : f = f_0, \text{ against } H_1 : f \in \mathcal{F} \text{ and } \|f - f_0\| > \rho, \tag{1.18}$$

where the parameter ρ measures in some sense the separability of the both hypotheses. The set $\mathcal{F}$ is assumed to be fixed (no dependency with respect to the data or the noise level).

A classical outcome in the (non-parametric) testing theory consists in investigating the lowest possible achievable separation radius ρ. More formally, given $\alpha, \beta \in]0,1[$ prescribed levels for the first and second kinds errors, and a testing procedure Φ_α, one can define the separation radius $\rho(\Phi_\alpha, \beta, \mathcal{F})$ associated to Φ_α as

$$\rho(\Phi_\alpha, \beta, \mathcal{F}) := \inf \left\{ \rho > 0 : \sup_{f \in \mathcal{F},\ \|f\| > \rho} P_f(\Phi_\alpha = 0) \leq \beta \right\}.$$

The quantity $\rho(\Phi_\alpha, \beta, \mathcal{F})$ corresponds to the smallest possible radius for which the second kind error can be controlled by β. The lowest possible separation radius $\rho(\alpha, \beta, \mathcal{F})$ is then defined as

$$\rho(\alpha, \beta, \mathcal{F}) = \inf_{\Phi_\alpha} \rho(\Phi_\alpha, \beta, \mathcal{F}),$$

where the infimum is taken over all possible level-α tests. This term $\rho(\alpha, \beta, \mathcal{F})$ is called the minimax separation rate (radius) over the set $\mathcal{F}$.

Given a smoothness constraint $\mathcal{F}$, one of the main goal is to establish a lower bound for this separation radius. Then, one might want to build testing procedures that will achieve this bound.

Signal detection for the sequence space model

Consider the sequence space model (1.4). The particular case where $A = \mathrm{Id}$, i.e. $b_k = 1$ for all $k \in \mathbb{N}$, has been widely investigated in the litterature: see for instance [60]-[62] or [3] in a non-asymptotic setting. In an inverse problem framework, few investigations have been proposed. Most of these contributions are concerned with the particular mildly ill-posed case: the sequence

$(b_k^2)_{k\in\mathbb{N}}$ is polynomially decreasing (see for instance [47]).

In Laurent et al. (2012), we propose a precise (non-asymptotic) study of the separation radii in this setting. We consider various kind of inverse problems (mildly and severely ill-posed) and different smoothness constraints (sparsity, ellipsoids, l^p-balls). For the sake of brevity, we will only discuss in this report smoothness constraints described as ellipsoids

$$\mathcal{F} = \mathcal{E}_{a,2}(R) := \left\{ \nu : \sum_{j=1}^{+\infty} a_k^2 \nu_k^2 \leq R \right\}, \tag{1.19}$$

for some non-decreasing sequence $a = (a_k)_{k\in\mathbb{N}}$ and a positive constant R. Without loss of generality, we deal with the case $f_0 = 0$.

We start with the study of the lower bound. We are interested in the value of

$$\beta(\mathcal{E}_{a,2}(R), \alpha) := \inf_{\Phi_\alpha} \sup_{f\in\mathcal{F},\ \|f\|>\rho} P_f(\Phi_\alpha = 0) \in [0; 1-\alpha],$$

In particular, an interesting outcome is to investigate the smallest possible value ρ for which the previous quantity can be, following the setting, lower bounded by a constant, equal to a prescribed level β or tends to 0.

A possible way to achieve this goal is to consider a probability measure π on the set $\mathcal{F}[\rho] := \{f \in \mathcal{F},\ \|f\| > \rho\}$. Denote by P_0 (resp. P_π) the measure associated to the observations vector Y when the sequence θ is equal to 0 (resp. follows the measure π). Then, following [4], we prove that

$$\beta(\mathcal{E}_{a,2}(R), \alpha) \geq 1 - \alpha - \frac{1}{2}\left(\mathbb{E}_0[L_\pi^2(Y)] - 1\right)^{1/2},$$

where $L_\pi(Y)$ denotes the likelihood ratio between the two measures P_0 and P_π. The construction of the lower bound can be reduced to the study of this likelihood ratio. In particular, two different regimes can be considered:

- The likelihood ratio tends to 1. In such a case, $\beta(\mathcal{E}_{a,2}(R), \alpha) \to 1 - \alpha$ as $\epsilon \to 0$ which means that both hypotheses H_0 and H_1 are not separable.

- The likelihood ratio can be bounded by a constant. In this case, the second kind error is also lower bounded by a constant. An interesting situation corresponds to the case where $\mathbb{E}_0[L_\pi^2(Y)]$ is (asymptotically) bounded by $1 + 4(1 - \alpha - \beta)$ for some $\beta \in]0, 1[$. Then, $\beta(\mathcal{E}_{a,2}(R)$ is (asymptotically) lower bounded by β. This last case is of first interest if one want to establish a radius that guarantee the separation of both hypotheses with prescribed error levels.

The main conclusion of the discussion above is that the construction of the lower bound heavily rely to the construction of a prior π on the set $\mathcal{F}$. In the following, we consider the symmetric prior π defined as

$$\pi = \prod_{k\in\mathbb{N}} \pi_k, \text{ where } \pi_k = \frac{1}{2}(\delta_{-b_k\theta_k} + \delta_{b_k\theta_k})\ \forall k \in \mathbb{N},$$

for some sequence $\theta = (\theta_k)_{k\in\mathbb{N}}$ which will be made explicit below. Since the ξ_k are Gaussian random variable, we obtain after some technical algebra

$$\mathbb{E}_0[L_\pi^2(Y)] = \prod_{k\in\mathbb{N}} \cosh(b_k^2\theta_k^2/\epsilon^2) \leq \exp\left(\frac{1}{2\epsilon^4}\sum_{k\in\mathbb{N}} b_k^4\theta_k^4\right) := \exp(u_\epsilon^2(\theta)). \tag{1.20}$$

Then, we have to find an explicit sequence θ^0 for which the previous quantity is bounded by $1 + 4(1-\alpha-\beta)$ where β denotes a prescribed level for the second kind error. To this end, we have considered in Laurent et al. (2011) the sequence θ^0 defined as

$$\theta_k^0 := \begin{cases} \dfrac{\rho\epsilon^2 b_k^{-2}}{\left(\epsilon^4 \displaystyle\sum_{j=1}^{D} b_k^{-4}\right)^{1/2}}, \ \forall k \in \{1,\dots,D\}, \\ \\ 0, \quad \forall k > D, \end{cases} \tag{1.21}$$

for some parameter D. In particular, we get from (1.20), (1.21) that $\|\theta^0\|^2 = \rho^2$ and

$$\mathbb{E}_0[L_\pi^2(Y)] \leq \exp\left(\frac{1}{2\epsilon^4}\sum_{k\in\mathbb{N}} b_k^4\theta_k^4\right) = \exp\left[\frac{\rho^4}{\epsilon^4\sum_{j=1}^{D} b_k^{-4}}\right] \leq 1 + 4(1-\alpha-\beta)^2,$$

as soon as

$$\rho^2 = \rho_D^2 := c(\alpha,\beta)\epsilon^2\sqrt{\sum_{j=1}^{D} b_j^4}, \tag{1.22}$$

for some constant $c(\alpha,\beta)$ which can be explicitly computed. In order to conclude, it remains to choose an appropriate D such that $\theta^0 \in \mathcal{E}_{a,2}(R)$. To this end, remark that

$$\sum_{k\in\mathbb{N}} a_k^2(\theta_k^0)^2 \leq a_D^2\sum_{j=1}^{D}(\theta_k^0)^2 = a_D^2\rho_D^2 \leq R \text{ as soon as } \rho_D^2 \leq a_D^{-2}R.$$

Hence, if we define

$$\rho_{inf}^2 := \sup_{D\in\mathbb{N}}\left[c(\alpha,\beta)\epsilon^2\sqrt{\sum_{j=1}^{D} b_j^{-4}} \wedge R^2 a_D^{-2}\right], \tag{1.23}$$

we get

$$\inf_{\Phi_\alpha}\sup_{f\in\mathcal{F}[\rho_{inf}]} P_f(\Phi_\alpha = 0) > \beta,$$

which means that

$$\rho(\mathcal{E}_{a,2}(R),\alpha,\beta) \geq \rho_{inf}.$$

This corresponds to a non-asymptotic lower bound. The main advantage of such a bound is that it provides a precise and intuitive description of the limitation of a given test. Indeed, the lower bound (1.23) can be seen as a trade-off between an approximation term $R^2a_D^{-2}$ and a standard-deviation term $\epsilon^2\sqrt{\sum_{j=1}^{D} b_j^{-4}}$ related to the estimation of $\|\theta\|^2$ by $\sum_{j\leq D} b_k^{-2}y_k^2$ for a given dimension D. Bellow, we propose a corresponding upper bound on this separation rate.

In order to take a decision coherent with the problem (1.18), a possible way is to construct an estimator of $\|f\|^2$. Indeed, if one find a large enough value for this estimator (w.r.t. a prescribed threshold: see below for more details), we are possibly observing (at least) a small amount of signal. In order to estimate $\|f\|^2$, we will use a projection (spectral cut-off) scheme

$$T_D = \sum_{j=1}^{D} b_j^{-2}(y_j^2 - \epsilon^2), \tag{1.24}$$

for some bandwidth $D \in \mathbb{N}$. Then, we define the test $\Phi_{\alpha,D}$ as

$$\Phi_{\alpha,D} = \mathbf{1}_{\{T_D > t_{\alpha,D}\}}, \tag{1.25}$$

where $t_{\alpha,D}$ denotes the $1-\alpha$ quantile of T_D under H_0. Now, one want to express conditions on $\|f\|$ for which the second kind error can be controlled, namely condition for which

$$P_f(\Phi_{\alpha,D} = 0) = P_f(T_D \leq t_{\alpha,D}) \leq \beta. \tag{1.26}$$

For all $f \in \mathcal{X}$, define $t_{\beta,D}(f)$ as the β-quantile of the variable T_D. In particular $P_f(T_D \leq t_{\beta,D}(f)) \leq \beta$. Hence, in order to verify (1.26), it suffices to find condition for which

$$t_{\alpha,D} \leq t_{\beta,D}(f).$$

In Laurent et al. (2012), we have established non-asymptotic upper and lower bounds for respectively $t_{\alpha,D}$ and $t_{\beta,D}(f)$. Both bounds are summarized in the following proposition.

Proposition 2 *Let $t_{\alpha,D}$ and $t_{\beta,D}(\theta)$ the two quantiles defined above. There exists a constant $C(\alpha)$ such that*

$$t_{\alpha,D} \leq \epsilon^2 \sum_{j=1}^{D} b_j^{-2} + C(\alpha)\epsilon^2 \left(\sum_{j=1}^{D} b_j^{-4} \right)^{1/2},$$

and

$$t_{\beta,D} \geq \sum_{j=1}^{D} \theta_j^2 + \epsilon^2 \sum_{j=1}^{D} b_j^{-2} + 2\sqrt{\ln(1/\beta)} \sqrt{\epsilon^4 \sum_{j=1}^{D} b_j^{-4} + 2\epsilon^2 \sum_{j=1}^{D} b_j^{-2}\theta_j^2}.$$

The previous proposition provide a pertinent upper bound. It matches the lower bound proposed above up to constants. As a matter of fact,

$$\sum_{j=1}^{D} \theta_j^2 \quad \geq \quad C(\alpha)\epsilon^2 \left(\sum_{j=1}^{D} b_j^{-4} \right)^{1/2} - 2\sqrt{\ln(1/\beta)} \sqrt{\epsilon^4 \sum_{j=1}^{D} b_j^{-4} - 2\epsilon^2 \sum_{j=1}^{D} b_j^{-2}\theta_j^2}$$

implies that

$$t_{\alpha,D} \leq t_{\beta,D}(f).$$

A precise investigation of the previous inequality leads to an upper bound for the separation radius associated to the test $\Phi_{\alpha,D}$ similar to (1.23) . In this study, the constants do not match, but are all explicit. For a precise study of the optimal constant associated to the minimax separation radii, we refer to [63]. These results can be summarized in the following theorem.

Theorem 1 *Let α, β be fixed and denote by $\rho_2(\mathcal{E}_{a,2}(R), \alpha, \beta)$ the minimax rate of testing over $\mathcal{E}_{a,2}(R)$ with respect to the l_2 norm. Then*

$$\rho_2^2(\mathcal{E}_{a,2}(R), \alpha, \beta) \geq \sup_{D \in \mathbb{N}} (\rho_D^2 \wedge R^2 a_D^{-2}),$$

where ρ_D^2 has been introduced (1.22). Moreover, for all $D \in J$,

$$\sup_{\theta \in \mathcal{E}_{a,2}(R), \|\theta\|_2^2 \geq C\rho_D^2 + R^2 a_D^{-2}} P_\theta(\Phi_\alpha = 0) \leq \beta,$$

where $C = C(\alpha, \beta)$ is a positive constant depending only on α and β and Φ_α denotes the test introduced in (1.25). Hence,

$$\rho_2^2(\mathcal{E}_{a,2}(R), \alpha, \beta) \leq \inf_{D \in \mathbb{N}} (C\rho_D^2 + R^2 a_D^{-2}),$$

Remark that the result presented in this theorem is non-asymptotic: we do not require that $\epsilon \to 0$ in order to characterize the behavior of the separation radius. Nevertheless, this result can also be used in order to get asymptotic minimax separation rates, as soon as we set specific constraints on the sequences $a = (a_k)_{k \in \mathbb{N}}$ and $b = (b_k)_{k \in \mathbb{N}}$. The following table summarizes the asymptotic rates that we have obtained in this setting. They are similar to those obtained by [63] or [17] in an error-in-variable framework.

	Mildly ill-posed	**Severely ill-posed**
	$b_k \sim k^{-t}$	$b_k \sim \exp(-\gamma k^r)$
$a_k \sim k^s$	$\sigma^{\frac{4s}{2s+2t+1/2}}$	$\left(\log(\sigma^{-2})\right)^{-2s/r}$
$a_k \sim \exp(\nu k^s)$	$\sigma^2 \left(\log(\sigma^{-2})\right)^{(2t+1/2)/s}$	$e^{-2\nu \tilde{D}^s}$ $(s \leq 1)$

Figure 1.1: *Asymptotic minimax separation rates for the l^2-norm. Here $\tilde{D}$ denotes the integer part of the solution of $\rho_D^2 = R^2 a_D^{-2}$.*

In Laurent et al. (2012), we have investigated the separation radii in various settings (sparse signal, ellipsoids, l^p balls) and derived corresponding asymptotic testing rates. We refer to the corresponding paper for more details.

Multi-dimensional case

The previous study was only concerned with the uni-dimensional setting. In Ingster et al. (2013) we have investigated the separation rates related to the sequence model (1.4) in a multidimensional case. In particular, index are $\mathbb{N}^d$ valued for some $d \geq 1$, which allows to model more general situations.

In such a setting, we have adopted an asymptotic point of view. Smoothness constraints in the alternative are expressed via ellipsoids $\mathcal{E}_{a,2}(R)$ defined as

$$\mathcal{E}_{a,2}(R) = \left\{ \nu \in l^2(\mathbb{N}^d), \ \sum_{l \in \mathbb{N}^d} a_l^2 \nu_l^2 \leq R \right\},$$

for some sequence $(a_l)_{l\in\mathbb{N}^d}$. We have considered two different possible behaviors for this sequence a. The ellipsoid $\mathcal{E}_{a,2}(R)$ corresponds to a *Tensor product space* when

$$a_l^2 = a_{l_1\ldots l_d}^2 = \prod_{j=1}^d |l_j|^{2s_1}, \ \forall l \in \mathbb{N}^d,$$

and a so-called *Sobolev space* when

$$a_l^2 = a_{l_1\ldots l_d}^2 = \sum_{j=1}^d |l_j|^{2s_1}, \ \forall l \in \mathbb{N}^d,$$

for some $s = (s_1, \ldots, s_d)$.

In this setting, we investigate minimax asymptotic separation rates. Concerning the lower bound, we start from (1.20) where we have seen (the generalization to the multi-dimensional case is straightforward) that

$$\mathbb{E}_0[L_\pi^2(Y)] \leq \exp\left(\frac{1}{2\epsilon^4}\sum_{k\in\mathbb{N}^d} b_k^4\theta_k^4\right) := \exp(u_\epsilon^2(\theta)).$$

In Laurent et al. (2012) (see also (1.21)), we have constructed an explicit sequence $(\theta_k^0)_{k\in\mathbb{N}}$ in order to guarantee a prescribed error for the second kind error. An alternative idea consists in finding the smallest possible value of $u_\epsilon^2(\theta)$ for which $\theta \in \mathcal{E}_{a,2}(R)$. In other words, one choose the sequence $\theta = \theta(\rho_\epsilon)$ such that

$$\theta = \theta(\rho_\epsilon) = \arg\inf_\theta \left\{ u_\epsilon^2(\theta) := \frac{1}{2\epsilon^4}\sum_{k\in\mathbb{N}} b_k^4\theta_k^4, \ \text{s.t. } \|\theta\| = \rho_\epsilon \text{ and } \theta \in \mathcal{E}_{a,2}(R) \right\}. \tag{1.27}$$

This principle has been in particular developed in the series of papers [60], [61], [62], or more recently in [63] in an inverse problem framework. The optimization problem (1.27) is often called *extremal problem* in the aforementioned literature.

Concerning the upper bound, we do not use a spectral cut-off regularization scheme. We deal instead with general filters $\lambda = (\lambda_k)_{k\in\mathbb{N}^d}$ where $\lambda_k \in [0,1]^d$ for all $k \in \mathbb{N}^d$. The related test is defined as

$$\Phi_\alpha = \mathbf{1}_{\{\sum_{k\in\mathbb{N}^d} \lambda_k^2 y_k^2 > t_{\alpha,\lambda}\}},$$

for some threshold $t_{\alpha,\lambda}$. For the sake of brevity, we will not dwell into details. It it is nevertheless possible to prove that finding an appropriate sequence λ can be explicitly related to the extremal problem (1.27). Hence, all the analysis reduces to the investigation of the behavior of this sequence. In particular, we are looking for situations where $u_\epsilon^2(\theta)$ is constant and we provide the associated value for the separation radius ρ_ϵ. We refer for Ingster et al. (2013) for more details and an exhaustive list of related separation rates.

