# Some (statistical) applications of Ockham's principle

Erwan Le Pennec

UNIVERSITÉ PARIS SUD

Habilitation à Diriger des Recherches

# Some (statistical) applications of Ockham's principle

Erwan Le Pennec

Soutenue le 19 mars 2013

Devant le jury constitué de:

| | | |
|---|---|---|
| G. Celeux, | Directeur de recherche, | Select - Inria / LMO - Université Paris Sud |
| A. Juditsky, | Professeur, | Laboratoire J. Kuntzmann - Université Joseph Fourier (rapporteur) |
| S. Mallat, | Professeur, | DI - ENS Ulm |
| P. Massart, | Professeur, | LMO - Université Paris Sud / Select - Inria (rapporteur) |
| D. Picard, | Professeur, | LPMA - Université Paris Diderot |
| A. Trouvé, | Professeur, | CMLA - ENS Cachan (rapporteur) |

Numquam ponenda est pluralitas sine necessitate
Plurality must never be posited without necessity

**William of Ockham** (c. 1288 – c. 1348)

William of Ockham was a English Franciscan friar of the $14^{\text{th}}$ century. He was also a scholastic philosopher which is often credited for the law of parsimony that states roughly that among good explanations one should favor the simplest ones. Although he was not the first to postulate this principle of simplicity, this idea can be traced back to Aristotle, his name is now associated to this idea which is often called Ockham's principle or Ockham's razor.

I would not consider this principle as a philosophical doctrine but rather as a loose heuristic principle used in sciences: a *good solution* is often obtained by balancing its *apparent goodness* and its *complexity*. Although very naive, this principle turns out to be the cornerstone of my scientific contributions. For sure, we will have to specify definitions for *good solution*, *apparent goodness* and *complexity*, but it is nevertheless amusing to me that *up to some technicalities* this simple idea summarizes most of my works.

# Contents

# Chapitre 0

# Survol

Ce manuscrit décrit mon travail scientifique de ces dix dernières années. Celui peut-être découpé autour des dix thèmes suivants :

### Thème 0. Bandelettes et approximations (2000-2005)

Dans ce travail, qui correspond à ma thèse de doctorat réalisé sous la direction de S. Mallat, j'ai construit une nouvelle représentation d'image adaptée aux images géométriques : la représentation en bandelettes. Nous avons démontré que les propriétés d'approximation de ce frame adaptatif conduisait pour les images géométriques à des vitesses d'approximation non linéaires adaptatives à des termes logarithmique près. Ces résultats ont été transcrits en des résultats de compression à la fois théoriquement et numériquement.

### Thème 1. Bandelettes et estimations (2002-2011)

Profitant des propriétés d'approximations des bandelettes, nous avons proposé un premier algorithme de débruitage d'image basé sur le principe MDL (Comprimer c'est presque estimer). En utilisant ensuite les bases de bandelettes de seconde génération de G. Peyré, nous avons en collaboration avec Ch. Dossal utilisé les techniques de L. Birgé et P. Massart pour proposer un algorithme de sélection de modèles de bandelettes dont nous avons pu prouver la quasi optimalité minimax pour des images géométriques.

### Thème 2. Maxiset (2004-2009)

Une question naturelle est de se demander pour quelles fonctions l'estimateur précédent est-il efficace. Dans ce travail réalisé en collaboration avec F. Autin, J.-M. Loubes et V. Rivoirard, on montre que, sous des hypothèses faibles de structure, les fonctions bien estimées sont exactement celles bien approchées par des modèles de faibles dimensions de la collection.

### Thème 3. Dantzig (2006-2010)

En partant d'une question que nous a posée S. Tsybakov, j'ai étudié avec K. Bertin et V. Rivoirard une variation du Lasso dans un cadre d'estimation de densité. Nous avons pu calibrer de manière fine la pénalité de cet estimateur de Dantzig et vérifier théoriquement et numériquement ses performances.

### Thème 4. Radon (2007-2012)

En partant de la superbe construction de la représentation en needlet de P. Petrushev et des coauteurs, j'ai proposé avec G. Kerkyacharian et D. Picard une stratégie de seuillage en needlet pour inverser une transformée de Radon d'un objet dans le cadre du modèle du bruit blanc. Nous avons prouvé que cet estimateur est adaptatif et presque minimax pour une large gamme d'espace de Besov. Ces résultats ont été confirmés numériquement. En utilisant une construction similaire, j'ai obtenu en collaboration avec

M. Bergounioux et E. Trélat des résultats de type inégalité oracle pour la transformée de Radon d'objet axisymétrique.

### Thème 5. Copule (2008-2009)

Les stratégies de seuillage en ondelettes sont connues pour être efficaces. Dans ce travail avec F. Autin et K. Tribouley, nous montrons que c'est également le cas pour l'estimation des copules. Nous avons obtenu des estimateurs adaptatifs et quasi minimax pour les espaces de Besov. Ces résultats théoriques ont été confirmés par nos expériences numériques.

### Thème 6. NL-Means (2008-2009)

Ce travail correspond au début de la thèse de J. Salmon. Nous avons essayé d'étudier le lien entre l'une des méthodes numériques les plus efficaces de débruitage, les NL-means, et le principe statistique de pondération exponentielles. Notre analyse a conduit à des améliorations numériques publiées par J. Salmon.

### Thème 7. Modèle du choix binaire (2010-2011)

Dans ce travail avec E. Gautier, nous étudions un modèle classique économétrique : le modèle du choix binaire. Il s'agit d'un problème inverse qui se combine à une estimation de densité conditionnelle. Nous proposons une technique basé sur les ondelettes permettant d'obtenir des estimations adaptatives.

### Thème 8. Segmentation d'image hyperspectrale (2010-2012)

Une des techniques les plus classiques de classification non supervisée est basée sur des mélanges de gaussiennes dont les composantes sont associées à des classes. Avec S. Cohen, nous avons proposé une extension de ce modèle permettant de prendre en compte une covariable importante dans les images hyperspectrales, la position. Nous avons proposé un schéma numérique efficace qui nous a permis de tester avec succès notre algorithme sur des jeux de données réels venant de la plateforme IPANEMA du synchrotron Soleil.

### Thème 9. Estimation de densité conditionnelle (2010-2012)

L'algorithme précédent repose sur une estimation de densité par sélection de modèles. Nous avons cherché, et réussi, à le justifier théoriquement. Nous avons ainsi donné une nouvelle technique d'estimation de densité conditionnelle par maximum de vraisemblance pénalisé dont on contrôle l'efficacité sous des hypothèses faibles. L. Montuelle travaille sous notre direction sur des extensions du modèle de mélanges de gaussienne rentrant dans ce cadre.

Après une introduction aux questions abordées à travers l'exemple jouet de l'estimation de la moyenne d'un vecteur gaussien, le manuscrit reprend, en anglais, la description de ces 10 thèmes. Il se concentre autour des thèmes de 1 à 9 en les organisant autour des modèles statistiques utilisés.

# Chapter 1

# Estimating the mean of a Gaussian: a toy example

We focus in this first chapter on a very simple statistic problem: estimation of the mean of a Gaussian random vector of covariance matrix a known multiple of the identity. Let $X \in \mathbb{R}^N$ be the unknown mean, this problem can be rewritten as the observation of

$$Y = S + \sigma W$$

where $W$ is a standard Gaussian vector and $\sigma$ the known standard deviation. Our goal is now to estimate $S$ from the observation $Y$.

## 1.1 Trivial estimate and James-Stein's one

The most natural estimate is the trivial one: one estimates $S$ by $\widehat{S} = Y$ itself. This is indeed an unbiased estimate which seems the best possible without any further assumptions. Using the quadratic risk, which corresponds here to the Kullback-Leibler risk up to the variance factor, as a *good solution* criterion, we obtain a goodness of

$$\mathbb{E}\left[\|S - \widehat{S}\|^2\right] = \mathbb{E}\left[\|S - Y\|^2\right] = \sigma^2 N.$$

Is this the best one can do? No, as proved first by James and Stein [JS61]. They proposed to *shrink* toward 0 the previous estimate using

$$\widehat{S}^{JS} = \left(1 - \frac{(N-2)\sigma^2}{\|Y\|^2}\right)_+ Y$$

and have shown that as soon as $n \geq 3$ this estimate always has a smaller quadratic risk than the trivial one. Heuristically, the shrinkage reduces the variance but augments the bias. The factor $\frac{(N-2)\sigma^2}{\|Y\|^2}$ is such that the gain is greater is than the loss. Note that this estimate is very close to the trivial one when the norm of $Y$ is large but much close to 0 when the norm of $Y$ is small.

## 1.2 Projection based estimator

Instead of being applied globally, this strategy could be applied locally. Indeed, if for any subset $I$ of $\{1, \ldots, N\}$, we define $\widehat{S}_I$ as the following projection estimator

$$\widehat{S}_I[k] = \begin{cases} Y[k] & \text{if } k \in I \\ 0 & \text{otherwise} \end{cases}$$

The quadratic risk of such an estimator is given by

$$\mathbb{E}\left[\|S - \widehat{S}_I\|^2\right] = \|S - S_I\|^2 + \sigma^2|I|$$

where $|I|$ denotes the cardinal of $I$ and

$$S_I[k] = \begin{cases} S[k] & \text{if } k \in I \\ 0 & \text{otherwise} \end{cases}.$$

Obviously, as soon as the deterministic quantity $\|S - S_I\|^2$ is smaller than $\sigma^2|I|$, this estimator outperforms the trivial one. The best subset $I_O$ is the one that minimizes over all subsets $I$

$$\|S - S_I\|^2 + \sigma^2|I|.$$

This *good solution* is thus obtained by an application of Ockham's principle where the *apparent goodness* is the quadratic norm of the bias and the *complexity* the variance term, that is proportional to the size of $I$. A straightforward computation shows that $I_O$ can be expressed as the set of index of coefficients larger, in absolute value, than the standard deviation $\sigma$:

$$I_O = \{k \in \{1, \dots, n\} | |S_k| \geq \sigma\}.$$

The best subset $I_O$ depends on the unknown mean $S$ and thus $\widehat{S}_{I_O}$ is not an acceptable estimator of $S$. It is however often called an oracle estimator.

## 1.3 Model selection and thresholding

Two strategies, strictly equivalent in the setting, can be used to obtain an estimator almost as efficient as the previous oracle one. In the model selection approach one starts from the definition of $I_O$ as the minimizer of the quadratic risk while in the thresholding approach one capitalizes on its definition as the set of indices of the largest coefficients.

As observed by Akaike [Ak74] and Mallows [Ma73],

$$\|Y - \widehat{S}_I\|^2 + \sigma^2(2|I| - N)$$

is an unbiased estimator of the quadratic risk of $\widehat{S}_I$. This suggest as in the $C_p$ and AIC approach to replace $I_O$ by the subset minimizing the previous quantity or equivalently the subset minimizing

$$\|Y - \widehat{S}_I\|^2 + 2\sigma^2|I|.$$

This *good solution* is again obtained as a tradeoff between an *apparent goodness* measured by the quadratic distance between the observation $Y$ and the proposed estimate $\widehat{S}_I$, and a *complexity* measured by the size of $I$ times $2\sigma^2$. Our hope is to prove that if $\widehat{I}^{AIC}$ is the subset minimizing the previous quantity then the risk of the estimate $\widehat{S}_{\widehat{I}^{AIC}}$ is small. One way to prove this is to obtain than

$$\mathbb{E}\left[\|S - \widehat{S}_{\widehat{I}^{AIC}}\|^2\right] \leq C \min_{I \subset \{1,\dots,n\}} \|S - S_I\|^2 + \sigma^2|I| = C \min_{I \subset \{1,\dots,n\}} \mathbb{E}\left[\|S - \widehat{S}_I\|^2\right].$$

Such an inequality, called an oracle inequality, means that the estimate $\widehat{S}_{\widehat{I}^{AIC}}$ is almost as efficient as the best fixed one in the family $\widehat{S}_I$. It turns out that the complexity term $2\sigma^2$, often called penalty, is not large enough to ensure this behavior when $I$ can be chosen among all subsets.

Starting from the observation that

$$\widehat{I}^{AIC} = \left\{k \in \{1, \dots, N\} \middle| |Y[k] \geq \sqrt{2}\sigma\right\}$$

is almost a good replacement for $I_O$, Donoho and Johnstone [DJ94b] propose to replace it with

$$\widehat{I}^{Th} = \{k \in \{1, \dots, N\} | |Y[k]| \geq T\}$$

with $T = \sigma\sqrt{2\log N}$. They are then able to prove that

$$\mathbb{E}\left[\|S - \widehat{S}_{\widehat{I}^{Th}}\|^2\right] \leq (2\log n + 2.4)\left(\min_{I\subset\{1,\dots,N\}} \|S - S_I\|^2 + \sigma^2|I| + \sigma^2\right).$$

The adaptive estimator $\widehat{S}_{\widehat{I}^{Th}}$ achieves up to a logarithmic factor the best risk among the family considered.

A very similar result can be obtained following the model selection approach of Barron, Birgé, and Massart [BBM99], they explicitly modified the *complexity* term of AIC and define $\widehat{I}^{MS}$ as the minimizer of

$$\|Y - \widehat{S}_I\|^2 + T^2|I|$$

with $T = \kappa\sigma(1 + \sqrt{2\log N})$ and $\kappa > 1$. They are then able to prove

$$\begin{aligned}\mathbb{E}\left[\|S - \widehat{S}_{\widehat{I}^{MS}}\|^2\right] &\leq C(\kappa)\left(\min_{I\subset\{1,\dots,N\}} \|S - S_I\|^2 + T^2|I| + \sigma^2\right)\\ &\leq \left(C(\kappa)\kappa^2(1 + \sqrt{2\log N})^2\right)\left(\min_{I\subset\{1,\dots,n\}} \|S - S_I\|^2 + \sigma^2|I| + \sigma^2\right)\end{aligned}$$

where $C(\kappa) > 1$. Obviously those estimators coincide for the same threshold $T$ and thus Donoho and Johnstone's result, which allows a smaller threshold and yields a smaller constant in the oracle inequality, is better in this case.

Both principles can be related to Ockham's principle. In thresholding, each coefficient is considered individually: if it is large enough and thus useful for the *apparent goodness* and one should estimate it in a *good solution* and thus pay its price in term of *complexity*, otherwise not. In model selection, coefficients, or more precisely spaces, are considered jointly: the *good solution* is obtained by explicitly balancing an *apparent goodness*, the contrast, and a *complexity*, a multiple of the dimension. Although this approach coincide when dealing with orthonormal bases, we will see that this is not always the case.

## 1.4 Importance of the basis

We have considered so far estimators that are projection of the observation on spaces spanned by the axes of the canonical basis. As we are working in a quadratic norm the very same results holds after any change of orthonormal basis. If we denote by $\Phi$ the matrix in which each column is an element of the new basis, our estimator $\widehat{S}_I$ becomes $\widehat{S}_{\Phi,I} = \Phi(\Phi' S)_I$ where

$$(\Phi' S)_I[k] = \begin{cases}(\Phi' S)[k] & \text{if } k \in I\\ 0 & \text{otherwise}\end{cases}.$$

From both the thresholding and the model selection approaches, we obtain

$$\mathbb{E}\left[\|S - \widehat{S}_{\Phi,\widehat{I}}\|^2\right] \leq C(n)\left(\min_{I\subset\{1,\dots,N\}} \|(S'\Phi) - (S'\Phi)_I\|^2 + \sigma^2|I| + \sigma^2\right)$$

with only a slightly different $C(n)$. The performance of those methods depends thus on the basis used.

The oracle risk

$$\min_{I\subset\{1,\dots,N\}} \|(S'\Phi) - (S'\Phi)_I\|^2 + \sigma^2|I|$$

corresponds to a *good solution* obtained by an optimal balance between an *apparent goodness*, the approximation error using only the coefficients in $I$, and a *complexity* measured by the number of coefficients used multiplied by a factor $\sigma^2$. It is small when the object of interest $S$ can be well approximated with few elements in the basis $\Phi$. A natural question is thus the existence of a best basis. This question is hopeless if one does not specify a set, or a collection of set, to which our object is known to belong. In that case, we are indeed in the setting of classical approximation theory.

Not surprisingly, those good approximation properties can also be translated into good compression properties. Indeed it is known since the 50's (see for instance Kramer and Mathews [KM56]), that transforming linearly a signal can help its coding and this is through this angle that I encountered first these approximation issues. In a basis, transform coding corresponds simply to the quantification of the basis coefficients and their lossless coding using an entropic coder. When there is a large number of coefficients quantized to 0, the performance of the code can be explicitly linked to the approximation performance.

## 1.5 Basis choice

It is then natural to ask whether one can choose the basis $\Phi$ among a family, in order to use the best of each for each object of interest. It turns out that Donoho and Johnstone's approach can no longer be used: their proofs are valid only when the estimator is obtained by projection on a subset $I$ of its coordinates in any fixed orthogonal basis and that subset $I$ is chosen among all possible subsets. Only Barron, Birgé, and Massart's one can be applied, as noted by Donoho and Johnstone [DJ94a], to select an orthogonal basis amongst a dictionary onto which to project the observation $Y$. Note also that the first inequality obtained following Barron, Birgé, and Massart [BBM99] is slightly finer that the second one as the factor in front of the bias term is much smaller. Although those types of inequalities linking the estimation risk with a deterministic quantity depending on the unknown $S$ that is valid for every $S$ are not oracle inequalities in the original sense, I will call them so in the sequel.

## 1.6 What's next?

A lot of questions are raised by such a simple example. Among those, I have considered the following ones

- Can we provide a theoretical guaranty on the performance of these estimators? In the minimax sense? In an oracle way?
- How precisely the performance of such an estimator is related to approximation theory? How does this constrain the choice or the design of the basis?
- Is there a way to extend those types of result to select a basis among a family? To cope with frames or arbitrary dictionary?
- Can we go beyond the quadratic loss case?
- Can this type of result be extended to inverse problem? To density estimation?
- Can we implement efficiently those estimators?

# Chapter 2

# Overview

The next three chapters are devoted to the description of the results I (with my coauthors) have obtained so far. I have organized them around 10 themes presented here in approximate chronological order of beginning:

**Theme 0.** Bandlets and geometrical image compression

**Theme 1.** Bandlets and geometrical image denoising

**Theme 2.** Maxiset of penalized model selection in the white noise model

**Theme 3.** Dictionary and adaptive $\ell_1$ penalization for density estimation

**Theme 4.** Inverse tomography and Radon needlet thresholding

**Theme 5.** Copula estimation with wavelet thresholding

**Theme 6.** NL-Means and statistical aggregation

**Theme 7.** Adaptive thresholding for the binary choice model

**Theme 8.** Unsupervised hyperspectral image segmentation

**Theme 9.** Conditional density estimation by penalized maximum likelihood

## 2.1 Brief overview

### Theme 0. Bandlets and geometrical image compression

**Publications:** [*Proc-LPM00*; *Proc-LPM01a*; *Proc-LPM01b*; *Proc-LPM03a*; *Proc-LPM03b*; *Art-LPM05a*; *Art-LPM05b*]
**Coauthor:** S. Mallat (Supervisor)

In this work, which corresponds to my PhD thesis under the supervision of S. Mallat, I have constructed a novel image representation adapted to geometrical image, the bandlets. We have been able to prove that the approximation properties of this adaptive frame representation leads to an adaptive optimal non linear approximation rate up to a logarithmic factor for geometrical images. This result has been translated into compression performance both theoretically and numerically.

### Theme 1. Bandlets and geometrical image denoising

**Publications:** [*Art-DLPM11*; *Proc-LePe+07*; *Art-LPM05a*; *Proc-Pe+07*]
**Coauthors:** Ch. Dossal, S. Mallat and G. Peyré

Capitalizing on the approximation properties of the bandlets, we have proposed a first bandlet image denoising algorithm based on the MDL approach: it suffices to use the coding algorithm to denoise efficiently geometrical image. Using the second generation bandlet basis construction of G. Peyré, with S. Mallat and Ch. Dossal, using techniques from L. Birgé and P. Massart [BM97], we have been able to propose a model selection based bandlet image denoising, for which the (quasi) minimax optimality for geometrical images can be proved.

### Theme 2. Maxiset of penalized model selection in the white noise model

**Publication:** [*Art-Au+10*]
**Coauthors:** F. Autin, J.-M. Loubes and V. Rivoirard

A natural question arising is the previous work is which are the functions well estimated by the model selection bandlet estimator. In this work, with F. Autin, J.-M. Loubes and V. Rivoirard, we show under mild assumption on the model structure, that one can estimate well the function that can be approximated well by the model collections and only those functions.

### Theme 3. Dictionary and adaptive $\ell_1$ penalization for density estimation

**Publication:** [*Art-BLPR11*]
**Coauthors:** K. Bertin and V. Rivoirard

Trying to answer a question asked by A. Tsybakov, I have considered, with K. Bertin and V. Rivoirard, a density estimation algorithm based on a variation of the Lasso, the Dantzig estimator. In this setting, we have shown how to fully exploit the density framework to calibrate accurately and automatically the Dantzig constraints from the data both theoretically and numerically.

### Theme 4. Inverse tomography and Radon needlet thresholding

**Publications:** [*Unpub-BLPT12*; *Art-Ke+10*; *Art-KLPP12*]
**Coauthors:** G. Kerkyacharian and D. Picard / M. Bergounioux and E. Trélat

Capitalizing on the beautiful needlet representation construction of P. Petrushev and his coauthors [NPW06a; NPW06b; PX08], we have proposed, with G. Kerkyacharian and D. Picard, a needlet thresholding strategy to inverse the fanbeam tomography operator, which is a case of Radon type transform, in the white noise model. The resulting estimator is proved to be adaptive and almost minimax for a large range of Besov spaces. Numerical experiments confirm this good behavior. A similar construction and analysis has been conducted with M. Bergounioux and E. Trélat for an axisymmetric objet. We have obtained oracle type inequalities in this setting.

### Theme 5. Copula estimation with wavelet thresholding

**Publication:** [*Art-ALPT10*]
**Coauthors:** F. Autin and K. Tribouley

Wavelet thresholding strategies have been proved to be a versatile technique. In this work with F. Autin and K. Tribouley, we apply this technique to non parametric estimation of copulas. Our estimator is proved to be adaptive and almost minimax for standard Besov bodies. Numerical experiments on artificial and real datasets are conducted.

### Theme 6. NL-Means and statistical aggregation

**Publications:** [*Proc-LPS09a*; *Proc-LPS09b*; *Proc-SLP09*]
**Coauthor:** S. Salmon (PhD Student)

Exponential weighting schemes have proved to be very efficient in statistic estimation. In this work, which corresponds to the beginning of J. Salmon's thesis under my supervision, we have tried to link this scheme with the NL-Means image denoising technique. This has leads to some improvements published by J. Salmon.

### Theme 7. Adaptive thresholding for the binary choice model

**Publications:** [*Unpub-GLP11*]
**Coauthor:** E. Gautier

In this work with E. Gautier, we tackle a classical econometric model, the binary choice model, whose geometry corresponds to a half hypersphere, using adapted needlet construction as well as adapted thresholds.

### Theme 8. Unsupervised hyperspectral image segmentation

**Publications:** [*Art-Be+11*; *Proc-CLP11b*; *Art-CLP12b*; *Unpub-CLP12c*]
**Coauthor:** S. Cohen

Unsupervised classification is often perform using Gaussian Mixture Model, whose components are associated to classes. With S. Cohen, we have proposed an extension of this model in which the mixture proportions depend of a covariate, the position. We have proposed an efficient numerical scheme that leads to an unsupervised hyperspectral image segmentation used with real datasets at IPANEMA, an ancient material study platform located at Synchrotron Soleil.

### Theme 9. Conditional density estimation by penalized maximum likelihood

**Publications:** [*Unpub-CLP11a*; *Unpub-CLP12a*; *Art-CLP12b*; *Unpub-MCLP12*]
**Coauthors:** S. Cohen and L. Montuelle (PhD Student)

The algorithm of the previous section relies on a model selection principle to select a suitable number of classes. In this work, with S. Cohen, we prove that, more generally, conditional density estimation can be performed by a penalized maximum likelihood principle under weak assumptions on the model selection. This analysis is exemplified by two piecewise constant with respect to the covariate partition-based conditional density strategy, one combined with piecewise polynomial density and the other with Gaussian Mixture densities. L. Montuelle is beginning a PhD on this theme under our supervision.

## 2.2 Manuscript organization

### 2.2.1 A chronological point of view?

Ordering the manuscript in chronological order would have been a natural point of view. As illustrated in Figure 2.1, my contributions can be roughly be decomposed into 4 periods: 1998-2002 in which I worked on geometrical approximation (Theme 0), 2002-2005 in which I considered some statistical extension of this question (Themes 1 and 2), 2006-2010 in which I have considered statistical problem in which approximation theory plays a central role (Themes 3, 4, 5, 6 and 7) and 2010-2011 in which I started to look at some extension of Gaussian Mixture Models for unsupervised segmentation and its conditional density estimation counterpart (Themes 8 and 9). Not surprisingly, those periods coincides, up to some inertia, with the different positions I had so far:

- PhD student at the CMAP (1998-2002), where I have learned from S. Mallat image processing and approximation theory,
- Post-Doc at LetItWave (2002-2004), during which I have focused on denoising while working on more *industrial* math,
- Assistant Professor (Maitre de Conférence) at the university Paris Diderot in the statistical team of the LPMA (2004-2009), during it which, under the guidance of D. Picard, I have really discovered the (mathematical) statistical world and

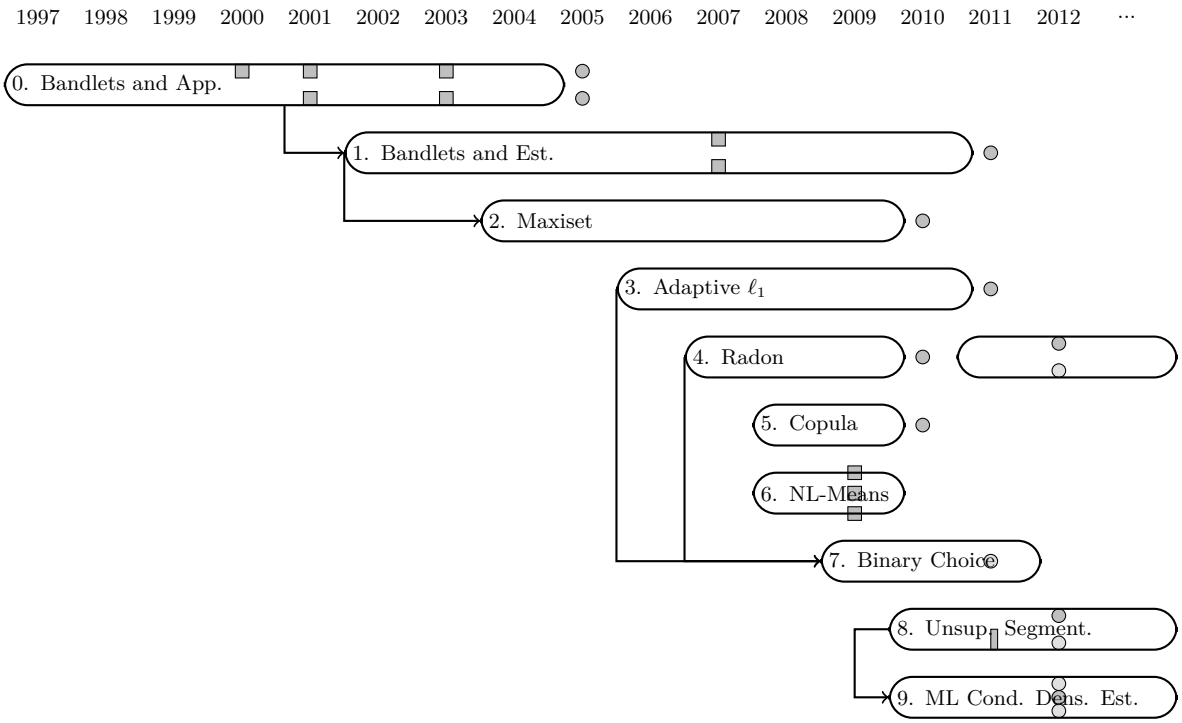



Figure 2.1: Chronological view

- Research Associate (Chargé de Recherche) at Inria Saclay (2009-) in the SELECT project, in which I have learned from P. Massart and G. Celeux how to combine model selection and mixtures.

This ordering is however too rough to be used as there is too many overlaps as soon as one looks carefully...

### 2.2.2 An approximation theory point of view?

As hinted in the previous chapter, approximation theory plays a central role in most of my contributions. In all my works but the one on NL-means (and even that could be discussed), there is the idea that the object of interest can be approximated, with respect to a certain distance, by a *simple* parametric model. As always, the more complex the model, the higher the cost. For instance, in estimation estimating within a complex model will lead to a large variance term, while in compression the storage cost of a very complex model may become prohibitive. A good model is one that realizes a good balance between this complexity and the model proximity with the object of interest. Of course, as this object is unknown, the best possible model, the oracle one, is unknown and one has to rely on a proxy of it. This best model is always obtained by a tradeoff between a complexity term and an empirical model proximity term related to the distance used. Those terms depend on the kind of model used as well on the observation model considered.

The representations I have used the most are the bandlets with Ch. Dossal, S. Mallat and G. Peyré (Themes 0 and 1), the needlets with G. Kerkyacharian, D. Picard and E. Gautier (Themes 4 and 7) and, of course, the wavelets with F. Autin, J-M. Loubes, V. Rivoirard and K. Tribouley (Themes 2 and 5). I have also considered some more abstract approximation setting with F. Autin, J-M. Loubes, V. Rivoirard, S. Cohen and L. Montuelle (Themes 2 and 9). Recently, I have considered generalization of the Gaussian Mixture Model with S. Cohen and L. Montuelle (Themes 8 and 9). I should also stress than even in the NL-Means study with J. Salmon (Theme 6) approximation theory plays a central role as we

were looking for an oracle type inequality. Figure 2.2 summarizes this point of view. While being a key tool, approximation theory does not provide a satisfactory ordering. In turns out that the observation model and the related goal is a much more interesting guideline.

### 2.2.3 A Statistical point of view!

The aim of my PhD thesis was the compression of geometrical images. We have proposed a novel image representation, the bandlet representation, and an efficient algorithm to compute a best $M$-term approximation of a given image. The performance of this method has been analyzed by proving non linear approximation property of the bandlet representation (Theme 0). Following the folk's theorem *Well approximating is well estimating*, with S. Mallat and then Ch. Dossal and G. Peyré, we have used a thresholding principle in this representation to obtain an image denoising algorithm. Image denoising is naturally modeled in a stochastic world and our analysis has been performed in this framework (Theme 1).

More precisely, we have considered the white noise model, which is probably the simplest stochastic model and, at least, the one I have more studied. Indeed, a question raised by our bandlet estimation study is whether one can slightly reverse the previous folk's theorem: *Well estimating a function in a representation means that it is well approximated.* With F. Autin, J.-M. Loubes and V. Rivoirard, who were working on a similar issue, we have studied this question for a large class of penalized model selection estimators (Theme 2). With G. Kerkyacharian and D. Picard, we have used the same white noise model to analyze the performance of a needlet based inversion of the Radon transform (Theme 4). This is also the model used with J. Salmon in the PAC-Bayesian analyze of the NL-Means image estimator.

The most classical model arises probably in the density estimation problem I have considered with K. Bertin and V. Rivoirard as well as with F. Autin and K. Tribouley. In the first work, we have studied a specific $\ell_1$ type density estimator (Theme 3). In the second one, we have considered a related issue, which is the estimation of the copula, which can be seen as the density of the *uniformized* observations and propose a wavelet based approach (Theme 5).

Finally, working on an unsupervised classification algorithm for hyperspectral image with S. Cohen (Theme 8), we have discovered that the most natural framework for our problem was to rephrase as a conditional density estimation problem. We have then studied a quiet general penalized maximum likelihood estimator adapted to conditional density estimation (Theme 9). It also turns out that the work in progress with E. Gautier on the binary choice model which can be seen at first as a density estimation problem is much closer to a conditional density problem (Theme 7).

For the manuscript, I have eventually settled for this ordering illustrated in Figure 2.3.

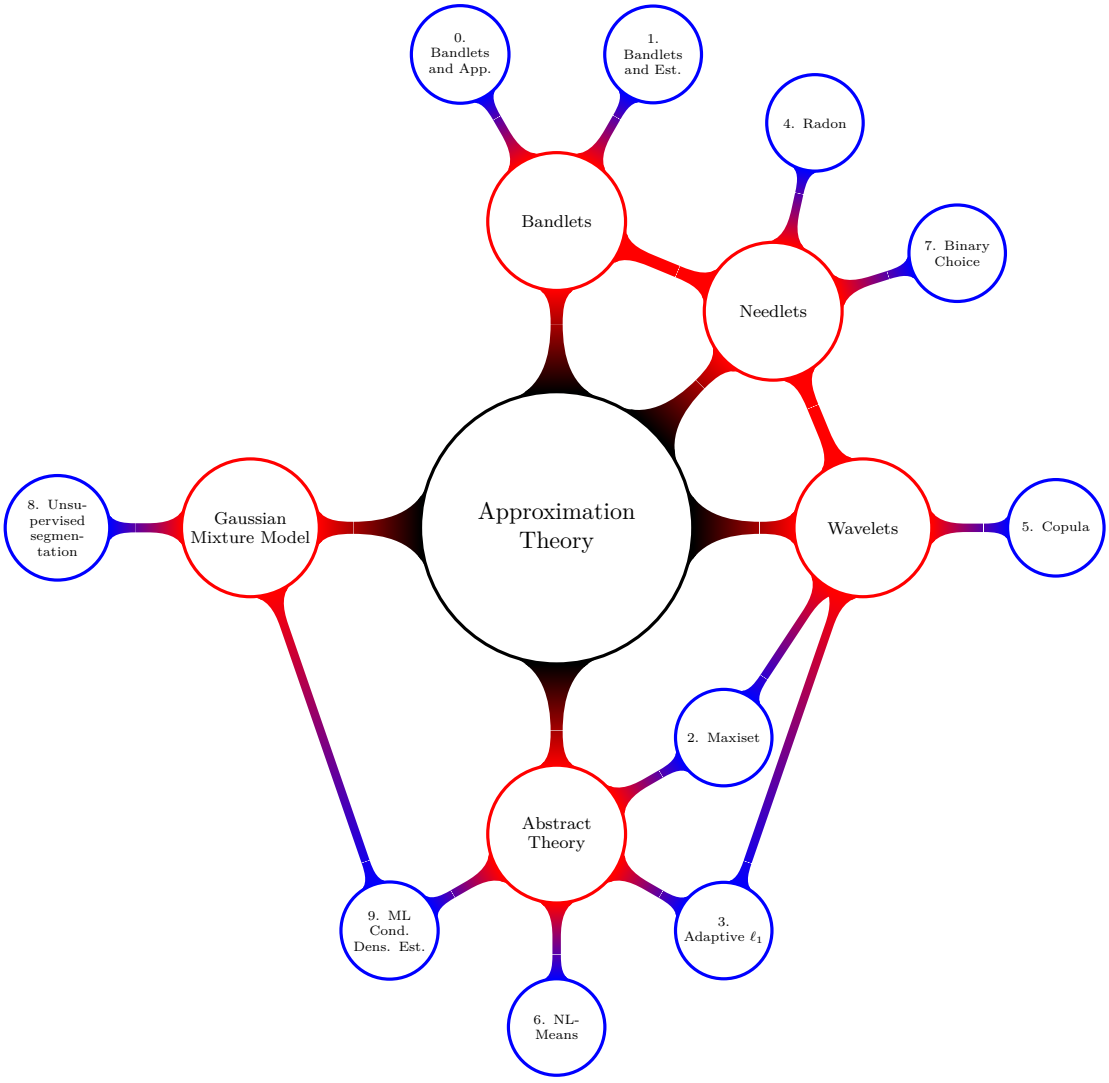



Figure 2.2: An approximation theory point of view on my contributions

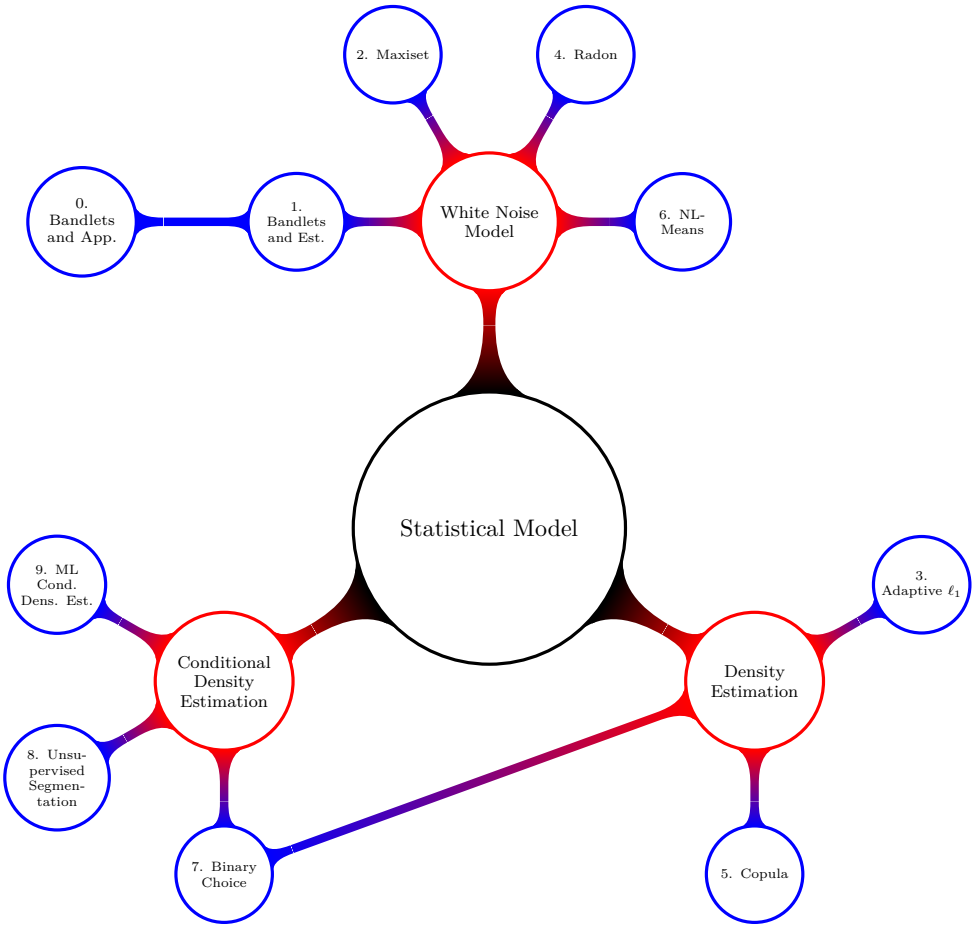



Figure 2.3: A Statistical point of view on my contributions

# Chapter 3

# Estimation in the white noise model

## 3.1 Image acquisition and white noise model

The first statistical I have encountered is the white noise model, seen as a noisy image acquisition model. Indeed, during the digital acquisition process, a camera measures an analog image $s_0$ with a filtering and sampling process corrupted by some noise. More precisely, if we denote the "noisy" measurement of a camera with $N$ pixels by $Y_{b_n}$, where $b_n$ belongs to a family of $N$ impulse responses of the photo-sensors, those "noisy" measurements are often modeled as sums of ideal noiseless measurements and Gaussian noises:

$$Y_{b_n} = \langle s_0, b_n \rangle + \sigma W_{b_n} \text{ for } 0 \leq n < N$$

where $(W_{b_n})_{0 \leq n < N}$ is a centered Gaussian vector and $\sigma$ is a known noise level parameter. When the family $(b_n)_{0 \leq n < N}$ is an orthonormal family, the Gaussian vector $(W_{b_n})_{0 \leq n < N}$ is often assumed to be white; its components are assumed independents. For a general family of impulse responses $(b_n)_{0 \leq n < N}$, this assumption is relaxed and the correlation between two measures is linked to the correlation between the two correspond impulse responses: more precisely, the covariance matrix of the Gaussian vector $(W_{b_n})_{0 \leq n < N}$ is assumed to be the following Gramm matrix $(\langle b_n, b_{n'} \rangle)_{0 \leq n, n' \leq N}$.

This situation corresponds to the (classical) white noise statistical model which is formally described as the observation of a process $Y$ that satisfies

$$\mathrm{d}Y_x = s_0(x)\mathrm{d}x + \sigma \mathrm{d}W_x,$$

where $W_x$ is now the Wiener process. This equation means that one is able to observe a Gaussian field $Y_g$ indexed by functions $g \in L^2([0,1]^2)$ of mean $E(Y_g) = \langle f, g \rangle$ and covariance $E[Y_g Y_{g'}] = \langle g, g' \rangle$. It generalizes the model of the previous paragraph in which the Gaussian field $Y_g$ can only be observed for function $g$ in the space $V_N$ generated by the family of impulse responses $(b_n)_{0 \leq n < N}$. Using a more abstract model that allows to state the statistical problem in the continuous framework will be important to consider asymptotics over the noise level $\sigma$: smaller noise level will require a better resolution for the camera measurement process than larger one.

Indeed, this white noise model allows us to define for any space $V$ spanned by some functions $\{b_n\}_{n \in I}$ a "projection" $P_V Y$ of the same observation $\mathrm{d}Y$ on $V$. When the family $\{b_n\}_{n \in I}$ is orthonormal, $P_V$ can be written as

$$P_V Y = \sum_{n \in I} Y_{b_n} b_n$$

whereas the decomposition coefficients are slightly more involved in the general case. Nevertheless, this projection depends only on the space $V$ spanned by the functions $\{b_n\}_{n \in I}$ and not on the functions themselves. In the following, we will work mainly in term of spaces and thus may assume, with no loss of generality, that $\{b_n\}_{n \in I}$ is an orthogonal family so that the decomposition $P_V Y = \sum_{n \in I} Y_{b_n} b_n$ holds. Following ideas I discovered with Donoho [Do93], I have studied projection estimator in which estimates of $s_0$ are obtained by *projecting* the observation $Y$ onto a space $V$ chosen adaptively.

## 3.2 Projection Estimator and Model Selection

More precisely, we follow here the setting proposed with Ch. Dossal and S. Mallat [*Art-DLPM11*] which have been implemented with the help of G. Peyré [*Proc-LePe+07*; *Proc-Pe+07*]. We have considered projection estimators that are decomposed in two steps. First, a linear projection reduces the dimensionality of the problem by projecting the noisy observation into a finite dimensional space. In signal processing, this first projection is typically performed by the digital acquisition device. Then, a non-linear projection estimator refines this projector by reprojecting the resulting finite dimensional observation into a space that is chosen depending upon this observation. In the setting we have considered, this non-linear projection can be termed as a thresholding in a best basis selected from a dictionary of orthonormal bases. Best basis algorithms for noise removal have been introduced by Coifman and Wickerhauser [CW92]. As recalled by Candès [Ca06], their risks have already been studied by Donoho and Johnstone [DJ94a] and are a special case of the general framework of model selection proposed by Birgé and Massart [BM97] Note that Kolaczyk and Nowak [KN04] have studied a similar problem in a slightly different setting. We recall in this section the model selection estimators principle and its relation with thresholding estimation when using orthonormal basis.

### 3.2.1 Approximation space $V_N$ and further projection

The first step of our estimators is a projection in a finite dimension space $V_N$ spanned by an orthonormal family $\{b_n\}_{0\leq n<N}$. The choice of the dimension $N$ and of the space $V_N$ depends on the noise level $\sigma$ but should not depend on the function $f$ to be estimated. Assume for now that $V_N$ is fixed and thus that we observe $P_{V_N}X$. This observation can be decomposed into $P_{V_N}s_0 + \sigma W_{V_N}$ where $W_{V_N}$ is a finite dimensional white noise on $V_N$.

Our final estimator is a reprojection of this observation $P_{V_N}Y$ onto a subspace $\mathcal{M} \subset V_N$ which may (and will) depend on the observation: the projection based estimator $P_{\mathcal{M}}P_{V_N}Y = P_{\mathcal{M}}X$. The overall quadratic error can be decomposed in three terms:

$$\|s_0 - P_{\mathcal{M}}Y\|^2 = \|s_0 - P_{V_N}s_0\|^2 + \|P_{V_N}s_0 - P_{\mathcal{M}}s_0\|^2 + \sigma^2\|P_{\mathcal{M}}W\|^2.$$

The first term is a bias term corresponding to the first linear approximation error due to the projection on $V_N$, the second term is also a bias term which corresponds to the non linear approximation of $P_{V_N}s_0$ on $\mathcal{M}$ while the third term is a "variance" term corresponding to the contribution of the noise on $\mathcal{M}$.

The dimension $N$ of $V_N$ has to be chosen large enough so that with high probability, for reasonable $\mathcal{M}$, $\|s_0 - P_{V_N}s_0\|^2 \leq \|P_{V_N}s_0 - P_{\mathcal{M}}s_0\|^2 + \|P_{\mathcal{M}}W\|^2$. From the practical point of view, this means that the acquisition device resolution is set so that the first linear approximation error due to discretization is smaller than the second non linear noise related error. Engineers often set $N$ so that both terms are of the same order of magnitude, to limit the cost in terms of storage and computations. In our white noise setting, we will explain how to chose $N$ depending on $\sigma$.

For a fixed $V_N$, in order to obtain a small error, we need to balance between the two remaining terms. A space $\mathcal{M}$ of large dimension may reduce the second bias term but will increase the variance term, a space $\mathcal{M}$ of small dimension does the opposite. It is thus necessary to find a trade-off between these two trends, and select a space $\mathcal{M}$ to minimize the sum of those two terms.

### 3.2.2 Model selection in a single orthonormal basis

We consider a (not that) specific situation in which the space $\mathcal{M}$ is spanned by some vectors from an orthonormal basis of $V_N$. More precisely, let $\mathcal{B} = \{\Phi_n\}_{0\leq n<N}$ be an orthonormal basis of $V_N$, that may be different from $\{b_n\}$, we consider spaces $\mathcal{M}$ spanned by a sub-family $\{\Phi_{n_k}\}_{1\leq k\leq M}$ of $M = \dim(\mathcal{M})$ vectors and the projections of our observation on those spaces

$$P_{\mathcal{M}}Y = \sum_{k=1}^{M} Y_{\Phi_{n_k}}\,\Phi_{n_k}.$$

Note that this projection, or more precisely its decomposition in the basis $\{b_n\}$, can be computed easily from the decomposition of $P_{\mathcal{M}}Y$ in the same basis.

As a projection estimator yields an estimation error

$$\|s_0 - P_{\mathcal{M}} Y\|^2 = \|f - P_{V_N}\|^2 + \|P_{V_N} - P_{\mathcal{M}} s_0\|^2 + \|P_{\mathcal{M}} W\|^2 = \|f - P_{\mathcal{M}} s_0\|^2 + \|P_{\mathcal{M}} W\|^2,$$

the expected error of such an estimator is given by

$$E\left[\|s_0 - P_{\mathcal{M}} Y\|^2\right] = \|f - P_{\mathcal{M}} s_0\|^2 + \sigma^2 \dim(\mathcal{M}).$$

The best subspace for this criterion is the one that realizes the best trade-off between the approximation error $\|f - P_{\mathcal{M}} s_0\|^2$ and the complexity of the models measured by $\sigma^2 \dim(\mathcal{M})$.

This expected error cannot be computed in practice since we have a single realization of $\mathrm{d}Y$ (or of $P_{V_N} Y$) . To (re)derive the classical model selection procedure of Birgé and Massart [BM97] or the thresholding theorems of Donoho and Johnstone [DJ94b], we first slightly modify our problem by searching for a subspace $\mathcal{M}$ such that the estimation error obtained by projecting $P_{V_N} Y$ on this subspace is small with an overwhelming probability. As in most model selection analysis, we use an upper bound of the estimation error obtained from an upper bound of the energy of the noise projected on $\mathcal{M}$. Each of the $K_N$ projections of the noise on the $K_N$ different vectors in the bases of the dictionary $\mathcal{D}_N$ is thus $W_{\Phi_k}\, \Phi_k$. Its law is a Gaussian random variable of variance $\sigma^2$ along the vector $\Phi_k$. A standard large deviation result proves that the norms of $K_N$ such Gaussian random variables are bounded simultaneously by $T = \sigma\sqrt{2 \log K_N}$ with a probability that tends to 1 when $N$ increases. Since the noise energy projected in $\mathcal{M}$ is the sum of $\dim(\mathcal{M})$ squared dictionary noise coefficients, we get $\|P_{\mathcal{M}} W\|^2 \leq \dim(\mathcal{M})\ T^2$. It results that

$$\|s_0 - P_{\mathcal{M}} Y\|^2 \leq \|s_0 - P_{\mathcal{M}} s_0\|^2 + \dim(\mathcal{M})\ T^2. \tag{3.1}$$

over all subspaces $\mathcal{M}$ with a probability that tends to 1 as $N$ increases. The estimation error is small if $\mathcal{M}$ is a space of small dimension $\dim(\mathcal{M})$ which yields a small approximation error $\|s_0 - P_{\mathcal{M}} s_0\|$. We denote by $\mathcal{M}_O \in \mathcal{C}_N$ the space that minimizes the estimation error upper bound (3.1)

$$\mathcal{M}_O = \underset{\mathcal{M} \in \mathcal{C}_N}{\operatorname{argmin}} (\|s_0 - P_{\mathcal{M}} s_0\|^2 + \dim(\mathcal{M})\ T^2).$$

Note that this optimal space cannot be determined from the observation $Y$ since $s_0$ is unknown. It is called the oracle space , hence the $O$ in the notation, to remind this fact.

To obtain an estimator, it is thus necessary to replace this oracle space by a *best* space obtained only from the observation $P_{V_N} Y$ that yields (hopefully) a small estimation error. A first step toward this goal is to notice that since all the spaces $\mathcal{M}$ are included into $V_N$, minimizing

$$\|s_0 - P_{\mathcal{M}} s_0\|^2 + \dim(\mathcal{M})\ T^2$$

is equivalent to minimizing

$$\|P_{V_N} s_0 - P_{\mathcal{M}} s_0\|^2 + \dim(\mathcal{M})\ T^2.$$

A second step is to consider the crude estimation of $\|P_{V_N} s_0 - P_{\mathcal{M}} s_0\|^2$ given by the empirical norm

$$\|P_{V_N} Y - P_{\mathcal{M}} Y\|^2 = \|P_{V_N} Y\|^2 - \|P_{\mathcal{M}} Y\|^2.$$

This may seem naive because estimating $\|P_{V_N} s_0 - P_{\mathcal{M}} s_0\|^2$ with $\|P_{V_N} Y - P_{\mathcal{M}} Y\|^2$ yields a large error

$$\|P_{V_N} Y - P_{\mathcal{M}} Y\|^2 - \|P_{V_N} s_0 - P_{\mathcal{M}} s_0\|^2 = (\|P_{V_N} Y\|^2 - \|P_{V_N} s_0\|^2) + (\|P_{\mathcal{M}} s_0\|^2 - \|P_{\mathcal{M}} Y\|^2),$$

whose expected value is $(N - \dim(\mathcal{M}))\sigma^2$, with typically $\dim(\mathcal{M}) \ll N$. However, most of this error is in the first term on the right hand-side, which has no effect on the choice of space $\mathcal{M}$. This choice depends only upon the second term and is thus only influenced by noise projected in the space $\mathcal{M}$ of lower dimension $\dim(\mathcal{M})$. The bias and the fluctuation of this term, and thus the choice of the basis, are controlled by increasing the parameter $T$.

We define the best empirical projection estimator $P_{\widehat{\mathcal{M}}}$ as the estimator that minimizes the resulting empirical penalized risk:

$$\widehat{\mathcal{M}} = \underset{\mathcal{M} \in \mathcal{C}_N}{\operatorname{argmin}} \|P_{V_N} Y - P_{\mathcal{M}} Y\|^2 + \dim(\mathcal{M})\ T^2. \tag{3.2}$$

As the spaces $\mathcal{M}$ are spanned by subsets of the same orthogonal basis, this minimization can be performed coefficientwise and $\widehat{\mathcal{M}}$ is the space spanned by the basis elements corresponding to the observed coefficients larger than $T$ in absolute value. Donoho and Johnstone [DJ94b] have shown that this estimator is efficient:

$$(1+o(1))\operatorname*{argmin}_{\mathcal{M}\in\mathcal{C}_N}(\|s_0 - P_{\mathcal{M}}s_0\|^2 + \dim(\mathcal{M})\ T^2)$$
$$\leq \mathbb{E}\left[\| P_{\widehat{\mathcal{M}}}Y - s_0\|^2\right] \leq (2\log N + 1)\operatorname*{argmin}_{\mathcal{M}\in\mathcal{C}_N}(\|s_0 - P_{\mathcal{M}}s_0\|^2 + \dim(\mathcal{M})\ T^2).$$

### 3.2.3 Model Selection in a dictionary of orthonormal bases

Now, instead of choosing a specific single orthonormal basis $\mathcal{B}$, we define a dictionary $\mathcal{D}_N$ which is a collection of orthonormal bases in which we choose adaptively the basis used. Note that some bases of $\mathcal{D}_N$ may have vectors in common. This dictionary can thus also be viewed as set $\{\Phi_n\}$ of $P \geq N$ different vectors, that are regrouped to form many different orthonormal bases. Any collection of $M$ vectors from the same orthogonal basis $\mathcal{B} \in \mathcal{D}_N$ generates a space $\mathcal{M}$ of dimension $M$ that defines a possible estimator $P_{\mathcal{M}}Y$ of $s_0$. Let $\mathcal{C}_N = \{\mathcal{M}_\gamma\}_{\Gamma_N}$ be the family of all such projection spaces. Ideally we would like to find the space $\mathcal{M} \in \mathcal{C}_N$ which minimizes $\|s_0 - P_{\mathcal{M}}Y\|$. We want thus to choose a "best" model $\mathcal{M}$ among a collection that is we want to perform a model selection task.

Notice that the analysis of the previous section has never really used the fact that the spaces are spanned by subsets of a single orthonormal basis. We could have repeated the previous analysis up to the search for the best estimator. Indeed, finding the minimizer of (3.2) may seem computationally now untractable because the number of possible spaces $\mathcal{M} \in \mathcal{C}$ is typically an exponential function of the number $P$ of vectors in $\mathcal{D}_N$ and there is no way to reduce this question to a coefficientwise estimation. We show that this best estimator may however still be found with a thresholding strategy, but one in a *best* basis. as soon as one that $\mathcal{M}$ are generated by a subset of vectors from a basis $\mathcal{B} \in \mathcal{D}_N$. One can verify that this implies that the best projection estimator is necessarily a thresholding estimator in some basis. Minimizing $\|P_{V_N}Y - P_{\mathcal{M}}Y\|^2 + \dim(\mathcal{M})\ T^2$ over $\mathcal{M} \in \mathcal{C}$ is thus equivalent to find the basis $\widehat{\mathcal{B}}$ of $V_N$ which minimizes the thresholding penalized empirical risk:

$$\widehat{\mathcal{B}} = \operatorname*{argmin}_{\mathcal{B}\in\mathcal{D}_N} \|P_{V_N}Y - P_{\mathcal{M}_{\mathcal{B},Y,T}}Y\|^2 + \dim(\mathcal{M})\ T^2.$$

The best space which minimizes the empirical penalized risk in (3.2) is derived from a thresholding in the best basis $\widehat{\mathcal{M}} = \mathcal{M}_{\widehat{\mathcal{B}},T}$. The following theorem, similar to the one obtained first by Barron, Birgé, and Massart [BBM99] proves that the thresholding estimation error in the best basis is bounded by the estimation error by projecting in the oracle space $\mathcal{M}_O$, up to a multiplicative factor. Note that a similar result can be found in an earlier article of Donoho and Johnstone [DJ94a].

**Theorem 1.** *There exists an absolute bounded function $\lambda_0(P) \geq \sqrt{2}$ and some absolute constants $\epsilon > 0$ and $\kappa > 0$ such that if we denote $\mathcal{C}_N = \{\mathcal{M}_\gamma\}_\Gamma$ the family of projection spaces generated by some vectors in an orthogonal basis of a dictionary $\mathcal{D}_N$ and denote $P$ be the number of different vectors in $\mathcal{D}_N$. Then for any $\sigma > 0$, if we let $T = \lambda\sqrt{\log(P)}\sigma$ with $\lambda \geq \lambda_0(P)$, then for any $s_0 \in L^2$, the thresholding estimator $F = P_{\mathcal{M}_{\widehat{\mathcal{B}},X,T}}Y$ in the best basis*

$$\widehat{\mathcal{B}} = \operatorname*{argmin}_{\mathcal{B}\in\mathcal{D}_N} \|P_{V_N}Y - P_{\mathcal{M}_{\mathcal{B},Y,T}}Y\|^2 + \dim(\mathcal{M}_{\mathcal{B},Y,T})\ T^2$$

*satisfies*

$$E\left[\|s_0 - \hat{s}\|^2\right] \leq (1+\epsilon)\left(\min_{\mathcal{M}\in\mathcal{C}_N} \|s_0 - P_{\mathcal{M}}s_0\|^2 + \dim(\mathcal{M})\ T^2\right) + \frac{\kappa}{P}\sigma^2. \tag{3.3}$$

For the sake of completion, in our paper [*Art-DLPM11*], we propose a simple proof of Theorem 1, inspired by Birgé and Massart [BM97], which requires only a concentration lemma for the norm of the noise in all the subspaces spanned by the $P$ generators of $\mathcal{D}_N$ but with worse constants: $\lambda_0(P) =$

$\sqrt{32+\frac{8}{\log(P)}}$, $\epsilon = 3$ and $\kappa = 64$. Note that this Theorem can be deduced from Massart [Ma07] with different (better) constant (and for roughly $\lambda_0(P) > \sqrt{2}$) using a more complex proof based on subtle Talagrand's inequalities. It results that any bound on $\min_{\mathcal{M}\in\mathcal{C}_N} \|s_0 - P_{\mathcal{M}} s_0\|^2 + \dim(\mathcal{M})\, T^2$, gives a bound on the risk of the best basis estimator $\hat{s}$.

To obtain a computational estimator, the minimization

$$\widehat{\mathcal{B}} = \underset{\mathcal{B}\in\mathcal{D}_N}{\operatorname{argmin}} \|P_{V_N} Y - P_{\mathcal{M}_{\mathcal{B},Y,T}} Y\|^2 + \dim(\mathcal{M}_{\mathcal{B},Y,T})\, T^2 \quad ,$$

should be performed with a number of operations typically proportional to the number $K_N$ of vectors in the dictionary. This requires to construct appropriate dictionaries of orthogonal bases. Examples of such dictionaries have been proposed by Coifman and Wickerhauser [CW92] with wavelet packets or by Coifman and Meyer [CM91] with local cosine bases for signals having localized time-frequency structures.

## 3.3 Best basis image estimation and bandlets

Given those theoretical results, a natural question arise: are those types of estimators efficient? This will depend heavily on the representation used and the bandlets, introduced in my thesis [*Proc-LPM00*; *Proc-LPM01a*; *Proc-LPM01b*; *Proc-LPM03a*; *Proc-LPM03b*; *Art-LPM05a*; *Art-LPM05b*] and further enhanced by Peyré and Mallat [PM08], will prove to be a very valuable tool when one wants to estimate geometrically regular images.

### 3.3.1 Minimax risk and geometrically regular images

Indeed, we study the maximum risk of estimators for images $f$ in a given class with respect to $\sigma$. Model classes are often derived from classical regularity spaces ($\mathbf{C}^\alpha$ spaces, Besov spaces,...). This does not take into account the existence of geometrically regular structures such as edges. Here, we use a geometric image model appropriate for edges, but not for textures, where images are considered as piecewise regular functions with discontinuities along regular curves in $[0,1]^2$. This geometrical image model has been proposed by Korostelev and Tsybakov [KT93] in their seminal work on image estimation. It is used as a benchmark to estimate or approximate images having some kind of geometric regularity (Donoho [Do99], Shukla, Dragotti, Do, and Vetterli [Sh+05],...). An extension of this model that incorporates a blurring kernel $h$ has been proposed by Le Pennec and Mallat [*Art-LPM05a*] to model the various diffraction effects. The resulting class of images, the one studied here, is the set of $\mathbf{C}^\alpha$ geometrically regular images specified by the following definition.