General regularization schemes

In the two previous sections, we were concerned with the sequence space model (1.4). Nevertheless, there exists several situations for which this sequence is not available, or at least up to a large amount of computation time. This is in particular the case when the bases $(\phi_k)_k$ and $(\psi_k)_k$ are unknown or difficult to handle.

In Marteau and Mathé (2013), our aim was to construct and study tests that do not necessary use this sequence model. Our testing procedures are based on regularization families (see Definition 1). Let g_τ a regularization method. The term $\hat{f}_\tau := g_\tau(A^\star A)A^\star Y$ is an estimator of f. In particular, we have

$$\mathbb{E}[\hat{f}_\tau] = \mathbb{E}[g_\tau(A^\star A)A^\star(Af+\epsilon\xi)] = g_\tau(A^\star A)A^\star Af := f_\tau.$$

Following the approach presented in (1.24), one can estimate the norm $\|f\|^2$ by $\|g_\tau(A^\star A)A^\star Y\|^2$. In particular

$$\begin{aligned} \mathbb{E}_f\|g_\tau(A^\star A)A^\star Y\|^2 &= \|g_\tau(A^\star A)A^\star Af\|^2 + \sigma^2\mathbb{E}\|g_\tau(A^\star A)A^\star\xi\|^2, \\ &:= \|f_\tau\|^2 + \sigma^2\mathrm{tr}(g_\tau(A^\star A)A^\star A), \end{aligned}$$

where $\mathrm{tr}(B)$ denotes the trace of a given operator B. Thanks to the properties introduced in Definition 1, $\|f_\tau\|^2 \to \|f\|^2$ as $\tau \to 0$. The term $\|f_\tau\|^2$ can hence be seen as an approximation of $\|f\|^2$.

In the following, we will introduce the terms S_τ and v_τ defined as follows

$$S_\tau^2 = \sigma^2\mathrm{tr}(g_\tau(A^\star A)A^\star A) \text{ and } v_\tau^2 = \sigma^2\|g_\tau(A^\star A)A^\star\|^2.$$

In fact, theses terms respectively correspond to the variance and weak variance of $g_\tau(A^\star A)A^\star Y$. Then, we consider the testing procedure $\Phi_{\alpha,\tau}$ defined as

$$\Phi_{\alpha,\tau} = \mathbf{1}_{\{\|g_\tau(A^\star A)A^\star Y\|^2 - S_\tau^2 > t_{\alpha,\tau}\}}, \tag{1.28}$$

for some threshold $t_{\alpha,\tau}$. Using standard tools in operator theory, one can generalize the construction of separation radii presented above. We obtain the following result.

Proposition 3 *Consider the test $\Phi_{\alpha,\tau}$ as introduced in (1.28), and let*

$$r^2(\Phi_{\alpha,\tau},\beta) := C^*_{\alpha,\beta}\epsilon^2\frac{\sqrt{\mathcal{N}(\tau)}}{\tau} + (4x_\alpha + 8x_\beta)\frac{\epsilon^2}{\tau}, \tag{1.29}$$

where $C^\star_{\alpha,\beta}$ denotes a positive constant and

$$\mathcal{N}(\tau) := \mathrm{tr}\left[(A^\star A + \lambda I)^{-1}A^\star A\right].$$

Then

$$\sup_{f,\|f_R\|^2 \geq r^2(\Phi_{\alpha,\tau},\beta)} P_f(\Phi_{\alpha,\tau} = 0) \leq \beta.$$

For all $\tau > 0$, the term $\mathcal{N}(\tau)$ is called *effective dimension*. In particular, it is easy to see that, since A is a compact operator, $\mathcal{N}(\tau) \to +\infty$ as $\tau \to 0$. Hence, the main dominating term in (1.29) is $\sqrt{\mathcal{N}(\tau)}/\tau$. This result can be compared to the one obtained in the sequence model, with projection (spectral cut-off) regularization scheme. As presented in (1.23), the radius is of order

$$\sigma^2 \sqrt{\sum_{j=1}^{D^\star} b_j^4} \leq \sigma^2 b_{D^\star}^{-2} \sqrt{D^\star},$$

for some $D^\star > 0$. In this particular case, one can compare respectively $\mathcal{N}(\tau)$ to $D^\star$, and $b_{D^\star}$ to $1/\tau$, where $D^\star$ corresponds to the *dimension* on which the estimation is performed.

As in the aforementioned contributions, we can then rely these radii to smoothness constraints on the target f. To this end, we use in Marteau and Mathé (2013) source conditions and investigate properties of regularization schemes trough there qualification. Some attention is also paid to the adaptation issue, i.e. to the construction of a data driven algorithm for the choice of τ.

Testing as a direct problem

In the testing problem (1.18), our aim is to determine whether there is signal in our observations, namely if $f = 0$ or not. First notice that a compact operator is always injective. Hence, both assertions $f = 0$ and $Af = 0$ are equivalent. Concerning testing theory, two problems are in fact at hand. One might want to consider the inverse problem, as described above, where one test

$$H_0^{IP} : f = 0, \text{ against } H_1^{IP} : f \in \mathcal{F}, \ \|f\| > \rho^{IP}, \tag{1.30}$$

or consider the direct problem where one want to test

$$H_0^{DP} : Af = 0, \text{ against } H_1^{DP} : f \in \mathcal{F}, \ \|Af\| > \rho^{DP}. \tag{1.31}$$

These hypotheses are not equivalent. Indeed, alternatives are not expressed in the same way. This question has been adressed for the first time in [57]. A theoretical comparison of these two points of view has been provided in Laurent et al. (2011) for the sequence space model. In particular, when $\mathcal{F}$ corresponds to an ellipsoid, we have proved that

- A test minimax for the testing problem (1.31) is always minimax for the testing problem (1.30),
- There exists tests minimax for the testing problem (1.30) that are not minimax for (1.31).

Such a results hence indicates that both direct and indirect testing problesm are not equivalent. Moreover, the regularization of the problem does not appear to be necessary for a signal detection purpose.

We will not dwell into details since this discussion is developed in Chapter 2. We just mention that related discussions are proposed in Marteau and Mathé (2013) for general regularization schemes or in Loubes and Marteau (2013) in an error-in-variable model.

1.2.2 Supervised classification with error-in-variables

Binary supervised classification has been widely investigated over last years. Since a wide literature is available, we will only mention [16] or [38] among others for a complete introduction to this topic.

We will focus here on a particular framework: the smooth discriminant analysis which has been popularized by [76]. The classical setting is the following: we have at our disposal two samples $\mathcal{S}_1 = (X_1^{(1)}, \dots, X_n^{(1)})$ and $\mathcal{S}_2 = (X_1^{(2)}, \dots, X_n^{(2)})$ where the $X_i^{(1)}$ (resp. $X_i^{(2)}$) admit a probability density f (resp. g) w.r.t. a given measure Q on $\mathbb{R}^d$. This reference measure Q is assumed to be σ-finite w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Given a new incoming observation X, the goal is to determine whether $X \sim f$ or $X \sim g$. In this context, a decision rule (namely a classifier) is associated to a set $G \subset \mathbb{R}^d$ where we attribute the density f to X if $X \in G$, and g in the other case. The performances of a given classifier can then be measured via its Bayes risk $R_K(G)$ defined as

$$R_K(G) = \int_{K/G} f(x)dQ(x) + \int_G g(x)dQ(x).$$

In this context, the best possible classifier $G_K^\star$ is

$$G_K^\star = \arg\min_{G \subset K} R_K(G) = \{x \in K : \ f(x) > g(x)\},$$

where the infimum is taken over all possible subsets of K. In practice, $G_K^\star$ can be considered as an oracle: it corresponds to the best possible strategy. Nevertheless, it is not available since the underlying densities f and g are unknown. One of the main goal in smooth discriminant analysis is then to provide an estimator $\hat{G}_{n,m}$ of this oracle $G_K^\star$: we are in fact faced to a non-parametric set estimation problem. In general, $R_K(G_K^\star)$ does not tends to 0 as $n, m \to +\infty$. The performances of a given *estimator* $\hat{G}_{n,m}$ are hence measured in terms of its associated excess risk

$$R_K(\hat{G}_{n,m}) - R_K(G_K^\star).$$

In particular, since for all $G \subset K$, $f - g$ and $\mathbf{1}_{G_K^\star} - \mathbf{1}_G$ have the same sign, we get

$$\begin{aligned} R_K(\hat{G}_{n,m}) - R_K(G_K^\star) &= \int (f-g)(\mathbf{1}_{G_K^\star} - \mathbf{1}_{\hat{G}_{n,m}})dQ, \\ &= \int |f-g||\mathbf{1}_{G_K^\star} - \mathbf{1}_{\hat{G}_{n,m}}|dQ, \\ &:= d_{f,g}(G_K^\star, \hat{G}_{n,m}). \end{aligned}$$

This pseudo-distance $d_{f,g}$ will be of first interest in the following. We will also sometimes use $d_\Delta(.,.)$ defined as

$$d_\Delta(G_1, G_2) = \int |\mathbf{1}_{G_1} - \mathbf{1}_{G_2}|dQ := \int_{G_1 \Delta G_2} dQ, \ \forall G_1, G_2 \subset K$$

where $G_1 \Delta G_2$ denotes the symmetric difference between G_1 and G_2. In this setting, asymptotics of minimax excess risk over specified conditions have been investigated. We refer to [76] for a seminal study.

Few years ago, we were wondering with S. Loustau whether classical algorithms in supervised binary classification could take into account measurements errors as described in Section 1.1.2. In order to begin our investigation in that direction, we have decided to start our study with this smooth discriminant analysis model. We assume that we only have at our disposal noisy samples $\mathcal{S}_1 = (Z_1^{(1)}, \dots, Z_n^{(1)})$ and $\mathcal{S}_2 = (Z_1^{(2)}, \dots, Z_m^{(2)})$, where

$$Z_i^{(j)} = X_i^{(j)} + \epsilon_i^{(j)}, \ \forall j \in \{1, 2\}, \tag{1.32}$$

and the $\epsilon_i^{(j)}$ denotes i.i.d. random variables having a *known* density η w.r.t. the Lebesgue measure. In this context, our aim is to provide estimators for $G_K^\star$ and to investigate related performances.

This question has been addressed in Loustau and Marteau (2013a) and Loustau and Marteau (2013b) for a general measure Q. For the sake of convenience, we will assume in this manuscript that Q is the Lebesgue measure.

A deconvolution classifier

The first step consists in proposing an estimator for $G_K^\star$. In the free-noise (direct) case, the performances of ERM (empirical risk minimizer) algorithms have been for instance investigated in [76]. The main idea is to construct an estimator of the risk $R_K(.)$ and then to minimize this estimator over all possible subsets $G \subset K$. Given $G \subset K$, the corresponding risk $R_K(G)$ can be estimated by

$$\tilde{R}_{n,m}(G) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i^{(1)} \in K/G\}} + \frac{1}{m} \sum_{j=1}^m \mathbf{1}_{\{X_j^{(2)} \in G\}}.$$

Indeed, we have for instance

$$\mathbb{E}\left[\mathbf{1}_{\{X_i^{(1)} \in G\}}\right] = \int_G f(x)dx, \ \forall i \in \{1, \dots, n\}, \text{ and } \forall G \subset K.$$

Hence, $\tilde{R}_{n,m}(G)$ is an unbiased estimator of $R_K(G)$.

In the case where observations are described by the model (1.32), we can not just replace the $X_i^{(j)}$ by the $Z_i^{(j)}$. Indeed, the density associated to the $Z_i^{(1)}$ (resp. $Z_i^{(2)}$) corresponds to the convolution product $f * \eta$ (resp. $g * \eta$). Hence

$$R_{n,m}(G) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{Z_i^{(1)} \in K/G\}} + \frac{1}{m} \sum_{j=1}^m \mathbf{1}_{\{Z_j^{(2)} \in G\}} \xrightarrow{P} \frac{1}{2} \int_{K/G} f * \eta(x)dx + \frac{1}{2} \int_G g * \eta(x)dx,$$

as $n, m \to +\infty$. The right-hand side of the previous equation corresponds to the *prediction* risk, and is in general different from $R_K(G)$.

In order to get round of this problem, the idea is to construct a function $h_G(.)$ such that for all $G \subset K$, $h_G(Z_i^{(j)})$ will be close (in a sense which will be made precise later on) of $\mathbf{1}_{\{X_i^{(j)} \in G\}}$. In other words, we have to introduce a deconvolution step in the ERM algorithm.

Let $\mathcal{K}$ be a kernel and $\lambda \in [0,1]^d$ a bandwidth. We can then denote by $\mathcal{K}_\eta$ the deconvolution kernel defined as (see also Section 1.1.2)

$$\mathcal{F}[\mathcal{K}_\eta](t) = \frac{\mathcal{F}[\mathcal{K}](t)}{\mathcal{F}[\eta](t/\lambda)},\ \forall t \in \mathbb{R}^d. \tag{1.33}$$

Then, for all $G \subset K$, we can define the function $h_G(.)$ as

$$h_G(x) = \int_G \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx,\ \forall x \in \mathbb{R}^d. \tag{1.34}$$

The following lemma sheds light on the behavior of this functions. The proof is a straightforward extension of the one proposed in (1.10).

Lemma 1 *Let $G \subset K$. Then, for all $j \in \{1,2\}$, $i \in \{1,\dots,n_j\}$ we have*

$$\mathbb{E}\left[h_G(Z_i^{(j)})/X_i^{(j)}\right] = \int_G \frac{1}{\lambda}\mathcal{K}\left(\frac{X_i^{(j)}-x}{\lambda}\right)dx := \mathcal{K}_\lambda * \mathbf{1}_G(X_i^{(j)}),\ \forall x \in \mathbb{R}^d,$$

where the $Z_i^{(j)}$ are defined in (1.32) and $\mathcal{K}_\lambda = \lambda^{-1}\mathcal{K}(./\lambda)$.

The main conclusion of Lemma 1 is that the function $h_G(.)$ may be useful in order to construct an estimation of the risk. Indeed, it allows to remove the noise $\epsilon_i^{(j)}$. Moreover, the term $\mathcal{K}_\lambda * \mathbf{1}_G$ provides a good approximation of the indicatrice function $\mathbf{1}_G$ (see also Figure 2.1 for an illsutration in a simple case). We will see bellow that the price to pay for the noise removal is a larger variance than in the classical direct case, while we introduce bias when using a smoothed indicatrice function.

At this step, we are now able to propose a classifier. For all $G \subset K$, define

$$R_{n,m}(G) = \frac{1}{2n}\sum_{i=1}^n h_{K/G}(Z_i^{(1)}) + \frac{1}{2m}\sum_{j=1}^m h_G(Z_j^{(2)}). \tag{1.35}$$

Then, given a family $\mathcal{G}$ of subsets of K, we define $\hat{G}_{n,m}$ as

$$\hat{G}_{n,m} = \arg\min_{G\in\mathcal{G}} R_{n,m}(G).$$

Remark that for all $G \subset K$, we have

$$\begin{aligned}
\mathbb{E}\left[R_{n,m}(G)\right] &= \frac{1}{2}\int_{K/G}\frac{1}{\lambda}\mathcal{K}\left(\frac{X_1^{(1)}-x}{\lambda}\right)dx + \frac{1}{2}\int_G \frac{1}{\lambda}\mathcal{K}\left(\frac{X_1^{(2)}-x}{\lambda}\right)dx,\\
&= \frac{1}{2}\int_{\mathbb{R}^d}\int_{K/G}\frac{1}{\lambda}\mathcal{K}\left(\frac{y-x}{\lambda}\right)dxf(y)dy + \frac{1}{2}\int_{\mathbb{R}^d}\int_G\frac{1}{\lambda}\mathcal{K}\left(\frac{y-x}{\lambda}\right)dxg(y)dy,\\
&:= \frac{1}{2}\int_{\mathbb{R}^d} f(y)\mathcal{K}_\lambda * \mathbf{1}_{K/G}(y)dy + \frac{1}{2}\int_{\mathbb{R}^d} g(y)\mathcal{K}_\lambda * \mathbf{1}_G(y)dy,\\
&:= R_K^\lambda(G).
\end{aligned}$$

In this context, we hence use a biased ERM approach. The expectation of $R_{n,m}(G)$ is not $R_K(G)$, but instead $R_K^\lambda(G)$ as described in formula above. One of the main difficulty is then to control this bias in an *optimal* way.

Bellow, we introduce some assumptions on the model.

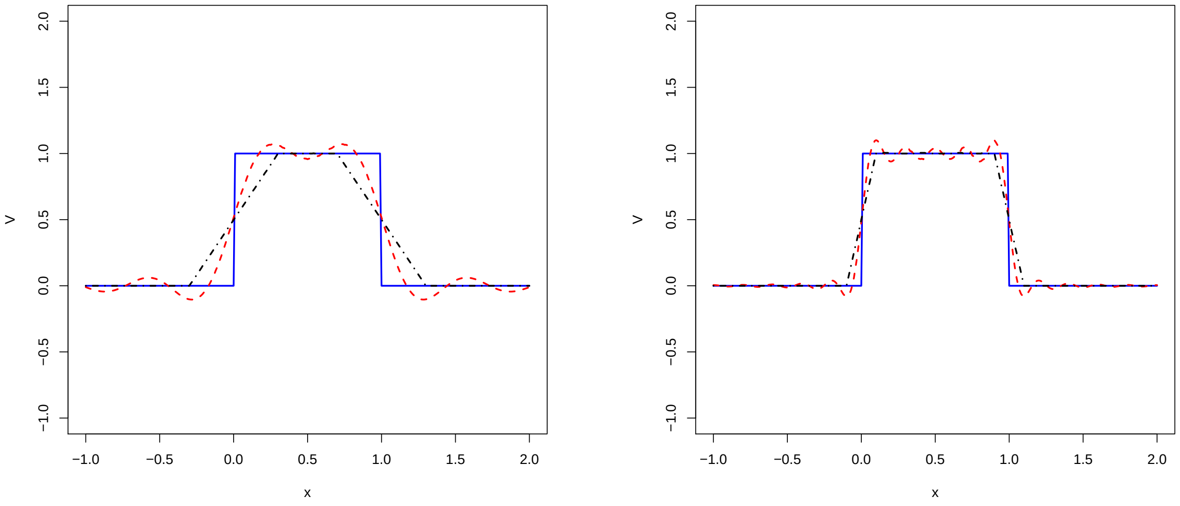

Figure 1.2: *Expectation of the function h_G in one dimension, when $G = [0,1]$ (dashed line for the sinc kernel and dotted lines for the indicatrice kernel), for two different bandwidth ($\lambda = 0.3$ on the left hand side, $\lambda = 0.1$ on the right hand side) and comparison with the indicatrice function (continuous line).*

Assumptions

In order to investigate the asymptotic of the excess risk, we will require some assumptions. First of all, we have to characterize the *difficulty* of the inverse problem related to the presence of noise measurements. In formula (1.33), we have already implicitly assumed that $\mathcal{F}[\eta](t) \neq 0$. The following assumption is more restrictive and characterizes the decay of this Fourier transform.