**Definition 1.** *A function $s \in L^2([0,1]^2)$ is $\mathbf{C}^\alpha$ geometrically regular over $[0,1]^2$ if*

- *$s = \tilde{s}$ or $s = \tilde{s} \star h$ with $\tilde{s} \in \mathbf{C}^\alpha(\Lambda)$ for $\Lambda = [0,1]^2 - \{\mathcal{C}_\gamma\}_{1\leq\gamma\leq G}$,*
- *the blurring kernel $h$ is $\mathbf{C}^\alpha$, compactly supported in $[-s,s]^2$ and $\|h\|_{\mathbf{C}^\alpha} \leq s^{-(2+\alpha)}$,*
- *the edge curves $\mathcal{C}_\gamma$ are $\mathbf{C}^\alpha$ and do not intersect tangentially if $\alpha > 1$.*

Korostelev and Tsybakov [KT93] have built an estimator that is asymptotically minimax for geometrically regular functions $s_0$, as long as there is no blurring and hence that $s_0 = \tilde{s_0}$. With a detection procedure, they partition the image in regions where the image is either regular or contains a "boundary fragment", a subpart of a single discontinuity curve. In each region, they use either an estimator tailored to this "boundary fragments" or a classical kernel estimator adapted to regular regions. This yields a global estimate $\hat{s}$ of the image $s_0$. If the $s_0$ is $\mathbf{C}^\alpha$ outside the boundaries and if the parametrization of the curve is also $\mathbf{C}^\alpha$ then there exists a constant $C$ such that

$$\forall \sigma \quad , \quad E\left[\|s_0 - \hat{s}\|^2\right] \leq C\sigma^{\frac{2\alpha}{\alpha+1}} \quad .$$

This rate of convergence achieves the asymptotic minimax rate for uniformly $\mathbf{C}^\alpha$ functions and thus the one for $\mathbf{C}^\alpha$ geometrically regular functions that includes this class. This means that sharp edges do not alter the rate of asymptotic minimax risk. However, this estimator is not adaptive relatively to the

Holder exponent $\alpha$ that must be known in advance. Furthermore, it uses an edge detection procedure that fails when the image is blurred or when the discontinuity jumps are not sufficiently large.

Donoho [Do99] and Shukla, Dragotti, Do, and Vetterli [Sh+05] reuse the ideas of "boundary fragment" under the name "horizon model" to construct a piecewise polynomial approximation of images. They derive efficient estimators optimized for $\alpha \in [1, 2]$. These estimators use a recursive partition of the image domain in dyadic squares, each square being split in two parts by an edge curve that is a straight segment. Both optimize the recursive partition and the choice of the straight edge segment in each dyadic square by minimizing a global function. This process leads to an asymptotically minimax estimator up to a logarithmic factor which is adaptive relatively to the Holder exponent as long as $\alpha \in [1, 2]$.

Korostelev and Tsybakov [KT93] as well as Donoho [Do99] and Shukla, Dragotti, Do, and Vetterli [Sh+05] rely on the sharpness of image edges in their estimators. In both cases, the estimator is chosen among a family of images that are discontinuous across parametrized edges, and these estimators are therefore not appropriate when the image edges are blurred. We will consider estimators that do not have this restriction: they project the observation on adaptive subspaces in which blurred as well as sharp edges are well represented. They rely on two ingredients: the existence of bases in which geometrical images can be efficiently approximated and the existence of a mechanism to select, from the observation, a good basis and a good subset of coefficients onto which it suffices to project the observation to obtain a good estimator. We focus first on the second issue.

### 3.3.2 Estimation in a single basis

When the dictionary $\mathcal{D}_N$ is reduced to a single basis $\mathcal{B}$, and there is thus no basis choice, Theorem 1 clearly applies and reduces to the classical thresholding Theorem of Donoho and Johnstone [DJ94b]. The corresponding estimator is thus the classical thresholding estimator which quadratic risk satisfies

$$E\left[\|s - P_{\mathcal{M}_{\mathcal{B},Y,T}} Y\|^2\right] \leq (1+\epsilon)\left(\min_{\mathcal{M}\in\mathcal{C}_N} \|s - P_{\mathcal{M}} s\|^2 + \dim(\mathcal{M})\ T^2\right) + \frac{\kappa}{N}\sigma^2$$

It remains "only" to choose which basis to use and how to define the space $V_N$ with respect to $\sigma$.

Wavelet bases provide a first family of estimators used commonly in image processing. Such a two dimensional wavelet basis is constructed from two real functions, a one dimensional wavelet $\psi$ and a corresponding one dimensional scaling function $\phi$, which are both dilated and translated:

$$\psi_{j,k}(x) = \frac{1}{2^{j/2}}\psi\left(\frac{x - 2^j k}{2^j}\right) \text{ and } \phi_{j,k}(x) = \frac{1}{2^{j/2}}\phi\left(\frac{x - 2^j k}{2^j}\right) \quad .$$

Note that the index $j$ goes to $-\infty$ when the wavelet scale $2^j$ decreases. For a suitable choice of $\psi$ and $\phi$, the family $\{\psi_{j,k}(x)\}_{j,k}$ is an orthogonal basis of $L^2([0,1])$ and the following family constructed by tensorization

$$\left\{\begin{array}{l} \psi^V_{j,k}(x) = \psi^V_{j,k}(x_1,x_2) = \phi_{j,k_1}(x_1)\,\psi_{j,k_2}(x_2), \\ \psi^H_{j,k}(x) = \psi^H_{j,k}(x_1,x_2) = \psi_{j,k_1}(x_1)\,\phi_{j,k_2}(x_2), \\ \psi^D_{j,k}(x) = \psi^D_{j,k}(x_1,x_2) = \psi_{j,k_1}(x_1)\,\psi_{j,k_2}(x_2) \end{array}\right\}_{(j,k_1,k_2)}$$

is an orthonormal basis of the square $[0,1]^2$. Furthermore, each space

$$V_j = \text{Span}\{\phi_{j,k_1}(x_1)\phi_{j,k_2}(x_2)\}_{k_1,k_2},$$

called approximation space of scale $2^j$, admits $\{\psi^o_{l,k}\}_{o,l\geq j,k_1,k_2}$ as an orthogonal basis. The approximation space $V_N$ of the previous section coincides with the classical wavelet approximation space $V_j$ when $N = 2^{-j/2}$.

A classical approximation result ensures that for any function $s \in \mathbf{C}^\alpha$, as soon as the wavelet has more than $\lfloor\alpha\rfloor + 1$ vanishing moments, there is a constant $C$ such that, for any $T$, $\min_{\mathcal{M}\in\mathcal{C}_N} \|P_{V_N} s - P_{\mathcal{M}} s\|^2 + \dim(\mathcal{M})\ T^2 \leq C(T^2)^{\frac{\alpha}{\alpha+1}}$, and, for any $N$, $\|P_{V_N} s - s\|^2 \leq CN^{-\alpha}$. For $N = 2^{-j/2}$ with $\sigma^2 = [2^j, 2^{j+1}]$, Theorem 1 thus implies

$$E[\|s - \hat{s}\|^2] \leq C(|\log(\sigma)|\sigma^2)^{\frac{\alpha}{\alpha+1}} \ .$$

This is up to the logarithmic term the best possible rate for $\mathbf{C}^\alpha$ functions. Unfortunately, wavelets bases do not provide such an optimal representation for the $\mathbf{C}^\alpha$ geometrically regular functions specified by Definition 1. Wavelets fail to capture the geometrical regularity of edges: near them, the wavelets coefficients remain large. As explained in Mallat [Ma08], by noticing that those edges contribute at scale $2^j$ to $O(2^{-j})$ coefficients of order $O(2^{j/2})$, one verifies that the rate of convergence in a wavelet basis decays like $(|\log(\sigma)|\sigma^2)^{1/2}$, which is far from the asymptotically minimax rate.

### 3.3.3 Estimation in a fixed frame

No known basis seems able to capture the geometric regularity, however a remarkably efficient representation was introduced by Candès and Donoho [CD99]. Their curvelets are not isotropic like wavelets but are more elongated along a preferential direction and have two vanishing moments along this direction. They are dilated and translated like wavelets but they are also rotated. The resulting family of curvelets $\mathcal{C} = \{c_n\}_n$ is not a basis of $L^2([0,1]^2)$ but a tight normalized frame of $L^2(\mathbb{R}^2)$. This means that for any $s \in L^2([0,1]^2)$

$$\sum_{c_n \in \mathcal{C}} |\langle s, c_n \rangle|^2 = \|s\|^2$$

which implies

$$s = \sum_{c_n \in \mathcal{C}} \langle s, c_n \rangle c_n.$$

Although this is not an orthonormal basis, the results of Section 3.2 can be extended to this setting by replacing the thresholding operator by the search of the space $\mathcal{M}$ spanned by a subset of $(c_n)_{0 \leq n < N}$, which spans $V_N$, that minimizes

$$\|P_{V_N} Y - P_{\mathcal{M}} Y\|^2 + T^2 \dim(\mathcal{M})$$

with $N = \sigma^{-1/2}$. The error rate for $\mathbf{C}^\alpha$ geometrically regular function with $\alpha \in [1,2]$ is

$$E\left[\sum_n \|f - F\|^2\right] \leq C(|\log \sigma|\sigma^2)^{\frac{\alpha}{\alpha+1}}$$

which is up to the logarithmic factor the minimax rate. Unfortunately, computing this estimator is complex as it requires to compute all the projections $P_{\mathcal{M}} Y$ which is not an easy task. This difficulty may be overcome by working in the coefficient domain. Projecting the data on the first $N = \sigma^{-1/2}$ curvelets with significant intersection with the unit square and thresholding the remaining coefficients with a threshold $\lambda = \sqrt{\log N}\sigma$ yields an estimator $\widehat{\langle s_0, c_n \rangle}$ of the coefficients $\langle s_0, c_n \rangle$. Those estimated coefficients are such that

$$E\left[\sum_n (\langle s_0, c_n \rangle - \widehat{\langle s_0, c_n \rangle})^2\right] \leq C(|\log \sigma|\sigma^2)^{\frac{\alpha}{\alpha+1}}$$

with a constant $C$ that depends only on $s_0$. Using the inverse frame operator as defined by Christensen [Ch03], one obtains an estimator $\hat{s}$ not necessarily equal to $\sum_n \widehat{\langle s_0, c_n \rangle} c_n$ that nevertheless satisfies

$$E\left[\sum_n \|s_0 - \hat{s}\|^2\right] \leq C(|\log \sigma|\sigma^2)^{\frac{\alpha}{\alpha+1}}$$

for $\mathbf{C}^\alpha$ geometrically regular functions with $\alpha \in [1,2]$.

While the two error bounds of those two estimators are similar, they are deduced from two different kinds of control. The first one is obtained by a synthesis control: a control on the error of the best approximation with a given number of coefficients. The second one is obtained by an analysis control: a control on the number of coefficients above a threshold. Although the first (synthesis) approach and the second (analysis) approach are equivalent for orthonormal basis, they are very different for frames.

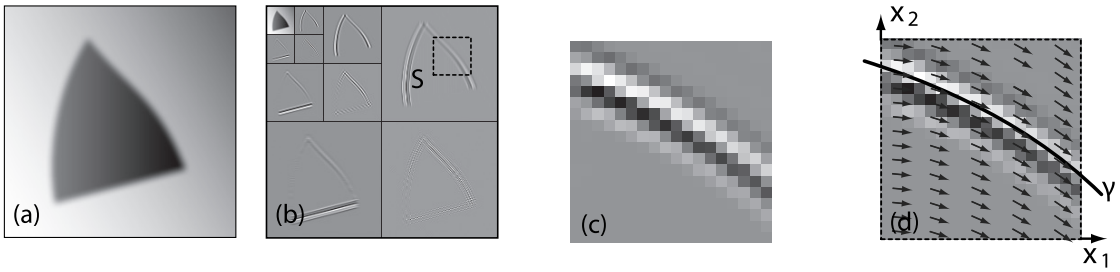



Figure 3.1: a) a geometrically regular image, b) the associated wavelet coefficients, c) a close-up of wavelet coefficients in a detail space $W_j^o$ that shows their remaining regularity, d) the geometrical flow adapted to this square of coefficients, here it is vertically constant and parametrized by a polynomial curve $\gamma$

Other fixed representations, such as the shearlets of Labate, Lim, Kutyniok, and Weiss [La+05], achieve this optimal rate for $\alpha = 2$ by being able to approximate $\mathbf{C}^2$ curve with anisotropic elements approximately aligned with their tangent and having 2 vanishing moments. Unfortunately, no fixed representation is known to achieve a similar result for $\alpha$ larger than 2; More adaptivity seems required.

### 3.3.4 Dictionary of orthogonal bandlet bases

To cope with higher regularity, S. Mallat and I [*Art-LPM05a*; *Art-LPM05b*] and then Peyré and Mallat [PM08], inspired by the curvelets and the shearlets that are optimal for $\mathbf{C}^2$ geometrically regular functions, have searched basis elements with a more "curvy" geometry and more anisotropy to follow $\mathbf{C}^\alpha$ edges efficiently, and with more vanishing moments. Arandiga, Cohen, Donat, Dyn, and Matei [Ar+08] has proposed a very different approach: a ENO-EA wavelet type lifting scheme in which the "wavelets" are defined only through the computation of the corresponding coefficients. Although well understood in the noiseless case as shown by Matei [Ma05], the mathematical analysis of those schemes in presence of noise remains a challenge.

We will thus use the bandlet bases of Peyré and Mallat [PM08] that are orthogonal bases whose elements have the required anisotropy, directionality and vanishing moments. Their construction is based on the observation that even if the wavelet coefficients are large in the neighborhood of an edge, these wavelets coefficients are regular along the direction of the edge as illustrated by Fig 3.1.

To capture this geometric regularity, the key tool is a local orthogonal transform, inspired by the work of Alpert [Al92], that combines locally the wavelets along the direction of regularity, represented by arrows in the rightmost image of Fig 3.1, to produce a new orthogonal basis, a bandlet basis. By construction, the bandlets are elongated along the direction of regularity and have the vanishing moments along this direction. The (possibly large) wavelets coefficients are thus locally recombined along this direction, yielding more coefficients of small amplitudes than before.

More precisely, the construction of a bandlet basis of a wavelet multiresolution space $V_j = \mathrm{Span}\{\phi_{j,k_1,k_2}\}_{k_1,k_2}$ starts by decomposing this space into detail wavelet spaces

$$V_j = \bigoplus_{o,l>j} W_l^o \quad \text{with} \quad W_l^o = \mathrm{Span}\{\psi_{l,k_1,k_2}^o\}_{k_1,k_2} \ .$$

For any level $l$ and orientation $o$, the detail space $W_l^o$ is a space of dimension $(2^{-l})^2$. Its coefficients are recombined using the Alpert transform induced by some directions of regularity. This geometry is specified by a local geometric flow, a vector field meant to follow the geometric direction of regularity. This geometric flow is further constrained to have a specific structure as illustrated in Fig. 3.2. It is structured by a partition into dyadic squares in which the flow, if there exists, is vertically or horizontally constant. In each square of the partition, the flow being thus easily parametrized by its tangent.

For each choice of geometric flow, a specific orthogonalization process given by Peyré and Mallat [PM08] yields an orthogonal basis of bandlets that have vanishing moments along the direction of the

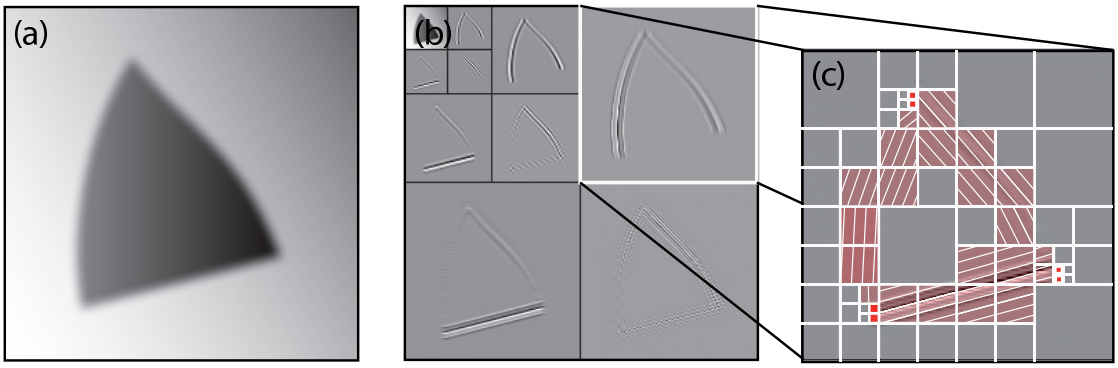



Figure 3.2: a) a geometrically regular image b) the corresponding wavelet coefficients c) the quadtree associated to the segmentation of a detail space $W_j^o$. In each square where the image is not uniformly regular, the flow is shown.

geometric flow. This geometry should obviously be adapted to each image: the partition and the flow direction should match the image structures. This choice of geometry can be seen as an ill posed problem of estimation of the edges or of the direction of regularity. To avoid this issue, the problem is recasted as a best basis search in a dictionary. The geometry chosen is the one of the best basis.

The first step is to define a dictionary $\mathcal{D}_{(2^{-j})^2}$ of orthogonal bandlet bases of $V_j$ or equivalently a dictionary of possible geometric flows. Obviously this dictionary should be finite and this requires a discretization of the geometry. As proved by Peyré and Mallat [PM08], this is not an issue: the flow does not have to follow exactly the direction of regularity but only up to a sufficient known precision. It is indeed sufficient to parametrize the flow in any dyadic square by the tangent of a polynomial of degree $p$ (the number of vanishing moments of the wavelets). The coefficients of this polynomial can be further quantized. The resulting family of geometric flow in a square is of size $O(2^{-jp})$.

A basis of the dictionary $\mathcal{D}_{(2^{-j})^2}$ is thus specified by a set of dyadic squares partitions for each details spaces $W_l^o$, $l > j$, and, for each square of the partition, a flow parametrized by a direction and one of these $O(2^{-jp})$ polynomials. The number of bases in the dictionary $\mathcal{D}_{(2^{-j})^2}$ grows exponentially with $2^{-j}$, but the total number of different bandlets $P$ grows only polynomially like $O(2^{-j(p+4)})$. Indeed the bandlets in a given dyadic square with a given geometry are reused in numerous bases. The total number of bandlets in the dictionary is thus bounded by the sum over all $O(2^{-2j})$ dyadic squares and all $O(2^{-jp)})$ choices for the flow of the number of bandlets in the square. Noticing that $(2^{-j})^2$ is a rough bound of the number of bandlets in any subspaces of $V_j$, we obtain the existence of a constant $C_K$ such that $2^{-j(p+4)} \leq P \leq C_K 2^{-j(p+4)}$.

### 3.3.5 Approximation in bandlet dictionaries

The key property of the bandlet basis dictionary is that it provides an asymptotically optimal representation of $C^\alpha$ geometrically regular functions. Indeed Peyré and Mallat [PM08] proved

**Theorem 2.** *Let $\alpha < \alpha_0$ where $\alpha_0$ in the number of wavelet vanishing moments, for any $s_0$ $\mathbf{C}^\alpha$ geometrically regular function, there exists a real number $C$ such that for any $T > 0$ and $2^j \leq T$*

$$\min_{\mathcal{B} \in \mathcal{D}_{(2^{-j})^2}} \|s_0 - P_{\mathcal{M}_{\mathcal{B},s,T}} s_0\|^2 + \dim\left(\mathcal{M}_{\mathcal{B},s,T}\right) T^2 \leqslant C T^{2\alpha/(\alpha+1)}$$

*where the subspace $\mathcal{M}_{\mathcal{B},s,T}$ is the space spanned by the vectors of $\mathcal{B}$ whose inner product with $s_0$ is larger than $T$.*

This Theorem gives the kind of control we require in Theorem 1.

For practical applications the possibility to compute efficiently the above minimization is as important as the bound $CT^{2\alpha/(\alpha+1)}$ itself. It turns out that a fast algorithm can be used to find the best basis that

minimizes $\|s_0 - P_{\mathcal{M}_{\mathcal{B},s,T}} s_0\|^2 + \dim(\mathcal{M}_{\mathcal{B},s,T})\, T^2$ or equivalently $\|P_{V_j} s_0 - P_{\mathcal{M}_{\mathcal{B},s,T}} s_0\|^2 + \dim(\mathcal{M}_{\mathcal{B},s,T})\, T^2$. We use first the additive structure with respect to the subband $W_l^o$ of this "cost" $\|P_{V_j} s_0 - P_{\mathcal{M}_{\mathcal{B},s,T}} s_0\|^2 + \dim(\mathcal{M}_{\mathcal{B},s,T})\, T^2$ to split the minimization into several independent minimizations on each subbands. A bottom-top fast optimization of the geometry (partition and flow) similar to the one proposed by Coifman and Wickerhauser [CW92], and Donoho [Do97] can be performed on each subband thanks to two observations. Firstly, for a given dyadic square, the limited number of possible flows is such that the best flow can be obtained with a simple brute force exploration. Secondly, the hierarchical tree structure of the partition and the additivity of the cost function with respect to the partition implies that the best partition of a given dyadic square is either itself or the union of the best partitions of its four dyadic subsquares. This leads to a bottom up optimization algorithm once the best flow has been found for every dyadic squares. Note that this algorithm is adaptive with respect to $\alpha$: it does not require the knowledge of the regularity parameter to be performed.

More precisely, the optimization algorithm goes as follows. The brute force search of the best flow is conducted independently over all dyadic squares and all detail spaces with a total complexity of order $O(2^{-j(p+4)})$. This yields a value of the penalized criterion for each dyadic squares. It remains now to find the best partition. We proceed in a bottom up fashion. The best partition with squares of width smaller than $2^{j+1}$ is obtained from the best partition with squares of width smaller than $2^j$: inside each dyadic square of width $2^{j+1}$ the best partition is either the partition obtained so far or the considered square. This choice is made according to the cost computed so far. Remark that the initialization is straightforward as the best partition with square of size 1 is obviously the full partition. The complexity of this best partition search is of order $O(2^{-2j})$ and thus the complexity of the best basis is driven by the best flow search whose complexity is of order $O(2^{-j(p+4)})$, which nevertheless remains polynomial in $2^{-j}$.

### 3.3.6 Bandlet estimators

Estimating the edges is a complex task on blurred function and becomes even much harder in presence of noise. Fortunately, the bandlet estimator proposed by Peyré, Le Pennec, Dossal, and Mallat [*Proc-Pe+07*] do not rely on such a detection process. The chosen geometry is obtained with the best basis selection of the previous section. This allows one to select an efficient basis even in the noisy setting.

Indeed, combining the bandlet approximation result of Theorem 2 with the model selection results of Theorem 1 proves that the selection model based bandlet estimator is near asymptotically minimax for $\mathbf{C}^\alpha$ geometrically regular images.

For a given noise level $\sigma$, one has to select a dimension $N = (2^{-j})^2$ and a threshold $T$. The best basis algorithm selects then the bandlet basis $\widehat{\mathcal{B}}$ among $\mathcal{D}_N = \mathcal{D}_{(2^{-j})^2}$ that minimizes

$$\|P_{V_N} Y - P_{\mathcal{M}_{\mathcal{B},Y,T}} Y\|^2 + T^2 \dim(\mathcal{M}_{\mathcal{B},Y,T})$$

and the model selection based estimate is $F = P_{\mathcal{M}_{\mathcal{B},Y,T}} Y$. We should now specify the choice of $N = (2^{-j})^2$ and $T$ in order to be able to use Theorem 1 and Theorem 2 to obtain the near asymptotic minimaxity of the estimator. On one hand, the dimension $N$ should be chosen large enough so that the unknown linear approximation error $\|s_0 - P_{V_N} s_0\|^2$ is small. One the other hand, the dimension $N$ should not be too large so that the total number of bandlets $P$, which satisfies $\sqrt{N}^{(p+4)} \leq K_N \leq C_K \sqrt{N}^{(p+4)}$, imposing a lower bound on the value of the threshold remains small. For the sake of simplicity, as we consider an asymptotic behavior, we assume that $\sigma$ is smaller than $1/4$. This implies that it exists $j < 0$ such that $\sigma \in (2^{j-1}, 2^j]$ The following theorem proves that choosing $N = 2^{-2j}$ and $T = \widetilde{\lambda}\sqrt{|\log \sigma|}\sigma$ with $\widetilde{\lambda}$ large enough yields a nearly asymptotically minimax estimator.

**Theorem 3.** *Let $\alpha < p$ where $p$ in the number of wavelet vanishing moments and let $K_0 \in \mathcal{N}^*$ and $\widetilde{\lambda} \geq \sqrt{2(p+4)} \sup_{K \geq K_0} \lambda_0(K)$. For any $\mathbf{C}^\alpha$ geometrically regular function $s_0$, there exists $C > 0$ such that for any*

$$\sigma \leq \min(\frac{1}{4}, \max(C_K, K_0/2)^{-1/(p+4)}),$$

| Image | Noise | Wavelet | Curvelet | Bandlet |
|---|---|---|---|---|
| Polygons | 22 | 32.73 | 32.36 | 34.56 |
| Lena | 22 | 28.15 | 28.29 | 28.7 |
| Barbara | 22 | 26.57 | 27.49 | 28.14 |
| Peppers | 22 | 27.85 | 27.74 | 28.49 |

Table 3.1: PSNR for the wavelet, curvelet and bandlet estimators for a geometrical image (Polygons given in Figure 3.3) and three classical images (Lena, Barbara and Peppers) with a noise level of $22\,dB$.

*if we let* $N = 2^{-2j}$ *with* $j$ *such that* $\sigma \in (2^{j-1}, 2^j]$ *and* $T = \widetilde{\lambda}\sqrt{|\log\sigma|}\sigma$*, the estimator* $\hat{s} = P_{\mathcal{M}_{\widehat{\mathcal{B}},Y,T}} Y$ *obtained by thresholding* $P_{V_N} Y$ *with a threshold* $T$ *in the basis* $\widehat{\mathcal{B}}$ *of* $\mathcal{D}_N$ *that minimizes*

$$\|P_{V_N} Y - P_{\mathcal{M}_{\mathcal{B},Y,T}} Y\|^2 + T^2 \dim\left(\mathcal{M}_{\mathcal{B},Y,T}\right)$$

*satisfies*

$$E\left[\|s_0 - \hat{s}\|^2\right] \leq C(|\log\sigma|\sigma^2)^{\frac{\alpha}{\alpha+1}}.$$

Theorem 3 is a direct consequence of Theorem 1 and Theorem 2.

The estimate $F = P_{\mathcal{M}_{\widehat{\mathcal{B}},T}} Y$ is computed efficiently by the same fast algorithm used in the approximation setting without requiring the knowledge of the regularity parameter $\alpha$. The model selection based bandlet estimator is thus a tractable adaptive estimator that attains, up to the logarithmic term, the best possible asymptotic minimax risk decay for $\mathbf{C}^\alpha$ geometrically regular function.

Although Theorem 3 applies only to $\mathbf{C}^\alpha$ geometrically regular functions, one can use the bandlet estimator with many kinds of images. Indeed for any function for which a theorem similar to Theorem 2 exists, the proof of Theorem 3 yields a control on the estimation risk. An important case is the Besov bodies. As among the bandlets bases there is the classical wavelet basis, any Besov function can be approximated optimally in this specific "bandlet" basis. The bandlet estimate will thus provide, up to a logarithmic term, an optimal asymptotic minimax rate.

To illustrate the good numerical behavior of the bandlet estimator, we show some experiments extracted from [*Proc-Pe+07*] and completed by a comparison with a (translation invariant) curvelet estimator. Table 3.1 shows the improvement due to the bandlet representation by comparing the PSNR for an optimized thresholding method in a wavelet representation, a curvelet representation and a bandlet representation. As expected, the bandlet estimator yields the best results. This quantitative improvement translates into a better visual quality as illustrated in Figure 3.3. Both curvelets and bandlets preserve much more geometric structures than wavelets. Curvelets are even better than bandlets to preserve the geometry of true edges but at the price of introducing some geometric artifacts mostly parallel to true edges as visible in the Polygons example but also in random direction due to noise shaping as visible in the top part of Lena's hat. This effect is nevertheless less visible in natural images than in artificial ones because of texture masking effect.

## 3.4 Maxiset of model selection

So far, I had shown that the bandlets are useful to estimate geometrically regular images. A natural question is then to ask whether they are other functions *well* estimated by such an estimator or more precisely to characterize those functions. It turns out that this problem had already been addressed by Kerkyacharian and Picard [KP00] and Cohen, De Vore, Kerkyacharian, and Picard [Co+01] for wavelet basis and termed the *maxiset* approach. With F. Autin, V. Rivoirard and J.-M. Loubes [*Art-Au+10*], I have studied this question for general model selection estimators. Our purpose is not to build new model selection estimators but to determine thoroughly the functions for which well known model selection procedures achieve good performances. Of course, approximation theory plays a crucial role in our setting but surprisingly its role is even more important than the one of statistical tools. This statement will be emphasized by the use of the *maxiset approach*, which illustrates the well known fact that *well estimating is well approximating.*

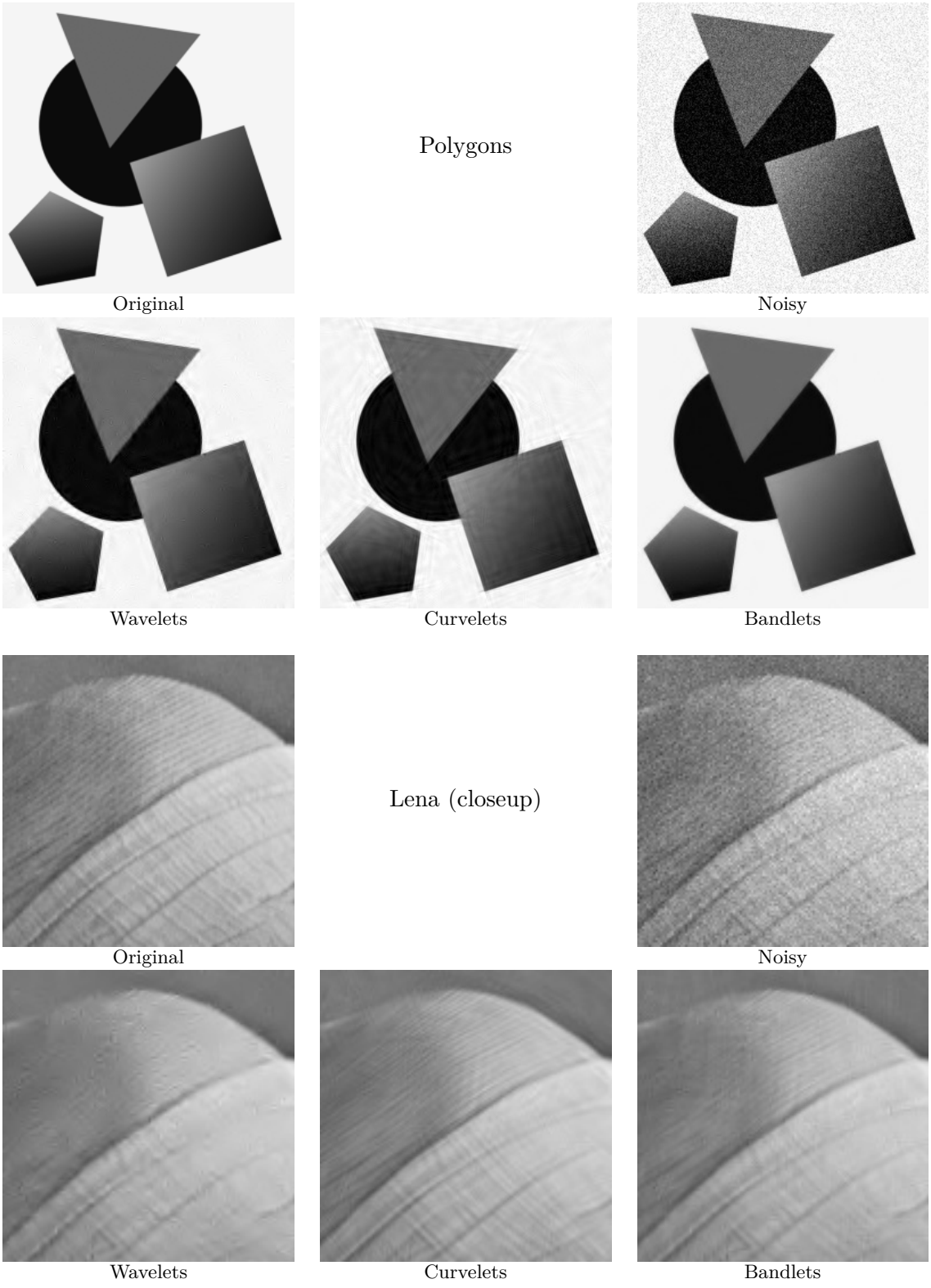



Figure 3.3: Visual comparison of the different estimators

### 3.4.1 Model selection procedures

The model selection methodology of the previous sections seems restricted to the proposed two step projection strategy. This is not the case and we have considered a more general case. Roughly speaking, the model selection methodology consists in choosing a collection of set, called model collection, construction for each set an estimator by minimizing an empirical contrast $\gamma$ and finally selecting a best model within this collection. The pioneer work in model selection goes back in the 1970's with Mallows [Ma73] and Akaike [Ak73]. Birgé and Massart, with the help of Barron, develop the whole modern theory of model selection in [BM00; BM01; BM07] or [BBM99] for instance. Estimation of a regression function with model selection estimators is considered by Baraud in [Ba00; Ba02], while inverse problems are tackled by Loubes and Ludeña [LL08; LL10]. Finally model selection techniques provide nowadays valuable tools in statistical learning (see Boucheron, Bousquet, and Lugosi [BBL05]).

We continue here with the quadratic loss and hence use the following empirical contrast:

$$\gamma(s) = -2Y_s + \|s\|^2.$$

For any model (set) $S_m$, we define $\hat{s}_m$ as the minimizer of the contrast over this set:

$$\hat{s}_m = \operatorname*{argmin}_{s \in S_m} \gamma(s).$$

Note that when $S_{m'} \subset S_m$ then one also has

$$\hat{s}_{m'} = \operatorname*{argmin}_{s \in S_{m'}} \|\hat{s}_m - s\|^2.$$

Finally, when $S_m$ is a linear space $V$, as in the previous section, one verifies immediately that

$$\hat{s}_m = P_V Y.$$

Combining those two observations allows to slightly generalize the setting of the previous sections.

From now on, we assume we are given a dictionary of functions of $\mathbb{L}_2$, denoted by $\mathcal{D} = (\Phi_p)_{p \in \mathcal{I}}$ where $\mathcal{I}$ is a countable set. We consider then a collection of model $\mathcal{M}_N$ in which every model $S_m$ is spanned by some functions of the dictionary. For any $S_m \in \mathcal{M}_n$, we denote by $\mathcal{I}_m$ the subset of $\mathcal{I}$ such that

$$model_m = \mathrm{Span}\{\varphi_p : \quad p \in \mathcal{I}_m\}$$

and $D_m \leq |\mathcal{I}_m|$ the dimension of $S_m$. As observed previously, as the model $S_m$ are linear space, one has

$$\hat{s}_m = P_{S_m} Y.$$

Let $\{b_1^m, \ldots, b_{D_m}^m\}$ an orthonormal basis (not necessarily related to $\Phi$) of $S_m$ then

$$\hat{s}_m = \sum_{p \in \mathcal{I}_m} Y_{b_p^m} b_p^m, \quad \text{and} \quad \gamma(\hat{s}_m) = -\sum_{p \in \mathcal{I}_m} (Y_{b_p^m})^2.$$

Now, the issue is the selection of the best model $\hat{m}$ from the data which gives rise to the *model selection estimator* $\hat{s}_{\hat{m}}$. For this purpose, a penalized rule is considered, which aims at selecting an estimator, close enough to the data, but still lying in a small space to avoid overfitting issues. Let $\mathrm{pen}_N(m)$ be a penalty function, the model $\hat{m}$ is selected using the following penalized criterion

$$\hat{m} = \arg \min_{m \in \mathcal{M}_N} \left\{ \gamma(\hat{s}_m) + \mathrm{pen}_n(m) \right\}. \tag{3.4}$$

The choice of the model collection and the associated penalty are then the key issues handled by model selection theory. We point out that the choices of both the model collection and the penalty function should depend on the noise level $\sigma$. This is emphasized by the subscript $N$ for $\mathcal{M}_N$ and $\mathrm{pen}_N(m)$ which has been already used in the previous sections. For sake of simplicity, we will use the calibration $\sigma = 1/\sqrt{N}$ (the one that corresponds to the asymptotic equivalence between the white noise model and the regression one). For sake of simplicity, we will consider only integer values of $N$, but this can be easily overcome.

The asymptotic behavior of model selection estimators has been studied by many authors. We refer to Massart [Ma07] for general references and recall hereafter the main oracle type inequality. Such an oracle inequality provides a non asymptotic control on the estimation error with respect to a bias term $\|s_0 - s_m\|$, where $s_m$ stands for the best approximation (in the $\mathbb{L}_2$ sense) of the function $s_0$ by a function of $S_m$. In other words $s_m$ is the orthogonal projection of $s_0$ onto $S_m$, defined by

$$S_m = \sum_{p \in \mathcal{I}_m} \beta_p^m b_p^m, \quad \beta_i^m = \int b_i^m(t) s_0(t) dt.$$

Theorem 1 was indeed a special case of

**Theorem 4** (Theorem 4.2 of [Ma07])**.** *Let $N$ be fixed and let $(x_m)_{m \in \mathcal{M}_N}$ be some family of positive numbers such that*

$$\sum_{m \in \mathcal{M}_N} \exp(-x_m) = \Sigma_N < \infty. \tag{3.5}$$

*Let $\kappa > 1$ and assume that*

$$\mathrm{pen}_N(m) \geq \frac{\kappa}{N} \left( \sqrt{D_m} + \sqrt{2x_m} \right)^2. \tag{3.6}$$

*Then, almost surely, there exists some minimizer $\hat{m}$ of the penalized least-squares criterion*

$$\gamma(\hat{s}_m) + \mathrm{pen}_N(m)$$

*over $m \in \mathcal{M}_n$. Moreover, the corresponding penalized least-squares estimator $\hat{s}_{\hat{m}}$ is unique and the following inequality holds:*

$$\mathbb{E}\left[ \|\hat{s}_0 - s_{\hat{m}}\|^2 \right] \leq C \left[ \inf_{m \in \mathcal{M}_n} \left\{ \|s_0 - s_m\|^2 + \mathrm{pen}_N(m) \right\} + \frac{1 + \Sigma_N}{n} \right], \tag{3.7}$$

*where $C$ depends only on $\kappa$.*

As shown in the previous sections, an equation of type (3.7) is a key result to establish optimality of penalized estimators under oracle or minimax points of view. We focus now on an alternative to these approaches: the maxiset point of view.

### 3.4.2 The maxiset point of view

Before describing the maxiset approach, let us briefly recall that for a given procedure $s^* = (s_N^*)_N$, which may differ for different noise level, the minimax study of $s^*$ consists in comparing the rate of convergence of $s^*$ achieved on a given functional space $\mathcal{F}$ with the best possible rate achieved by any estimator. More precisely, let $\mathcal{F}(R)$ be the ball of radius $R$ associated with $\mathcal{F}$, the procedure $s^* = (s_N^*)_N$ achieves the rate $\rho^* = (\rho_N^*)_N$ on $\mathcal{F}(R)$ if

$$\sup_N \left\{ (\rho_N^*)^{-2} \sup_{s \in \mathcal{F}(R)} \mathbb{E}\left[ \|s_N^* - s_0\|^2 \right] \right\} < \infty.$$

To check that a procedure is optimal from the minimax point of view (said to be minimax), it must be proved that its rate of convergence achieves the best rate among any procedure on each ball of the class. This minimax approach is extensively used and many methods cited above are proved to be minimax in different statistical frameworks.

However, the choice of the function class is subjective and, in the minimax framework, statisticians have no idea whether there are other functions well estimated at the rate $\rho^*$ by their procedure. A different point of view is to consider the procedure $s^*$ as given and search all the functions $s$ that are well estimated at a given rate $\rho^*$: this is the *maxiset* approach, which has been proposed by Kerkyacharian and Picard [KP00]. The maximal space, or maxiset, of the procedure $s^*$ for this rate $\rho^*$ is defined as the set of all these functions. Obviously, the larger the maxiset, the better the procedure. We set the following definition.

**Definition 2.** *Let $\rho^* = (\rho_N^*)_N$ be a decreasing sequence of positive real numbers and let $s^* = (s_N^*)_N$ be an estimation procedure. The maxiset of $s^*$ associated with the rate $\rho^*$ is*

$$MS(s^*, \rho^*) = \left\{ s_0 \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_N \left\{ (\rho_N^*)^{-2} \mathbb{E}\left[ \|s_N^* - s_0\|^2 \right] \right\} < \infty \right\},$$

*the ball of radius $R > 0$ of the maxiset is defined by*

$$MS(s^*, \rho^*)(R) = \left\{ s_0 \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_N \left\{ (\rho_N^*)^{-2} \mathbb{E}\left[ \|s_N^* - s_0\|^2 \right] \right\} \leq R^2 \right\}.$$

Of course, there exist connections between maxiset and minimax points of view: $s^*$ achieves the rate $\rho^*$ on $\mathcal{F}$ if and only if

$$\mathcal{F} \subset MS(s^*, \rho^*).$$

In the white noise setting, the maxiset theory has been investigated for a wide range of estimation procedures, including kernel, thresholding and Lepski procedures, Bayesian or linear rules. We refer to Autin [Au08]; Autin, Picard, and Rivoirard [APR06]; Bertin and Rivoirard [BR09]; Cohen, De Vore, Kerkyacharian, and Picard [Co+01]; Kerkyacharian and Picard [KP00]; Rivoirard [Ri04; Ri05] for general results. Maxisets have also been investigated for other statistical models, see Autin [Au06]; Rivoirard and Tribouley [RT07].

Our goal was to investigate maxisets of model selection procedures. Following the classical model selection literature, we only use penalties proportional to the dimension $D_m$ of $m$:

$$\mathrm{pen}_N(m) = \frac{\lambda_N}{N} D_m, \tag{3.8}$$

with $\lambda_N$ to be specified. Our main result characterizes these maxisets in terms of approximation spaces. More precisely, we establish an equivalence between the statistical performance of $\hat{s}_{\hat{m}}$ and the approximation properties of the model collections $\mathcal{M}_N$. With

$$\rho_{N,\alpha} = \left( \frac{\lambda_N}{N} \right)^{\frac{\alpha}{1+2\alpha}} \tag{3.9}$$

for any $\alpha > 0$, Theorem 5, combined with Theorem 4 proves that, for a given function $s$, the quadratic risk $\mathbb{E}[\|s_0 - \hat{s}_{\hat{m}}\|^2]$ decays at the rate $\rho_{N,\alpha}^2$ if and only if the deterministic quantity

$$Q(s, N) = \inf_{m \in \mathcal{M}_N} \left\{ \|s_0 - s_m\|^2 + \frac{\lambda_N}{N} D_m \right\} \tag{3.10}$$

decays at the rate $\rho_{N,\alpha}^2$ as well. This result holds with mild assumptions on $\lambda_N$ and under an embedding assumption on the model collections ($\mathcal{M}_N \subset \mathcal{M}_{N+1}$). Once we impose additional structure on the model collections, the deterministic condition can be rephrased as a linear approximation property and a non linear one as stated in Theorem 6.

We illustrate these results for three different model collections based on wavelet bases. The first one deals with sieves in which all the models are embedded, the second one with the collection of all subspaces spanned by vectors of a given basis. For these examples, we handle the issue of calculability and give explicit characterizations of the maxisets. In the third example, we provide an intermediate choice of model collections and use the fact that the embedding condition on the model collections can be relaxed. Finally performances of these estimators are compared and discussed.

As explained, our goal is to investigate maxisets associated with model selection estimators $\hat{s}_{\hat{m}}$ where the penalty function is defined in (3.8) and with the rate $\rho_\alpha = (\rho_{N,\alpha})_N$ where $\rho_{n,\alpha}$ is specified in (3.9). Observe that $\rho_{N,\alpha}$ depends on the choice of $\lambda_N$ that defines the penalty. It can be for instance polynomial, or can take the classical form

$$\rho_{N,\alpha} = \left( \frac{\log N}{N} \right)^{\frac{\alpha}{1+2\alpha}}.$$

So we wish to determine

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_n \left\{ \rho_{n,\alpha}^{-2} \mathbb{E}\left[ \|\hat{s}_{\hat{m}} - s\|^2 \right] \right\} < \infty \right\}.$$

In the sequel, we use the following notation: if $\mathcal{F}$ is a given space

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) =: \mathcal{F}$$

means that for any $R > 0$, there exists $R' > 0$ such that

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha)(R) \subset \mathcal{F}(R') \tag{3.11}$$

and for any $R' > 0$, there exists $R > 0$ such that

$$\mathcal{F}(R') \subset MS(\hat{s}_{\hat{m}}, \rho_\alpha)(R). \tag{3.12}$$

### 3.4.3 Abstract maxiset results

**The case of general dictionaries**

In this section, we make no assumption on $\Phi$. Theorem 4 is a non asymptotic result while maxisets results deal with rates of convergence (with asymptotics in $n$). Therefore obtaining maxiset results for model selection estimators requires a structure on the sequence of model collections. We first focus on the case of nested model collections ($\mathcal{M}_N \subset \mathcal{M}_{N+1}$). Note that this does not imply a strong structure on the model collection for a given $N$. In particular, this does not imply that the models are nested. Identifying the maxiset $MS(\hat{s}_{\hat{m}}, \rho_\alpha)$ is a two-step procedure. We need to establish inclusion (3.11) and inclusion (3.12). Recall that we have introduced previously

$$Q(s, N) = \inf_{m \in \mathcal{M}_N} \left\{ \|s_0 - s_m\|^2 + \frac{\lambda_n}{n} D_m \right\}.$$

Roughly speaking, Theorem 4 established by Massart proves that any function $s$ satisfying

$$\sup_N \left\{ \rho_{N,\alpha}^{-2} Q(s, n) \right\} \leq (R')^2$$

belongs to the maxiset $MS(\hat{s}_{\hat{m}}, \rho_\alpha)$ and thus provides inclusion (3.12). The following theorem establishes inclusion (3.11) and highlights that $Q(s, N)$ plays a capital role.

**Theorem 5.** *Let $0 < \alpha_0 < \infty$ be fixed. Let us assume that the sequence of model collections satisfies for any $N$*

$$\mathcal{M}_N \subset \mathcal{M}_{N+1}, \tag{3.13}$$

*and that the sequence of positive numbers $(\lambda_N)_N$ is non-decreasing and satisfies*

$$\lim_{N \to +\infty} N^{-1} \lambda_n = 0, \tag{3.14}$$

*and there exist $N_0 \in \mathbb{N}^*$ and two constants $0 < \delta \leq \frac{1}{2}$ and $0 < p < 1$ such that for $N \geq N_0$,*

$$\lambda_{2n} \leq 2(1 - \delta)\lambda_N, \tag{3.15}$$

$$\sum_{m \in \mathcal{M}_N} e^{-\frac{(\sqrt{\lambda_n} - 1)^2 D_m}{2}} \leq \sqrt{1 - p} \tag{3.16}$$

*and*

$$\lambda_{N_0} \geq \Upsilon(\delta, p, \alpha_0), \tag{3.17}$$

*where $\Upsilon(\delta, p, \alpha_0)$ is a positive constant only depending on $\alpha_0$, $p$ and $\delta$. Then, the penalized rule $\hat{s}_{\hat{m}}$ is such that for any $\alpha \in (0, \alpha_0]$, for any $R > 0$, there exists $R' > 0$ such that for $s_0 \in \mathbb{L}_2(\mathcal{D})$,*

$$\sup_N \left\{ \rho_{N,\alpha}^{-2} \mathbb{E}\left[ \|s_0 - \hat{s}_{\hat{m}}\|^2 \right] \right\} \leq R^2 \Rightarrow \sup_N \left\{ \rho_{N,\alpha}^{-2} Q(s, n) \right\} \leq (R')^2.$$

Technical Assumptions (3.14), (3.15), (3.16) and (3.17) are very mild and could be partly relaxed while preserving the results. Assumption (3.14) is necessary to deal with rates converging to 0. Note that the classical cases $\lambda_N = \lambda_0$ or $\lambda_N = \lambda_0 \log(N)$ satisfy (3.14) and (3.15). Furthermore, Assumption (3.17) is always satisfied when $\lambda_N = \lambda_0 \log(N)$ or when $\lambda_N = \lambda_0$ with $\lambda_0$ large enough. Assumption (3.16) is very close to Assumptions (3.5)-(3.6). In particular, if there exist two constants $\kappa > 1$ and $0 < p < 1$ such that for any $n$,

$$\sum_{m \in \mathcal{M}_N} e^{-\frac{(\sqrt{\kappa^{-1}\lambda_N}-1)^2 D_m}{2}} \leq \sqrt{1-p} \tag{3.18}$$

then, since

$$\mathrm{pen}_N(m) = \frac{\lambda_N}{N} D_m,$$

Conditions (3.5), (3.6) and (3.16) are all satisfied. The assumption $\alpha \in (0, \alpha_0]$ can be relaxed for particular model collections, which will be highlighted in Proposition 2 of Section 3.4.4. Finally, Assumption (3.13) can be removed for some special choice of model collection $\mathcal{M}_N$ at the price of a slight overpenalization as it shall be shown in Proposition 1 and Section 3.4.4.

Combining Theorems 4 and 5 gives a first characterization of the maxiset of the model selection procedure $\hat{s}_{\hat{m}}$:

**Corollary 1.** *Let $\alpha_0 < \infty$ be fixed. Assume that Assumptions (3.13), (3.14), (3.15) (3.17) and (3.18) are satisfied. Then for any $\alpha \in (0, \alpha_0]$,*

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_n \left\{ \rho_{n,\alpha}^{-2} Q(s,n) \right\} < \infty \right\}.$$

The maxiset of $\hat{s}_{\hat{m}}$ is characterized by a deterministic approximation property of $s$ with respect to the models $\mathcal{M}_n$. It can be related to some classical approximation properties of $s$ in terms of approximation rates if the functions of $\Phi$ are orthonormal.

**The case of orthonormal bases**

From now on, $\mathcal{D} = \{\Phi_i\}_{i \in \mathcal{I}}$ is assumed to be an orthonormal basis (for the $\mathbb{L}_2$ scalar product). We also assume that the model collections $\mathcal{M}_n$ are constructed through restrictions of a single model collection $\mathcal{M}$. Namely, given a collection of models $\mathcal{M}$ we introduce an increasing sequence $\mathcal{J}_n$ of collection of indice subsets, and define the intermediate collection $\mathcal{M}'_n$ as

$$\mathcal{M}'_N = \{ S'_m = \mathrm{Span}\{\Phi_i : \quad i \in \mathcal{I}_m \cap \mathcal{J}_N\} : \quad S_m \in \mathcal{M}\} \tag{3.19}$$

where the sets $\mathcal{I}_m$ do not depend on $N$ The model collections $\mathcal{M}'_n$ do not necessarily satisfy the embedding condition (3.13). Thus, we define

$$\mathcal{M}_N = \bigcup_{k \leq N} \mathcal{M}'_k$$

so $\mathcal{M}_n \subset \mathcal{M}_{n+1}$. The assumptions on $\mathcal{D}$ and on the model collections allow to give an explicit characterization of the maxisets. We denote $\widetilde{\mathcal{M}} = \cup_N \mathcal{M}_n = \cup_N \mathcal{M}'_n$. Remark that without any further assumption $\widetilde{\mathcal{M}}$ can be a larger model collection than $\mathcal{M}$. Now, let us denote by $V = (V_N)_N$ the sequence of approximation spaces defined by

$$V_N = \mathrm{Span}\{\phi_i : \quad i \in \mathcal{J}_N\}$$

and consider the corresponding approximation space

$$\mathcal{L}_V^\alpha = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_n \left\{ \rho_{N,\alpha}^{-1} \| P_{V_N} s - s \| \right\} < \infty \right\},$$

where $P_{V_N} s$ is the projection of $s$ onto $V_N$. Define also another kind of approximation sets:

$$\mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}} = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_{M>0} \left\{ M^{\alpha} \inf_{\{m \in \widetilde{\mathcal{M}}:\, D_m \leq M\}} \|s_m - s\| \right\} < \infty \right\}.$$

The corresponding balls of radius $R > 0$ are defined, as usual, by replacing $\infty$ by $R$ in the previous definitions. We have the following result.

**Theorem 6.** *Let $\alpha_0 < \infty$ be fixed. Assume that (3.14), (3.15), (3.17) and (3.18) are satisfied. Then, the penalized rule $\hat{s}_{\hat{m}}$ satisfies the following result: for any $\alpha \in (0, \alpha_0]$,*

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}} \cap \mathcal{L}^{\alpha}_V.$$

The result pointed out in Theorem 6 links the performance of the estimator to an approximation property for the estimated function. This approximation property is decomposed into a linear approximation measured by $\mathcal{L}^{\alpha}_V$ and a non linear approximation measured by $\mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}}$. The linear condition is due to the use of the reduced model collection $\mathcal{M}_n$ instead of $\mathcal{M}$, which is often necessary to ensure either the calculability of the estimator or Condition (3.18). It plays the role of a minimum regularity property that is easily satisfied.

Observe that if we have one model collection, that is for any $k$ and $k'$, $\mathcal{M}_k = \mathcal{M}_{k'} = \mathcal{M}$, $\mathcal{J}_n = \mathcal{I}$ for any $n$ and thus $\widetilde{\mathcal{M}} = \mathcal{M}$. Then

$$\mathcal{L}^{\alpha}_V = \text{span}\left\{ \varphi_i : \quad i \in \mathcal{I} \right\}$$

and Theorem 6 gives

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \mathcal{A}^{\alpha}_{\mathcal{M}}.$$

The spaces $\mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}}$ and $\mathcal{L}^{\alpha}_V$ highly depend on the models and the approximation space. At first glance, the best choice seems to be $V_n = \mathbb{L}_2(\mathcal{D})$ and

$$\mathcal{M} = \left\{ m : \quad \mathcal{I}_m \subset \mathcal{I} \right\}$$

since the infimum in the definition of $\mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}}$ becomes smaller when the collection is enriched. There is however a price to pay when enlarging the model collection: the penalty has to be larger to satisfy (3.18), which deteriorates the convergence rate. A second issue comes from the tractability of the minimization (3.4) itself which will further limit the size of the model collection.

To avoid considering the union of $\mathcal{M}'_k$, that can dramatically increase the number of models considered for a fixed $n$, leading to large penalties, we can relax the assumption that the penalty is proportional to the dimension. Namely, for any $n$, for any $m \in \mathcal{M}'_n$, there exists $\tilde{m} \in \mathcal{M}$ such that

$$S_m = \text{Span}\left\{ \Phi_i : \quad i \in \mathcal{I}_{\tilde{m}} \cap \mathcal{J}_n \right\}.$$

Then for any model $m \in \mathcal{M}'_n$, we replace the dimension $D_m$ by the larger dimension $D_{\tilde{m}}$ and we set

$$\widetilde{\text{pen}}_n(m) = \frac{\lambda_n}{n} D_{\tilde{m}}.$$

The minimization of the corresponding penalized criterion over all model in $\mathcal{M}'_n$ leads to a result similar to Theorem 6. Mimicking its proof, we can state the following proposition that will be used in Section 3.4.4:

**Proposition 1.** *Let $\alpha_0 < \infty$ be fixed. Assume (3.14), (3.15) (3.17) and (3.18) are satisfied. Then, the penalized estimator $\hat{s}_{\tilde{m}}$ where*

$$\tilde{m} = \arg\min_{m \in \mathcal{M}'_n} \left\{ \gamma(\hat{s}_m) + \widetilde{\text{pen}}_n(m) \right\}$$

*satisfies the following result: for any $\alpha \in (0, \alpha_0]$,*

$$MS(\tilde{s}_{\tilde{m}}, \rho_\alpha) :=: \mathcal{A}^{\alpha}_{\mathcal{M}} \cap \mathcal{L}^{\alpha}_V.$$

Remark that $\mathcal{M}_n$, $\mathcal{L}^{\alpha}_V$ and $\mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}}$ can be defined in a similar fashion for any arbitrary dictionary $\mathcal{D}$. However, one can only obtain the inclusion $MS(\hat{s}_{\hat{m}}, \rho_\alpha) \subset \mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}} \cap \mathcal{L}^{\alpha}_V$ in the general case.

### 3.4.4 Comparisons of model selection estimators

The aim of this section is twofold. Firstly, we propose to illustrate our previous maxiset results to different model selection estimators built with wavelet methods by identifying precisely the spaces $\mathcal{A}^{\alpha}_{\underset{\mathcal{M}}{\sim}}$ and $\mathcal{L}^{\alpha}_{V}$. Secondly, comparisons between the performances of these estimators are provided and discussed.

For sake of simplicity, we work with periodic functions on the interval $[0,1]$ and will use the associated periodic wavelet base construction (see Daubechies [Da92] for instance). We recall the characterization of Besov spaces using wavelets. Such spaces will play an important role in the following. In this section we assume that the multiresolution analysis associated with the basis $\Psi$ is $r$-regular with $r \geq 1$ as defined by Meyer [Me90]. In this case, for any $0 < \alpha < r$ and any $1 \leq p, q \leq \infty$, the periodic function $s$ belongs to the Besov space $\mathcal{B}^{\alpha}_{p,q}$ if and only if $|\alpha_{00}| < \infty$ and

$$\sum_{j=0}^{\infty} 2^{jq(\alpha+\frac{1}{2}-\frac{1}{p})} \|\beta_{j.}\|^q_{\ell_p} < \infty \quad \text{if } q < \infty,$$

$$\sup_{j\in\mathbb{N}} 2^{j(\alpha+\frac{1}{2}-\frac{1}{p})} \|\beta_{j.}\|_{\ell_p} < \infty \quad \text{if } q = \infty$$

where $(\beta_{j.}) = (\beta_{jk})_k$. This characterization allows to recall the following embeddings:

$$\mathcal{B}^{\alpha}_{p,q} \subsetneq \mathcal{B}^{\alpha'}_{p',q'} \text{ as soon as } \alpha - \frac{1}{p} \geq \alpha' - \frac{1}{p'},\ p < p' \text{ and } q \leq q'$$

and

$$\mathcal{B}^{\alpha}_{p,\infty} \subsetneq \mathcal{B}^{\alpha}_{2,\infty} \text{ as soon as } p > 2.$$

#### Collection of Sieves

We consider first a single model collection corresponding to a class of nested models

$$\mathcal{M}^{(s)} = \{m = \text{span}\{\phi_{00}, \psi_{jk} :\ j < N_m, 0 \leq k < 2^j\} :\ N_m \in \mathbb{N}\}.$$

For such a model collection, Theorem 6 could be applied with $V_N = \mathbb{L}_2$. One can even remove Assumption (3.17) which imposes a minimum value on $\lambda_{n_0}$ that depends on the rate $\rho_\alpha$:

**Proposition 2.** *Let $0 < \alpha < r$ and let $\hat{s}^{(s)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(s)}$. Then, under Assumptions* (3.14), (3.15) *and* (3.18),

$$MS(\hat{s}^{(s)}_{\hat{m}}, \rho_\alpha) :=: \mathcal{B}^{\alpha}_{2,\infty}.$$

Remark that it suffices to choose $\lambda_N \geq \lambda_0$ with $\lambda_0$, independent of $\alpha$, large enough to ensure Condition (3.18).

It is important to notice that the estimator $\hat{s}^{(s)}_{\hat{m}}$ cannot be computed in practice because to determine the best model $\hat{m}$ one needs to consider an infinite number of models, which cannot be done without computing an infinite number of wavelet coefficients. To overcome this issue, we specify a maximum resolution level $j_0(N)$ for estimation where $n \mapsto j_0(N)$ is non-decreasing. This modification is also in the scope of Theorem 6: it corresponds to

$$V_N = \text{span}\{\phi_{00}, \psi_{jk} :\ 0 \leq j < j_0(N),\ 0 \leq k < 2^j\}$$

and the model collection $\mathcal{M}^{(s)}_N$ defined as follows:

$$\mathcal{M}^{(s)}_N \quad = \quad \mathcal{M}'^{(s)}_N = \{S_m \in \mathcal{M}^{(s)} :\ N_m < j_0(n)\}.$$

For the specific choice

$$2^{j_0(n)} \leq N\lambda_N^{-1} < 2^{j_0(n)+1}, \tag{3.20}$$

we obtain:

$$\mathcal{L}^{\alpha}_{V} = \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}.$$

Since $\mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{B}^{\alpha}_{2,\infty}$ reduces to $\mathcal{B}^{\alpha}_{2,\infty}$, arguments of the proofs of Theorem 6 and Proposition 2 give:

**Proposition 3.** *Let $0 < \alpha < r$ and let $\hat{s}^{(st)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(s)}_n$. Then, under Assumptions* (3.14), (3.15) *and* (3.18)

$$MS(\hat{s}^{(st)}_{\hat{m}}, \rho_\alpha) :=: \mathcal{B}^\alpha_{2,\infty}.$$

This tractable procedure is thus as efficient as the original one. We obtain the maxiset behavior of the non adaptive linear wavelet procedure pointed out by Rivoirard [Ri04] but here the procedure is completely data-driven.