Noise assumption *The density η satisfies $\eta = \prod_{i=1}^{n} \eta_i$, where $\eta_i : \mathbb{R} \to \mathbb{R}$ are probability density functions. Moreover, there exist $\beta = (\beta_1, \ldots, \beta_d)'$ and C_1, C_2 positive constants such that*

$$\frac{C_1}{|t|^{\beta_i}} \leq |\mathcal{F}[\eta_i](t)| \leq \frac{C_2}{|t|^{\beta_i}},\ \forall i \in \{1, \ldots, d\},\ \forall t \in \mathbb{R}.$$

In other words, we require that the noise measurements are independent in each direction. This is only a technical assumption that could be certainly removed, up to some technical algebra. Concerning the behavior of the Fourier transforms of the marginal distributions, we assume a polynomial decay. The related problem is then said to be mildly ill-posed.

The second assumption concerns the behavior of the function $f - g$ at the boundary of $G_K^\star$. It has been introduced for the first time in a *classification* context by [76], but it finds its origin in the level-set estimation theory.

Margin assumption *There exists positive constants t_0, c_2, α such that for all $0 < t < t_0$*

$$Q\left(\{x \in K : \ |f(x) - g(x)| \leq t\}\right) \leq c_2 t^\alpha,$$

where Q denotes the Lebesgue measure on $\mathbb{R}^d$.

This margin assumption characterizes the difficulty to distinguish one observation having label 0 to another having label 1 close to the boundary of $G_K^\star$. Large values of α (in particular $\alpha = +\infty$) correspond to the most favorable situations. On the other hand, small values of α are related to difficult problems since both densities f and g are almost equal in a large neighborhood of the boundary of $G_K^\star$.

Typically, the last assumption in such a setting is a complexity assumption: it should measure the difficulty to find the good set (classifier) in a given family of candidates. We have considered two different kind of complexity assumptions. The first one is related the behavior of $f-g$ and can be considered as a smoothness assumption. To this end, we will introduce the class $\Sigma_d(\gamma, L)$ of isotropic Hölder continuous functions ν, having continuous partial derivatives up to order $\lceil \gamma \rceil$, the maximal integer strictly less than γ, and such that

$$|\nu(y) - p_{\nu,x}(y)| \leq L\|x-y\|^\gamma, \ \forall x, y \in \mathbb{R}^d,$$

where $p_{\nu,x}(y)$ is the Taylor polynomial of ν of order $\lfloor \gamma \rfloor$ and point x and $\|.\|$ stands for the Euclidean norm on $\mathbb{R}^d$.

Plug-in assumption *There exists positive constants γ and L such that $f-g \in \Sigma_d(\gamma, L)$.*

This assumption is of *plug-in* type since it is often used to measure performances of classifier of the form

$$\hat{G} = \left\{ x \in K : \ \hat{f}(x) \geq \hat{g}(x) \right\},$$

where $\hat{f}, \hat{g}$ are preliminary estimators of f and g. Such an assumption appears to be quite natural if one look at the expression of the estimator $R_{n,m}$ introduced in (1.35). Indeed, for all $G \subset K$, we have

$$\begin{aligned}
R_{n,m}(G) &:= \frac{1}{n}\sum_{i=1}^n h_{K/G}(Z_i^{(1)}) + \frac{1}{m}\sum_{j=1}^m h_G(Z_j^{(2)}), \\
&= \frac{1}{n}\sum_{j=1}^n \int_{K/G} \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_j^{(1)}-x}{\lambda}\right)dx + \frac{1}{m}\sum_{j=1}^m \int_G \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_j^{(2)}-x}{\lambda}\right)dx, \\
&= \int_{K/G}\left\{\frac{1}{n}\sum_{j=1}^n \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_j^{(1)}-x}{\lambda}\right)\right\}dx + \int_G \left\{\frac{1}{m}\sum_{j=1}^m \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_j^{(2)}-x}{\lambda}\right)\right\}dx, \\
&= \int_{K/G}\hat{f}_\lambda(x)dx + \int_G \hat{g}_\lambda(x)dx,
\end{aligned}$$

where for all $x \in \mathbb{R}^d$,

$$\hat{f}_\lambda(x) := \frac{1}{n}\sum_{j=1}^n \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_j^{(1)}-x}{\lambda}\right), \text{ and } \hat{g}_\lambda(x) := \frac{1}{m}\sum_{j=1}^m \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_j^{(2)}-x}{\lambda}\right).$$

The both terms $\hat{f}_\lambda$ and $\hat{g}_\lambda$ correspond to the classical estimators of the densities f and g when dealing with error-in-variable models (see [49], [36], [81] or Section 1.1.2 for more details). Hence, this classification problem can be seen as a density estimation problem. Nevertheless, we will see below that the minimax behavior of the bandwidth λ is quite different compared to usual non-parametric problems (see for instance [88] or [17]). In a classification context, we will take advantage of the margin assumption. Hence, the construction of an estimator of $G_K^\star$ in this context does not simply reduce to a density estimation problem.

In the following, we will denote $\mathcal{G}_{plug}$ as the set of all $G \subset K$ of the form

$$G = \{x \in K : \ \nu(x) \geq 0\},$$

where $\nu \in \Sigma_d(\gamma, L)$. Hence, each set is associated to a function $\nu \in \Sigma_d(\gamma, L)$, which plays the role of a possible candidate for $f - g$. By the same way, we call $\mathcal{F}_{plug}$ the set of all pairs (f, g) satisfying both the margin and plug-in assumptions.

The second complexity assumption that we have considered concerns geometric properties of the set of interest $G_K^\star$. For the sake of simplicity, we will concentrate on the case where $K = [0,1]^d$. Then, we introduce the set of boundary fragments $\mathcal{G}_{frag}$ defined as

$$\mathcal{G}_{frag} = \left\{ x \in [0,1]^d, \ 0 \leq x_d \leq b(x_1, \dots, x_{d-1}), \ b \in \Sigma_{d-1}(\gamma_b, L) \right\}, \tag{1.36}$$

where $\Sigma_{d-1}(\gamma_b, L)$ denotes the set of isotropic Hölder functions on $\mathbb{R}^{d-1}$. We also denote by $\mathcal{F}(\alpha, \gamma)$ the set of all pairs (f, g) satisfying the margin assumption and such that

$$\{x \in K : \ f(x) \geq g(x)\} \subset \mathcal{G}_{frag}.$$

In such a setting, the considered problem appears to be a non-parametric set estimation problem. In particular, we will see that the control of the different involved quantities (bias in particular) requires a specific treatment.

Excess risk with the plug-in assumption

For all $\delta > 0$, using the notion of entropy (see for instance [77] or [90]) for Hölderian function on compact sets, we can construct a δ-network $\mathcal{N}_\delta$ on $\Sigma_d(\gamma, L)$ restricted to $[0,1]^d$ such that

- $\log(\text{card}(\mathcal{N}_\delta)) \leq A\delta^{-d/\gamma}$,
- For all $h_0 \in \Sigma_d(\gamma, L)$, we can find $h \in \mathcal{N}_\delta$ such that $\|h - h_0\|_\infty \leq \delta$.

In Loustau and Marteau (2013), we have associated to each $\nu := f - g \in \Sigma_d(\gamma, L)$, a set $G_\nu = \{x : \nu(x) \geq 0\}$ and defined the ERM estimator as:

$$\hat{G}_{n,m}^\lambda = \arg\min_{\nu \in \mathcal{N}_\delta} R_{n,m}^\lambda(G_\nu), \tag{1.37}$$

where $\delta = \delta_{n,m}$ has to be chosen carefully (see Theorem 2 below). This procedure has been introduced in the direct case by [76] and referred as an hybrid Plug-in/ERM procedure. The following theorem describes the performances of $\hat{G}_{n,m}^\lambda$.

Theorem 2 *Let $\hat{G}^\lambda_{n,m}$ the set introduced in (1.37) with*

$$\lambda_j = (n \wedge m)^{-\frac{1}{\gamma(2+\alpha)+2\sum_{i=1}^d \beta_i + d}},\ \forall j \in \{1,\dots,d\},\ \text{and } \delta = \delta_{n,m} = \left(\frac{\prod_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n \wedge m}} \right)^{\frac{2}{2/\gamma+2+\alpha}}.$$

Suppose $f-g \in \Sigma(\gamma, L)$ and that the ***noise assumption*** *is satisfied with $\beta_i > 1/2$, $\forall i = 1,\dots d$. Consider the deconvolution kernel $\mathcal{K}_\eta$ defined in (1.8) where $\mathcal{K} = \Pi_{j=1}^d \mathcal{K}_j$ is a kernel of order $\lfloor \gamma \rfloor$ with respect to Q, with compact supported Fourier transform. Then, for all $\alpha \geq 0$*

$$\lim_{n,m\to+\infty} \sup_{(f,g)\in\mathcal{F}_{\text{plug}}(Q)} (n \wedge m)^{\tau_d(\alpha,\beta,\gamma)} \mathbb{E}_{f,g} d_\square(\hat{G}_n, G^\star_K) < +\infty,$$

where

$$\tau_d(\alpha,\beta,\gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha)+d+2\sum_{i=1}^d \beta_i} & \textit{for } d_\square = d_\Delta \\[2ex] \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha)+d+2\sum_{i=1}^d \beta_i} & \textit{for } d_\square = d_{f,g}. \end{cases}$$

As γ increases, the associated rate becomes faster: classification is easier for smooth densities. In the same spirit, large values of α provides faster rates while small values for the margin leads to very difficult problems. Compared to the free noise case, we can see that the price to pay for having error measurements is an additional term of the form $\sum_{j=1}^d \beta_j$ that increases the difficulty of the problem.

This upper bound has been validated by a corresponding lower bound in Loustau and Marteau (2013). This result is obtained in a slightly different framework, where we deal with an other measure than the usual Lebesgue measure. For technical constraints, this lower bound is valid only for small values of α.

In order to conclude this discussion, it is important to note that in some specific situations, the above result provides fast rates for classification (i.e. faster that $1/\sqrt{n}$). Up to my knowledge, these fast rates where obtained for the first time in [76]. In our situation, fast rates occur when

$$\alpha\gamma > d + 2\sum_{j=1}^d \beta_j. \tag{1.38}$$

In particular, the fact that we deal with an error-in-variable model is compatible with such kind of results. Nevertheless, this property has to be slightly nuanced. Indeed, the condition (1.38) occurs for very restrictive situations. Finding very smooth function associated to large values for α appears to be contradictory. Such examples can be provided when α is close to one. We also mention that large values for α under the plug-in assumption requires particular behavior for the measure Q... which are not satisfied in our paper (see Loustau and Marteau (2013a) for more details in the case where Q does not correspond to the Lebesgue measure).

Proof. We will not provide a complete proof for the above result, but only describe the most important steps. For the sake of convenience, we assume that $m = n$ in the following. By the way, we will write $\hat{G}_{n,m} = \hat{G}_n$. First of all, remark that

$$\begin{aligned}
&d_{f,g}(\hat{G}_n, G_K^\star) \\
&\quad = \int_{\hat{G}_n \Delta G_K^\star} |f-g|, \\
&\quad = \int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_{\hat{G}_n} - \mathcal{K}_\lambda * \mathbf{1}_{G_K^\star)}\right] + \int_{\hat{G}_n \Delta G_K^\star} |f-g| - \int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_{\hat{G}_n} - \mathcal{K}_\lambda * \mathbf{1}_{G_K^\star)}\right], \\
&\quad = \int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_{\hat{G}_n} - \mathcal{K}_\lambda * \mathbf{1}_{G_K^\star)}\right] + \int (f-g)\left[\mathcal{K}_\lambda * \{\mathbf{1}_{\hat{G}_n} - \mathbf{1}_{G_K^\star}\} - \{\mathbf{1}_{\hat{G}_n} - \mathbf{1}_{G_K^\star}\}\right], \\
&\quad := T_1 + T_2.
\end{aligned}$$

First, we can study the behavior of the first term. Remark that thanks to (1.37)

$$\int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_{\hat{G}_n} - \mathcal{K}_\lambda * \mathbf{1}_{G_K^\star)}\right] \leq R_n(G_K^\star) - R_n(\hat{G}_n) + \int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_{\hat{G}_n} - \mathcal{K}_\lambda * \mathbf{1}_{G_K^\star)}\right],$$

where we use the definition of $\hat{G}_n$. Then, the right hand side of the previous equation appears to be the sum of two independent and centered empirical processes. Using entropy properties of the set $\Sigma(\gamma, L)$ and the Bernstein inequality, we can control (with a great probability) this process by a term depending on n and λ. In particular, we get that

$$T_1 \leq C\left(\frac{\lambda_1^{-\beta_1}\lambda_2^{-\beta_2}}{\sqrt{n}}\right)^{-\frac{2(1+\alpha)}{2/\gamma+2+\alpha}} V_n, \tag{1.39}$$

where V_n has controlled moments and C denotes a positive constant. This term can be seen as a variance term.

Then, we have to provide a bound for the bias term T_2. In the following, we set $\nu = f - g$. For all $G_1, G_2 \in \mathcal{G}$, we get

$$\begin{aligned}
&\int (f-g)\left[\mathcal{K}_\lambda * \{\mathbf{1}_{G_1} - \mathbf{1}_{G_2}\} - \{\mathbf{1}_{G_1} - \mathbf{1}_{G_2}\}\right] \\
&\quad \leq \left| \int \left[\int \frac{1}{\lambda}\mathcal{K}\left(\frac{z-x}{\lambda}\right) f(z)dz - f(x)\right]\left[\mathbf{1}(x \in K/G_1) - \mathbf{1}(x \in K/G_2)\right] dx \right. \\
&\qquad \left. + \int \left[\int \frac{1}{\lambda}\mathcal{K}\left(\frac{z-x}{\lambda}\right) g(z)dz - g(x)\right]\left[\mathbf{1}(x \in G_1) - \mathbf{1}(x \in G_2)\right] dx \right|, \\
&\quad \leq \int_{G_1 \Delta G_2} |\mathcal{K}_\lambda * \nu(x) - \nu(x)|\, dx, \\
&\quad \leq \|\mathcal{K}_\lambda * \nu - \nu\|_\infty d_\Delta(G_1, G_2), \\
&\quad \leq C d_\Delta(G_1, G_2)\left[\lambda_1^\gamma + \lambda_2^\gamma\right],
\end{aligned}$$

for some $C > 0$. Indeed, provided that $\nu \in \Sigma(\gamma, L)$ and $\mathcal{K}$ is a kernel of order $l = \lfloor \gamma \rfloor$,

$$\|\mathcal{K}_\lambda * \nu - \nu\|_\infty \leq C\left[\lambda_1^\gamma + \lambda_2^\gamma\right]. \tag{1.40}$$

Using the Young inequality

$$xy^r \leq ry + (1-r)x^{1/(1-r)},$$

with $r = \alpha/(\alpha+1)$, we get

$$\begin{aligned}&\int (f-g)\left[\mathcal{K}_\lambda * \{\mathbf{1}_{G_1} - \mathbf{1}_{G_2}\} - \{\mathbf{1}_{G_1} - \mathbf{1}_{G_2}\}\right] \\ \leq \quad & (1-r)\gamma^{1/(1-r)}\left[\lambda_1^2 + \lambda_2^2\right]^{\frac{\gamma(1+\alpha)}{2}} + \gamma^{-1/r} d_{f,g}(G_1, G_2). \qquad (1.41)\end{aligned}$$

The end of the proof consists in finding a trade-off between the bias (1.41) and the 'variance' (1.39) terms, which leads to the desired result.

□

An extended version of this result, with related discussions, is available in Loustau and Marteau (2013a).

Excess risk with the Boundary fragments assumption

In this part, we deal with geometric constraints on the set $G_K^\star$ of interest: we assume that $G_K^\star \in \mathcal{G}_{frag}$ which is defined in (1.36). In such a case, the complexity of the considered problem is different from the one of the previous (plug-in) assumption. The following theorem proposes a minimax lower bound on the excess risk.

Theorem 3 *Let $K = [0,1]^d$ and $\mathcal{G} = \mathcal{G}_{frag}$. Suppose that the noise assumption is satisfied for some β. Then we have:*

$$\liminf_{n\to+\infty} \inf_{\hat{G}_{n,m}} \sup_{(f,g)\in\mathcal{F}(\alpha,\gamma)} (n\wedge m)^{\tau_d(\alpha,\beta,\gamma)} \mathbb{E} d_\square(\hat{G}_{n,m}, G_K^\star) > 0,$$

where the infinimum is taken over all possible estimators of the set $G_K^\star$ and

$$\tau_d(\alpha,\beta,\gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha) + (d-1)\alpha + 2\alpha\displaystyle\sum_{i=1}^{d-1}\beta_i + 2\alpha\beta_d\gamma} & \text{for } d_\square = d_\Delta \\[4ex] \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha) + (d-1)\alpha + 2\alpha\displaystyle\sum_{i=1}^{d-1}\beta_i + 2\alpha\beta_d\gamma} & \text{for } d_\square = d_{f,g}. \end{cases}$$

Clearly, the rates are different from the one obtained in the plug-in case. In particular, the way where the coefficient β and α interact is different. We refer to Loustau and Marteau (2013b) for a complete discussion on this result.

The control of the excess risk associated to the ERM classifier follows the same lines than in the plug-in case, except some important details. Indeed, using the same algebra as in the first part, we get

$$\begin{aligned}
&d_{f,g}(\hat{G}_n, G^\star_K)\\
&\quad\leq \int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_{\hat{G}_n} - \mathcal{K}_\lambda * \mathbf{1}_{G^\star_K)}\right] - \int (f-g)\left[\mathcal{K}_\lambda * \{\mathbf{1}_{\hat{G}_n} - \mathbf{1}_{G^\star_K}\} - \{\mathbf{1}_{\hat{G}_n} - \mathbf{1}_{G^\star_K}\}\right],\\
&\quad\leq \int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_{\hat{G}_n} - \mathcal{K}_\lambda * \mathbf{1}_{G^\star_K)}\right] + 2\sup_{G\in\mathcal{G}_{bf}}\left|\int (f-g)\left[\mathcal{K}_\lambda * \mathbf{1}_G - \mathbf{1}_G\right]\right|,\\
&\quad:= S_1 + S_2.
\end{aligned}$$

Using a control of the entropy related to the set $\mathcal{G}_{frag}$, we can prove that

$$S_1 \leq C\left(\frac{\Pi_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}}\right)^{\frac{2\gamma(\alpha+1)}{\gamma(\alpha+2)+(d-1)\alpha}} \tilde{V}_n, \tag{1.42}$$

where $\tilde{V}_n$ has controlled moments, and C denotes a positive constant. Remark that in this setting, the bias is slightly different. Indeed, we can not use Fubini and try to take advantage on the regularity of $f-g$ since we only have constraint on the geometry of the set of interest. As in the previous part, we have to control this bias term and to find a trade-off. To this end, we have investigated different possible strategies.

We first provide a rough bound for the bias term. Then, we discuss some heuristic that may provide a more convenient control for this quantity. For the sake of convenience, we deal in the following with the case where $d=2$. This discussion can easily be extended to the d-dimensional case.

In a first time, we can simply remark that for all $G\subset K$,

$$\begin{aligned}
\int (f-g)(\mathcal{K}_\lambda * \mathbf{1}_G - \mathbf{1}_G) &= \int (f-g)(x)\left(\int_{\mathbb{R}^2}\frac{1}{\lambda}\mathcal{K}\left(\frac{x-z}{\lambda}\right)[\mathbf{1}_G(z)-\mathbf{1}_G(x)]\,dz\right)dx,\\
&= \int (f-g)(x)\left(\int_{\mathbb{R}^2}\mathcal{K}(z)[\mathbf{1}_G(x-\lambda z)-\mathbf{1}_G(x)]\,dz\right)dx,\\
&= \int_{\mathbb{R}^2}\mathcal{K}(z)\int (f-g)(x)[\mathbf{1}_G(x-\lambda z)-\mathbf{1}_G(x)]\,dxdz,\\
&\leq \mathcal{Q}((G+\lambda)\Delta G) \leq C(\lambda_1+\lambda_2),
\end{aligned} \tag{1.43}$$

for some constant C, assuming for simplicity that $\mathcal{K}$ is compactly supported. If we equilibrate the previous bound with the one obtained in (1.42), we get

$$\mathbb{E}_\theta d_\Delta(\hat{G}_{n,m}, G^*_K) \leq Cn^{-\kappa_d(\alpha\beta,\gamma)},$$

where

$$\kappa_d(\alpha,\beta,\gamma) = \frac{\gamma\alpha}{\gamma(\alpha+2)+\alpha+2\gamma(\alpha+1)(\beta_1+\beta_2)}.$$

We get the following result which can be found in Loustau and marteau (2013b).

Theorem 4 *Let $\hat{G}_n$ the classifier introduced above where $\gamma > d-1$. Assume that the kernel $\mathcal{K}$ satisfies*

$$\sup_{t\in\mathbb{R}^d} |\mathcal{F}[\mathcal{K}_\eta](t)| \leq C\prod_{i=1}^d \lambda_i^{-\beta_i}, \text{and } \|\mathcal{K}_\eta\|^2 \leq C\prod_{i=1}^d \lambda_i^{-2\beta_i}.$$

Then, there exists a positive constant C such that

$$\mathbb{E}_\theta d_\Delta(\hat{G}_{n,m}, G_K^*) \leq Cn^{-\kappa_d(\alpha\beta,\gamma)},$$

where

$$\kappa_d(\alpha,\beta,\gamma) = \frac{\gamma\alpha}{\gamma(\alpha+2)+(d-1)\alpha+2\gamma(\alpha+1)\sum_{i=1}^d \beta_i}.$$

This result does not match with the lower bound presented above. The main reason is that we did not take advantage of the smoothness of the boundary b in the bound (1.43). Moreover, we have provided a crude bound for $f-g$. Up to now, providing minimax rates of convergence for the excess risk in this context remains for us an open problem. We are convinced that the lower bound is optimal. In particular, we have not yet used the smoothness propetry of b.

Bellow, we present some heuristics that may improve the control of this bias term. An investigation of the upper bound above indicates that an *optimal* control of the bias would require an upper bound of order

$$\left[\left(\sum_{i=1}^{d-1}\lambda_i\right)^\gamma + \lambda_d\right]^{1+\frac{1}{\alpha}} \quad \text{instead of} \quad \sum_{i=1}^d \lambda_i.$$

Using simple algebra (in dimension 2 for the sake of convenience), we can re-write the bias as

$$\begin{aligned}&\int_{\mathbb{R}^d}(f-g)(x)\left(\mathcal{K}_\lambda * \mathbf{1}_G(x) - \mathbf{1}_G(x)\right)dx\\ &=\int_{\mathbb{R}^2}\mathcal{K}(z)\int_{\mathbb{R}}\left[\int_0^{(b(x_1+\lambda_1 z_1)-\lambda_2 z_2)_+}(f-g)(x)\mathbf{1}_{\{x_1+\lambda_1 z_1\in[0,1]\}}dx_2 - \int_0^{b(x_1)}(f-g)(x)dx_2\right]dx_1 dz.\end{aligned}$$

In order to get a satisfying bound, different pathes could be investigated:

- The first problem concerns the localization of our data. Indeed, the bias term above contains control on the couple (x_1,x_2) that significantly slow down the rate. Since our data are noisy, we have indeed to estimate the fact that the variables $X_i^{(j)}$ belongs or not to the set of interest K. This problem could be avoided, assuming for instance that the unobserved $X_i^{(j)}$ belongs to K. All the previous analysis could hence be managed conditionally on this event.

- Up to some technical constraints related to such an assumption, one may construct a Taylor expansion of $b(.+\lambda_1 z_1)$ and try to control the behavior of the difference between the two integrals. In this context, there is no hope of improvement unless we have some smoothness constraint on $f-g$ in the 2^{nd} direction. Provided such a property holds, one may control

$$\int \mathcal{K}(z) \int_{b(x_1)}^{b(x_1+\lambda_1 z_1)+\lambda_2 z_2} (f-g)(x) dx_2 d_z \qquad \text{by a term of order} \qquad \lambda_1^{\gamma} + \lambda_2,$$

taking advantage on both behavior of b and the kernel $\mathcal{K}$. Even in this case, the upper bound will not be optimal.

- Finally, we have in mind stronger assumption on the behavior of $f-g$ at the boundary of $G_K^\star$. The exponent $1/\alpha$ seems indeed strongly related to the fact that $f-g$ is close to 0 in the neighborhood of the boundary of the Bayes set. Nevertheless, the margin assumption stated in [76] does not guarantee such a behavior. To this end, one may consider couple (f,g) of the form
$$(f-g)(x) = \nu(x_1)(x_2 - b(x_1))^{1/\alpha},$$
for some unknown function ν. The margin is *uniform* in such a case. One may then investigate the behavior of the minimax excess risk in such a context.

All this discussion presents a way that may allow to verify whether the lower bound is optimal or not. Nevertheless, it requires strong additional assumptions. This is not reasonable from a mathematical point of view.

Conclusion

The case where we deal with geometric constraint (boundary fragment assumption) is far from being completely understood. In particular, the control of the bias appears to be very difficult. At this step, lower and upper bounds do not match, even with additional assumptions. Hence, providing the minimax excess risk in this context remains an open problem.

We strongly believe that we did not find a convenient control of this bias. Our conjecture is that the lower bound is true and that the considered ERM algorithm is not optimal in this setting. In particular, as mentioned above, the deconvolution classifier can be considered as some kind of plug-in classifier. Hence, there is few hope to obtain a control on the excess risk with an hypothesis on the boundary of $G_K^\star$ instead of $f-g$.

1.2.3 Estimation for inverse problems

I have discovered statistical inverse problems through an estimation purpose. At the beginning of my PhD, Laurent Cavalier proposed me to investigate performances of different adaptive methods on the sequence space model (1.4). Since I am in Toulouse, I get the opportunity to extend these results on different models. The section bellow provides a short overview of my different contributions in this topic.

Adaptive algorithms

In this part, we deal again with the sequence space model (1.4) defined as

$$y_k = b_k \theta_k + \epsilon \xi_k, \ \forall k \in \mathbb{N},$$

where the sequence of interest is $\theta = (\theta_k)_{k\in\mathbb{N}}$. Given a family of filters Λ, $\lambda \in \Lambda$ and an associated linear estimator $\hat{\theta}_\lambda$ (as described in (1.5)), we can compute the related quadratic risk

$$R(\theta,\lambda) = \mathbb{E}_\theta\|\hat{\theta}_\lambda - \theta\|^2 = \mathbb{E}_\theta \sum_{k=1}^{+\infty}(\lambda_k b_k^{-1} y_k - \theta_k)^2 = \sum_{k=1}^{+\infty}(1-\lambda_k)^2\theta_k^2 + \epsilon^2 \sum_{k=1}^{+\infty}\lambda_k^2 b_k^{-2}.$$

In an ideal way, we would like to use the best possible filter among Λ, namely the filter λ^0 defined as

$$\lambda^0 = \arg\min_{\lambda\in\Lambda} R(\theta,\lambda). \tag{1.44}$$

In practice, this filter (called *oracle*) is not available since it strongly depend on the unknown parameter of interest. Nevertheless, we would like to approximate this term.

If we have an *a priori* knowledge on the underlying function f (for instance a smoothness constraint), we can propose a choice for the filter. Indeed, assume for instance that $\theta \in \mathcal{E}_{a,2}(R)$, where the ellipsoid $\mathcal{E}_{a,2}(R)$ has been introduced in (1.19). Then

$$R(\theta,\lambda) = \sum_{k=1}^{+\infty}(1-\lambda_k)^2 a_k^{-2} a_k^2 \theta_k^2 + \epsilon^2 \sum_{k=1}^{+\infty}\lambda_k^2 b_k^{-2} \leq \sup_{k\in\mathbb{N}}\{(1-\lambda_k)^2 a_k^{-2}\} + \epsilon^2 \sum_{k=1}^{+\infty}\lambda_k^2 b_k^{-2}.$$

Then, choosing λ^1 such that

$$\lambda^1 = \arg\min_{\lambda\in\Lambda}\left\{\sup_{k\in\mathbb{N}}(1-\lambda_k)^2 a_k^{-2} + \epsilon^2 \sum_{k=1}^{+\infty}\lambda_k^2 b_k^{-2}\right\},$$

we get

$$R(\theta,\lambda^1) \leq \inf_{\lambda\in\Lambda}\sup_{\theta\in\mathcal{E}_{a,2}(R)} R(\theta,\lambda).$$

This is the minimax point of view, which have been used in a slightly different form in Section 1.2.1 above. The main problem related to such kind of approaches is that it requires the knowledge of the sequence a. This is strong assumption that is not satisfied in many practical problems. To this end, it is necessary to consider adaptive algorithms, i.e. procedures that do not use this sequence.

The idea that we will use is very close to the ERM scheme studied in Section 1.2.2 in a smooth discriminant analysis context. Our aim is to approximate the oracle λ^0 ... which is unknown since it explicitly depends on θ. Nevertheless, this risk can be estimated by a given random term $U(\lambda, Y)$ (whose construction will be specified later on) as described in the following scheme.

$$\begin{tikzcd}
\lambda^0 \arrow[r, "\text{minimizes}"] & R(\theta,\lambda) \\
\lambda^\star \arrow[u, "\text{stable?}"] \arrow[r, "\text{minimizes}"] & U(Y,\lambda) \arrow[u, "\text{estimates}"']
\end{tikzcd}$$

At this step, we have to choose a family of filters and an estimator for the quadratic risk. I have been interested in the following approaches.

- We can deal with a given regularization scheme: for instance we consider a spectral cut-off or Tikhonov regularization $\lambda = \lambda(t)$ for which we want to select an appropriate regularization parameter t. In some sense, we have a structure on the family Λ. In such a case, we want to approach the best possible risk using such a procedure. Concerning the estimation of the risk, one can consider a penalized estimator $U_\theta(Y,t)$ defined as

  $$U_\theta(Y,t) = \sum_{k=1}^{+\infty} (\lambda_k^2 - 2\lambda_k) b_k^{-2} y_k^2 + 2 \sum_{k=1}^{+\infty} \lambda_k b_k^{-2} + \mathrm{pen}(t),$$

  where $\mathrm{pen}(t)$ is a penalization term.

  Different kind of penalization have been proposed in the literature. For instance, [25] consider the case where $\mathrm{pen}(t) = 0$, which corresponds to the URE (Unbiased Risk Estimation) approach. Alternative penalties are proposed in [12] or [1] for instance. In Marteau (2010), I have been interested in the RHM (Risk Hull Minimization) approach. This method has been introduced in [26] for spectral cut-off filters. It is based on the following heuristic. When there is no signal in the observation, then the URE method ($\mathrm{pen}(t) = 0$) leads to an estimator of the risk defined as

  $$U_0(Y,t) = \epsilon^2 \sum_{k=1}^{+\infty} (\lambda_k(t)^2 - 2\lambda_k(t)) b_k^{-2} \xi_k^2 + 2\epsilon^2 \sum_{k=1}^{+\infty} \lambda_k(t) b_k^{-2}.$$

  In particular, it is possible to see that

  $$\mathrm{Var}(U_0(Y,t)) = \epsilon^4 \sum_{k=1}^{+\infty} (\lambda_k(t)^2 - 2\lambda_k(t))^2 b_k^{-4},$$

  which explodes for small values of t. Indeed, the sequence b_k tends to 0 as $k \to +\infty$. The main consequence is that the related data-driven parameter $t^\star$ defined as

  $$t^\star = \arg\min_{t>0} U(Y,t),$$

  is very unstable. The alternative proposed by [26] was to construct a hull for the l^2 loss in order to contain this variability, namely we define a term $S(\theta,\lambda)$ such that

  $$\mathbb{E}_\theta \sup_{\lambda\in\Lambda} \left[\|\hat{\theta}_\lambda - \theta\|^2 - S(\theta,\lambda) \right] \leq 0.$$

  Then, we construct an estimator for S, which is equivalent to penalize the quadratic risk. Performances for this approach have been investigated in Marteau (2010) for general spectral regularization schemes.

- The problem with the previous approach is that we are limited by the regularization scheme that we have decided to use. For instance, if we deal with Tikhonov regularization, we can prove (see [25] for more details) that the proposed estimator do as well as the oracle: in other words, we mimic the behavior of the best possible Tikhonov estimator. In an ideal way, we would like to get an oracle inequality over a wider family of filters: for instance over all possible linear filters. In order to obtain such a result, the idea is to consider the family of blockwise constant filters $\Lambda^\star$ defined as

  $$\Lambda^\star = \left\{ \lambda : \ \lambda_k \in [0,1], \ \lambda_k = \lambda_{K_j} \ \forall k \in [K_j; K_{j+1}], j = 0, \dots, J-1 \text{ and } \lambda_k = 0, \ k > N \right\},$$

  for some parameters $J, (K_j)_{j=1..J}$ and N, that describe the shape of the considered blocks. It has been proved (see for instance [29]) that the oracle over $\Lambda^\star$ is very close to the one over the class of monotone filters. Hence, the family $\Lambda^\star$ is a good candidate for the construction of an adaptive estimator.

  In [29] and [30], the authors apply the URE estimation method with a penalized estimator. They obtain a filter of the form

  $$\lambda_j^{I,\star} = \begin{cases} \left(1 - \frac{\sigma_j^2(1+\varphi_j)}{\|\tilde{y}\|^2_{(j)}}\right)_+, \text{ if } j = 1, \dots, J, \\ 0 \text{ if } j > J, \end{cases} \tag{1.45}$$

  for some penalty $(\varphi_j)_{j=1..J}$. Then, they obtain (under reasonable conditions) the following oracle inequality

  $$\mathbb{E}_\theta \|\theta^{I,\star} - \theta\|^2 \leq (1 + \varphi(\epsilon)) \inf_{\lambda \in \Lambda^\star} R^I(\theta, \lambda) + 8c_1\epsilon^2,$$

  for some constant c_1 and a function $\varphi(.)$ verifying $\varphi(\epsilon) \to 0$ as $\epsilon \to 0$. In Marteau (2010) , I have investigated the link existing between such an approach and the RHM method. In particular, the expectation of the considered estimator for the quadratic risk appears to be a risk hull, up to some restrictions over the penalty φ.

Inference with noise in the operator

All the regularization algorithms presented above in different situations, heavily depend on the structure of the operator. Nevertheless, in some particular cases, this operator appears to be difficult to handle. Regularization with a completely unknown operator appears to be quite difficult. In a slightly different setting, it has been proved by [80] that only logarithmic rates of convergence can be expected in such a situation, whatever the smoothness assumptions set on the function of interest. Nevertheless, some interesting results can be obtained in the cases where the operator is partially known.

We start with the error in variable model considered above. Recall that in such a setting, given observation

$$Z_i = X_i + \epsilon_i, \ \forall i \in \{1, \dots, n\},$$

where the X_i (resp. the ϵ_i) admits a density f (resp. η) w.r.t. the Lebesgue measure, our aim is to provide some inference on the density f of interest. As discussed in a previous section, the density of the observation Z_i corresponds to the convolution product $f * \eta$: we are hence faced to a deconvolution problem. *But what appends if η is not exactly known?* Different setting have been considered in the literature.

We can for instance mention the contribution of [19] where estimation of the density f, of a related quadratic functional or goodness-of-fit testing is considered. In this paper, they use a semi-parametric model: the Fourier transform of the noise density is assumed to verify

$$\mathcal{F}[\eta](t) = \exp\left(-|\gamma t|^s\right),$$

In particular, the authors assume that the parameter s is unknown, but that it belongs to a known grid S over $\mathbb{R}^+$. In this setting, they provide algorithms that take into account this uncertainty, and prove that the related performances do not significantly deteriorate. In particular, *classical* rates of convergence are reached, up to a logarithmic loss.

Other approaches have been proposed in order to circumvent the lack of knowledge of the density η. For instance, it is assumed in [34] that an additional sample of noise if available. One observe both the $(Z_i)_{i=1\dots M}$ and the sample $(\tilde{\epsilon}_j)_{j=1\dots M}$ where the $\tilde{\epsilon}_j$ are i.i.d. random variables having common density η. Using this additional sample, the authors construct a preliminary estimator of η and then plug this estimator in the deconvolution kernel. Then, they investigate the performances of the related procedure. In particular, it is proved that up to some constraint on the sample size M, the (adaptive) estimator proposed in [34] attains *classical* rates of convergence for estimation with l^2 losses.