**The largest model collections**

In this paragraph we enlarge the model collections in order to obtain much larger maxisets. We start with the following model collection

$$\mathcal{M}^{(l)} = \{S_m = \text{Span}\{\phi_{00}, \psi_{jk} : \ (j,k) \in \mathcal{I}_m\} : \mathcal{I}_m \in \mathcal{P}(\mathcal{I})\}$$

where

$$\mathcal{I} = \bigcup_{j \geq 0} \{(j,k) : \quad k \in \{0, 1, \ldots, 2^j - 1\}\}$$

and $\mathcal{P}(\mathcal{I})$ is the set of all subsets of $\mathcal{I}$. This model collection is so rich that whatever the sequence $(\lambda_N)_n$, Condition (3.18) (or even Condition (3.5)) is not satisfied. To reduce the cardinality of the collection, we restrict the maximum resolution level to the resolution level $j_0(N)$ defined in (3.20) and consider the collections $\mathcal{M}^{(l)}_N$ defined from $\mathcal{M}^{(l)}$ by

$$\mathcal{M}^{(l)}_N \quad = \quad \mathcal{M}'^{(l)}_N = \left\{m \in \mathcal{M}^{(l)} : \quad \mathcal{I}_m \in \mathcal{P}(\mathcal{I}^{j_0})\right\}$$

where

$$\mathcal{I}^{j_0} = \bigcup_{0 \leq j < j_0(N)} \{(j,k) : \quad k \in \{0, 1, \ldots, 2^j - 1\}\}.$$

The classical logarithmic penalty

$$\text{pen}_N(m) = \frac{\lambda_0 \log(N) D_m}{N},$$

which corresponds to $\lambda_N = \lambda_0 \log(N)$, is sufficient to ensure Condition (3.18) as soon as $\lambda_0$ is a constant large enough (the choice $\lambda_N = \lambda_0$ is not sufficient). The identification of the corresponding maxiset focuses on the characterization of the space $\mathcal{A}^\alpha_{\mathcal{M}^{(l)}}$ since, as previously, $\mathcal{L}^\alpha_V = \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}$. We rely on sparsity properties of $\mathcal{A}^\alpha_{\mathcal{M}^{(l)}}$. In our context, sparsity means that there is a *small* proportion of *large* coefficients of a signal. Introduce for, for $N \in \mathbb{N}^*$, the notation

$$|\beta|_{(N)} = \inf\left\{u : \quad \text{card}\left\{(j,k) \in \mathbb{N} \times \{0, 1, \ldots, 2^j - 1\} : \ |\beta_{jk}| > u\right\} < N\right\}$$

to represent the non-increasing rearrangement of the wavelet coefficient of a periodic signal $s$:

$$|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(N)} \geq \cdots .$$

As the best model $S_m \in \mathcal{M}^{(l)}$ of prescribed dimension $M$ is obtained by choosing the subset of index corresponding to the $M$ largest wavelet coefficients, a simple identification of the space $\mathcal{A}^\alpha_{\mathcal{M}^{(l)}}$ is

$$\mathcal{A}^\alpha_{\mathcal{M}^{(l)}} = \left\{ s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j - 1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \quad \sup_{M \in \mathbb{N}^*} M^{2\alpha} \sum_{i=M+1}^{\infty} |\beta|^2_{(i)} < \infty \right\}.$$

Theorem 2.1 of Kerkyacharian and Picard [KP00] provides a characterization of this space as a weak Besov space:

$$\mathcal{A}^\alpha_{\mathcal{M}^{(l)}} = \mathcal{W}_{\frac{2}{1+2\alpha}}$$

with for any $q \in ]0,2[$,

$$\mathcal{W}_q = \left\{ s = \alpha_{00}\phi_{00} + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \quad \sup_{n\in\mathbb{N}^*} n^{1/q}|\beta|_{(n)} < \infty \right\}.$$

Following their definitions, the larger $\alpha$, the smaller $q = 2/(1+2\alpha)$ and the sparser the sequence $(\beta_{jk})_{j,k}$. We obtain thus the following proposition.

**Proposition 4.** *Let $\alpha_0 < r$ be fixed, let $0 < \alpha \leq \alpha_0$ and let $\hat{s}^{(l)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(s)}_N$. Then, under Assumptions* (3.14), (3.15), (3.17) *and* (3.18)*:*

$$MS\left(\hat{s}^{(l)}_{\hat{m}}, \rho_\alpha\right) =: \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{W}_{\frac{2}{1+2\alpha}}.$$

Observe that the estimator $\hat{s}^{(l)}_{\hat{m}}$ is easily tractable from a computational point of view as one easily verify that the best subset $\mathcal{I}_{\hat{m}}$ is the set $\{(j,k) \in \mathcal{I}^{j_0} : \quad |\hat{\beta}_{jk}| > \sqrt{\lambda_n/n}\}$ and $\hat{s}^{(l)}_{\hat{m}}$ corresponds to the well-known hard thresholding estimator,

$$\hat{s}^{(l)}_{\hat{m}} = \hat{\alpha}_{00}\phi_{00} + \sum_{j=0}^{j_0(n)-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{jk}\mathbf{1}_{|\hat{\beta}_{jk}|>\sqrt{\frac{\lambda_N}{N}}}\,\psi_{jk}.$$

Proposition 4 corresponds thus to the maxiset result established by Kerkyacharian and Picard [KP00].

**A special strategy for Besov spaces**

We consider now the model collection proposed by Massart [Ma07]. This collection can be viewed as an hybrid collection between the two previous collections. This strategy turns out to be minimax for all Besov spaces $\mathcal{B}^\alpha_{p,\infty}$ when $\alpha > \max(1/p - 1/2, 0)$ and $1 \leq p \leq \infty$.

More precisely, for a chosen $\theta > 2$, define the model collection by

$$\mathcal{M}^{(h)} = \{m = \text{span}\{\phi_{00}, \psi_{jk} : \ (j,k) \in \mathcal{I}_m\} : J \in \mathbb{N},\, \mathcal{I}_m \in \mathcal{P}_J(\mathcal{I})\},$$

where for any $J \in \mathbb{N}$, $\mathcal{P}_J(\mathcal{I})$ is the set of all subsets $\mathcal{I}_m$ of $\mathcal{I}$ that can be written

$$\begin{aligned}\mathcal{I}_m = &\left\{(j,k) : \quad 0 \leq j < J, 0 \leq k < 2^j\right\} \\ &\bigcup \cup_{j\geq J}\left\{(j,k) : \quad k \in A_j, |A_j| = \lfloor 2^J(j-J+1)^{-\theta}\rfloor\right\}\end{aligned}$$

with $\lfloor x \rfloor := \max\{n \in \mathbb{N} : \ n \leq x\}$.

As remarked by Massart [Ma07], for any $J \in \mathbb{N}$ and any $\mathcal{I}_m \in \mathcal{P}_J(\mathcal{I})$, the dimension $D_m$ of the corresponding model $m$ depends only on $J$ and is such that

$$2^J \leq D_m \leq 2^J\left(1 + \sum_{n\geq 1} n^{-\theta}\right).$$

We denote by $D_J$ this common dimension. Note that the model collection $\mathcal{M}^{(h)}$ does not vary with $n$. Using Theorem 6 with $V_n = \mathbb{L}_2$, we have the following proposition.

**Proposition 5.** *Let $\alpha_0 < r$ be fixed, let $0 < \alpha \leq \alpha_0$ and let $\hat{s}^{(h)}_{\hat{m}}$ be the model selection estimator associated with the model collection $\mathcal{M}^{(h)}$. Then, under Assumptions* (3.14), (3.15), (3.17) *and* (3.18)*:*

$$MS\left(\hat{s}^{(h)}_{\hat{m}}, \rho_\alpha\right) =: \mathcal{A}^\alpha_{\mathcal{M}^{(h)}},$$

*with*

$$\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}} = \left\{ s = \alpha_{00}\phi_{00} + \sum_{j\geq 0}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \right.$$
$$\left. \sup_{J\geq 0} 2^{2J\alpha} \sum_{j\geq J} \sum_{k\geq \lfloor 2^J (j-J+1)^{-\theta} \rfloor} |\beta_j|^2_{(k)} < \infty \right\},$$

*where* $(|\beta_j|_{(k)})_k$ *is the reordered sequence of coefficients* $(\beta_{jk})_k$*:*

$$|\beta_j|_{(1)} \geq |\beta_j|_{(2)} \cdots |\beta_j|_{(k)} \geq \cdots \geq |\beta_j|_{(2^j)}.$$

Remark that, as in Section 3.4.4, as soon as $\lambda_n \geq \lambda_0$ with $\lambda_0$ large enough, Assumption (3.18) holds.

This large set cannot be characterized in terms of classical spaces. Nevertheless it is undoubtedly a large functional space, since for every $\alpha > 0$ and every $p \geq 1$ satisfying $p > 2/(2\alpha + 1)$ we get

$$\mathcal{B}^{\alpha}_{p,\infty} \quad \subsetneq \quad \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}. \tag{3.21}$$

This new procedure is not computable since one needs an infinite number of wavelet coefficients to perform it. The problem of calculability can be solved by introducing, as previously, a maximum scale $j_0(n)$ as defined in (3.20). We consider the class of collection models $(\mathcal{M}^{(h)}_n)_n$ defined as follows:

$$\mathcal{M}^{(h)}_n = \{m = \text{span}\{\phi_{00}, \psi_{jk} : \ (j,k) \in \mathcal{I}_m, j < j_0(n)\} : \\ J \in \mathbb{N},\, \mathcal{I}_m \in \mathcal{P}_J(\mathcal{I})\}.$$

This model collection does not satisfy the embedding condition $\mathcal{M}^{(h)}_n \subset \mathcal{M}^{(h)}_{n+1}$. Nevertheless, we can use Proposition 1 with

$$\widetilde{\text{pen}}_n(m) = \frac{\lambda_n}{n} D_J$$

if $m$ is obtained from an index subset $\mathcal{I}_m$ in $\mathcal{P}_J(\mathcal{I})$. This slight over-penalization leads to the following result.

**Proposition 6.** *Let* $\alpha_0 < r$ *be fixed, let* $0 < \alpha \leq \alpha_0$ *and let* $\hat{s}^{(ht)}_{\tilde{m}}$ *be the model selection estimator associated with the model collection* $\mathcal{M}^{(h)}_n$*. Then, under Assumptions* (3.14)*,* (3.15)*,* (3.17) *and* (3.18)*:*

$$MS\left(\hat{s}^{(ht)}_{\tilde{m}}, \rho_\alpha\right) =: \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}.$$

Modifying Massart's strategy in order to obtain a practical estimator changes the maxiset performance. The previous set $\mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}$ is intersected with the strong Besov space $\mathcal{B}^{\alpha/(1+2\alpha)}_{2,\infty}$. Nevertheless, the maxiset $MS\left(\hat{s}^{(ht)}_{\tilde{m}}, \rho_\alpha\right)$ is still a large functional space. Indeed, for every $\alpha > 0$ and every $p$ satisfying $p \geq \max(1, 2\left(\frac{1}{1+2\alpha} + 2\alpha\right)^{-1})$

$$\mathcal{B}^{\alpha}_{p,\infty} \quad \subseteq \quad \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty} \cap \mathcal{A}^{\alpha}_{\mathcal{M}^{(h)}}. \tag{3.22}$$

**Comparisons of model selection estimators**

In this paragraph, we compare the maxiset performances of the different model selection procedures described previously. For a chosen rate of convergence let us recall that the larger the maxiset, the better the estimator. To begin, we propose to focus on the model selection estimators which are tractable from the computational point of view. Gathering Propositions 3, 4 and 6 we obtain the following comparison.

**Proposition 7.** *Let* $0 < \alpha < r$.

- *If for every $n$, $\lambda_n = \lambda_0 \log(n)$ with $\lambda_0$ large enough, then*

$$MS(\hat{s}_{\hat{m}}^{(st)}, \rho_\alpha) \subsetneq MS(\hat{s}_{\tilde{m}}^{(ht)}, \rho_\alpha) \subsetneq MS(\hat{s}_{\hat{m}}^{(l)}, \rho_\alpha). \tag{3.23}$$

- *If for every $n$, $\lambda_n = \lambda_0$ with $\lambda_0$ large enough, then*

$$MS(\hat{s}_{\hat{m}}^{(st)}, \rho_\alpha) \subsetneq MS(\hat{s}_{\tilde{m}}^{(ht)}, \rho_\alpha). \tag{3.24}$$

It means the followings.

- If for every $n$, $\lambda_n = \lambda_0 \log(n)$ with $\lambda_0$ large enough, then, according to the maxiset point of view, the estimator $\hat{s}_{\hat{m}}^{(l)}$ strictly outperforms the estimator $\hat{s}_{\tilde{m}}^{(ht)}$ which strictly outperforms the estimator $\hat{s}_{\hat{m}}^{(st)}$.

- If for every $n$, $\lambda_n = \lambda_0$ or $\lambda_n = \lambda_0 \log(n)$ with $\lambda_0$ large enough, then, according to the maxiset point of view, the estimator $\hat{s}_{\tilde{m}}^{(ht)}$ strictly outperforms the estimator $\hat{s}_{\hat{m}}^{(st)}$.

The hard thresholding estimator $\hat{s}_{\hat{m}}^{(l)}$ appears as the best estimator when $\lambda_n$ grows logarithmically while estimator $\hat{s}_{\tilde{m}}^{(ht)}$ is the best estimator when $\lambda_n$ is constant. In both cases, those estimators perform very well since their maxiset contains all the Besov spaces $\mathcal{B}_{p,\infty}^{\frac{\alpha}{1+2\alpha}}$ with $p \geq \max\left(1, \left(\frac{1}{1+2\alpha} + 2\alpha\right)^{-1}\right)$.

We forget now the calculability issues and consider the maxiset of the original procedure proposed by Massart. Propositions 4, 5 and 6 lead then to the following result.

**Proposition 8.** *Let $0 < \alpha < r$.*

- *If for any $n$, $\lambda_n = \lambda_0 \log(n)$ with $\lambda_0$ large enough then*

$$MS(\hat{s}_{\hat{m}}^{(h)}, \rho_\alpha) \not\subset MS(\hat{s}_{\hat{m}}^{(l)}, \rho_\alpha) \quad \textit{and} \quad MS(\hat{s}_{\hat{m}}^{(l)}, \rho_\alpha) \not\subset MS(\hat{s}_{\hat{m}}^{(h)}, \rho_\alpha). \tag{3.25}$$

- *If for any $n$, $\lambda_n = \lambda_0$ or $\lambda_n = \lambda_0 \log(n)$ with $\lambda_0$ large enough then*

$$MS(\hat{s}_{\tilde{m}}^{(ht)}, \rho_\alpha) \subsetneq MS(\hat{s}_{\hat{m}}^{(h)}, \rho_\alpha). \tag{3.26}$$

Hence, within the maxiset framework, the estimator $\hat{s}_{\hat{m}}^{(h)}$ strictly outperforms the estimator $\hat{s}_{\tilde{m}}^{(ht)}$ while the estimators $\hat{s}_{\hat{m}}^{(h)}$ and $\hat{s}_{\hat{m}}^{(l)}$ are not comparable. Note that we did not consider the maxisets of the estimator $\hat{s}_{\hat{m}}^{(s)}$ in this section as they are identical to the ones of the tractable estimator $\hat{s}_{\hat{m}}^{(st)}$. We summarize all those embeddings in Figure 3.4 and Figure 3.5: Figure 3.4 represents these maxiset embeddings for the choice $\lambda_n = \lambda_0 \log(n)$, while Figure 3.5 represents these maxiset embeddings for the choice $\lambda_n = \lambda_0$.

## 3.5 Inverse problem, needlet and thresholding

With D. Picard and G. Kerkyacharian [*Art-Ke+10*; *Art-KLPP12*], I have worked on a natural extension of the previous white noise model to inverse problem. Let $A$ be a compact linear operator from one Hilbert space $\mathcal{H}$ to another $\mathfrak{H}$, we assume now we observe

$$dY_x = As_0(x)dx + \sigma dW_x$$

and want to estimate $s_0$ from $Y$.

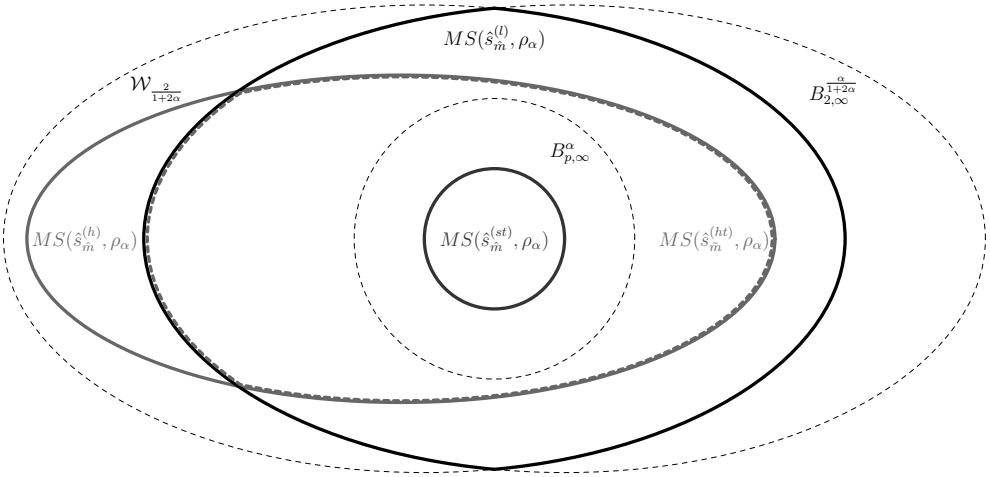



Figure 3.4: Maxiset embeddings when $\lambda_n = \lambda_0 \log(n)$ and $\max(1, 2\left(\frac{1}{1+2\alpha} + 2\alpha\right)^{-1}) \leq p \leq 2$.

### 3.5.1 Inverse problem, SVD and needlets

#### SVD and smoothed inversion

The key property of such a compact linear operator is the existence of the Singular Value Decomposition (SVD), i.e. the existence of two orthonormal bases $\{b_n\}_{n\in\mathbb{N}}$ and $\{\mathfrak{b}_n\}_{n\in\mathbb{N}}$ of respectively $\mathcal{H}$ and Im$A$ and of a sequence $\{\mu_n^2\}_{n\in\mathbb{N}}$ of decreasing positive number vanishing to 0 such that

$$Ab_n = \mu_n\, \mathfrak{b}_n \qquad \text{and} \qquad A^*\mathfrak{b}_n = \mu_n\, b_n$$

where $A^*$ denotes the adjoint of $A$. Indeed, thanks to this representation, the white noise model becomes a much simpler sequential model: it is equivalent to the observation of

$$\begin{aligned} Y_{\mathfrak{b}_n} &= \langle As_0, \mathfrak{b}_n\rangle + \sigma W_{\mathfrak{b}_n} \\ &= \mu_n \langle s_0, b_n\rangle + \sigma W_{\mathfrak{b}_n} \end{aligned}$$

where $(W_{\mathfrak{b}_n})$ is nothing but an i.i.d. sequence of standard Gaussian variables. As

$$s_0 = \sum_{n\in\mathbb{N}} \langle s_0, b_n\rangle\, b_n,$$

this paves the way for SVD based estimator of type

$$\hat{s} = \sum_{n\in\mathbb{N}} \theta_n \left(\frac{Y_{\mathfrak{b}_n}}{\mu_n}\right) b_n$$

where the $\theta_n$ are functions that may depends on the whole observation. As described by Cavalier [Ca11], the most classical choice for $\theta_n$ is a simple multiplication by a factor $\gamma_n$ ($\theta_n(y) = \gamma_n y$). Classical choices for $\gamma_n$ range from a simple cut-off ($\gamma_n = \mathbf{1}_{\{n\leq N\}}$ with $N$ to be defined) to implicit definition using iterative scheme through subtle Pinsker weighting scheme. Several procedures (including model selection procedures) have been proposed to select automatically the parameters of those methods yielding efficient adaptive estimation procedure. This SVD method is very attractive theoretically and can be shown to be asymptotically optimal in many situations (see Dicken and Maass [DM96], Mathé and Pereverzev [MP03] together with their nonlinear counterparts Cavalier and Tsybakov [CT02], Cavalier, Golubev, Picard, and Tsybakov [Ca+02], Tsybakov [Ts00], Goldenshluger and Pereverzev [GP03], Efromovich and Koltchinskii [EK01]). It also has the big advantage of performing a quick and stable inversion of

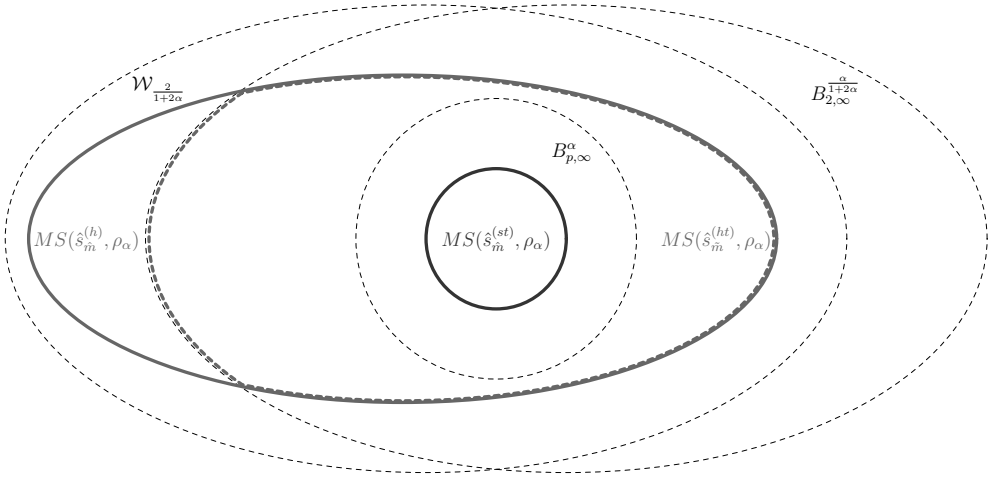



Figure 3.5: Maxiset embeddings when $\lambda_n = \lambda_0$ and $\max(1, 2\left(\frac{1}{1+2\alpha} + 2\alpha\right)^{-1}) \leq p \leq 2$.

the operator $A$. As explained by Loubes and Rivoirard [LR09], when considering the quadratic loss, regularity spaces associated to these methods, whether in the minimax approach or the maxiset one, are of Sobolev scales type, i.e. of type

$$\left\{ s | \sum_{n\in\mathbb{N}} \beta_n \left| \langle s, b_n \rangle \right|^2 < +\infty \right\}$$

where $b_n$ is the basis associated to the SVD decomposition and $\beta_n$ are some increasing weights specifying the space. Unfortunately, those spaces are not necessarily adapted to the function $s_0$ of interest. For instance, if one consider a (periodic) image deconvolution problem, the SVD basis is the usual Fourier basis and those spaces are classical Sobolev spaces, which are known to be not well suited for natural images. Along the same line, the Sobolev space are intimately related to the quadratic loss and are not adapted to other losses.

**Generic needlet construction**

As discovered by Petrushev and Xu [PX05] and generalized by Coulhon, Kerkyacharian, and Petrushev [CKP12], a much better well localized representation, the needlet representation, can often be constructed from the SVD basis. We refer to Coulhon, Kerkyacharian, and Petrushev [CKP12] for the general construction or to Kerkyacharian, Kyriazis, Le Pennec, Petrushev, and Picard [*Art-Ke+10*]; Kerkyacharian, Petrushev, Picard, and Willer [Ke+07]; Narcowich, Petrushev, and Ward [NPW06b]; Petrushev and Xu [PX05; PX08] for some specific ones. For sake of completeness, we present here a rough sketch of the construction. With a slight change of notation, we let now $(b_{n,k})_{(n,k)}$ be a SVD basis such that $(b_{n,k})$ is a basis of the eigenspace of $A^*A$ associated to $\mu_n^2$. For any positive function $a$ bounded by 1, equal to 1 on $[-1/2, 1/2]$ and equal to 0 outside of $[-1, 1]$, one can define a smoothed projector:

$$P_{a,N} s = \sum_{n\in\mathbb{N}} a\left(\frac{n}{N}\right) \sum_k \langle s, b_{n,k} \rangle b_{n,k}.$$

As soon as $a$ is regular, one can often obtained that the associated kernel

$$A_{a,N}(x, x') = \sum_{n\in\mathbb{N}} a\left(\frac{n}{N}\right) \sum_k b_{n,k}(x) b_{n,k}(x'),$$

which is always well localized spectrally, is well localized spatially. Capitalizing on this result, one obtains that those smoothed projectors are continuous for all $L_p$ norm. This exact construction has

been proposed Petrushev and Xu [PX05] when the SVD basis is a Jacobi polynomial basis but can be traced back to the work of Gottlieb [GS97] in the Fourier case. In inversion method, this amounts to choose the multipliers $\gamma_{n,k}$ as $a(n/N)$. A better insight on this method is obtained with the *needlets* constructed by Petrushev and Xu [PX05]. Let

$$B_{a,N}(x,x') = \sum_{n\in\mathbb{N}} \sqrt{a\left(\frac{n}{N}\right)} \sum_k b_{n,k}(x) b_{n,k}(x')$$

a straightforward computation shows that

$$\langle B_{a,N}(x,x''), B_{a,N}(x'',x')\rangle = A_{a,N}(x,x').$$

Assume now that there is a cubature scheme $(\xi,\omega_\xi)_{\xi\in\Xi_N}$ associated to the SVD basis $(b_{n,k})$ with $n \le N$ then the previous equality can be rewritten as a sum

$$\sum_{\xi\in\Xi_N} \omega_\xi B_{a,N}(x,\xi) B_{a,N}(\xi,x') = A_{a,N}(x,x').$$

If the $\omega_\xi$ are positive, one can define the father needlets

$$\phi_{N,\xi}(x) = \sqrt{\omega_\xi} B_{a,N}(x,\xi) = \sqrt{\omega_\xi} \sum_{n\in\mathbb{N}} \sqrt{a\left(\frac{n}{N}\right)} \sum_k b_{n,k}(x) b_{n,k}(\xi)$$

which are such that

$$P_{a,N} s = \sum_{\xi\in\Xi_N} \langle s, \phi_{N,\xi}\rangle \phi_{N,\xi}.$$

A multiscale representation can be deduced from this one by letting $d(x) = a(x/2) - a(x)$ and defining the needlets

$$\psi_{N,\xi}(x) = \sqrt{\omega_\xi} \sum_{n\in\mathbb{N}} \sqrt{d\left(\frac{n}{N}\right)} \sum_k b_{n,k}(x) b_{n,k}(\xi).$$

One easily verifies then that

$$\{\phi_{1,\xi}\}_{\xi\in\Xi_1} \cup \bigcup_{j\ge 0} \left\{\psi_{2^j,\xi}\right\}_{\xi\in\Xi_{2^j}}$$

is a tight frame. Under some regularity assumptions on the cubature, those well localized spectrally functions are well localized spatially around cubature points $\xi$ with a support of size of order $2^{-j}$. This frame is such that one has a reconstruction formula

$$s = \sum_{\xi\in\Xi_1} \langle s, \phi_{1,\xi}\rangle \, \phi_{1,\xi} + \sum_{j\ge 0} \sum_{\xi\in\Xi_{2^j}} \langle s, \psi_{2^j,\xi}\rangle \, \psi_{2^j,\xi}.$$

Furthermore, thanks to the good localization properties, $L^p$ spaces and more generally Besov spaces can be characterized in term of needlet coefficients. Finally as

$$\begin{aligned}\langle s, \psi_{2^j,\xi}\rangle &= \sqrt{\omega_\xi} \sum_n \sqrt{d\left(\frac{n}{2^j}\right)} \sum_k \langle s, b_{n,k}\rangle \\ &= \sqrt{\omega_\xi} \sum_n \frac{\sqrt{d\left(\frac{n}{2^j}\right)}}{\mu_n} \sum_k \langle As, \mathfrak{b}_{n,k}\rangle,\end{aligned}$$

$\langle s_0, \psi_{2^j,\xi}\rangle$ is naturally estimated (without bias) by

$$\sqrt{\omega_\xi}\sum_n \frac{\sqrt{d\left(\frac{n}{2^j}\right)}}{\mu_n}\sum_k Y_{\mathfrak{b}_{n,k}}$$

leading to efficient estimators. Finally, one should stress that the classical wavelets share these good localization properties and, as stressed already by Donoho, Johnstone, Kerkyacharian, and Picard [Do+96], one can prove that the thresholding procedure of the previous sections is also efficient for all $L^p$ losses.

We examplify this construction with two instances of the Radon transform: the fan beam Radon transform on the sphere [*Art-Ke+10*; *Art-KLPP12*] and the Radon transform of axially symmetric objects [*Unpub-BLPT12*]. In both examples, in 2D, we will try to estimate a function from its integral along lines.

### 3.5.2 A needlet based inversion for Radon transform

With G. Kerkyacharian and D. Picard, we have focused our analysis on this particular inverse problems.

#### SVD of Radon transform

We recall the definition and some basic facts about the Radon transform (cf. Helgason [He99], Natterer [Na01], Logan and Shepp [LS75]). Denote by $B^d$ the unit ball in $\mathbb{R}^d$ and by $\mathbb{S}^{d-1}$ the unit sphere in $\mathbb{R}^d$. The Lebesgue measure on $B^d$ will be denoted by $dx$ and the usual surface measure on $\mathbb{S}^{d-1}$ by $d\sigma(x)$

The Radon transform of a function $s$ is defined by

$$Rs(\theta,t) = \int_{\substack{y\in\theta^\perp \\ t\theta+y\in B^d}} s(t\theta+y)dy, \quad \theta\in\mathbb{S}^{d-1},\ t\in[-1,1],$$

where $dy$ is the Lebesgue measure of dimension $d-1$ and $\theta^\perp = \{x\in\mathbb{R}^d : \langle x,\theta\rangle = 0\}$. It is easy to see (cf. e.g. Natterer [Na01]) that the Radon transform is a bounded linear operator mapping $\mathbb{L}^2(B^d, dx)$ into $\mathbb{L}^2\left(\mathbb{S}^{d-1}\times[-1,1], d\mu(\theta,t)\right)$, where

$$d\mu(\theta,t) = d\sigma(\theta)\frac{dt}{(1-t^2)^{(d-1)/2}}.$$

The SVD of the Radon transform was first established by Cormack [Co64]; Davison [Da81]; Louis [Lo84]. In this regard we also refer the reader to Natterer [Na01]; Xu [Xu07].

The Radon SVD bases are defined in terms of Jacobi and Gegenbauer polynomials. The Jacobi polynomials $P_n^{(\alpha,\beta)}$, $n\geq 0$, constitute an orthogonal basis for the space $\mathbb{L}^2([-1,1], w_{\alpha,\beta}(t)dt)$ with weight $w_{\alpha,\beta}(t) = (1-t)^\alpha(1+t)^\beta$, $\alpha,\beta > -1$. They are standardly normalized by $P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$ and then, following Andrew, Askey, and Roy [AAR06]; Erdélyi, Magnus, Oberhettinger, and Tricomi [Er+81]; Szegö [Sz75],

$$\int_{-1}^1 P_n^{(\alpha,\beta)}(t)P_m^{(\alpha,\beta)}(t)w_{\alpha,\beta}(t)dt = \delta_{n,m}h_n^{(\alpha,\beta)},$$

where

$$h_n^{(\alpha,\beta)} = \frac{2^{\alpha+\beta+1}}{(2n+\alpha+\beta+1)}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)}. \tag{3.27}$$

The Gegenbauer polynomials $C_n^\lambda$ are a particular case of Jacobi polynomials, traditionally defined by

$$C_n^\lambda(t) = \frac{(2\lambda)_n}{(\lambda+1/2)_n}P_n^{(\lambda-1/2,\,\lambda-1/2)}(t), \quad \lambda > -1/2,$$

where $(a)_n = a(a+1)\dots(a+n-1) = \frac{\Gamma(a+n)}{\Gamma(a)}$

Let $\Pi_n(\mathbb{R}^d)$ be the space of all polynomials in $d$ variables of degree $\leq n$. We denote by $\mathcal{P}_n(\mathbb{R}^d)$ the space of all homogeneous polynomials of degree $n$ and by $\mathcal{V}_n(\mathbb{R}^d)$ the space of all polynomials of degree $n$ which are orthogonal to lower degree polynomials with respect to the Lebesgue measure on $B^d$. $\mathcal{V}_0$ is the set of constants. We have the following orthogonal decomposition:

$$\Pi_n(\mathbb{R}^d) = \bigoplus_{k=0}^{n} \mathcal{V}_k(\mathbb{R}^d).$$

Also, denote by $\mathbb{H}_n(\mathbb{R}^d)$ the subspace of all harmonic homogeneous polynomials of degree $n$ and by $\mathbb{H}_n(\mathbb{S}^{d-1})$ the restriction of the polynomials from $\mathbb{H}_n(\mathbb{R}^d)$ to $\mathbb{S}^{d-1}$. Let $\Pi_n(\mathbb{S}^{d-1})$ be the space of restrictions to $\mathbb{S}^{d-1}$ of polynomials of degree $\leq n$ on $\mathbb{R}^d$. As is well known

$$\Pi_n(\mathbb{S}^{d-1}) = \bigoplus_{m=0}^{n} \mathbb{H}_m(\mathbb{S}^{d-1})$$

(the orthogonality is with respect of the surface measure $d\sigma$ on $\mathbb{S}^{d-1}$).

Let $\boldsymbol{Y}_{l,i}$, $1 \leq i \leq N_{d-1}(l)$, be an orthonormal basis of $\mathbb{H}_l(\mathbb{S}^{d-1})$, i.e.

$$\int_{\mathbb{S}^{d-1}} \boldsymbol{Y}_{l,i}(\xi)\overline{\boldsymbol{Y}_{l,i'}(\xi)}d\sigma(\xi) = \delta_{i,i'}.$$

Then the natural extensions of $\boldsymbol{Y}_{l,i}$ on $B^d$ are defined by $\boldsymbol{Y}_{l,i}(x) = |x|^l \boldsymbol{Y}_{l,i}\big(\frac{x}{|x|}\big)$ and satisfy

$$\int_{B^d} \boldsymbol{Y}_{l,i}(x)\overline{\boldsymbol{Y}_{l,i'}(x)}dx = \delta_{i,i'}\frac{1}{2l+d}.$$

Assume that $\{\boldsymbol{Y}_{l,i} : 1 \leq i \leq N_{d-1}(l)\}$ is an orthonormal basis for $\mathbb{H}_l(\mathbb{S}^{d-1})$. The SVD basis of the Radon operator is given by

$$b_{k,l,i}(x) = (2k+d)^{1/2} P_j^{(0,\, l+d/2-1)}(2|x|^2 - 1)\boldsymbol{Y}_{l,i}(x), \;\; 0 \leq l \leq k, \; k-l = 2j, \; 1 \leq i \leq N_{d-1}(l),$$

as the orthonormal basis of $\mathcal{V}_k(B^d)$,

$$\mathfrak{b}_{k,l,i}(\theta,t) = [h_k^{(d/2)}]^{-1/2}(1-t^2)^{(d-1)/2} C_k^{d/2}(t)\boldsymbol{Y}_{l,i}(\theta), \quad k \geq 0, \; l \geq 0, \; 1 \leq i \leq N_{d-1}(l),$$

as the orthonormal basis of $\mathbb{L}^2(\mathbb{S}^{d-1} \times [-1,1], d\mu(\theta,s))$, while the eigenvalues

$$\mu_k^2 = \frac{2^d \pi^{d-1}}{(k+1)(k+2)\ldots(k+d-1)} = \frac{2^d \pi^{d-1}}{(k+1)_{d-1}} \sim k^{-d+1}. \tag{3.28}$$

For more details we refer the reader to Dunkl and Xu [DX01], Natterer [Na01] and Xu [Xu07].

**Needlet and Besov spaces**

Following the generic construction, one defines the smoothed projection

$$P_{2^j} s = \sum_k a\left(k/2^j\right) \sum_{l,i} \langle s, b_{k,l,i}\rangle b_{k,l,i}$$

and, for an existing collection of cubatures (see [*Art-Ke+10*] for the details), the father needlets and the needlets

$$\phi_{2^j,\xi} = \sqrt{\omega_\xi} \sum_k \sqrt{a\left(k/2^j\right)} \sum_{l,i} b_{k,l,i}$$

$$\psi_{2^j,\xi} = \sqrt{\omega_\xi} \sum_k \sqrt{d\left(k/2^j\right)} \sum_{l,i} b_{k,l,i}.$$

Thanks to Petrushev and Xu [PX08], one knows that

$$|\phi_{2^j,\xi}(x)|, |\psi_{2^j,\xi}(x)| \leq C_M \frac{2^{jd/2}}{\sqrt{W_j(\xi)}(1+2^j d(x,\xi))^M} \quad \forall M > 0$$

where $W_j(x) = 2^{-j} + \sqrt{1-|x|^2}$, $|x|^2 = |x|_d^2 = \sum_{i=1}^d x_i^2$, and

$$d(x,y) = \mathrm{Arccos}(\langle x, y\rangle + \sqrt{1-|x|^2}\sqrt{1-|y|^2}).$$

Nontrivial lower bounds for the norms of the needlets can be deduced. More precisely, Kyriazis, Petrushev, and Xu [KPX08] show that for $0 < p \leq \infty$

$$\|\psi_{2^j,\xi}\|_p \sim \|\phi_{2^j,\xi}\|_p \sim \Big(\frac{2^{jd}}{W_j(\xi)}\Big)^{1/2-1/p}, \quad \xi \in \Xi_{2^j}.$$

In order to introduce the Besov spaces of positive smoothness on the ball as spaces of $L^p$-approximation from algebraic polynomial, using the notations from Kyriazis, Petrushev, and Xu [KPX08], we will denote by $E_n(s,p)$ the best $L^p$-approximation of $s \in \mathbb{L}^p(B^d)$ from $\Pi_n$.

**Definition 3.** *Let $0 < t < \infty$, $1 \leq p \leq \infty$, and $0 < q \leq \infty$. The space $B_{p,q}^{t,0}$ on the ball is defined as the space of all functions $s \in \mathbb{L}^p(B^d)$ such that*

$$|s|_{B_{p,q}^{t,0}} = \Big(\sum_{n\geq 1}(n^t E_n(s,p))^q \frac{1}{n}\Big)^{1/q} < \infty \quad \text{if } q < \infty,$$

*and $|s|_{B_{p,q}^{t,0}} = \sup_{n\geq 1} n^t E_n(f,p) < \infty$ if $q = \infty$. The norm on $B_{p,q}^{s,0}$ is defined by*

$$\|s\|_{B_{p,q}^{t,0}} = \|s\|_p + |s|_{B_{p,q}^{t,0}}.$$

From the monotonicity of $\{E_n(s,p)\}$ it readily follows that

$$\|f\|_{B_{p,q}^{t,0}} \sim \|f\|_p + \Big(\sum_{j\geq 0}(2^{jt}E_{2^j}(s,p))^q\Big)^{1/q}$$

with the obvious modification when $q = \infty$. There are several different equivalent norms on the Besov space $B_{p,q}^{s,0}$.

**Theorem 7.** *With indexes $t, p, q$ as in the above definition the following norms are equivalent to the Besov norm $\|s\|_{B_{p,q}^{t,0}}$:*

$$\begin{aligned}
&(i) \quad \mathcal{N}_1(s) = \|s\|_p + \|(2^{jt}\|P_{2^j}s\|_p)_{j\geq 0}\|_{l^q},\\
&(ii) \quad \mathcal{N}_2(s) = \|s\|_p + \|(2^{jt}\|P_{2^{j+1}}s - P_{2^j}\|_p)_{j\geq 1}\|_{l^q},\\
&(iii) \quad \mathcal{N}_3(s) = \|s\|_p + \|(2^{jt}\sum_{\xi\in\chi_j}|\langle s,\psi_{j,\xi}\rangle|^p\|\psi_{j,\xi}\|_p^p)_{j\geq -1}\|_{l^q}.
\end{aligned}$$

Note that this space is not the classical Besov space $B_{p,q}^t$ but a slightly larger one.

**Linear needlet estimator**

We consider first linear estimator related to the smoothed projection operator. Namely, we define our estimator by

$$\hat{s}_J = \sum_k \frac{a\left(\frac{k}{2^J}\right)}{\mu_k}\sum_{l,i} Y_{\mathfrak{b}_{k,l,i}} b_{k,l,i}$$

or equivalently by

$$\hat{s}_J = \sum_{\xi \in \Xi_{2^J}} \widehat{\alpha_{2^J,\xi}} \phi_{2^J,\xi}$$

with

$$\begin{aligned}\widehat{\alpha_{2^J,\xi}} &= \sqrt{\omega_\xi} \sum_k \frac{\sqrt{a\left(\frac{k}{2^J}\right)}}{\mu_k} \sum_{l,i} Y_{\mathfrak{b}_{k,l,i}} \\ &= \sum_k \frac{1}{\mu_k} \sum_{l,i} \langle \phi_{2^J,\xi}, b_{k,l,i} \rangle Y_{\mathfrak{b}_{k,l,i}}.\end{aligned}$$

Remark that the needlets are not necessary to define the estimator here, they are nevertheless a very valuable tool in its analysis.

We obtain:

**Theorem 8.** *Let $1 \leq p \leq \infty, 0 < t < \infty$, and assume that $s_0 \in B^{t,0}_{p,\infty}$ with $\|s_0\|_{B^{t,0}_{p,\infty}} \leq M$. Let $\widehat{s}_J$ be the needlet estimator defined above, assume $J$ is selected depending on the parameters as described below*

1. *If $M 2^{-J(t+d)} \sim \sigma$ when $p = \infty$, then*

$$\mathbb{E}\left[\|s_0 - \widehat{s_J}\|_\infty\right] \leq c_\infty M^{\frac{d}{t+d}} \sigma^{\frac{t}{t+d}} \sqrt{\log M/\sigma}.$$

2. *If $M 2^{-Jt} \sim \sigma 2^{J(d-2/p)}$ when $4 \leq p < \infty$, then*

$$\mathbb{E}\left[\|s_0 - \widehat{s_J}\|_p^p\right] \leq c_p M^{\frac{(d-2/p)p}{t+d-2/p}} \sigma^{\frac{tp}{t+d-2/p}},$$

   *where when $p = 4$ there is an additional factor $\log(M/\sigma)$ on the right.*

3. *If $M 2^{-Js} \sim \sigma 2^{J(d-1/2)}$ when $1 \leq p < 4$, then*

$$\mathbb{E}\left[\|s_0 - \widehat{s_J}\|_p^p\right] \leq c_p M^{\frac{(d-1/2)p}{t+d-1/2}} \sigma^{\frac{tp}{t+d-1/2}}.$$

- As shown in the next section, the following rates of convergence are, in fact, minimax, i.e. there exist positive constants $c_1$ and $c_2$ such that

$$\sup_{\|s_0\|_{B^{t,0}_{p,\infty}} \leq M} \inf_{\tilde{s} \text{ estimator}} \mathbb{E}\left[\|s_0 - \tilde{s}\|_p^p\right] \geq c_1 \max\{\sigma^{\frac{tp}{t+d-2/p}},\ \sigma^{\frac{tp}{t+d-1/2}}\},$$

$$\sup_{\|s_0\|_{B^{t,0}_{\infty,\infty}} \leq M} \inf_{\tilde{s} \text{ estimator}} \mathbb{E}\left[\|s_0 - \tilde{s}\|_\infty\right] \geq c_2 \sigma^{\frac{s}{s+d}} \sqrt{\log 1/\sigma}.$$

- The case $p = 2$ above corresponds to the standard SVD method which involves Sobolev spaces. In this setting, minimax rates have already been established (cf. Cavalier, Golubev, Picard, and Tsybakov [Ca+02]; Cavalier and Tsybakov [CT02]; Dicken and Maass [DM96]; Efromovich and Koltchinskii [EK01]; Goldenshluger and Pereverzev [GP03]; Mathé and Pereverzev [MP03]; Tsybakov [Ts00]); these rates are $\sigma^{\frac{2t}{t+d-1/2}}$. Also, it has been shown that the SVD algorithms yield minimax rates. These results extend (using straightforward comparisons of norms) to $L^p$ losses for $p < 4$, but still considering the Sobolev ball $\{\|s\|_{B^{t,0}_{2,\infty}} \leq M\}$ rather than the Besov ball $\{\|s\|_{B^{t,0}_{p,\infty}} \leq M\}$. Therefore, our results can be viewed as an extension of the above results, allowing a much wider variety of regularity spaces.

- The Besov spaces involved in our bounds are in a sense well adapted to our method. However, one verify that the bounds from Theorem 8 hold in terms of the standard Besov spaces as well. This means that in using the Besov spaces described above, our results are but stronger.

- In the case $p \geq 4$ we exhibit here new minimax rates of convergence, related to the ill posedness coefficient of the inverse problem $\frac{d-1}{2}$ along with edge effects induced by the geometry of the ball. These rates have to be compared with similar phenomena occurring in other inverse problems involving Jacobi polynomials (e.g. Wicksell problem), see Kerkyacharian, Petrushev, Picard, and Willer [Ke+07].

**Needlet inversion of a noisy Radon transform and minimax performances**

We propose here an adaptive method based on a thresholding of needlet coefficients. Starting from the observation that

$$s_0 = \sum_{\xi \in \Xi_1} \langle s_0, \phi_{1,\xi} \rangle \, \phi_{1,\xi} + \sum_{j \geq 0} \sum_{\xi \in \Xi_{2^j}} \langle s_0, \psi_{2^j,\xi} \rangle \, \psi_{2^j,\xi}$$

and

$$\langle s_0, \psi_{2^j,\xi} \rangle = \sqrt{\omega_\xi} \sum_k \frac{\sqrt{d\left(\frac{k}{2^j}\right)}}{\mu_k} \sum_{l,i} \langle A s_0, \mathfrak{b}_{k,l,i} \rangle,$$

we let

$$\widehat{\alpha_{2^j,\xi}} = \sqrt{\omega_\xi} \sum_k \frac{\sqrt{a\left(\frac{k}{2^j}\right)}}{\mu_k} \sum_{l,i} Y_{\mathfrak{b}_{k,l,i}}$$

$$\widehat{\beta_{2^j,\xi}} = \sqrt{\omega_\xi} \sum_k \frac{\sqrt{d\left(\frac{k}{2^j}\right)}}{\mu_k} \sum_{l,i} Y_{\mathfrak{b}_{k,l,i}}$$

and define our estimator by

$$\hat{s} = \sum_{\xi \in \Xi_1} \widehat{\alpha_{1,\xi}} \, \phi_{1,\xi} + \sum_{0 \leq j \leq J_\sigma} \sum_{\xi \in \Xi_{2^j}} \rho_{T_{2^j,\xi}} \left( \widehat{\beta_{2^j,\xi}} \right) \psi_{2^j,\xi}.$$

where $\rho_T$ is the hard threshold function $\rho_T(\beta) = \beta \mathbf{1}_{\{|\beta| \geq T\}}$.

The tuning parameters of this estimator are

- The range $J_\sigma$ of resolution levels will be taken such that

$$2^{J_\sigma(d-\frac{1}{2})} \leq \left(\sigma \sqrt{\log 1/\sigma}\right)^{-1} < 2^{(J_\sigma+1)(d-\frac{1}{2})}.$$

- The thresholds $T_{2^j,\xi}$ that will be chosen in our theoretical analysis as

$$T_{2^j,\xi} = \kappa 2^{j\nu} c_\sigma$$

where

  - The threshold constant $\kappa$ is an important tuning of our method. The theoretical point of view asserts that, for $\kappa$ above a constant (for which our evaluation is probably not optimal), the minimax properties hold.
  - $c_\sigma$ is a constant depending on the noise level. We shall see that the following choice is appropriate:

$$c_\sigma = \sigma \sqrt{\log 1/\sigma}.$$

  - Notice that the threshold function for each coefficient contains $2^{j\nu}$. This is due to the inversion of the Radon operator and the concentration relative to the $b_{k,l,i}$'s of the needlets.

It is important to remark here that, unlike the (linear) procedures proposed in the previous section, this one does not require the knowledge of the regularity while, as will be seen in the sequel, it attains bounds that are as good as the linear ones and even better since they are handling much wider ranges for the parameters of the Besov spaces.

**Theorem 9.** *For $0 < r \leq \infty$, $\pi \geq 1$, $1 \leq p < \infty$, there exist some constant $c_p = c_p(t, \pi, r, M)$, $\kappa_0$ such that if $\kappa \geq \kappa_0$, $t > (d+1)(\frac{1}{\pi} - \frac{1}{p})_+$ and, in addition, if $\pi < p$, $t > \frac{d+1}{\pi} - \frac{1}{2}$:*

- *If $\frac{1}{p} < \frac{d}{d+1}$,*

$$\begin{aligned}\sup_{s_0 \in B^s_{\pi,r}(M)} &(\mathbb{E}\left[\|\hat{s} - s_0\|_p^p\right])^{\frac{1}{p}} \\ &\leq c_p (\log 1/\sigma)^{\frac{p}{2}} \\ &\quad \times \left(\sigma\sqrt{\log 1/\sigma}\right)^{\frac{t-(d+1)(1/\pi-1/p)}{t+d-(d+1)/\pi} \wedge \frac{t}{t+d-1/2} \wedge \frac{t-2(1/\pi-1/p)}{t+d-2/\pi}}.\end{aligned}$$

- *If $\frac{d}{d+1} \leq \frac{1}{p}$ and $d > 2$ or $p > 1$,*

$$\sup_{s_0 \in B^s_{\pi,r}(M)} (\mathbb{E}\left[\|\hat{s} - s_0\|_p^p\right])^{\frac{1}{p}} \leq c_p (\log 1/\sigma)^{\frac{p}{2}} \left(\sigma\sqrt{\log 1/\sigma}\right)^{\frac{t}{t+d-1/2} \wedge \frac{t-2(1/\pi-1/p)}{t+d-2/\pi}}.$$

- *If $d = 2$ and $p = 1$,*

$$\sup_{f \in B^s_{\pi,r}(M)} (\mathbb{E}\left[\|\hat{f} - f\|_1\right]) \leq c_1 (\log 1/\sigma)^{\frac{1}{2}} \left(\sigma\sqrt{\log 1/\sigma}\right)^{\frac{t}{t+2-1/2}}.$$

Up to logarithmic terms, the rates observed here are minimax, as will appear in the following theorem. It is known that in this kind of estimation, full adaptation yields unavoidable extra logarithmic terms. The rates of the logarithmic terms obtained in these theorems are, most of the time, suboptimal (for instance, for obvious reasons, the case $p = 2$ yields fewer logarithmic terms). A more detailed study could lead to optimized rates, which we decided not to include here for the sake of simplicity.

The cumbersome comparisons of the different rates of convergence are summarized in Figures 3.6 and 3.7 for the case $0 < \frac{1}{p} < \frac{d}{d+1}$. These figures illustrate and highlight the differences between the cases $p > 4$ and $p < 4$. We put $\frac{1}{p}$ as the horizontal axis and the regularity $t$ as the vertical axis. As explained later, after the lower-bound results, zones I and II correspond to two different types of the so called "dense" case, whereas zone III corresponds to the "sparse" case.

For the case of an $\mathbb{L}_\infty$ loss function, we have a slightly different result since the thresholding depends on the $\mathbb{L}_\infty$ norm of the local needlet. Let us consider the following estimate:

$$\begin{aligned}\hat{s}_\infty &= \sum_{j=-1}^{J_\sigma} \sum_{\xi \in \Xi_j} \rho_{\kappa 2^{jd} c_\sigma / \|\psi_{2^j,\xi}\|}\left(\widehat{\beta}_{2^j,\xi}\right) \psi_{j,\xi}, \\ 2^{J_\sigma d} &= \left(\sigma\sqrt{\log 1/\sigma}\right)^{-1}.\end{aligned}$$

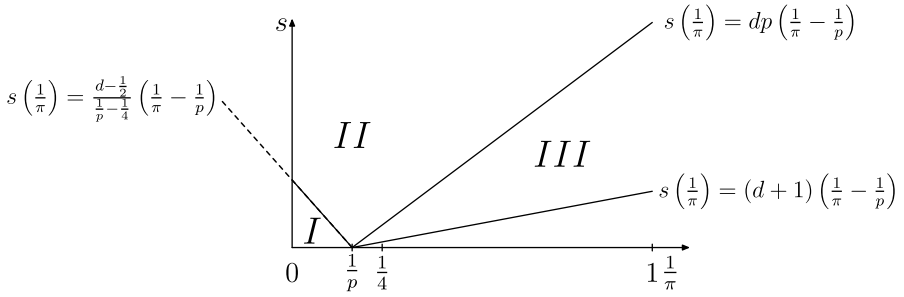
Then, for this estimate, we have the following results:

**Theorem 10.** *For $0 < r \leq \infty$, $\pi \geq 1$, $t > \frac{d+1}{\pi}$, there exist some constants $c_\infty = c_\infty(t, \pi, r, M)$ such that if $\kappa^2 \geq 4\tau_\infty$, where $\tau_\infty := \sup_{j,\xi} 2^{-j\frac{d+1}{2}} \|\psi_{j,\xi}\|_\infty$,*

$$\sup_{f \in B^t_{\pi,r}(M)} \mathbb{E}\|\hat{s}_\infty - s\|_\infty \leq c_\infty \left(\sigma\sqrt{\log 1/\sigma}\right)^{\frac{t-(d+1)/\pi}{t+d-(d+1)/\pi}}.$$

The following theorem states lower bounds for the minimax rates over Besov spaces in this model.

**Theorem 11.** *Let $\mathcal{E}$ be the set of all estimators, for $0 < r \leq \infty$, $\pi \geq 1$, $t > \frac{d+1}{\pi}$.*




$$I : \frac{s}{s+d-\frac{1}{2}} \qquad II : \frac{s-2\left(\frac{1}{\pi}-\frac{1}{p}\right)}{s+d-\frac{2}{\pi}} \qquad III : \frac{s-(d+1)\left(\frac{1}{\pi}-\frac{1}{p}\right)}{s+d-\frac{d+1}{\pi}}$$

Figure 3.6: The three different minimax rate type zones are shown with respect to the Besov space parameters $s$ and $\pi$ for a fixed loss norm $L^p$ with $0 < \frac{1}{p} < \frac{1}{4}$.

1. *There exists some constant* $C_\infty = C_\infty(t, \pi, r, M)$ *such that,*

$$\inf_{s^\star \in \mathcal{E}} \sup_{s_0 \in B^s_{\pi,r}(M)} \mathbb{E}\left[\|s^\star - s_0\|_\infty\right] \geq C_\infty \left(\sigma \sqrt{\log 1/\sigma}\right)^{\frac{t-(d+1)/\pi}{t+d-(d+1)/\pi}}.$$

2. *For* $1 \leq p < \infty$, *there exists some constant* $C_p = C_p(t, \pi, r, M)$ *such that if* $t > (\frac{d+1}{\pi} - \frac{d+1}{p})_+$,

   (a) *If* $\frac{1}{p} < \frac{d}{d+1}$

$$\inf_{s^\star \in \mathcal{E}} \sup_{s_0 \in B^t_{\pi,r}(M)} \left(\mathbb{E}\left[\|s^\star - s_0\|_p^p\right]\right)^{\frac{1}{p}} \geq C_p \sigma^{\frac{t-(d+1)(1/\pi-1/p)}{t+d-(d+1)/\pi} \wedge \frac{t}{t+d-1/2} \wedge \frac{t-2(1/\pi-1/p)}{t+d-2/\pi}}.$$

   (b) *If* $\frac{d}{d+1} \leq \frac{1}{p}$ *and* $d > 2$ *or* $p > 1$

$$\inf_{s^\star \in \mathcal{E}} \sup_{s_0 \in B^t_{\pi,r}(M)} \left(\mathbb{E}\left[\|s^\star - s_0\|_p^p\right]\right)^{\frac{1}{p}} \geq C_p \sigma^{\frac{t}{t+d-1/2} \wedge \frac{t-2(1/\pi-1/p)}{t+d-2/\pi}}.$$

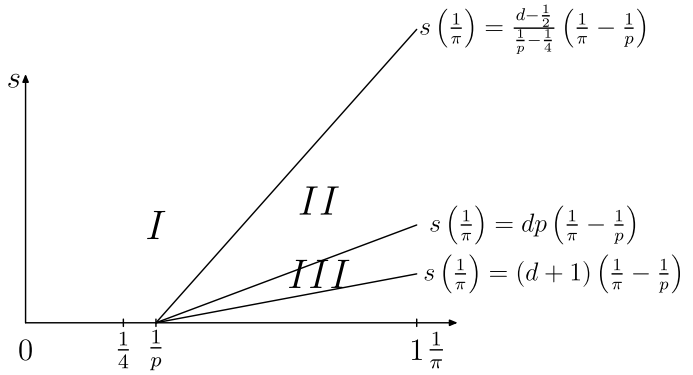
   (c) *If* $d = 2$ *and* $p = 1$

$$\inf_{s^\star \in \mathcal{E}} \sup_{s_0 \in B^t_{\pi,r}(M)} (\mathbb{E}\|s^\star - s_0\|_1) \geq C_p \sigma^{\frac{t}{t+2-1/2}}.$$

A careful look at the proof shows that the different rates observed in the two preceding theorems can be *explained* by geometrical considerations. In fact, depending on the cubature points around which they are centered, the needlets do not behave the same way. In particular, their $\mathbb{L}_p$ norms differ. This leads us to consider two different regions on the sphere, one near the pole and one closer to the equator. In these two regions, we considered dense and sparse cases in the usual way. This yielded four rates. Then it appeared that one of them (sparse) is always dominated by the others.

### 3.5.3 A needlet based inversion for Radon transform of axially symmetric objects

M. Bergounioux and E. Trélat [BT10] has been working on this special case of Radon transform thanks to a contract funded by the CEA. They asked me if I could provide them a numerical implementation of a wavelet based inversion in order to compare its performance on their model. Instead of this, we have worked on a new needlet based inversion method [*Unpub-BLPT12*].




$$I : \frac{s}{s+d-\frac{1}{2}} \qquad II : \frac{s-2\left(\frac{1}{\pi}-\frac{1}{p}\right)}{s+d-\frac{2}{\pi}} \qquad III : \frac{s-(d+1)\left(\frac{1}{\pi}-\frac{1}{p}\right)}{s+d-\frac{d+1}{\pi}}$$

Figure 3.7: The three different minimax rate type zones are shown with respect to the Besov space parameters $s$ and $\pi$ for a fixed loss norm $p$ with $\frac{1}{4} < \frac{1}{p} < \frac{d}{d+1}$.