Now, we turn to the study of the sequence space model (1.1) when the involved operator is partially observed. Even in this setting, there exists several situations for which the operator involved is the observation is not completely known. Nevertheless, as considered in [27], estimation is possible when only partial information on A is available. More precisely, they deal with the model

$$\begin{cases} y_k = b_k\theta_k + \epsilon\xi_k, \\ x_k = b_k + \sigma\eta_k \end{cases} \quad \forall k \in \mathbb{N}, \tag{1.46}$$

where the sequence $(x_k)_{k\in\mathbb{N}}$ corresponds to observation on the unknown sequence of eigenvalues of the operator A, σ is a noise level and η_k a sequence of i.i.d. standard Gaussian random variables (independent of the ξ_k).

In the model (1.46), we implicitly assume that the eigenvectors are known, which is sometimes a strong assumption. Nevertheless, such a situation can for instance holds in the case of a convolution operator

$$Af(.) = \int_0^1 g(t)f(t-.)dt,$$

for some convolution filter g. The eigenvectors of this operator correspond to the Fourier basis, whatever the behavior of g. In particular, the model (1.46) is encountered when additional and independent observations on the function g are available.

Several non-parametric estimators of the function f (i.e. the sequence θ) are of the form

$$\hat{f}_\lambda = \sum_{k=1}^{+\infty} \lambda_k b_k^{-1} y_k \phi_k, \tag{1.47}$$

for some filter λ. Such an estimator is not available for the model (1.46) for which the sequence $(b_k)_k$ is unknown (remark that this sequence is also often involved in the construction of the filters, see examples bellow for instance). This sequence can be replaced by $(x_k)_k$, but some modifications are necessary.

First, since the sequence $(b_k)_k$ tends to 0 as $k \to +\infty$ (A is a compact operator). The terms x_k will not be a *good* estimator for b_k as soon as k is large enough. In order to get round of this problem, the idea is to compare for all integer k, x_k to a threshold depending on σ. In particular, we will only consider index for which $k \leq M$ where

$$M = \inf \left\{ k : |x_k| \geq \sigma \log^{\tau}(1/\sigma) \right\}, \tag{1.48}$$

for some $\tau \geq 1$. If $k \leq M$, this means that $|x_k| \geq \sigma \log(1/\sigma)$, the r.h.s. term corresponding to the standard deviation of a Gaussian random variable with variance σ: in such a case, there is a great probability that x_k contains some *signal* and not only noise. In the other hand, when $k > M$, estimation might be complicated since the problem will not be regularized with the good sequence. Hence, we use in practice the following estimator

$$\hat{f}_\lambda = \sum_{k=1}^{+\infty} \lambda_k x_k^{-1} y_k \mathbf{1}_{\{k \leq M\}} \phi_k.$$

Such an approach corresponds more or less to using a hard thresholding rule on the diagonal of the (noisy) representation matrix M.

In a second time, the influence of the noise in the operator has to be precisely quantified: this is in particular of first importance when considering model selection approaches (see the discussion above for more details). In particular, one can remark that for all $k \in \mathbb{N}$,

$$\begin{aligned} x_k^{-1} y_k &= x_k^{-1} b_k \theta_k + \epsilon x_k^{-1} \xi_k, \\ &= \theta_k + \epsilon x_k^{-1} \xi_k + \left(\frac{b_k}{x_k} - 1 \right) \theta_k. \end{aligned}$$

Using a Taylor expansion of the last term in the right hand side of the previous equation, we get

$$\frac{b_k}{x_k} = \frac{b_k}{b_k + \sigma \eta_k} = b_k \times \left[\frac{1}{1 + \sigma b_k^{-1} \eta_k} \right] = b_k \times \left[1 + \sigma b_k^{-1} \eta_k + \sigma^2 b_k^{-2} \zeta_k^2 \right],$$

for some random variable ζ_k which can be controlled as soon as $k \leq M$ with a great probability. On such event, it is possible to prove that the last term is negligible w.r.t. the first one. Therefore, with a great probability, we get that

$$x_k^{-1} y_k \simeq \theta_k + \epsilon x_k^{-1} \xi_k + \sigma b_k^{-1} \theta_k \eta_k, \ \forall k \in \mathbb{N}. \tag{1.49}$$

The inverse problem model with noise in the operator written under the form (1.49) can hence be seen as a generalization of the sequence space model.

Inverse problems with noise in the operator, and more specifically the model (1.46) have been widely investigated in the literature. In [44], properties of projection method on general bases are investigated: in the considered setting, the operator is identified with a matrix, whose entries are independently observed with additional noise. [56] provides a similar study with a wavelet-based Galerkin regularization approach. In the sequence space model, [27] study the properties of the URE method in this context and derive an oracle inequality. Similar refined contributions are proposed in [64] where the case $\sigma >> \epsilon$ is also considered.

In this context, I have investigated the performances of the penalized blockwise Stein's rule and the RHM method in respectively Marteau (2009) and Marteau (2010). The following contributions have also been proposed in related models.

- As described in a Section 1.1.3, the shifted curves model can be written under the form (1.14), namely, we observe

$$\gamma_k^{-1} y_k = \theta_k + \frac{\epsilon}{\sqrt{n}} \gamma_k^{-1} \xi_k + \left(1 - \frac{\tilde{\gamma}_k}{\gamma_k}\right) \theta_k, \ k \in \mathbb{Z}.$$

  This equation can be compared with (1.49). In this case, the control of $\gamma_k^{-1} y_k$ is explicit: it is not necessary to use a Taylor expansion since the noise appears in the numerator in such a case. Indeed, we deal with a random inverse problems, for which the regularization is performed via a deterministic operator. In particular, the risk associated to a given linear filter can be explicitly computed.

  Lemma 2 *For any given nonrandom filter λ, the risk of the estimator $\hat{\theta}(\lambda)$ can be decomposed as*

$$R(\theta, \lambda) = \underbrace{\sum_{k \in \mathbb{Z}} (\lambda_k - 1)^2 |\theta_k|^2}_{Bias} + \underbrace{\frac{1}{n} \sum_{k \in \mathbb{Z}} \lambda_k^2 \frac{\epsilon^2}{|\gamma_k|^2}}_{V_1} + \underbrace{\frac{1}{n} \sum_{k \in \mathbb{Z}} \left[\lambda_k^2 |\theta_k|^2 \left(\frac{1}{|\gamma_k|^2} - 1 \right) \right]}_{V_2} \quad (1.50)$$

  In equation (1.50) above, the risk is decomposed in three terms. The two first one correspond to the classical risk, while the third is associated to the fact that we do not regularize with the good operator. In this context, we have proposed in Bigot et al. (2010) an extension of the URE method, and we have derived a related oracle inequality.

- Similar investigations are proposed in the case where we want to estimate intensities of non-homogeneous Poisson processes. In such a setting, we have proposed in Bigot et al. (2013) an estimator based on the wave-D algorithm (see [66] for more details). In particular, a precise study of the corresponding minimax rates of convergence is proposed, with related upper and lower bounds.

- As explained in Section 1.1.4, the instrumental variable regression model appears to be an interesting practical example of inverse problem with unknown operator. Indeed, recall that in this context, we observe $(Y_i, X_i, W_i)_{i=1 \dots n}$ with

$$Y_i = T\varphi(W_i) + V_i \ \forall i \in \{1, \dots, n\},$$

where the noise $(V_i)_{i=1\dots n}$ is centered conditionally to the design $(W_i)_{i=1\dots n}$ and T is defined as

$$\begin{aligned} T: \quad & L^2_X \to L^2_W \\ & \varphi \mapsto T\varphi(.) = \mathbb{E}[\varphi(X)/W = .], \end{aligned}$$

In Loubes and Marteau (2010), we have assumed that the eigenvectors related to the operator T are available (practical examples are provided in the corresponding paper). These eigenvectors $(\phi_j)_j$ and (ψ_j) corresponds to a basis of respectively L^2_X and L^2_W. In this context, our aim is to estimate the sequence $(\phi_j)_{j\in\mathbb{N}}$ where for all $j \in \mathbb{N}$,

$$\varphi_j = \langle \varphi, \phi_j \rangle_X = \mathbb{E}[\varphi(X)\phi_j(X)].$$

Since the triplet $(\phi_j, \psi_k, \lambda_j)$ corresponds to the singular value decomposition of T, we get that

$$\langle T\varphi, \psi_j \rangle_W = \langle \varphi, T^\star T\phi_j \rangle_X = \lambda_j \langle \varphi, \phi_j \rangle_X = \lambda_j \varphi_j.$$

Hence, for all $j \in \mathbb{N}$, the coefficients φ_j can be estimated by

$$\hat{\varphi}_j = \frac{1}{\hat{\lambda}_j} \times \frac{1}{n} \sum_{i=1}^n Y_i \psi_j(W_i),$$

where the $\hat{\lambda}_i$ are introduced in Section 1.1.4. We eventually use an estimator of the form

$$\hat{\varphi}_m = \sum_{j=1}^m \frac{1}{\hat{\lambda}_j} \times \frac{1}{n} \sum_{i=1}^n Y_i \psi_j(W_i) \mathbf{1}_{j\le M} \phi_j,$$

where m is a regularization parameter whose value has to be specified and M is a term similar to the one introduced in (1.48).

In Loubes and Marteau (2010), we propose an extension of the URE method and construct a related oracle inequality. In particular, we build an adaptive estimator which appears to be minimax up to log term for standard regularity constraints. This log is in our opinion due to the fact that we estimate both the operator and the *signal* φ from the same sample.

Chapter 2

Direct methods for inverse problems and some perspectives in nonparametric statistics

In this chapter I will not summarize recent contributions in the non-parametric statistical theory, but rather try to discuss how are these results related. In particular, both testing and supervised classification problems are not so far as it seems. The presented discussion has not been yet published in this form, but contain elements that could be developed in the forthcoming years.

In a first time, we prove that the regularization of the problem is not necessary in a testing purpose. In a second time, I will present some fields for which this principle could be extended. In particular, we will focus on supervised binary classification and on estimation in Gaussian white noise models. The last section of this chapter is devoted to a brief presentation of some possible perspectives in non-parametric statistics.

2.1 Signal detection without regularization

2.1.1 Motivation

For the sake of brevity, we will deal in this section with the sequence model (1.4), where we observe

$$y_k = b_k \theta_k + \epsilon \xi_k, \ \forall k \in \mathbb{N}.$$

Recall that the sequence $(b_k^2)_{k\in\mathbb{N}}$ corresponds to the eigenvalues of the operator $A^\star A$, the ξ_k are i.i.d. random variables and ϵ is a positive noise level.

In a signal detection purpose, our aim is to assess whether we are observing some signal or not. In other words, we want to test the hypothesis $H_0 : f = 0$ against a non parametric alternative. This question has been developed in Chapter 1, where separation rates were investigated.

Since the operator A is compact, it is injective. In particular, $b_k^2 > 0$ for all $k \in \mathbb{N}$. Hence, the two assertions "$f = 0$" and "$Af = 0$" are equivalent. It is then natural to ask whether there might exist an other way to deal with this detection problem. Indeed, setting $\nu = Af$, the inverse problem model $Y = Af + \epsilon\xi$ introduced in (1.2) becomes

$$Y = \nu + \epsilon\xi.$$

Providing inference on the function ν then appears to be a direct problem. Such kind of direct model has been widely considered in the literature, in particular in a testing purpose (see for instance [60] or [3]). The underlying question is then: *could this detection problem be considered as a direct signal detection problem*?.

More formally, two testing problems are at hand. As explained in Chapter 1, we can investigate the inverse testing problem where one want to test

$$H_0^{IP} : f = 0, \text{ against } H_1^{IP} : f \in \mathcal{E}_{a,2}(R), \ \|f\| \geq \rho^{IP}, \tag{2.1}$$

Alternatively, one can adopt a direct approach where we want to test

$$H_0^{DP} : Af = 0, \text{ against } H_1^{DP} : f \in \mathcal{E}_{a,2}(R), \ \|Af\| \geq \rho^{DP}. \tag{2.2}$$

Both problems are similar except that the alternatives are not expressed in the same way (see also [57]). The aim of this discussion is to establish some hierarchy between theses approaches. In particular, we will address two different questions

- Is a test minimax for the direct problem also minimax for the inverse problem?
- Is a test minimax for the inverse problem also minimax for the direct problem?

We will see that the the answer to the first question is *yes*, while we can construct counter-examples for the second one.

2.1.2 An heuristic discussion

Before providing formal results, we provide here a brief heuristic discussion. First of all, we will precise the behavior of the sequence $(b_k)_k$ and describe the considered smoothness in the alternative (we will only deal with ellipsoids $\mathcal{E}_{a,2}(R)$ as described in (1.19)).

For the sake of brevity, we will only discuss the case where both sequences b and a possess a polynomial behavior. Here and in the following, for a given sequence u, the notation $u_k \sim k^l$ means that there exists a positive constant $c_\star$ such that $c_\star^{-1} k^l \leq u_k \leq c_\star k^l$ for all $k \in \mathbb{N}$.

Assumption A1. *There exists $s > 0$ and $t > 0$ such that*

$$b_k \sim k^{-t}, \text{ and } a_k \sim k^s, \ \forall k \in \mathbb{N}. \tag{2.3}$$

By the way, we will only consider test based on a spectral cut-off regularization. All our results presented bellow have been extended in a more general framework (see Loubes and Marteau (2013) or Marteau and Mathé (2013) for more details).

Since we can alternatively address both testing problems (2.1) or (2.2) , we have the choice between two testing procedures $\Phi_{\alpha,D}^{IP}$ and $\Phi_{\alpha,D,}^{DP}$, defined as

$$\Phi_{\alpha,D}^{IP} = \mathbf{1}_{\left\{\sum_{k=1}^{D} b_k^{-2}(y_k^2 - \epsilon^2) > t_{\alpha,D}^{IP}\right\}} \text{ and } \Phi_{\alpha,D}^{DP} = \mathbf{1}_{\left\{\sum_{k=1}^{D} (y_k^2 - \epsilon^2) > t_{\alpha,D}^{DP}\right\}}, \tag{2.4}$$

where $t^{IP}_{\alpha,D}$ (resp. $t^{DP}_{\alpha,D}$) is the $1-\alpha$-quantile of the variable $\epsilon^2\sum_{k=1}^D b_k^{-2}(\xi_k^2-1)$ (resp. $\epsilon^2\sum_{k=1}^D(\xi_k^2-1)$). In this context, we can address the question of the best possible choice for the bandwidth D. Some answers are provided below.

- **Inverse case**: Consider the test $\Phi^{IP}_{\alpha,D}$ defined in (2.4). A described in Chapter 1, this test is powerful as soon as

$$\|\theta\|^2 \geq C_{\alpha,\beta}\left[R^2 a_D^{-2} + \epsilon^2\sqrt{\sum_{k=1}^D b_k^{-4}}\right].$$

Following Assumption A1, we can find a constant C such that

$$C_{\alpha,\beta}\left[R^2 a_D^{-2} + \epsilon^2\sum_{k=1}^D b_k^{-2}\right] \leq C\left[D^{-2s} + \epsilon^2 D^{2t+1/2}\right].$$

In order to provide the smallest possible detection radius, the idea is to choose D such that we minimize the right hand side of the previous inequality. The corresponding trade-off is then obtained when

$$D_{IP}^{-2s} \sim \epsilon^2 D_{IP}^{2t+1/2} \Rightarrow D_{IP} \sim \epsilon^{-\frac{2}{2s+2t+1/2}}. \tag{2.5}$$

Then, we eventually get that

$$\sup_{\theta\in\mathcal{E}_{a,2}(R),\|\theta\|>\rho^{IP}} P_\theta(\Phi^{IP}_{\alpha,D}=0) \leq \beta, \text{ where } \rho^{IP} = c_{\alpha,\beta}\epsilon^{\frac{2s}{2s+2t+1/2}}, \tag{2.6}$$

for some constant $c_{\alpha,\beta}$ which can be specified. The term ρ^{IP} is the minimax separation rate for H_0^{IP} on $\mathcal{E}_{a,2}(R)$.

- **Direct case**: We can use the same methodology for the direct test $\Phi^{DP}_{\alpha,D}$ introduced in (2.4). The main difference is related to the degree of ill-posedness (here 0 since we consider a direct problem) and the smoothness of the function Af. Concerning this issue, we can remark that

$$\sum_{k=1}^{+\infty} b_k^{-2}\theta_k^2 b_k^2 a_k^2 = \sum_{k=1}^{+\infty}\theta_k^2 a_k^2 \leq R,$$

since $\theta\in\mathcal{E}_{a,2}(R)$. In other words, $(b_k\theta_k)_k \in \mathcal{E}_{c,2}(R)$ where $c_k = b_k a_k$. Following Assumption A1, we can say that the smoothness indice of Af is $t+s$. Then, we can prove (see for instance [3]) that the test $\Phi^{DP}_{\alpha,D}$ is powerful as soon as

$$\|Af\|^2 \geq C_{\alpha,\beta}\left[R^2 c_D^{-2} + \epsilon^2\sqrt{D}\right],$$

for some constant C. The trade-off is obtained when

$$D_{DP}^{-2(s+t)} \sim \epsilon^2 D_{IP}^{1/2} \Rightarrow D_{IP} \sim \epsilon^{-\frac{2}{2s+2t+1/2}}. \tag{2.7}$$

Then, we eventually get that

$$\sup_{b\theta\in\mathcal{E}_{c,2}(R),\|f\|>\rho^{DP}} P_\theta(\Phi^{DP}_{\alpha,D}=0) \leq \beta, \text{ where } \rho^{DP} = \bar{c}_{\alpha,\beta}\epsilon^{\frac{2(s+t)}{2s+2t+1/2}}, \tag{2.8}$$

for some constant $\bar{c}_{\alpha,\beta}$. The term ρ^{IP} is the minimax separation rate for H_0^{DP} on $\mathcal{E}_{c,2}(R)$.

The main consequence of (2.6) and (2.8) is that the direct testing problem is easier than the inverse testing problem in the sense that the separation rate is smaller. Nevertheless, the good news is that the *optimal* (in the minimax sense) number of coefficients necessary to provide a satisfying test is the same in both frameworks (see (2.5) and (2.7))! In other words, whatever we will do with our sample, we will use the same amount of information. Hence, there is some chance that a test designed for one of the considered framework provide interesting behavior in the complementary case.