**A SVD type decomposition**

Assume, we observe an axially symmetric object defined by its *cylindrical* density $u(r, z)$ by measuring integrals along line orthogonal to the revolution axis and thus characterized by its distance $y$ to the axis and its height $z$. This tomographic observation $A$ is given by

$$As(y, z) = 2 \int_{|y|}^{+\infty} s(r, z) \frac{r}{\sqrt{r^2 - y^2}} dr.$$

This operator does nothing along the $z$ axis so we can focus on

$$As(y) = 2 \int_{|y|}^{+\infty} s(r) \frac{r}{\sqrt{r^2 - y^2}} dr.$$

This transform is related to the *classical* Abel integral transform $T_{1/2}$ defined by

$$T_{1/2} s(x) = \int_0^x \frac{s(t)}{(x-t)^{1/2}} dt.$$

as soon as we assume that $s(r) = 0$ for $r > 1$. Indeed, we have then

$$As(y) = \int_0^{1-|y|^2} s(\sqrt{1-t}) \frac{1}{\sqrt{1-y^2-t}} dt$$

so that if we let $\overline{s} : t \mapsto s(\sqrt{1-t})$

$$= T_{1/2} \overline{s}(1 - y^2).$$

A key result obtained by Ammari and Karoui [AK10] is a SVD type decomposition of the $T_\alpha$ operator. More precisely, let $P_n^{\alpha,\beta}$ be the classical $n$th degree Jacobi polynomial on $[-1, 1]$ and

$$Q_n^{\alpha,\beta}(x) = \sqrt{\frac{n!(2n+\alpha+\beta+1)\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}} P_n^{\alpha,\beta}(2x-1)$$

a rescaled version so that the $Q_n^{\alpha,\beta}$s yield an orthogonal basis of $L^2([0,1],(1-x)^\alpha x^\beta)$. Ammari and Karoui [AK10] prove

$$T_{1/2}(Q_n^{0,0})(x) = \beta_{n,1/2} x^\alpha Q_n^{-1/2,1/2}(x)$$

with

$$\beta_{n,1/2} = \Gamma(1/2)\sqrt{\frac{\Gamma(n+1/2)}{\Gamma(n+3/2)}} \sim \frac{\Gamma(1/2)}{(n+3/2)^{1/2}}.$$

Translating their result in term of the tomography operator $A$ yields

$$A\left(Q_n^{0,0}(1-r^2)\right)(y) = \beta_{n,1/2}\sqrt{1-y^2}Q_n^{-1/2,1/2}(1-y^2).$$

As by construction, for any $(n,n') \in \mathbb{N}^2$,

$$\int_0^1 Q_n^{0,0}(1-r^2)Q_{n'}^{0,0}(1-r^2)2rdr = \delta_{n,n'} \text{ and } \int_0^1 \sqrt{1-y^2}Q_n^{-1/2,1/2}(1-y^2)Q_n'^{-1/2,1/2}(1-y^2)2dy = \delta_{n,n'},$$

we have obtained a pseudo SVD decomposition of $A$ as an operator from $L^2([0,1],2rdr)$ to $L^2([0,1],2dr)$ by letting $b_n(r) = Q_n^{0,0}(1-r^2)$, $\mathfrak{b}_n(y) = \sqrt{1-y^2}Q_n^{-1/2,1/2}(1-y^2)$ and $\mu_n = \beta_{n,1/2}$. This decomposition is not a SVD basis as $(\mathfrak{b}_n)_n$ is not an orthonormal basis. However, $(\mathfrak{b}_n)_n$ is a basis for which $\left(\widetilde{\mathfrak{b}}_n(y) = Q_n^{-1/2,1/2}(1-y^2)\right)_n$ is a dual basis. Nevertheless, a SVD type scheme based on the not independent anymore observations of $Y_{\mathfrak{b}_n}$ can be derived. Observe that the needlet construction can be applied nevertheless to the orthonormal family $b_n$ leading to the Legendre needlet studied by Petrushev and Xu [PX05] and Kerkyacharian, Petrushev, Picard, and Willer [Ke+07] up to the change of variable $r$ to $1-r^2$. A line by line needlet inversion scheme using either a smoothed projection or a thresholding can then be considered.

**2D scheme**

This 1D scheme can be easily extended into a 2D scheme. We introduce first the Legendre orthonormal basis $\{B_{n'} = R_{n'}^{0,0}(z)\}_{n'\in\mathbb{N}}$ on the support $[-Z,Z]$ of the vertical axis and, following Ivanov, Petrushev, and Xu [IPX12], consider a tensorial construction. By construction, we have a pseudo SVD decomposition for the 2D transform

$$As(y,z) = 2\int_{|y|}^{+\infty} s(r,z)\frac{r}{\sqrt{r^2-y^2}}dr.$$

Indeed, if we let $\boldsymbol{b}_{n,n'}(r,z) = b_n(r)B_{n'}(z)$, $\mathfrak{b}_{n,n'}(y,z) = \mathfrak{b}_n(y)B_{n'}(z)$ and $\widetilde{\mathfrak{b}}_{n,n'}(y,z) = \widetilde{\mathfrak{b}}_n(y)B_{n'}(z)$ then we have

$$\begin{aligned} A\boldsymbol{b}_{n,n'} &= \beta_{n,1/2}\mathfrak{b}_{n,n'} \\ \int_{[0,1]\times[-Z,Z]} \boldsymbol{b}_{n,n'}\boldsymbol{b}_{l,l'}2rdrdz &= \delta_{n,l}\delta_{n',l'} \\ \int_{[0,1]\times[-Z,Z]} \mathfrak{b}_{n,n'}\widetilde{\mathfrak{b}}_{l,l'}2dydz &= \delta_{n,l}\delta_{n',l'}. \end{aligned}$$

With a construction similar to the one use to obtain the 2D tensorial wavelets, we can construct a 2D tensorial needlet basis. Along the horizontal axis, we use the needlet $\phi_{2^j,\xi}/\psi_{2^j,\xi}$ with cubature sets $\Xi_{2^j}$, as described quickly in the previous section, while we use Legendre needlets $\Phi_{2^j,\xi'}/\Psi_{2^j,\xi'}$ with cubature sets $\Xi'_{2^j}$, along the vertical axis. We define then

$$\begin{aligned} &\boldsymbol{\Phi}_{2^j,\xi,\xi'}(r,z) = \phi_{2^j,\xi}(r)\Phi_{2^j,\xi'}(r), && \boldsymbol{\Psi}^{ad}_{2^j,\xi,\xi'}(r,z) = \phi_{2^j,\xi}(r)\Psi_{2^j,\xi'}(r), \\ &\boldsymbol{\Psi}^{da}_{2^j,\xi,\xi'}(r,z) = \psi_{2^j,\xi}(r)\Phi_{2^j,\xi'}(r) \quad \text{and} && \boldsymbol{\Psi}^{dd}_{2^j,\xi,\xi'}(r,z) = \psi_{2^j,\xi}(r)\Psi_{2^j,\xi'}(r) \end{aligned}$$

or equivalently

$$\boldsymbol{\Psi}^o_{2^j,\boldsymbol{\xi}}(r,z) = \sqrt{\omega_\xi}\sqrt{\omega_{\xi'}} \sum_{n\leq 2^{j+1}} \sum_{n\leq 2^{j+1}} \sqrt{a^o\left(\frac{n}{2^j},\frac{n'}{2^j}\right)} Q_n^{0,0}(1-\xi^2) R_{n'}^{0,0}(\xi') Q_n^{0,0}(1-r^2) R_{n'}^{0,0}(z)$$

with $a^{ad}(w,w') = a(w)d(w')$, $a^{da}(w,w') = d(w)a(w')$ and $a^{dd}(w,w') = d(w)d(w')$. Those functions can be proved to be well localized around $(\xi,\xi')$ with support of size of order $2^- j$ and linear combination of the first $2^{j+1}\times 2^{j+1}$ polynomial tensor products. They are such that, if we let $\boldsymbol{\xi} = (\xi,\xi')$ and $\boldsymbol{\Xi}_{2^j} = \Xi_{2^j}\times\Xi'_{2^j}$,

$$\begin{aligned} s_0 = &\sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}_1} \left(\int_{[0,1]\times[-Z,Z]} s_0(r,z)\boldsymbol{\Phi}_{1,\boldsymbol{\xi}}(r,z)2rdrdz\right)\boldsymbol{\Phi}_{1,\boldsymbol{\xi}} \\ &+ \sum_{j\geq 0}\sum_{o\in\{ad,da,dd\}}\sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}_{2^j}} \left(\int_{[0,1]\times[-Z,Z]} u(r,z)\boldsymbol{\Psi}^o_{2^j,\boldsymbol{\xi}}(r,z)2rdrdz\right)\boldsymbol{\Psi}^o_{2^j,\boldsymbol{\xi}} \end{aligned}$$

while

$$\begin{aligned} \boldsymbol{P}_{a,2^J} s_0 = &\sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}_1} \left(\int_{[0,1]\times[-Z,Z]} s_0(r,z)\boldsymbol{\Phi}_{1,\boldsymbol{\xi}}(r,z)2rdrdz\right)\boldsymbol{\Phi}_{1,\boldsymbol{\xi}} \\ &+ \sum_{j=0}^{J-1}\sum_{o\in\{ad,da,dd\}}\sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}_{2^j}} \left(\int_{[0,1]\times[-Z,Z]} s_0(r,z)\boldsymbol{\Psi}^o_{2^j,\boldsymbol{\xi}}(r,z)2rdrdz\right)\boldsymbol{\Psi}^o_{2^j,\boldsymbol{\xi}}. \end{aligned}$$

Again it remains to estimate the needlet coefficients We rely on the decomposition of the $\boldsymbol{\Psi}$ into the *polynomial* basis which yields

$$\begin{aligned} \boldsymbol{c}^o_{2^j,\boldsymbol{\xi}} &= \int_{[0,1]\times[-Z,Z]} u(r,z)\boldsymbol{\Psi}^o_{2^j,\boldsymbol{\xi}}(r,z)2rdrdz \\ &= \sqrt{\omega_\xi}\sqrt{\omega_{\xi'}} \sum_{n\leq 2^{j+1}}\sum_{n\leq 2^{j+1}} \sqrt{d^o\left(\frac{n}{2^j},\frac{n'}{2^j}\right)} Q_n^{0,0}(1-\xi^2) R_{n'}^{0,0}(\xi') \frac{\int_{[0,1]\times[-Z,Z]} H_0(y,z)\widetilde{\mathbf{b}_{\boldsymbol{m},\boldsymbol{m'}}}(y,z)2dydz}{\beta_{n,1/2}} \end{aligned}$$

so that those coefficients can be estimated by

$$\widehat{\boldsymbol{c}^o_{2^j,\boldsymbol{xi}}} = \sqrt{\omega_\xi}\sqrt{\omega_{\xi'}} \sum_{n\leq 2^{j+1}}\sum_{n\leq 2^{j+1}} \sqrt{d^o\left(\frac{n}{2^j},\frac{n'}{2^j}\right)} Q_n^{0,0}(1-\xi^2) R_{n'}^{0,0}(\xi') \frac{Y_{\widetilde{\mathbf{b}_{\boldsymbol{m},\boldsymbol{m'}}}}}{\beta_{n,1/2}}.$$

Combined with the thresholding strategy, we obtain an estimator

$$\widehat{s} = \sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}_1} \widehat{\boldsymbol{c}^{aa}_{2^j,\boldsymbol{xi}}}\boldsymbol{\Phi}_{1,\boldsymbol{\xi}} + \sum_{0\leq j\leq J_\sigma}\sum_{o\in\{ad,da,dd\}}\sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}_{2^j}} \rho_{T_{o,2^j,\boldsymbol{\xi}}}\left(\widehat{\boldsymbol{c}^o_{2^j,\boldsymbol{xi}}}\right)\boldsymbol{\Psi}^o_{2^j,\boldsymbol{\xi}}$$

As in the previous section, the parameters are the maximum level $J_\sigma$ and the thresholds $T_{o,2^k,\boldsymbol{\xi}}$. We propose to use here a threshold proportional to the standard deviation $\sigma_{o,2^j,\boldsymbol{\xi}}$ of the coefficients: $T = \kappa\sigma_{o,2^j,\boldsymbol{\xi}}$. Note that this standard deviation is known in the white noise model.

Using the proofs of [*Unpub-GLP11*], which follows a slightly different path than the one used in the proof of the theorems of the previous section, we can obtain an oracle type inequality for this estimator.

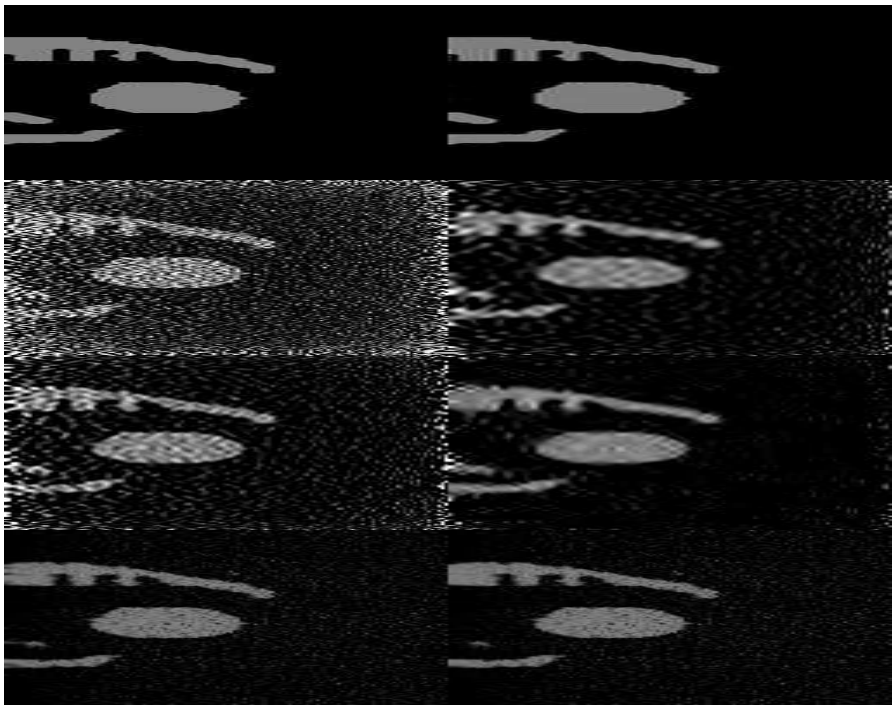


Figure 3.8: Estimation of $u$ with several methods: from top to bottom and left to right: original, projection of the original on Needlet basis, inversion without thresholding, smoothed projection, under-smoothed projection, thresholding estimation, two estimation obtain by Bergounioux and Trélat [BT10].

**Theorem 12.** *There exist some absolute constants $C_p$, $\boldsymbol{C_p}$, $C_{\boldsymbol{\xi}}$ depending only on the needlets used and the $L_p$ loss considered such that for $\kappa = \sqrt{2\gamma J_\sigma \log 2}$ with $\gamma > 1$, the previous estimator satisfies*

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\|\widehat{s} - s_0\|_p^p}{(3J_\sigma + 5)^p}\right] \\
&\quad \leq C_p \left(1 + \frac{\boldsymbol{C_p}}{(1/2)^p (2\kappa J_\sigma \log 2)^{p/2}}\right) \\
&\qquad \left[\sum_{\boldsymbol{\xi} \in \boldsymbol{\Xi_1}} \left( |c^{aa}_{1,\boldsymbol{\xi}}|^p \mathbf{1}_{\left\{|c^{aa}_{1,\boldsymbol{\xi}}| \leq 3/2\sqrt{2\kappa J_\sigma \log 2}\sigma_{aa,1,\boldsymbol{\xi}}\right\}} + \boldsymbol{C_p}\sigma^p_{1,\boldsymbol{\xi}} \mathbf{1}_{\left\{|c^{aa}_{1,\boldsymbol{\xi}}| \geq 3/2\sqrt{2\kappa J_\sigma \log 2}\sigma_{aa,1,\boldsymbol{\xi}}\right\}} \right)\right. \\
&\qquad \left. + \sum_{j=0}^{J_\sigma} 2^{j(p-2)} \sum_{o=ad,da,dd} \sum_{\boldsymbol{\xi} \in \boldsymbol{\Xi_{2^j}}} \left( |c^{o}_{2^j,\boldsymbol{\xi}}|^p \mathbf{1}_{\left\{|c^{o}_{2^j,\boldsymbol{\xi}}| \leq 3/2\sqrt{2\kappa J_\sigma \log 2}\sigma_{o,2^j,\boldsymbol{\xi}}\right\}} + \sigma^p_{o,2^j,\boldsymbol{\xi}} \mathbf{1}_{\left\{|c^{o}_{2^j,\boldsymbol{\xi}}| \geq 3/2\sqrt{2\kappa J_\sigma \log 2}\sigma_{o,2^j,\boldsymbol{\xi}}\right\}} \right)\right] \\
&\qquad + \|P_{\boldsymbol{2^{J_\sigma}}} u - u\|_p^p \\
&\qquad + 3C_p \boldsymbol{C_p} (1/2)^{p-1} \left(2\kappa J_\sigma \log 2\right)^{(p-1)/2} C_{\boldsymbol{\xi}} \sigma^p \frac{2^{J_\sigma(3/2(p-1/6\kappa))}}{1 - 2^{-(3/2p)}}
\end{aligned}
$$

Finally, we are also considering a more complex model in which one observe

$$dY_x = G \star As(x) + \sigma dW_x$$

where $G\star$ denotes the convolution with a known (Gaussian) kernel. Inspired by Neelamani, Choi, and Baraniuk [NCB04], we have implemented a numerical scheme based on a first smoothed inversion of $G\star$

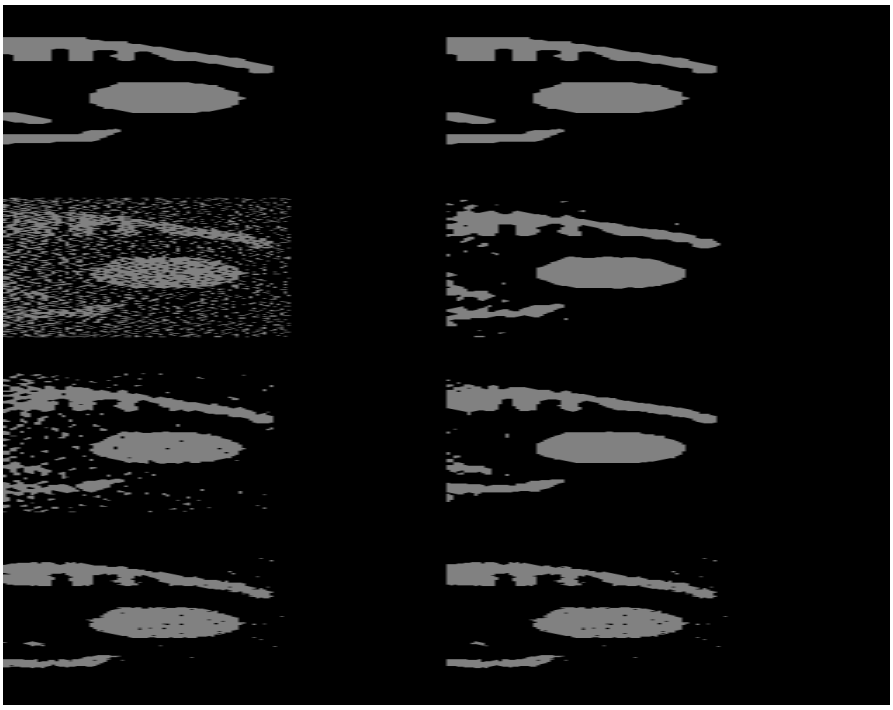


Figure 3.9: Binarized estimation of $u$ with several methods: from top to bottom and left to right: original, projection of the original on Needlet basis, inversion without thresholding, smoothed projection, undersmoothed projection, thresholding estimation, two estimation obtain by Bergounioux and Trélat [BT10].

followed by a needlet based inversion of $A$. The results are promising but no theoretical studies have been conducted yet.

## 3.6 Recreation: NL-Means and aggregation

I would like to conclude this section by the description of a very different approach followed with J. Salmon during his PhD thesis [*Proc-LPS09a*; *Proc-LPS09b*; *Proc-SLP09*]. It all started by the observation that some of the best denoising results are obtained by the patch based NL-means method proposed by Buades, Coll, and Morel [BCM05] or by some of its variants (for instance the one proposed by Kervrann and Boulanger [KB06]). These methods are based on a simple idea: consider the image not as a collection of pixels but as a collection of sub-images, the "patches", centered on those pixels and estimate each patch as a weighted average of patches. These weights take into account the similarities of the patches and are often chosen proportional to the exponential of the quadratic difference between the patches with a renormalization so they sum to 1. Understanding why these methods are so efficient is a challenging task.

In their seminal paper, Buades, Coll, and Morel [BCM05] show the consistency of their method under a strong technical $\beta$-mixing assumption on the image. NL-Means methods can also be seen as a smoothing in a patch space with a Gaussian kernel and their performances are related to the regularity of the underlying patch manifold (see for instance Peyré [Pe09] for a review). While intuitive and enlightening, those points of view have not yet permited to justify mathematically the performance of the NL-Means methods.

Inspired by Dalalyan and Tsybakov [DT07], we propose to look at those methods with a different eye so as to propose a different path to their mathematical justification. We consider them as special

instance of statistical aggregation. In this framework, one consider a collection of preliminary estimators and a noisy observation. We search then for a weighted average of those preliminary estimators. This *aggregate* estimate should be as close as possible to the unknown original signal. If one uses patches as preliminary estimators, a special case of a recent method inspired by PAC-Bayesian techniques considered by Dalalyan and Tsybakov [DT07] almost coincides with the NL-Means.

In the sequel, we describe this framework, propose some novel variants of patch based estimators and give some insights on their theoretical performances.

### 3.6.1 Image denoising, kernel and patch methods

We consider an image $S$ defined on a grid $(i,j)$, $1 \leq i \leq N$ and $1 \leq j \leq N$, of $N^2$ pixels and assume we observe a noisy version $Y$:

$$Y(i,j) = S(i,j) + \sigma W(i,j)$$

where $W$ is a white noise, an i.i.d. standard Gaussian sequence and $\sigma$ is a known standard deviation parameter. Our goal is to estimate the original image $I$ from the noisy observation $Y$.

Numerous methods have been proposed to fulfill this task. Most of them share the principle that the observed value should be replaced by a suitable local average, a local smoothing. Indeed all the kernel based methods, and even the dictionary based methods (thresholding for example), can be put in this framework. They differ in the way this local average is chosen. Those methods can be represented as a locally weighted sum

$$\hat{S}(i,j) = \sum_{k,l} \lambda_{i,j,k,l} Y(k,l)$$

where the weights $\lambda_{i,j,k,l}$ may depend in a complex way on both the indices and the values of $Y$. The weights $\lambda_{i,j,k,l}$ for a fixed pixel $(i,j)$ are nothing but the weights of a local smoothing kernel. The most famous weights are probably those of the Nadaraya-Watson estimator,

$$\lambda_{i,j,k,l} = \frac{K(i-k, j-l)}{\sum_{k',l'} K(i-k', j-l')} \quad ,$$

where $K$ is a fixed kernel (Gaussian for example). To make the estimator more efficient, the kernel and its scale can also vary depending on the local structure of the image such as in some locally adaptive method. Even if this is less explicit the representation based method can be put in this framework with a subtle dependency of the weights $i,j,k,l$ on the values of $Y$.

Patch based methods can be seen as extensions of such methods in which the image $f$ and the observation $Y$ are lifted in a higher dimensional space of patches. More precisely, for a fixed integer odd $S$, we define the patch $P(S)(i,j)$ as the sub-image of $I$ of size $K \times K$ centered on $(i,j)$ (for the sake of simplicity we assume here a periodic extension across the boundaries):

$$P(S)(i,j)(k,l) = S\left(i+k, j+l\right) \quad \text{for } -\frac{K-1}{2} \leq k,l \leq \frac{K-1}{2}.$$

An image $S$ belonging to $\mathbb{R}^{N^2}$ can thus be sent in a space of patch collection of dimension $\mathbb{R}^{N^2 \times S^2}$ through the application

$$S \mapsto P(S) = (P(S)(i,j))_{1 \leq i,j \leq N} \quad .$$

The denoising problem is reformulated as retrieving the original patch collection $\mathcal{P}(S)$ from the noisy patch collection $\mathcal{P}(Y)$. Note that an estimate $\hat{S}$ of the original image $S$ can be obtained from any estimate $\widehat{P(S)}$ of the original patch collection through a simple projection operator for example using the central values of the patches

$$\widehat{P(S)} \to \hat{S} = \left(\hat{S}(i,j) = \widehat{P(S)(i,j)}(0,0)\right)_{1 \leq i,j \leq N} \quad .$$

This simple projection can also be replaced by a more complex one, in which the value of a given pixel is obtained by averaging the values obtained for this pixel in different patches, as explored by Salmon and Strozecki [SS10; SS12].

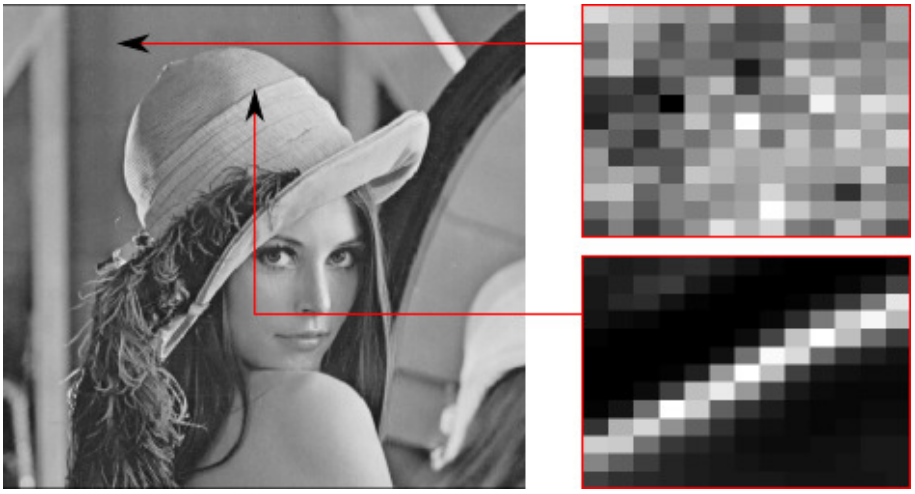


Figure 3.10: Adaptation of the NL-Means kernel to the local structures. The two right images show the kernel weights $(\lambda_{i,j,k,l})_{k,l}$ obtained for a patch in a uniformly regular zone and a patch centered on an edge.

Following the approach used for images, we consider here patch methods based on weighted sums

$$\widehat{P(S)(i,j)} = \sum_{k,l} \lambda_{i,j,k,l} P(Y)(k,l)$$

Note that when the $\lambda_{i,j,k,l}$ are chosen as in the Nadaraya-Watson estimator, the patch based estimator and the original pixel based estimator coincide. We will thus consider some other weight choices in which the weights for a given patch depends on the values of the other patches.

The method proposed by Buades, Coll, and Morel [BCM05] corresponds exactly to the use of the weights $\lambda_{i,j,k,l}$;

$$\lambda_{i,j,k,l} = \frac{e^{-\frac{1}{\beta}\|P(S)(i,j)-P(S)(k,l)\|^2}}{\sum_{k,l} e^{-\frac{1}{\beta}\|P(S)(i,j)-P(S)(k,l)\|^2}}$$

where $\|\cdot\|^2$ is the usual euclidean distance on patches. They have called this method Non Local Means (NL-Means from now on) as the weights depends only on the values of the patches and not on the distance between the patches (the distance between their centers). The influence of a patch on the reconstruction of another patch depends thus on their similarity so that the corresponding local smoothing kernels adapt themselves to the local structures in the image as illustrated in Figure 3.10.

They obtain the consistency of their method for stochastic process under some technical $\beta$-mixing conditions. Most of the other explanations of this method rely on the existence of a patch manifold of low dimension in which the patches live as for instance in the work of Peyré [Pe09]. The NL-Means method appears then as a local averaging using a Gaussian kernel in this patch space. Under the strong assumptions that the patches are evenly spaced on the patch manifold and that this manifold is flat enough to be approximated by an affine space, the performance of the NL-Means can be explained. Unfortunately, there is no guarantee this is the case.

Note that the strict non locality of this construction has been contradicted by a further study of J. Salmon [Sa10], which shows that using only patches in a neighborhood of the considered patch in the weights formula yields a significant improvement. The temperature parameter $\beta$ is also an important issue from both the theoretical and the practical point of view. We conclude this review by stressing that adding a localization term, such as a classical spatial kernel, renders the scheme close to a bilateral filtering in which the data dependent term is computed on the patch metric.

### 3.6.2 Aggregation and the PAC-Bayesian approach

We propose now a different point of view on this method: the aggregation point of view. In this setting, we consider a collection of preliminary estimates $\hat{s}_p$ of a given object $s_0$ and search for the best adaptive weighted combination

$$\hat{s}_\lambda = \sum_{p=1}^{P} \lambda_m \hat{s}_p$$

of those estimates from a noisy observation $Y = s_0 + \sigma W$ (or more generally in the white noise model). This setting has been introduced by Nemirovski [Ne00] and Yang [Ya00] and is the subject of a lot of studies since. This model is quite general as, for instance, both thresholding and estimator selection can be put in this framework. The key question is how to choose the aggregating weights.

We focus here on a special case in which the estimators are constructed for patches and the aggregation is based on the PAC-Bayesian approach considered by Catoni [Ca04]; Dalalyan and Tsybakov [DT07].

For any patch $P(S)(i,j)$, we assume we observe a noisy patch $P(Y)(i,j)$ and a collection of $P$ preliminary estimators $P_1$,...,$P_M$. We look then for an estimate

$$\widehat{P(S)(i,j)}_\lambda = \sum_{p=1}^{P} \lambda_p P_p$$

where $\lambda$ belongs to $\mathbb{R}^P$. The weights $\lambda_p$ are chosen, in the PAC Bayesian approach, in a very specific way from an arbitrary prior law $\pi$ on $\mathbb{R}^P$. The PAC-Bayesian aggregate $\widehat{P(S)(i,j)}^{\pi}$ is defined by the weighted "sum"

$$\widehat{P(S)(i,j)}^{\pi} = \int_{\mathbb{R}^P} \frac{e^{-\frac{1}{\beta}\|P(Y)(i,j) - \widehat{P(S)(i,j)}_\lambda\|^2}}{\int_{\mathbb{R}^P} e^{-\frac{1}{\beta}\|P(Y)(i,j) - \widehat{P(S)(i,j)}_{\lambda'}\|^2} d\pi(\lambda')} \widehat{P(S)}_\lambda d\pi(\lambda) \quad .$$

or equivalently by its weight components

$$\lambda_\pi = \int_{\mathbb{R}^P} \frac{e^{-\frac{1}{\beta}\|P(Y)(i,j) - \widehat{P(S)(i,j)}_\lambda\|^2}}{\int_{\mathbb{R}^P} e^{-\frac{1}{\beta}\|P(Y)(i,j) - \widehat{P(S)(i,j)}_{\lambda'}\|^2} d\pi(\lambda')} \lambda d\pi(\lambda) \quad .$$

Note that this estimator can be interpreted as a pseudo Bayesian estimator with a prior law $\pi$ in which the noise of variance $\sigma^2$ is replaced by a Gaussian noise of variance $\beta/2$.