2.1.3 Direct tests are always minimax for the inverse testing problem

In this section, the goal is to show that every test minimax for H_0^{DP} over $\mathcal{E}_{c,2}(R)$ is also minimax for H_0^{IP} over $\mathcal{E}_{a,2}(R)$. In particular, the regularization step does not appear to be necessary in a testing purpose. This result is essentially based on the following lemma which has been proposed in Laurent et al. (2011).

Lemma 3 *Let γ_ϵ a positive sequence such that $\gamma_\epsilon \to 0$ as $\epsilon \to 0$. The following embedding holds:*

$$\left\{ f \in \mathcal{E}_{a,2}(R),\ \|f\|^2 \geq \gamma_\epsilon \right\} \subset \left\{ f \in \mathcal{E}_{a,2}(R),\ \|Af\|^2 \geq \mu_\epsilon \right\},$$

where $\mu_\epsilon = b_{m(\epsilon)}^2 \gamma_\epsilon$ and $m(\epsilon)$ is such that $R^2 a_{m(\epsilon)}^{-2} \leq \gamma_\epsilon$.

Proof Let $m \in \mathbb{N}$ which will be chosen later and

$$f \in \left\{ \nu \in \mathcal{E}_{a,2}(R),\ \|\nu\|^2 \geq \gamma_\epsilon \right\}.$$

Then

$$\|Af\|^2 = \sum_{k\in\mathbb{N}} b_k^2 \theta_k^2 \geq \sum_{k\leq m} b_k^2 \theta_k^2 \geq b_m^2 \sum_{k\leq m} \theta_k^2 = b_m^2 \left(\|f\|^2 - \sum_{k>m} \theta_k^2 \right).$$

Since $f \in \mathcal{E}_{a,2}^{\mathcal{X}}(R)$

$$\sum_{k>m} \theta_k^2 \leq a_m^{-2} \sum_{k>m} a_k^2 \theta_k^2 \leq R^2 a_m^{-2}.$$

Hence

$$\|Af\|^2 \geq b_m^2 \left(\gamma_\epsilon - R^2 a_m^{-2} \right), \text{ as } \epsilon \to 0.$$

We conclude the proof choosing $m = m(\epsilon)$ such that $R^2 a_{m(\epsilon)}^{-2} \leq c\gamma_\epsilon$, for some $0 < c < 1$ independent of ϵ.

□

In the particular case case where

$$\gamma_\epsilon \sim (\rho_\epsilon^{IP})^2 \sim \epsilon^{\frac{4s}{2s+2t+1/2}},$$

we get

$$R^2 a_{m(\epsilon)}^{-2} \leq c\gamma_\epsilon \Leftrightarrow m(\epsilon) \sim \gamma_\epsilon^{-1/2s} \sim \epsilon^{-\frac{2}{2s+2t+1/2}},$$

and thus

$$\mu_\epsilon \sim m(\epsilon)^{-2\beta} \gamma_\epsilon \sim \epsilon^{\frac{4s+4t}{2s+2t+1/2}} \sim (\rho_\epsilon^{DP})^2.$$

The term μ_ϵ corresponds to the minimax separation rate on $\mathcal{E}_{c,2}^K(R)$. Hence, Lemma 3 provides in this setting the following embedding

$$\left\{f \in \mathcal{E}_{a,2}^H(R),\ \|f\|^2 \geq (\rho_\epsilon^{IP})^2\right\} \quad \subset \quad \left\{f \in \mathcal{E}_{a,2}^H(R),\ \|Af\|^2 \geq C(\rho_\epsilon^{DP})^2\right\}. \tag{2.9}$$

In this context, let Φ_α a level-α test minimax for H_0^{DP} on $\mathcal{E}_{c,2}(R)$. In particular, there exists a constant $C > 0$ depending on α and β such that

$$\sup_{f \in \mathcal{E}_{a,2}^H(R),\ \|Af\|^2 \geq C(\rho_\epsilon^{DP})^2} P_f(\Phi_\alpha = 0) \leq \beta.$$

Using the embedding (2.9) and the previous inequality, we can find a constant C such that

$$\sup_{f \in \mathcal{E}_{a,2}^H(R),\ \|f\|^2 \geq (\rho_\epsilon^{IP})^2} P_f(\Phi_\alpha = 0) \leq \sup_{f \in \mathcal{E}_{a,2}^H(R),\ \|Af\|^2 \geq C(\rho_\epsilon^{DP})^2} P_f(\Phi_\alpha = 0) \leq \beta.$$

In other words, the test Φ_α is minimax for H_0^{IP} on $\mathcal{E}_{c,2}(R)$. We have provided an optimal detection procedure without regularization step. This result can be summarized in the following proposition.

Proposition 4 *Every level-α test minimax for H_0^{DP} on $\mathcal{E}_{a,2}(R)$ is also minimax for H_0^{IP} on $\mathcal{E}_{c,2}(R)$ in the case where $a_j \sim j^s$ and $b_j \sim j^t$.*

There exists several extensions to this results, considering for instance super-smooth functions, severely ill-posed problems or alternative functional spaces (source conditions). In each corresponding case, the proof follows essentially the same lines and will not be reproduced here. For more details and extended discussions, we refer to Laurent et al. (2010), Loubes and Marteau (2013) or Marteau and Mathé (2013)

The main conclusion of the above result is that regularization is not necessary in a signal detection purpose. Indeed, in such a setting, the goal is not to describe as precisely as possible the signal contained in the observations, but rather to provide a qualitative information, namely to say if there is signal or not. In practice, such a result has to be slightly nuanced.

- This is clearly an asymptotic claim: we adopt a minimax point of view and the discussion only make sense if we assume that $\epsilon \to 0$.

- We do not provide a sharp control of the constants involved in the separation rates. We refer for instance to [60], [61] and [62] for more details on this topic.

Nevertheless, up to these restrictions, it appears that direct tests provide an interesting behavior in an inverse framework.

2.1.4 Inverse tests fail for signal detection in the direct case

In the previous section, we have seen that direct testing strategies appear to perform well in an inverse problem context. We will see in this part that the reverse is not true, namely that we can find testing procedure minimax for H_0^{IP} on $\mathcal{E}_{a,2}(R)$ but that are not minimax for H_0^{DP} on

$\mathcal{E}_{a,2}(R)$. In particular, we have to exhibit a testing procedure, $\epsilon_0 > 0$ and a function $f_1 \in \mathcal{E}_{a,2}(R)$ such that

$$\|Af\| \geq C^\star \rho_\epsilon^{DP} \ \forall \epsilon > 0, \ \text{but} \ P_f(\Phi_\alpha = 0) > \beta,$$

for all $0 < \epsilon < \epsilon_0$, whatever the value of the constant $C^\star$.

This question has been addressed in Laurent et al. (2011). Let $f_1 \in \mathcal{X}$ the function defined as

$$\langle f, \phi_k \rangle^2 := \theta_k^2 = \begin{cases} \mathcal{C}_1 \epsilon^2 \sqrt{D^\star}, \ \text{if } k = 1, \\ 0 \text{ else}, \end{cases}$$

where $D^\star = D^{IP} \sim D^{DP}$ is defined in (2.5) and (2.7). Recall that Lemma 3 assesses the following embedding

$$\left\{ f \in \mathcal{E}_{a,2}(R), \ \|f\|^2 \geq (\rho_\epsilon^{IP})^2 \right\} \subset \left\{ f \in \mathcal{E}_{a,2}(R), \ \|Af\|^2 \geq (\rho_\epsilon^{DP})^2 \right\}.$$

The reciprocal of this lemma does not work. Indeed, we have

$$\|Af_1\|^2 \sim \theta_1^2 \sim (\rho_\epsilon^{DP})^2,$$

but

$$\|f_1\|^2 \sim (\rho_\epsilon^{DP})^2 << (\rho_\epsilon^{IP})^2,$$

as $\epsilon \to 0$. Let

$$\Phi_{\alpha,D^\star}^{IP} = \mathbf{1}_{\{\sum_{j=1}^{D^\star} b_j^{-2}(y_j^2 - \sigma^2) > t_{\alpha,D^\star}^{IP}\}} := \mathbf{1}_{\{T_D^\star > t_{\alpha,D^\star}^{IP}\}},$$

the test introduced in (2.4). Thanks to results obtained in Section 2.1.2, this test is known to be minimax H_0^{IP} for $\mathcal{E}_{a,2}^H(R)$. In particular

$$P_\theta(\Phi_{\alpha,D^\star}^{IP} = 0) \leq \beta, \ \text{as soon as } \|\theta\| \geq C_{\alpha,\beta} \rho_\sigma^{IP}.$$

We will show that for all $\mathcal{C}_1$ and β, there exists ϵ_0 such that for all $\epsilon < \epsilon_0$

$$P_{f_1}(\Phi_{\alpha,D^\star}^{IP} = 0) > \beta.$$

We use the following upper bound on $P_{f_1}(\Phi_\alpha = 1)$:

$$\begin{aligned} P_{f_1}\left(\Phi_{\alpha,D^\star}^{IP} = 1\right) &= P_{f_1}\left(\sum_{j=1}^{D^\star} b_j^{-2}(y_j^2 - \sigma^2) > g_{\alpha,D^\star} \right), \\ &\leq P_{f_1}\left(T_D^\star - \mathbb{E}_{f_1}(T_D^\star) > g_{\alpha,D^\star} - \mathbb{E}_{f_1}(T_D^\star)\right), \\ &\leq P_{f_1}\left(T_D^\star - \mathbb{E}_{f_1}(T_D^\star) > g_{\alpha,D^\star} - \theta_1^2\right), \\ &\leq \left[\frac{\mathrm{Var}(T_D^\star)}{g_{\alpha,D^\star} - \theta_1^2} \right]^2, \end{aligned}$$

Then, recall from Proposition 2 that

$$g_{\alpha,D^\star} \simeq \mathcal{C}_\alpha \sigma^2 \left(\sum_{j=1}^{D^\star} b_j^{-4} \right)^{1/2} >> \theta_1^2 = \mathcal{C}_1 \sigma^2 \sqrt{D^\star}, \ \text{as } \epsilon \to 0.$$

Hence, we can find $\mathcal{C}_0$ such that

$$P_{f_1}\left(\Phi^{IP}_{\alpha,D^\star} = 1\right) \leq \left[\frac{\mathcal{C}_0\sigma^2(\sum_{j=1}^{D^\star} b_j^{-4})^{1/2}}{g_{\alpha,D^\star} - \theta_1^2}\right]^2 \leq 1-\beta$$

for $\mathcal{C}^\star$ and ϵ^{-1} large enough.

The test $\Phi^{IP}_{\alpha,D^\star}$ is not powerful for H_0^{DP} on $\mathcal{E}_{c,2}(R)$. It is constructed in order to contain observations with large variances and is too conservative for our problem. Once again, a similar discussion can be provided for different kind of smoothness constraints, degree of ill-posedness and functional spaces (source conditions). We refer to our results in Laurent et al. (2010), Loubes and Marteau (2013) and Marteau and Mathé (2013) for more details.

Conclusion

If we summarize the previous results, direct tests possess a good behavior for inverse problems but the reverse is not true. We can indeed construct testing procedures, minimax for the inverse problem but that will fail for the direct problem. Such procedures are often designed in order to control large variances. Indeed, the testing procedure $\Phi_{\alpha,D^\star}$ used above is associated to the statistics

$$T_D = \sum_{k=1}^{D} b_k^{-2}(y_k^2 - 2), \text{ where } \mathrm{Var}(T_D) = \epsilon^4 \sum_{k=1}^{D} b_k^{-4},$$

up to constant. Since the sequence $(b_k^2)_{k\in\mathbb{N}}$ tends to 0, this variance can explodes for large values fo D. Such kind of test hence appears to be too conservative in order to deal with the 'direct' problem where typically the variance is quite smaller.

As a conclusion, we can recommend to use direct test in the framework considered in this chapter. Indeed, even if there is a loss of accuracy in the constants, such procedures appears to be robust w.r.t. the way where we are measuring error.

2.2 Some strategies for binary supervised classification

2.2.1 Analogies with testing theory

The aim of this section is to show that there is strong analogies between signal detection for inverse problems and binary classification. At this step, this discussion is quite informal. No precise results will be presented but rather a perspective discussion. All this section will be the core of our future investigations in this topic.

Before discussing the way where the inverse problem can be considered in smooth discriminant analysis with error in variable, we give just few words on analogies between tests and classification. Recall that in the classical smooth discriminant analysis problem (noise free case), we deal with two samples $\mathcal{S}_1 = (X_1^{(1)}, \dots, X_n^{(1)})$ and $\mathcal{S}_2 = (X_1^{(2)}, \dots, X_n^{(2)})$ where the $X_i^{(1)}$ (resp. the $X_i^{(2)}$) are assumed to admit a density f (resp. g) w.r.t. the Lebesgue measure on $\mathbb{R}^d$.

Given a new incoming observation X, our aim is then to decide whether $X \sim f$ of $X \sim g$. This decision is associated to a classifier, i.e. a set $G \subset K$. The corresponding error is then measured by the risk $R_K(G)$ defined as

$$
\begin{aligned}
R_K(G) &= \frac{1}{2}\int_{K/G} f(x)dx + \frac{1}{2}\int_G g(x)dx, \\
&= \frac{1}{2}P_{X\sim f}(X \notin G) + \frac{1}{2}P_{X\sim g}(X \in G).
\end{aligned}
$$

Remark that this problem may be in fact translated to a testing problem where one want to test

$$
H_a : X \sim f, \text{ against } H_b : X \sim g.
$$

In this context, contrary to the testing problems investigated above, there is no hierarchy between the hypotheses H_a and H_b. Then, each classifier (set) G can in fact be associated to a testing procedure Φ_G where

$$
\Phi_G = \mathbf{1}_{X\in G}.
$$

The corresponding risk $R_K(G)$ being in such a case related to the sum of the first and second kind errors. Once again, we do not favor any hypothesis: each error is associated to the same weighting $1/2$.

In this setting, the main difference with the classical testing framework is that it is not possible to ensure a prescribe level for the risk. Indeed, the risk associated to the oracle (the Bayes classifier) does not depend on the sample size (and hence does not tends to 0). As described above, the main task in this context consists in constructing a set $\hat{G}_{n,m}$ whose corresponding risk will be as close as possible of the smallest possible one. In other words, we want to control the excess risk.

The analogy between classification and the test theory has already been briefly discussed in the literature. We mention for instance [72] or more recently to Laurent et al. (2013) in an unsupervised classification context (see also Section 2.4 below).

2.2.2 A direct or indirect problem?

Now, we turn to the case where we have at our disposal two noisy samples $\mathcal{S}_1 = (Z_1^{(1)}, \dots, Z_n^{(1)})$ and $\mathcal{S}_2 = (Z_1^{(2)}, \dots, Z_n^{(2)})$ with

$$
Z_i^{(j)} = X_i^{(j)} + \epsilon_i^{(j)}, \ \forall j \in \{1,2\},
$$

and where the $\epsilon_i^{(j)}$ are i.i.d. random variables, that admits a density η w.r.t. the Lebesgue measure.

In Loustau and Marteau (2013a) and Loustau and Marteau (2013b), we have implicitly assumed that the new incoming observation X was free of noise. In term of modeling, this is not the only possible approach. Indeed, three different point of views are at hand:

(a) The new incoming observation contains measurement error, namely we observe $Z = X + \epsilon$ where the density of X w.r.t. the measure Q belongs to $\{f, g\}$ and the variable ϵ admits the density η w.r.t. the Lebesgue measure on $\mathbb{R}^d$. We wish to provide the best possible classifier for Z. If we use the formalism introduced in Section 2.2.1 above, the corresponding problem amounts to test

$$H_a^{DP} : Z \sim f * \eta, \text{ against } H_b^{DP} : Z \sim g * \eta. \tag{2.10}$$

The testing problem is a direct problem, which has already been considered by [76]. In such a setting, assumptions (smoothness, margin) are set on the couple $(f * \eta, g * \eta)$.

(b) The new incomming observation is free of noise, namely we directly observe the variable X. In this case, this amounts to test

$$H_a^{IP} : X \sim f, \text{ against } H_b^{IP} : X \sim g.$$

This is an inverse problem, which require a deconvolution step. This case has been considered in Loustau and Marteau (2013a) and Loustau and Marteau (2013b) where minimax excess risks are provided in different settings.

(c) The new incoming observation is noisy as in the case (a), but we want to approximate $G_K^\star$. This can be motivated by the fact that we want to understand the link between the spatial position of a variable and its affiliation to one of the two available labels (i.e. $X \sim f$ or $X \sim g$). Once again, we consider in such a case the testing problem

$$H_a^{IP} : X \sim f, \text{ against } H_b^{IP} : X \sim g. \tag{2.11}$$

The main difference in this setting is related to the strategy that we can adopt in order to take our decision. Indeed, we can still estimate the optimal set (classifier) $G_K^\star$ by $\hat{G}_{n,m}$ as described in one of the section above. Nevertheless, we can not use the classifier $\Phi_G = \mathbf{1}_{\{X \in \hat{G}_{n,m}\}}$ since the variable X is not directly observed. Instead, on can try for instance to use a classifier based on the functions $h_{G,\lambda}$ introduced in (1.34) in the following way

$$\Phi_G = \mathbf{1}_{\left\{h_{\hat{G}_{n,m},\lambda}(Z) > 1/2\right\}}. \tag{2.12}$$

The corresponding risk of such a classifier is not $R_K(G)$ in this case and should be precisely investigated.

The point of view (b), which has up to now retain our attention can be considered as an intermediate case between the direct (a) and indirect (c) point of views. The main advantage of this model (b) is that we are allowed to take into account noisy measurement, but the error is measured in the same way (through the risk $R_K(G)$).

If one want to get round of this noisy classification problem, the next step is to provide a complete study of the model (c) which appears to be more realistic. A particular attention should be paid on the risk associated to the classifier (2.12), which appears to be

$$\tilde{R}_K(\hat{G}_{n,m}) = \frac{1}{2} P_f \left(h_{\hat{G}_{n,m},\lambda}(Z) \leq 1/2 \right) + \frac{1}{2} P_g \left(h_{\hat{G}_{n,m},\lambda}(Z) > 1/2 \right).$$

Such a quantity seems to be difficult to handle, since the function f involve a deconvolution step. Hence, specific algebra should be proposed in order to solve this problem.

Provided that this study can be achieved, the remaining question will be: *What is the best possible strategy?* Indeed, if one get two noisy sample and a noisy observation, should we try to invert the problem (i.e. to include different deconvolution steps in order to remove errors, or directly work with the noisy sample (in a [76] manner)? This is an interesting open question. By the way, we are faced to some new difficult theoretical problems. For instance, provided that the couple (f, g) satisfies the margin assumption, what can be said on the couple $(f * \eta, g * \eta)$. The answer is certainly not obvious since the margin assumption is a local (spatial) constraint, while convolution has effects on the frequency domain.

2.3 Alternative ways to measure errors

The main conclusion of Section 2.1 is that if a test is designed for the direct setting, then it will be minimax for the inverse setting. As mentioned above, using such a principle provides a robust method, that works both in direct and indirect setting. The most interesting property is that the reverse is not true.

Hence, a natural question arrises: *Can this principle be extended to other settings than testing theory?* The case of binary classification has been discussed in the previous section. As explained, there is a lot of work to do in order to get round of this question...

Nevertheless, there is a topic that could provide interesting properties: estimation in Gaussian white noise. Indeed, as a consequence of the results obtained in the testing theory, a change in the test design can lead to a more robust procedure. The following heuristic is very close to the one developed in the testing theory.

Minimax rates in the Gaussian white noise model

The same kind of methodology could be perhaps extended to the estimation/prediction task. Indeed, consider the Gaussian white noise model

$$y_k = b_k \theta_k + \epsilon \xi_k, \ \forall k \in \mathbb{N},$$

which can be re-written as

$$y_k = \nu_k + \epsilon \xi_k, \ \forall k \in \mathbb{N},$$

where $\nu_k = b_k \theta_k$. In this context, two different tasks can be achieved:

- **The estimation approach.** Our aim is to provide an estimator for the sequence θ (i.e. the function f). If one consider projection (spectral cut-off) regularization, we get estimator $\hat{\theta}_N$ of the form

$$\hat{\theta}_N = \sum_{k=1}^{N} b_k^{-1} \theta_k,$$

for some $N \in \mathbb{N}$, with corresponding quadratic risk $R(\theta, N)$ defined as

$$R(\theta, N) := \mathbb{E}_\theta \|\hat{\theta}_N - \theta\|^2 = \sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2}.$$

For the sake of convenience, we will assume in the following that $f \in \mathcal{E}_{a,2}(R)$ with $a_k \sim k^s$ and $b_k \sim k^{-t}$. In this context, we get

$$\sup_{\theta \in \mathcal{E}_{a,2}(R)} R(\theta, N_{IP}) \leq C \left(R N_{IP}^{-2s} + \epsilon^2 N_{IP}^{2t+1} \right) \sim \epsilon^{\frac{4s}{2s+2t+1}}, \text{ provided } N_{IP} \sim \epsilon^{-\frac{2}{2s+2t+1}}. \quad (2.13)$$

In particular, the term $\epsilon^{\frac{4s}{2s+2t+1}}$ corresponds to the minimax rates of convergence over the ellipsoid $\mathcal{E}_{a,2}(R)$. Now, the problem with the bandwidth N_{IP} introduced in (2.13) is that it explicitly depends on the smoothness index s which is unknown in several practical problems. As described in Section 1.2.3, it is necessary to use adaptive algorithm. For the sake of convenience, we will focus on the URE approach. In this context, one use

$$\hat{N} := \arg\min_{N \in \mathbb{N}} U(Y, N), \text{ where } U(Y, N) = -\sum_{k=1}^{N} b_k^{-2} y_k^2 + 2\epsilon^2 \sum_{k=1}^{N} b_k^{-2}. \quad (2.14)$$

Performances of this adaptation algorithm are for instance investigated in [25], and sharp oracle inequalities are provided.

- **The prediction approach.** In this context, our aim is to provide the best possible estimation of the parameter ν (i.e. of the function Af). This problem appears to be a direct problem, which has been widely investigated in the literature. If one consider projection estimator of the form

$$\hat{\nu}_N = \sum_{k=1}^{N} y_k \psi_k,$$

for some $N \in \mathbb{N}$, we get a corresponding quadratic risk defined as

$$R^{DP}(\theta, N) := \mathbb{E}_\theta \|\hat{\nu}_N - \nu\|^2 = \sum_{k>N} b_k^2 \theta_k^2 + \epsilon^2 N.$$

Then,

$$\sup_{\theta \in \mathcal{E}_{a,2}(R)} R^{DP}(\theta, N_{DP}) \leq C \left(R N_{DP}^{-2(s+t)} + \epsilon^2 N_{DP} \right) \sim \epsilon^{\frac{4(s+t)}{2s+2t+1}},$$

provided

$$N_{DP} \sim \epsilon^{\frac{2}{2s+2t+1}}, \quad (2.15)$$

as $\epsilon \to 0$. Once again, we can propose in this context a data driven parameter choice for the bandwidth N_{DP}. The URE algorithm leads to

$$\tilde{N} := \arg\min_{N \in \mathbb{N}} V(Y, N), \text{ where } V(Y, N) = -\sum_{k=1}^{N} y_k^2 + 2\epsilon^2 N.$$

Sharp oracle inequalities are also available in this context.

As in a testing context, it appears that the same number of coefficients is required whatever the kind of inference (i.e. estimation or prediction) is performed with our sample. In other words, we use exactly the same amount of information in a minimax setting. Below, we propose an heuristic algorithm, which is inspired by this property.

An heuristic algorithm and some numerical simulations

In Section 2.1, we have seen that it was necessary to *invert* the data in order to determine if there is signal or not in the observation. It may be perhaps possible to extend this idea to estimation. Since in this context, we want to provide an quantitative study, the inversion step is unavoidable if one want to recover the sequence θ. Nevertheless, the adaptation algorithm may be improved. Indeed, thanks to (2.13) and (2.15), the minimax bandwidth are of same order for both estimation and prediction. Moreover, the variance of the risk estimator $U(Y, N)$ is quite large:

$$\mathrm{Var}(U(Y,N)) = C\epsilon^4 \sum_{k=1}^{N} b_k^{-4} \ \forall N \in \mathbb{N},$$

for some constant C. Hence, in many cases, the URE algorithm (2.14), leads to very unstable results. On the other hand, we can remark that

$$\mathrm{Var}(V(Y,N)) = \epsilon^4 N << \mathrm{Var}(U(Y,N)),$$

for large values of N, since $b_k \to 0$ as $k \to +\infty$.

This motivates the following estimation procedures. We consider the adaptive estimator $f^\star$ defined as

$$f^\star = \sum_{k=1}^{\tilde{N}} b_k^{-2} y_k^2 \psi_k, \tag{2.16}$$

where $\tilde{N}$ is the *direct* adaptive bandwidth. In other words, given $A : \mathcal{X} \to \mathcal{Y}$ a compact operator, we estimate the function $f \in \mathcal{X}$ using a bandwidth designed on $\mathcal{Y}$.

Such an approach may certainly lead to pertinent adaptive estimation strategies, at least from a minimax point of view. From now on, minimax results in this context are not available. Some difficult theoretical problems are at hand. In particular, one has to prove that $\tilde{N}$ is close to $\hat{N}$ with a large probability. This is not obvious. Nevertheless, investigation on theoretical properties of $f^\star$ will be at the heart of our forthcoming investigations.

In order to motivate this algorithm, we provide a brief numerical comparison of the different methods presented above. Once again, we did not obtain yet theoretical results. The aim of this experimental study is only to show that adaptive inverse estimation with direct data-driven parameters could provide interesting perspectives.

Below, we consider the sequence θ^a defined as

$$\theta_k^a = \frac{0.1 \times a}{1 + (k/6)^6}, \ \forall k \in \mathbb{N}.$$

For all $a \in \{0, \dots, 50\}$, we investigate the performances of the URE algorithm and the estimator (2.16) when $\epsilon = 0.1$. When consider alternatively two different degrees of ill posedness: $b_k = k^{-1}$ and $b_k = k^{-2}$ for all $k \in \mathbb{N}$. Then, each considered estimator $\hat{f}$ is compared to the spectral cut-off oracle via the quantity

$$\frac{\inf_{N \in \mathbb{N}} \mathbb{E}\|\tilde{f}_N - f\|^2}{\mathbb{E}\|\hat{f} - f\|^2},$$

where $\tilde{f}_N$ denotes the projection estimator associated to a given $N \in \mathbb{N}$. If this ratio is close to 1, this means that $\hat{f}$ is close to the oracle. Small values for this quantity indicate poor performances.

Remark that the function θ^a has been introduced in [26], which was devoted to the investigation of the risk hull method. Since this method appears to be a benchmark in this setting, it it is included to our numerical study. Results are summarized in Figure 2.1 bellow.

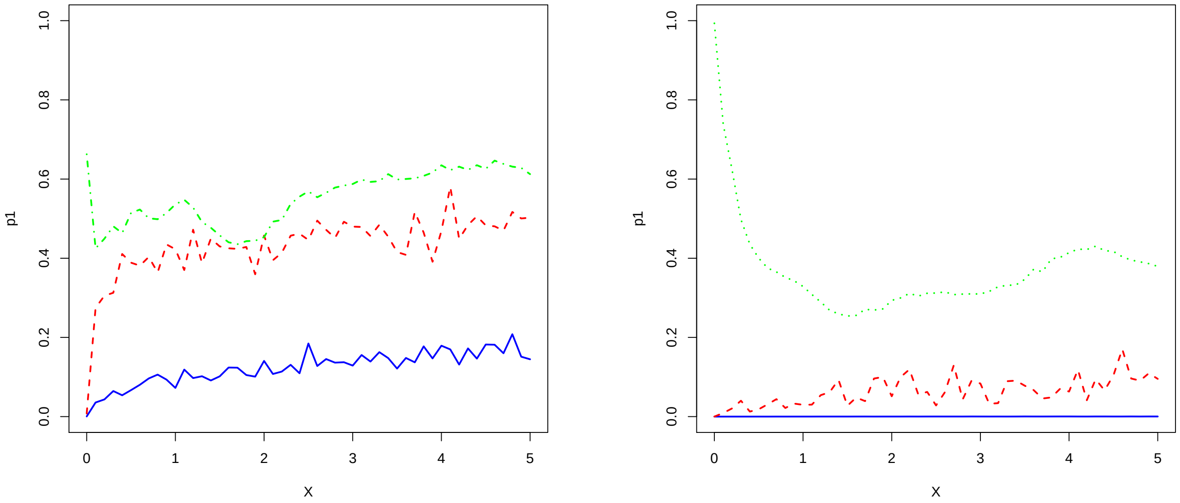

Figure 2.1: *Comparison of adaptive estimation algorithms (continuous line for the URE method, dashed line for 'direct' URE method and dotted lines for the RHM algorithm), for two different degrees of ill posedness ($b_k = k^{-1}$ on the left hand side, $b_k = k^{-2}$ on the right hand side).*

We can remark that in both cases, the URE algorithm does not provide satisfying performances: in most cases, it is far away from the oracle. As discussed above, this is essentially due to the fact that the variance of the risk estimator $U(Y, N)$ is too large, hence leading to very unstable recovery. Since the $f^\star$ is based on a direct approach, it provides a better behavior (for this numerical study). In particular, we can see that for large values of a, it is comparable to the RHM method introduced in [26]. The only cases where this algorithm is not pertinent is when there is a small amount of signal in the observation (a close to 0).

In order to provide an honest comparison, one should better penalize the risk estimator $V(Y, N)$ in order to avoide larges values of N. For instance, one could apply the direct RHM method that has also been considered in [26] in order to choose the bandwidth N and then plug this data-driven estimator in the inverse operator.

Conclusion

In this section, we have presented situations where direct methods may provide interesting outcomes. In particular, we have discussed cases where one want to provide some inference on a function $f \in H$, when dealing with an inverse problems associated to an operator

$$A : H \to K.$$

The main conclusion of this chapter is that it is possible to measure the error associated to our problem in an other space than H. This principle is used for tests, where the alternative is expressed over K, or in estimation where the bandwidth N is chosen in K.

At this step, there is several paths that could be investigated:

- investigation of the theoretical performances of the *direct URE* algorithm,
- error measurements in intermediate spaces,
- similar investigations in the error-in-variable model...

2.4 Some perspectives in non-parametric statistics

In this section, we discuss some additional topics that could be investigated in the forthcoming years.

Unsupervised classification and tests

In Laurent et al. (2013) we have addressed a mixture detection problem in dimension 1. In particular, given an univariate sample $\mathcal{S} = \{X_1, \dots, X_n\}$ having an unknown density f, our aim was to test

$$H_0 : f = \phi(. - \mu), \text{ against } H_1 : f = (1-\epsilon)\phi(. - \mu_1) + \epsilon\phi(. - \mu_2), \ (\epsilon, \mu_1, \mu_2) \in \mathcal{F}, \qquad (2.17)$$

where ϵ, μ, μ_1 and μ_2 are *unknown* parameters, while the common shape ϕ is assumed to be *known*. In other words, we want to determine whether the sample of interest is drawn from a single *population* having a density ϕ (up to a location parameter) or if it includes a *sub-population* having the same shape ϕ. In such a case, the parameters ϵ and μ_2 correspond respectively to size proportion and location parameters for this sub-population. We refer to [73] among others for practical examples of such a modeling.

The testing problem (2.17) has been widely considered last ten years. We mention for instance the contributions of [2], [50], [59], [39] or [21] in the particular case where $\mu = \mu_1$. In most of these papers, an asymptotic study of the power of the proposed tests is provided. In particular, asymptotic conditions are set on the triplet (ϵ, μ_1, μ_2), which ensure (optimal) separation of both hypotheses H_0 and H_1.

In Laurent et al. (2013), our aim was to describe as precisely as possible the contain of the alternative $\mathcal{F}$ in both asymptotic and non-asymptotic contexts. Moreover, we did not want to fix the mean μ under H_0. Hence, we have considered a translation model. We have proposed a testing procedure based on the ordered statistics $X_{(1)} < \cdots < X_{(n)}$. In particular, we can remark that

- the spacing of these order statistics are free with respect to the mean under H_0. For some $k < l \in \{1, \ldots, n\}$, the mean value affects the spatial position of a given $X_{(k)}$, but not $X_{(l)} - X_{(k)}$.
- the distribution of the variables $X_{(l)} - X_{(k)}$ is known under H_0.
- it has a different behavior under H_1, provided k and l are well-chosen.

Our testing procedure is based on theses properties. Assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \ldots, n/2\}$ defined as

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{X_{(n-k+1)} - X_{(k)} > q_{\alpha_n, k}} \right\}, \tag{2.18}$$

where, for all $u \in]0,1[$, $q_{u,k}$ is the $(1-u)$-quantile of $X_{(n-k+1)} - X_{(k)}$ under the null hypothesis and

$$\alpha_n = \sup \left\{ u \in]0,1[, \mathbb{P}_{H_0} \left(\exists k \in \mathcal{K}_n, X_{(n-k+1)} - X_{(k)} > q_{u,k} \right) \leq \alpha \right\}.$$

The terms $q_{\alpha_n,k}$ and α_n can be approximated (via Monte-Carlo simulations for instance) under the assumption that the X_i's have common density ϕ.

The following result highlights the non-asymptotic behaviour of the testing procedure Ψ_α. The corresponding proof and an extended non-asymptotic study can be found in Laurent et al. (2013).

Theorem 5 *Let $\alpha, \beta \in]0,1[$ be fixed and assume that $\mu_2 - \mu_1 \leq M$ for some constant $M > 0$. Then, there exists $C = C(\alpha, \beta, M) > 0$ such that*

$$\inf_{\psi_\alpha} \sup_{\epsilon(\mu_2-\mu_1)^2 > C/\sqrt{n}} P_f(\psi_\alpha = 0) \geq \beta.$$

Moreover, there exists $c = c(\alpha, \beta, M) > 0$ such that

$$\sup_{\epsilon(\mu_2-\mu_1)^2 > c\sqrt{\log\log(n)}/\sqrt{n}} P_f(\Psi_\alpha = 0) \leq \beta,$$

where the test Ψ is introduced in (2.18).

Following Theorem 5, testing is impossible as soon as $\mu_2 - \mu_1$ is bounded and $\epsilon(\mu_2 - \mu_1)^2 < C/\sqrt{n}$ for some constant C. By the same way, we have provided a corresponding upper bound for the test Ψ_α. Remark that upper and lower bounds match up to a log term. This is due to the adaptation scheme that we use in the construction of our test. This logarithm can be removed up to some slight technical modifications.

We now turn to the study of the asymptotic performances of our procedure. Two different asymptotic behaviors are often encountered in the literature (which can be partly deduced from Theorem 5), namely

- the *sparse regime* where

$$\varepsilon \underset{n\to+\infty}{\sim} n^{-\delta} \text{ and } \mu_2-\mu_1 \underset{n\to+\infty}{\sim} \sqrt{2r\log(n)}, \quad \text{with } \frac{1}{2}<\delta<1 \text{ and } 0<r<1.$$

- the *dense regime* where

$$\varepsilon \underset{n\to+\infty}{\sim} n^{-\delta} \text{ and } \mu_2-\mu_1 \underset{n\to+\infty}{\sim} n^{-r}, \quad \text{with } 0<\delta\leq\frac{1}{2} \text{ and } 0<r<\frac{1}{2}.$$

In this context, the following results sheds light on the asymptotic performances of our procedure. Once again, the proof and an extended discussion (with numerical simulations) can be found in Laurent et al. (2013).

Theorem 6 *In the* ***dense*** *regime, the detection boundary is $r^*(\delta)=\frac{1}{4}-\frac{\delta}{2}$: the detection is possible when $r<r^*(\delta)=\frac{1}{4}-\frac{\delta}{2}$ (for n large enough, the power of our test is greater than $1-\beta$). In the* ***sparse*** *regime, assume that $r>r^*(\delta)$ with*

$$r^*(\delta)=\begin{cases}\delta-\frac{1}{2} & \text{if } \frac{1}{2}<\delta<\frac{3}{4}\\ (1-\sqrt{1-\delta})^2 & \text{if } \frac{3}{4}\leq\delta<1\end{cases}.$$

Then, setting $f(.)=(1-\varepsilon)\phi_G(.-\mu_1)+\varepsilon\phi_G(.-\mu_2)$, we have, for n large enough,

$$\mathbb{P}_f(\Psi_\alpha=0)\leq\beta.$$

In the *dense* regime, we do not recover the existing results in the literature. This is essentially due to the fact that the mean under H_0 is unknown. In the **sparse** regime, the separation 'conditions' are the same when the mean μ under H_0 is unknown. In this case, our procedure provides a similar behavior compared to the state of the art (Higher-Criticism and Likelihood ratio tests in particular, see [39] for instance).

In this context, several possible extensions, motivated by some difficulties encountered in practical applications, are at hand:

- In a first time, a possible heteroscedasticity of the data should be considered. Indeed, the variance under H_0 and in the two component of the mixtures under H_1 are assumed to be equal to 1. In an ideal way, one want to deal instead with general variances σ,σ_1,σ_2 and to investigate the corresponding separation rules.

- The multivariate setting could be also a challenging problem. In particular, the simplicity of our procedure which is based on the spacing of the ordered statistics may allow such a generalization.

- Up to now, we have only considered the situation where we oppose two models: translation model against a two component mixture model. Generalization of the investigation to higher component degrees may be quite interesting, even if it requires to establish a precise methodology.

Shifted curves classification

In Section 1.1.3, we have investigated a binary classification problem (more precisely a problem of discriminant analysis). This study has been performed in finite dimensional case, in the sense that the variable of interest where assumed to belong to $\mathbb{R}^d$ for some fixed d.

Nevertheless, interesting practical applications concerns functional data: one would like to classify curves instead of vectors. For instance, in [9], the author considers real data, corresponding to ECG signals (a measure of the heart electrical activity). An example is provided where two different kind of *signal* are considered: signal related to healthy people, and arrhythmic ECG which are often associated to ill people. In this last paper, the estimation point of view is adopted. Nevertheless, one might want to discriminate different kind of ECG profiles, provided a learning sample is available: this is a curve classification problem.

More formally, one could deal with the following model. We observe two different groups of curves $(X_1^{(1)}, ..., X_n^{(1)})$ and $(X_1^{(1)}, ..., X_n^{(1)})$ where

$$dX_i^{(j)}(t) = f_j(t - \tau_i)dt + \sigma dW_{i,j}(t),\ \forall i \in \{1, \dots n\},\ j \in \{1, 2\},$$

where the $W_{i,j}$ denotes independent Brownian motions and f_1, f_2 unknown functions. Then, given a new observation,

$$dX(t) = f(t - \tau_i)dt + \sigma dW_{i,j}(t),$$

the goal would be to determine whether the corresponding curve belongs to the first or to the second group, i.e. whether f is equal to f_1 or f_2.

Different strategies could be investigated in this context. Following [8], one could project the data in a finite d-dimensional space and then apply classical classification rule on this projection. A particular attention should be payed to the control of the excess risk, and to a control of the bias following the asymptotics in n and σ.

Binary supervised classification with error in variable

In order to conclude this discussion, one could mention that there is many possible improvement to propose in supervised binary classification with error in variable (see Section 2.2). As discussed above, the ERM procedure that we have proposed appears to be sometimes outperformed in some specific situations. By the way, several algorithms provide interesting algorithmic behaviors: k-nearest neighborhoods, SVM, and so on... A special attention could hence be payed to the behavior of these procedures in this inverse problem context.

Summary of my different contributions

Published papers and preprints - Testing theory

B. Laurent, J.M. Loubes and C. Marteau (2011). Testing inverse problems: a direct or an indirect problem? *Journal of Statistical Planning and Inference*, 141, pp. 1849-1861.

B. Laurent, J.M. Loubes and C. Marteau (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electronic Journal of Statistics*, 6, pp 91-122.

J.M. Loubes and C. Marteau (2013). Goodness-of-fit testing strategies from indirect observations. *To appear in Journal of Nonparametric statistics.*

Preprints

Y. Ingster, B. Laurent and C. Marteau (2013). Signal detection for inverse problems in a multidimensional framework. *Submitted.*

C. Marteau and P. Mathé (2013). General regularization schemes for signal detection in inverse problems. *Submitted.*

Published papers and preprints - Supervised and unsupervised classification

S. Loustau and C. Marteau (2013a). Minimax fast rates for discriminant analysis with errors in variables. *To appear in Bernoulli.*

Preprints

B. Laurent, C. Marteau and C. Maugis-Rabusseau (2013). Non-asymptotic detection of mixtures with unknown mean. *Submitted.*

S. Loustau and C. Marteau (2013b). Noisy classification with boundary assumptions. *Submitted.*

Published papers - Estimation for statistical inverse problems

J. Bigot, S. Gadat, T. Klein and C. Marteau (2013). Intensity estimation of inhomogeneous Poisson processes from shifted trajectories. *Electronic Journal of Statistics*, 7, pp 881-931.

J.M. Loubes and C. Marteau (2012). Adaptive estimation for an inverse regression model with unknown operator. *Statistics and Risk Modeling*, 29, pp. 215-242.

J. Bigot, S. Gadat and C. Marteau (2010). Sharp template estimation in a shifted curves model. *Electronic Journal of Statistics*, 4, pp. 994-1021.

C. Marteau (2010b). The Stein hull. *Journal of Nonparametric Statistics*, 22, pp. 685-702.

C. Marteau (2010a). Risk hull method for spectral regularization in linear statistical inverse problems. *ESAIM P&S*, 14 (2010), pp. 409-434.

C. Marteau (2009). On the stability of the risk hull method for projection estimators. *Journal of Statistical Planning and Inference*, 139, pp. 1821-1835.

C. Marteau (2006). Regularization of inverse problems with unknown operator. *Mathematical Methods of Statistics*, 15, pp. 415-443.

Other

C. Marteau. Recherche d'inégalités oracles pour des problèmes inverses. PhD. Thesis, November 2007.

Short CV

Education and work experience

- **Since September 2008**: Assistant professor in Institut National des Sciences Appliquées (INSA) de Toulouse.
- **2007-2008**: Assistant temporaire d'enseignement et de recherche. Université de Provence.
- **2004-2007**: PhD. Thesis under the supervision of Professor L. CAVALIER, defended the 29th November 2007. Université de Provence.
- **2002-2004**: Master degree in Applied Mathematics. Université de Provence.

PhD. Student

Maikol Solís Chacón (since September 2010): Estimation non-paramétrique de fonctionnelles.

Scientific and administrative responsabilities

- Since 09/2012: International relations coordinator for the Mathematical Department (GMM) in INSA Toulouse.
- 2011: Member of the local organization comity of the Journées de statistiques du sud 2012 (INSA, Toulouse).
- From 09/2010 to 01/2013: Organisation of the statistical seminar at the IMT (Mathematical Institute of Toulouse).
- From 06/2009 to 12/2011: Local contact of the SMAI (Société de Mathématiques Appliquées et Industrielles) in Toulouse.
- 2009: Local organization comitee of EMS (Toulouse).
- From 01/2006 to 06/2007: Organisation of the PhD. Students seminar in Université de Provence.

Other

Referee's reports for the journals: *Annals of Statistics, Journal of Statistical planning and inference, Electronic Journal of Statistics, Communications in Statistics: Theory and methods.*

Member of the ANR project DEMOS (Deformable Models in Statistics).

Bibliography

[1] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[2] Jean-Marc Azaïs, Élisabeth Gassiat, and Cécile Mercadier. The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM Probab. Stat.*, 13:301–327, 2009.

[3] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.

[4] Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.

[5] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6 (2002). pp. 127-146.

[6] J. Baumeister. *Stable solution of inverse problems.* Advanced lectures in mathematics. Friedr. Vieweg and Sohn, Braunschweig, 1987.

[7] E.N. Belitser and B.Ya. Levit. On minimax filtering on ellipsoids. *Math. Meth. Statist.*, 4 (1995), pp. 259-273.

[8] G. Biau, F. Bunea, and W. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172, 2005.

[9] J. Bigot. Fréchet means of curves for signal averaging and application to ecg data analysis. *Preprint.*

[10] J. Bigot and S. Gadat. A deconvolution approach to estimation of a common shape in a shifted curves model. *Ann. Statist.*, 38(4):2422–2464, 2010.

[11] J. Bigot, F. Gamboa, and M. Vimond. Estimation of translation, rotation, and scaling between noisy images using the fouriermellin transform. *SIAM Journal on Imaging Sciences*, 2:614–645, 2009.

[12] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3 (2001). pp. 203-268.

[13] N. Bissantz, G. Claeskens, H. Holzmann, and A. Munk. Testing for lack of fit in inverse regression, with applications to biophotonic imaging. *J. Royal Statisti. Soc. Ser. B*, 71:25–48, 2009.

[14] N. Bissantz, T. Hohage, and A. Munk. Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, 20 (2004), pp. 1773-1789.

[15] N. Bissantz, T. Hohage, A. Munk, and F. Ryumgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numerical Analysis*, 45:2610–2636, 2007.

[16] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[17] C. Butucea. Goodness-of-fit testing and quadratic functionnal estimation from indirect observations. *Annals of Statistics*, 35:1907–1930, 2007.

[18] C. Butucea and F. Comte. Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli*, 15(1):69–98, 2009.

[19] C. Butucea, C. Matias, and C. Pouet. Adaptivity in convolution models with partially knwon noise distribution. *Electronic Journal of Statistics*, 2:1935–7524, 2008.

[20] Cristina Butucea. Deconvolution of supersmooth densities with smooth noise. *Canad. J. Statist.*, 32(2):181–192, 2004.

[21] T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(5):629–662, 2011.

[22] E. Candès. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15 (2006), pp. 257-325.

[23] M. Carrasco, J-P. Florens, and E. Renault. *Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization*, volume 6. North Holland, 2006.

[24] I. Castillo and J-M. Loubes. Estimation of the law of random shifts deformation. Preprint.

[25] L. Cavalier, G.K. Golubev, D. Picard, and A.B. Tsybakov. Oracle inequalities for inverse problems. *Annals of Statistics*, 30 (2002), pp. 843-874.

[26] L. Cavalier and Y. Golubev. Risk hull method and regularization by projections of ill-posed inverse problems. *Annals of Statistics*, 34 (2006). pp. 1653-1677.

[27] L. Cavalier and N.W. Hengartner. Adaptative estimation for inverse problems with noisy operators. *Inverse Problems*, 21 (2005), pp. 1345-1361.

[28] L. Cavalier and M. Raimondo. Wavelet deconvolution with noisy eigen-values. *IEEE Trans. on Signal Processing*, 55 (2007), pp. 2414-2424.

[29] L. Cavalier and A.B. Tsybakov. Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields*, 123 (2002), pp. 323-354.

[30] L. Cavalier and A.B. Tsybakov. Penalized blockwise Stein's method, monotone oracles and sharp adaptative estimation. *Mathematical methods of Statistics*, 3 (2001), pp. 247-282.

[31] Laurent Cavalier. Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, volume 203 of *Lect. Notes Stat. Proc.*, pages 3–96. Springer, Heidelberg, 2011.

[32] X. Chen and M. Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Cowles Foundation Discussion Paper No. 1626*, 2008.

[33] A. Cohen, M. Hoffmann, and M. Reiß. Adaptive wavelet galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, 42 (2004). pp. 1479-1501.

[34] F. Comte and C. Lacour. Data driven density estimation in presence of unknown convolution operator. *J. Royal Stat. Soc., Ser B.*, 73 (4):601–627, 2011.

[35] A.S. Dalalyan, G.K. Golubev, and A.B. Tsybakov. Penalized maximum likelihood and semiparametric second-order efficiency. *The Annals of Statistics*, pages 169–201, 2006.

[36] A. Delaigle and I. Gijbels. Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.*, 56:19–47, 2004.

[37] Aurore Delaigle and Alexander Meister. Nonparametric function estimation under Fourier-oscillating noise. *Statist. Sinica*, 21(3):1065–1092, 2011.

[38] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[39] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[40] D.L. Donoho. Nonlinear solutions of linear inverse problems by wavelet-vaguelette decomposition. *Journal of Applied and Computationnal Harmonic Analysis*, 2 (1995), pp. 101-126.

[41] D.L. Donoho and I.M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26 (1998), pp. 425-455.

[42] D.L. Donoho and I.M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *CR Acad. Sci. Paris Sr. I Math.*, 319 (1994), pp. 1317-1322.

[43] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81 (1994), pp. 879-921.

[44] S. Efromovich and V. Koltchinskii. On inverse problems with unknown operators. *IEEE Trans. Inform. Theory*, 47 (2001), pp. 2876-2893.

[45] H.W. Engl. On the choice of the regularization parameter for iterated Tikhonov regularization of ill-posed problems. *Journal of approximation theory*, 49 (1987), pp. 55-63.

[46] H.W. Engl, M. Hank, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer Academic Publishers Group, Dordrecht, 1996.

[47] M.S Ermakov. Minimax detection of a signal in the heteroscedastic gaussian white noise. *J. Math. Sci.*, 137, No. 1:4516–4524, 2006.

[48] M.S Ermakov. Minimax estimation of the solution of an ill-posed convolution type problem. *Problems of Information Transmission*, 25 (1989), pp. 191-200.

[49] J. Fan. On the optimal rates of convergence for nonparametric deconvolutions problems. *Annals of Statistics*, 19 (1991), pp. 1257-1272.

[50] Bernard Garel. Recent asymptotic results in testing for mixtures. *Comput. Statist. Data Anal.*, 51(11):5295–5304, 2007.

[51] A. Goldenshluger and S.V. Pereverzev. Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probab. Theory Relat. Fields*, 118 (2000), pp. 169-186.

[52] G.K Golubev and R.Z Khasminskii. Statistical approach to some inverse boundary problems for partial differential equations. *Problems of Information Transmission*, 35 (1999), pp. 51-66.

[53] P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.*, 33(6):2904–2929, 2005.

[54] M. Hanke. Accelerated Lanweber iterations for the solution of ill-posed equations. *Numerische mathematik*, 60 (1991). pp. 341-373.

[55] T. Hida. *Brownian motion.* Springer-Verlag, New-York, 1980.

[56] M. Hoffmann and M. Reiß. Nonlinear estimation for linear inverse problems with error in the operator. *Annals of Statistics*, 36:310–336, 2008.

[57] H. Holzmann, N. Bissantz, and A. Munk. Density testing in a contaminated sample. *J. Multivariate Anal.*, 98(1):57–75, 2007.

[58] I.A. Ibragimov and Y.A. Rosanov. Gaussian random processes. 1978.

[59] Y. Ingster. Minimax detection of a signal for l^n-balls. *Mathematical Methods of Statistics*, 7(4):401–428, 1999.

[60] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.*, 2(2):85–114, 1993.

[61] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Math. Methods Statist.*, 2(3):171–189, 1993.

[62] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Math. Methods Statist.*, 2(4):249–268, 1993.

[63] Yu.I. Ingster, T. Sapatinas, and I.A. Suslina. Minimax signal detection in ill-posed inverse problems. *Annals of Statistics*, 40:1524–1549, 2012.

[64] J. Johannes and M. Schwarz. Adaptive gaussian inverse regression with partially unknown operator. 2012.

[65] I.M. Johnstone. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statistica Sinica*, 9 (1999), pp. 51-83.

[66] I.M. Johnstone, G. Kerkyacharian, D. Picard, and M. Raimondo. Wavelet deconvolution in a periodic setting. *J. R. Statist. Soc. B*, 66, part 3 (2004), pp. 547-573.

[67] I.M. Johnstone and B.W. Silverman. Speed of estimation in positron emission tomography and related inverse problems. *Annals of Statistics*, 18 (1990), pp. 251-280.

[68] I.M. Johnstone and B.W. Silverman. Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B*, 59 (1997), pp. 319-345.

[69] G. Kerkyacharian, P. Petrushev, D. Picard, and T. Willer. Need-vd: a second-generation wavelet algorithm for estimation in inverse problems. *Electronic Journal of Statistics*, 1:30–76, 2007.

[70] G. Kerkyacharian and D. Picard. Minimax or maxisets? *Bernoulli*, 8, (2002), pp. 219-253.

[71] A. Kneip. Ordered linear smoother. *Annals of Statistics*, 22 (1994). pp. 835-866.

[72] B.W. Igl L. Dumbgen and A. Munk. P-values for classification. *Electronic Journal of Statistics*, 2:1935–7524, 2008.

[73] G. Mc Lachlan and D. Peel. *Finite Mixture Models.* Wiley series in Probability and Statistics, 2000.

[74] J.M Loubes and C. Ludena. Penalized estimators for non-linear inverse problems. Preprint.

[75] B. Mair and F.H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56 (1996), pp. 1424-1444.

[76] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.

[77] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9 (2):245–303, 2000.

[78] P. Mathé and S.V. Pereverzev. Discretization strategy for linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19 (2003), pp. 1263-1277.

[79] P. Mathé and S.V. Pereverzev. Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. *SIAM J. Numer. Anal.*, 38 (2001). pp. 1333-2021.

[80] C. Matias. Semiparametric deconvolution with unknown noise variance. *ESAIM P&S*, 6:271–292, 2002.

[81] A. Meister. *Deconvolution problems in nonparametric statistics.* Springer-Verlag, 2009.

[82] A. Nemirovski. *Topics in Non-Parametric Statistics.* Lectures on Probability Theory and Statistics. Ecole d'été de Probabilités de St.Flour XXVIII, 1998.

[83] W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[84] M. Nussbaum. Spline smoothing in regression models and asymptotical efficiency in l_2. *Annals of Statistics*, 13 (1985), pp. 984-997.

[85] P. Rigollet. Adaptive density estimation using stein's blockwise method. *Bernoulli*, 12 (2006), pp. 351-370.

[86] D.N. Ghosh Roy and L.S. Couchman. *Inverse problems and inverse scattering of plane waves.* Academic press San Diego, 2002.

[87] C.M. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9 (1981), pp. 1135-1151.

[88] A.B. Tsybakov. *Introduction à l'estimation non-paramétrique.* Springer-Verlag, 2004.

[89] A.B. Tsybakov. On the best rate of adaptative estimation in some inverse problems. *C.R Acad. Sci. Paris, ser. 1.*, 330 (2000), pp. 835-840.

[90] A. W. van der Vaart and J. A. Weelner. *Weak convergence and Empirical Processes. With Applications to Statistics.* Springer Verlag, 1996.

[91] M. Vimond. Efficient estimation for a subclass of shape invariant models. *The annals of Statistics*, 38:1885–1912, 2010.

[92] T. Willer. *Estimation non-paramétrique et problèmes inverses.* Thèse de l'Université Paris VII - Denis Diderot, 2006.