The formula defining the estimator in the PAC-Bayesian approach looks similar to the formula defining the weights of the NL-Means, they are indeed equivalent when the preliminary estimators $P_m$ span the set of the noisy patches $P(Y)(k,l)$ and the prior law $\pi$ is chosen as the discrete law

$$\pi = \frac{1}{N^2} \sum_{(k,l)} \delta_{e_{(k,l)}}$$

where the sum runs across all the patches and $\delta_{e_{(k,l)}}$ is the Dirac measure charging only the patch $P(Y)(k,l)$. This choice leads to the estimate

$$\widehat{P(S)(i,j)}^{\pi} = \sum_{(k,l)} \frac{e^{-\frac{1}{\beta}\|P(Y)(i,j) - P(Y)(k,l)\|^2}}{\sum_{(k',l')} e^{-\frac{1}{\beta}\|P(Y)(i,j) - P(Y)(k',l')\|^2}} P(Y)(k,l) \quad ,$$

that is exactly the NL-Means estimator.

A lot of other variants of patch based method can be obtained through a suitable choice for the prior $\pi$. For example, for any kernel $K$,

$$\pi = \sum_{(k,l)} \frac{K(i-k, j-l)}{\sum_{(k',l')} K(i-k', j-l')} \delta_{e_{k,l}}$$

yields the localized NL-Means often used in practice.

### 3.6.3 Stein Unbiased Risk Estimator and Error bound

The analysis of the risk of this family of estimator is based on a SURE (Stein Unbiased Risk Estimator) principle as explained by Dalalyan and Tsybakov [DT07]; Leung and Barron [LB06]. Indeed, assume that the preliminary estimators $P_m$ are independent of $P(Y)(i,j)$, a simple computation shows that

$$\hat{r}_\lambda = \|P(Y)(i,j) - \widehat{P(S)(i,j)}_\lambda\|^2 - S^2\sigma^2$$

is an unbiased estimate of the risk of the estimator $\widehat{P(S)(i,j)}_\lambda$, $\|P(S)(i,j) - \widehat{P(S)(i,j)}_\lambda\|^2$. As $S^2\sigma^2$ is a term independent of $\lambda$, the PAC-Bayesian estimate of the previous section can be rewritten as

$$\widehat{P(S)(i,j)}^\pi = \int_{\mathbb{R}^P} \frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^P} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}} d\pi(\lambda')} \widehat{P(S)(i,j)}_\lambda d\pi(\lambda)$$

Using Stein's formula, one is able to construct an unbiased estimate $\hat{r}$ of the risk of this estimator such that, as soon as $\beta \geq 4\sigma^2$,

$$\hat{r} \leq \int_{\mathbb{R}^P} \frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^P} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}} d\pi(\lambda')} \hat{r}_\lambda d\pi(\lambda) \quad .$$

The key is then to notice (see for instance Catoni [Ca04]) that this renormalized exponential weights are such that for any probability law $p$

$$\int_{\mathbb{R}^P} \frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^P} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}} d\pi(\lambda')} \hat{r}_\lambda d\pi(\lambda) + \beta\mathcal{K}\left(\frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^P} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}}}\pi, \pi\right) \leq \int_{\mathbb{R}^P} \hat{r}_\lambda dp(\lambda) + \beta\mathcal{K}(p,\pi)$$

where $\mathcal{K}(p,\pi)$ is the Kullback divergence between $p$ and $\pi$:

$$\mathcal{K}(p,\pi) = \begin{cases} \int_{\mathbb{R}^P} \log\left(\frac{dp}{d\pi}(\lambda)\right) dp(\lambda) & \text{if } p \ll \pi, \\ +\infty & \text{otherwise} \end{cases}$$

Thus, as $\mathcal{K}(p,\pi)$ is always a positive quantity,

$$\hat{r} \leq \inf_{p\in\mathcal{P}} \int_{\mathbb{R}^P} \hat{r}_\lambda dp(\lambda) + \beta\mathcal{K}(p,\pi) \quad .$$

Taking the expectation and interchanging the order of the expectation and the infimum yield

$$E(\hat{r}) \leq E\left(\inf_{p\in\mathcal{P}} \int_{\mathbb{R}^P} \hat{r}_\lambda dp(\lambda) + \beta\mathcal{K}(p,\pi)\right) \leq \inf_{p\in\mathcal{P}} \left(\int_{\mathbb{R}^P} E(\hat{r}_\lambda) dp(\lambda) + \beta\mathcal{K}(p,\pi)\right)$$

or more explicitly using the fact that the $\hat{r}$ are unbiased estimates of the risks

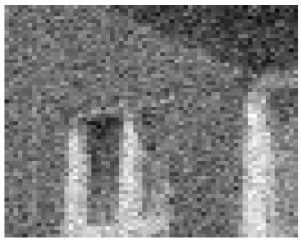
$$\mathbb{E}\left(\|P(S)(i,j) - \widehat{P(S)(i,j)}^\pi\|^2\right) \leq \inf_{p\in\mathcal{P}} \left(\int_{\mathbb{R}^P} \|P(S)(i,j) - \widehat{P(S)(i,j)}_\lambda\|^2 dp(\lambda) + \beta\mathcal{K}(p,\pi)\right) \quad .$$

The PAC-Bayesian aggregation principle is thus supported by a strong theoretical result when the preliminary estimators $P_m$ are independent of $P(Y)$, often call the frozen preliminary estimators case, and $\beta$ is larger than $4\sigma^2$. The quadratic error of the PAC-Bayesian estimate is bounded by the best trade-off between the average quadratic error of fixed $\lambda$ estimators under a law $p$ and an adaptation price corresponding to the Kullback distance between $p$ and the prior $\pi$. The optimal $p$ is thus one both concentrated around the best fixed $\lambda$ estimator and close to the prior law $\pi$.

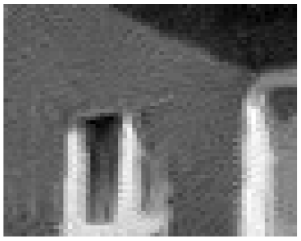
So far, these results have been proved only when the preliminary estimators are independent of the observation, which is obviously not the case when they are chosen as patches of the noisy images. We conjecture that the following similar inequality holds, up to a slight modification of the aggregation weights, with a $\gamma$ possibly larger than 1

$$\mathbb{E}\|P(S)(i,j) - \widehat{P(S)(i,j)}^\pi\|^2 \leq \inf_{p\in\mathcal{P}} \left(\int_{\mathbb{R}^P} \left(\|P(S)(i,j) - P(S)(i,j)_\lambda\|^2 + \gamma K^2\sigma^2\|\lambda\|^2\right) dp(\lambda) + \beta\mathcal{K}(p,\pi)\right)$$

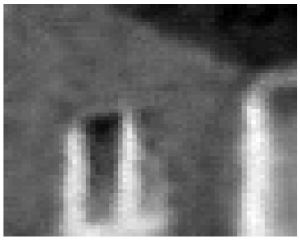
Original



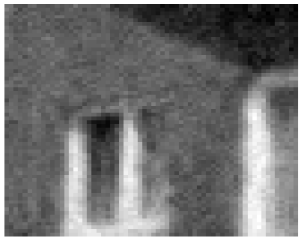
Noisy
22.0 dB



NL-Means
29.7 dB



Student
28.4 dB



Gaussian Mix.
29.5 dB

Figure 3.11: Numerical results on a small part of House for the 3 studied priors.

where the $K^2\sigma^2\|\lambda\|^2$ term appears as the variance of the estimator for a fixed $\lambda$, which is nothing but a classical kernel estimator. The trade-off for $p$ is thus between a concentration around the best linear kernel and a proximity with the prior law $\pi$. The aggregation point of view shows that this patch based procedure is close to a search for an optimal local kernel, which is one of the intuition behind the NL-Means construction.

We have obtained this result so far in three cases: when patches are computed on another noisy image, when all patches intersecting the central patch are removed from the collection and with a small modification of the weights when the image is split into two separate images with a quincunx grid. We are still working on a proof for the general case that requires some more modifications of the weights.

### 3.6.4 Priors and numerical aspects of aggregation

The most important parameter to obtain a good control of the error is thus the prior $\pi$. A good choice is one such that for any natural patch $P(i)(i,j)$ there is a probability law $p$ close to $\pi$ and concentrated around the best kernel weights. The goal is to prove that the penalty due to the Kullback divergence term is not too big compare to the best kernel performance. We have conducted numerical experiments, using a Langevin Monte Carlo method, in [*Proc-LPS09a*; *Proc-LPS09b*; *Proc-SLP09*].

Our numerical experiments can be summarized as follows:

- There is still a slight loss between our method using a Gaussian mixture prior and the optimized classical NL-Means. We have observed PAC-Bayesian aggregation is less sensitive to the parameters. The same parameter set yields good results for all our test images while for the NL-Means the temperature has to be tuned.
- The choice $\beta = 4\sigma^2$, recommended by the theory, does not lead to the best results: the choice $\beta = 2\sigma^2$ which corresponds to a classical Bayesian approach leads to better performances.

- The correction proposed for the central patch is effective as described by Salmon [Sa10].
- We have observed that the central point is responsible for more than .5 dB gain in the NL-Means approach and less in the PAC-Bayesian approach.
- We are still facing some convergence issues in our Monte Carlo scheme which could explain our loss of performances. We are working on a modified scheme to overcome this issue.

The PAC-Bayesian approach provides a novel point of view on patch based method. From the theoretical point of view, we have not yet however been able to control the performance of our method.

# Chapter 4

# Density estimation

## 4.1 Density estimation

I have also considered the classical density estimation framework. We observe $n$ random variables $((X_i))_{1\leq i\leq N}$ of random variables and are interested in estimating the law of this variable $X_i \in \mathcal{X}$. We further assume that the observations $X_i$'s are independent and identically distributed and follow a law of density $s_0(\cdot|X_i)$ with respect to a known measure $\mathrm{d}\lambda$. Our goal is to estimate this density function $s_0(\cdot)$ from the observations. We refer to the book of Tsybakov [Ts08] for an introduction as well as an analysis of kernel based estimators.

We will focus here on contrast based methods. When the losses considered are the quadratic loss or the Kullback-Leibler divergence, two natural contrasts appear: for the quadratic loss,

$$\gamma(s) = \frac{1}{2}\|s\|_2^2 - \frac{1}{N}\sum_{i=1}^{N} -s(X_i)$$

and for the Kullback-Leibler divergence

$$\gamma(s) = \frac{1}{N}\sum_{i=1}^{N} -\log\left(s(X_i)\right).$$

In this chapter, we will focus on the first one.

## 4.2 Dictionary, adaptive threshold and $\ell_1$ penalization

### 4.2.1 Dictionary, Lasso and Dantzig

Assume we have at hand a dictionary of functions $\mathcal{D} = (\varphi_p)_{p=1,\dots,P}$ a natural idea is to estimate $s_0$ by a weighted sum $s_\lambda$ of elements of $\mathcal{D}$

$$s_\lambda = \sum_{p=1}^{P} \lambda_p \varphi_p.$$

The quadratic loss contrast $\gamma(s_\lambda)$ can be rewritten in this case as

$$\begin{aligned}\gamma(s_\lambda) &= \frac{1}{2}\|s_\lambda\|^2 - \frac{1}{n}\sum_{i=1}^{n} s_\lambda(X_i) \\ &= \frac{1}{2}\lambda' G \lambda - A_X'\lambda\end{aligned}$$

with

$$G_{p,p'} = \langle \varphi_p, \varphi_{p'} \rangle \qquad \text{and} \qquad A_{x,p} = \frac{1}{n} \sum_{i=1}^{n} \phi_p(X_i)$$

whose minimizer are the solutions of $G\lambda = A_X$. As long as $G$ is invertible, i.e. the family $(\phi_p)$ is free, there is a unique solution but this assumption is strong. Furthermore, it could be interesting to *select* a few functions in the dictionary and estimate only the *projection* in this restricted model. The penalized model principle selection of the previous chapter can also be applied as described for instance by Massart [Ma07]. Unfortunately, the associated procedures may be hardly tractable from the numerical point of view as they often amount to a brute force exhaustive search algorithm. Various ideas has been proposed to replace this exhaustive search algorithm by a much more efficient one. A very succesful idea has been to replace the usual penalty proportional to the dimension by a penalty proportional to the $\ell_1$ norm of the coefficients. The resulting estimator can be computed by a *simple* convex function minimzation for which numerous *efficient* algorithms exist. The two most famous one are probably the LASSO procedure of Tibshirani [Ti96] and the Dantzig procedure of Candès and Tao [CT07]. The computational study has been performed notably by Efron, Hastie, Johnstone, and Tibshirani [Ef+04]; Osborne, Presnell, and Turlach [OPT00a; OPT00b]. Both has been widely considered in white noise and regression models, but only the Lasso estimator had been studied in the density model (see the works of Bunea, Tsybakov, and Wegkamp [BTW07b]; Bunea, Tsybakov, Wegkamp, and Barbu [Bu+10]; Geer [Ge08]) at the time I started this analysis with K. Bertin and V. Rivoirard [*Art-BLPR11*].

The Dantzig selector has been introduced by Candès and Tao [CT07] in the linear regression model. More precisely, given

$$Y = A\lambda_0 + \varepsilon,$$

where $Y \in \mathbb{R}^n$, $A$ is a $n$ by $M$ matrix, $\varepsilon \in \mathbb{R}^n$ is the noise vector and $\lambda_0 \in \mathbb{R}^M$ is the unknown regression parameter to estimate, the Dantzig estimator is defined by

$$\hat{\lambda}^D = \arg\min_{\lambda \in \mathbb{R}^M} \|\lambda\|_{\ell_1} \text{ subject to } \|A'(A\lambda - Y)\|_{\ell_\infty} \leq \eta,$$

where $\|\cdot\|_{\ell_\infty}$ is the sup-norm in $\mathbb{R}^M$, $\|\cdot\|_{\ell_1}$ is the $\ell_1$ norm in $\mathbb{R}^M$, and $\eta$ is a regularization parameter. A natural companion of this estimator is the Lasso procedure or more precisely its relaxed form

$$\hat{\lambda}^L = \arg\min_{\lambda \in \mathbb{R}^M} \left\{ \frac{1}{2} \|A\lambda - Y\|_{\ell_2}^2 \ + \eta \|\lambda\|_{\ell_1} \right\},$$

where $\eta$ plays exactly the exact same role as for the Dantzig estimator. This $\ell_1$ penalized method is also called *basis pursuit* in signal processing (see Chen, Donoho, and Saunders [CDS01]; Donoho, Elad, and Temlyakov [DET06]).

Candès and Tao [CT07] have obtained a bound for the $\ell_2$ risk of the estimator $\hat{\lambda}^D$, with large probability, under a global condition on the matrix $A$ (the Restricted Isometry Property) and a sparsity assumption on $\lambda_0$, even for $P \geq N$. Bickel, Ritov, and Tsybakov [BRT09] have obtained oracle inequalities and bounds of the $\ell_p$ loss for both estimators under weaker assumptions. Actually, Bickel, Ritov, and Tsybakov [BRT09] deal with the non parametric regression framework in which one observes

$$Y_i = s_0(X_i) + W_i, \quad i = 1, \ldots, N$$

where $s_0$ is an unknown function while $(X_i)_{i=1,\ldots,N}$ are known design points and $(W_i)_{i=1,\ldots,N}$ is a noise vector. There is no intrinsic matrix $A$ in this problem but for any dictionary of functions $\mathcal{D} = (\varphi_p)_{p=1,\ldots,P}$ one can search $s_0$ as a weighted sum $s_\lambda$ of elements of $\mathcal{D}$

$$s_\lambda = \sum_{p=1}^{P} \lambda_p \varphi_p$$

and introduce the matrix $\mathbf{\Phi} = (\varphi_p(X_i))_{i,p}$, which summarizes the information on the dictionary and on the design. Notice that if there exists $\lambda_0$ such that $s_0 = s_{\lambda_0}$ then the model can be rewritten exactly as

the classical linear model. However, if it is not the case and if a model bias exists, the Dantzig and Lasso procedures can be after all applied under similar assumptions on $A$. Oracle inequalities are obtained for which approximation theory plays an important role in the studies of Bickel, Ritov, and Tsybakov [BRT09]; Bunea, Tsybakov, and Wegkamp [BTW07a; BTW07c]; Geer [Ge08].

Let us also mention that in various settings, under various assumptions on the matrix $\boldsymbol{\phi}$ (or more precisely on the associated Gram matrix $G = \boldsymbol{\Phi}'\boldsymbol{\Phi}$), properties of these estimators have been established for subset selection (see Bunea [Bu09]; Lounici [Lo08]; Meinhausen and Yu [MY09]; Meinshausen and Bühlmann [MB06]; Yu and Zhao [YZ06]; Zhang and Huang [ZH08]) and for prediction (see Bickel, Ritov, and Tsybakov [BRT09]; Knight and Fu [KF00]; Lounici [Lo08]; Meinhausen and Yu [MY09]; Zou [Zo06]).

In the density framework, the Dantzig estimate $\hat{f}^D$ is then obtained by minimizing $\|\lambda\|_{\ell_1}$ over the set of parameters $\lambda$ satisfying the adaptive Dantzig constraint:

$$\forall\, p \in \{1, \ldots, P\}, \qquad |(G\lambda)_p - \hat{\beta}_p| \leq \eta_{\gamma,p}$$

where for $p \in \{1, \ldots, P\}$, $(G\lambda)_p$ is the scalar product of $s_\lambda$ with $\varphi_p$,

$$\eta_{\gamma,p} = \sqrt{\frac{2\tilde{\sigma}_p^2 \gamma \log M}{n}} + \frac{2\|\varphi_p\|_\infty \gamma \log P}{3N},$$

$\tilde{\sigma}_p^2$ is a sharp estimate of the variance of $\hat{\beta}_p$ and $\gamma$ is a constant to be chosen. Section 4.2.2 gives precise definitions and heuristics for using this constraint. We just mention here that $\eta_{\gamma,p}$ comes from sharp concentration inequalities to give tight constraints. Our idea is that if $s_0$ can be decomposed on $\mathcal{D}$ as

$$s_o = \sum_{p=1}^{P} \lambda_{0,m} \varphi_p,$$

then we force the set of feasible parameters $\lambda$ to contain $\lambda_0$ with large probability and to be as small as possible. Significant improvements in practice are expected.

Our goals is mainly twofold. First, we aim at establishing sharp oracle inequalities under very mild assumptions on the dictionary. Our starting point is that most of the papers in the literature assume that the functions of the dictionary are bounded by a constant independent of $P$ and $N$, which constitutes a strong limitation, in particular for dictionaries based on histograms or wavelets (see for instance Bunea [Bu09]; Bunea, Tsybakov, and Wegkamp [BTW06; BTW07a; BTW07b; BTW07c] or Geer [Ge08]). Such assumptions on the functions of $\mathcal{D}$ will not be considered here. Likewise, our methodology does not rely on the knowledge of $\|s_0\|_\infty$ that can even be infinite (as noticed by Birgé [Bi08] for the study of the integrated $\mathbb{L}_2$-risk, most of the papers in the literature typically assume that the sup-norm of the unknown density is finite with a known or estimated bound for this quantity). Finally, let us mention that, in contrast with what Bunea, Tsybakov, Wegkamp, and Barbu [Bu+10] did, we obtain oracle inequalities with leading constant 1, and furthermore these are established under much weaker assumptions on the dictionary than the ones of Bunea, Tsybakov, Wegkamp, and Barbu [Bu+10].

The second goal deals with the problem of calibrating the so-called *Dantzig constant* $\gamma$: how should this constant be chosen to obtain good results in both theory and practice? Most of the time, for Lasso-type estimators, the regularization parameter is of the form $a\sqrt{\frac{\log M}{n}}$ with $a$ a positive constant (see Bickel, Ritov, and Tsybakov [BRT09]; Bunea, Tsybakov, and Wegkamp [BTW06; BTW07b; BTW07c], Candès and Plan [CP09], Lounici [Lo08] or Meinhausen and Yu [MY09] for instance). These results are obtained with large probability that depends on the tuning coefficient $a$. In practice, it is not simple to calibrate the constant $a$. Unfortunately, most of the time, the theoretical choice of the regularization parameter is not suitable for practical issues. This fact is true for Lasso-type estimates but also for many algorithms for which the regularization parameter provided by the theory is often too conservative for practical purposes (see Juditsky and Lambert-Lacroix [JL04] who clearly explains and illustrates this point for their thresholding procedure). So, one of the main goals here is to fill the gap between the optimal parameter choice provided by theoretical results on the one hand and by a simulation study on the other hand. Only a few papers are devoted to this problem. In the model selection setting, the issue of calibration has been addressed by Birgé and Massart [BM07] who considered $\ell_0$-penalized

estimators in a Gaussian homoscedastic regression framework and showed that there exists a minimal penalty in the sense that taking smaller penalties leads to inconsistent estimation procedures. Arlot and Massart [AM09] generalized these results for non-Gaussian or heteroscedastic data and Reynaud-Bouret and Rivoirard [RR10] addressed this question for thresholding rules in the Poisson intensity framework.

Now, let us describe in a nutshell our results. By using the previous data-driven Dantzig constraint, oracle inequalities are derived under local conditions on the dictionary that are valid under classical assumptions on the structure of the dictionary. We extensively discuss these assumptions and we show their own interest. Each term of these oracle inequalities is easily interpretable. Classical results are recovered when we further assume:

$$\|\varphi_p\|_\infty^2 \leq c_1 \left(\frac{N}{\log P}\right) \|s_0\|_\infty,$$

where $c_1$ is a constant. This assumption is very mild and, unlike in classical works, allows to consider dictionaries based on wavelets. Then, relying on our Dantzig estimate, we build an adaptive Lasso procedure whose oracle performances are similar. This illustrates the closeness between Lasso and Dantzig-type estimates.

Our results are proved for $\gamma > 1$. For the theoretical calibration issue, we study the performance of our procedure when $\gamma < 1$. We show that in a simple framework, estimation of the straightforward signal $s_0 = \mathbf{1}_{[0,1]}$ cannot be performed at a convenient rate of convergence when $\gamma < 1$. This result proves that the assumption $\gamma > 1$ is thus not too conservative.

Finally, a simulation study illustrates how dictionary-based methods outperform classical ones. More precisely, we show that our Dantzig and Lasso procedures with $\gamma > 1$, but close to 1, outperform classical ones, such as simple histogram procedures, wavelet thresholding or Dantzig procedures based on the knowledge of $\|s_0\|_\infty$ and less tight Dantzig constraints.

### 4.2.2 The Dantzig estimator of the density $s_0$

As said in Introduction, our goal is to build an estimate of the density $f_0$ with respect to the measure $dx$ as a linear combination of functions of $\mathcal{D} = (\varphi_p)_{p=1,\dots,P}$, where we assume without any loss of generality that, for any $p$, $\|\varphi_p\|_2 = 1$:

$$s_\lambda = \sum_{p=1}^{P} \lambda_p \varphi_p.$$

For this purpose, we naturally rely on natural estimates of the $\mathbb{L}_2$-scalar products between $s_0$ and the $\varphi_p$'s. So, for $p \in \{1, \dots, P\}$, we set

$$\beta_{0,p} = \int \varphi_p(x) s_o(x) dx,$$

and we consider its empirical counterpart

$$\hat{\beta}_p = \frac{1}{N} \sum_{i=1}^{N} \varphi_p(X_i)$$

that is an unbiased estimate of $\beta_{0,p}$. The variance of this estimate is $\mathrm{Var}(\hat{\beta}_p) = \frac{\sigma_{0,p}^2}{N}$ where

$$\sigma_{0,p}^2 = \int \varphi_p^2(x) s_o(x) dx - \beta_{0,p}^2.$$

Note also that for any $\lambda$ and any $p$, the $\mathbb{L}_2$-scalar product between $f_\lambda$ and $\varphi_p$ can be easily computed:

$$\int \varphi_p(x) f_\lambda(x) dx = \sum_{p'=1}^{P} \lambda_{p'} \int \varphi_{p'}(x) \varphi_p(x) dx = (G\lambda)_p$$

where $G$ is the Gram matrix associated to the dictionary $\mathcal{D}$ defined for any $1 \leq p, p' \leq P$ by

$$G_{p,p'} = \int \varphi_p(x)\varphi_{p'}(x)dx.$$

Any reasonable choice of $\lambda$ should ensure that the coefficients $(G\lambda)_p$ are close to $\hat{\beta}_p$ for all $p$. Therefore, using Candès and Tao's approach, we define the Dantzig constraint:

$$\forall\, p \in \{1, \dots, P\}, \quad |(G\lambda)_p - \hat{\beta}_p| \leq \eta_{\gamma,p} \tag{4.1}$$

and the Dantzig estimate $\hat{s}^D$ by $\hat{s}^D = f_{\hat{\lambda}^{D,\gamma}}$ with

$$\hat{\lambda}^{D,\gamma} = \underset{\lambda \in \mathbb{R}^M}{\operatorname{argmin}} \|\lambda\|_{\ell_1} \quad \text{such that } \lambda \text{ satisfies the Dantzig constraint (4.1)},$$

where for $\gamma > 0$ and $p \in \{1, \dots, P\}$,

$$\eta_{\gamma,p} = \sqrt{\frac{2\tilde{\sigma}_p^2 \gamma \log P}{N}} + \frac{2\|\varphi_p\|_\infty \gamma \log P}{3N}, \tag{4.2}$$

with

$$\tilde{\sigma}_p^2 = \hat{\sigma}_p^2 + 2\|\varphi_p\|_\infty \sqrt{\frac{2\hat{\sigma}_p^2 \gamma \log P}{N}} + \frac{8\|\varphi_p\|_\infty^2 \gamma \log P}{N}$$

and

$$\hat{\sigma}_p^2 = \frac{1}{N(N-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\varphi_p(X_i) - \varphi_p(X_j))^2.$$

Note that $\eta_{\gamma,p}$ depends on the data, so the constraint (4.1) will be referred as the *adaptive Dantzig constraint* in the sequel. We now justify the introduction of the density estimate $\hat{s}^D$.

The definition of $\eta_{\lambda,\gamma}$ is based on the following heuristics. Given $p$, when there exists a constant $c_0 > 0$ such that $s_0(x) \geq c_0$ for $x$ in the support of $\varphi_p$ satisfying $\|\varphi_p\|_\infty^2 = o_N(N(\log P)^{-1})$, then, with large probability, the deterministic term of (4.2), $\frac{2\|\varphi_p\|_\infty \gamma \log P}{3N}$, is negligible with respect to the random one, $\sqrt{\frac{2\tilde{\sigma}_p^2 \gamma \log P}{N}}$. In this case, the random term is the main one and we asymptotically derive

$$\eta_{\gamma,p} \approx \sqrt{2\gamma \log P \, \frac{\tilde{\sigma}_p^2}{N}}. \tag{4.3}$$

Having in mind that $\tilde{\sigma}_p^2/N$ is a convenient estimate for $\mathrm{Var}(\hat{\beta}_p)$, the shape of the right hand term of the formula (4.3) looks like the bound proposed by Candès and Tao [CT07] to define the Dantzig constraint in the linear model. Actually, the deterministic term of (4.2) allows to get sharp concentration inequalities. As often done in the literature, instead of estimating $\mathrm{Var}(\hat{\beta}_p)$, we could use the inequality

$$\mathrm{Var}(\hat{\beta}_p) = \frac{\sigma_{0,p}^2}{N} \leq \frac{\|s_0\|_\infty}{N}$$

and we could replace $\tilde{\sigma}_p^2$ with $\|s_0\|_\infty$ in the definition of the $\eta_{\gamma,p}$. But this requires a strong assumption: $s_0$ is bounded and $\|s_0\|_\infty$ is known. In our study, $\mathrm{Var}(\hat{\beta}_p)$ is estimated, which allows not to impose these conditions. More precisely, we slightly overestimate $\sigma_{0,p}^2$ to control large deviation terms and this is the reason why we introduce $\tilde{\sigma}_p^2$ instead of using $\hat{\sigma}_p^2$, an unbiased estimate of $\sigma_{0,p}^2$. Finally, $\gamma$ is a constant that has to be suitably calibrated and plays a capital role in practice.

The following result justifies previous heuristics by showing that, if $\gamma > 1$, with high probability, the quantity $|\hat{\beta}_p - \beta_{0,p}|$ is smaller than $\eta_{\gamma,p}$ for all $p$. The parameter $\eta_{\gamma,p}$ with $\gamma$ close to 1 can be viewed as the *smallest* quantity that ensures this property.

**Theorem 13.** *Let us assume that $P$ satisfies*

$$N \leq P \leq \exp(N^\delta) \tag{4.4}$$

*for $\delta < 1$. Let $\gamma > 1$. Then, for any $\varepsilon > 0$, there exists a constant $C_1(\varepsilon, \delta, \gamma)$ depending on $\varepsilon$, $\delta$ and $\gamma$ such that*

$$\mathbb{P}\left\{\exists p \in \{1, \dots, P\}, \quad |\beta_{0,p} - \hat{\beta}_p| \geq \eta_{\gamma,p}\right\} \leq C_1(\varepsilon, \delta, \gamma) P^{1 - \frac{\gamma}{1+\varepsilon}}.$$

*In addition, there exists a constant $C_2(\delta, \gamma)$ depending on $\delta$ and $\gamma$ such that*

$$\mathbb{P}\left\{\forall p \in \{1, \dots, P\}, \quad \eta^{(-)}_{\gamma,p} \leq \eta_{\gamma,p} \leq \eta^{(+)}_{\gamma,p}\right\} \geq 1 - C_2(\delta, \gamma) P^{1-\gamma}$$

*where, for $p \in \{1, \dots, P\}$,*

$$\eta^{(-)}_{\gamma,p} = \sigma_{0,p} \sqrt{\frac{8\gamma \log P}{7N}} + \frac{2\|\varphi_p\|_\infty \gamma \log P}{3N}$$

*and*

$$\eta^{(+)}_{\gamma,p} = \sigma_{0,p} \sqrt{\frac{16\gamma \log p}{N}} + \frac{10\|\varphi_p\|_\infty \gamma \log P}{N}.$$

The first part is a sharp concentration inequality proved by using Bernstein type controls. The second part of the theorem proves that, up to constants depending on $\gamma$, $\eta_{\gamma,p}$ is of order $\sigma_{0,p}\sqrt{\frac{\log P}{N}} + \|\varphi_p\|_\infty \frac{\log P}{N}$ with high probability. Note that the assumption $\gamma > 1$ is essential to obtain probabilities going to 0.

Finally, let $\lambda_0 = (\lambda_{0,p})_{p=1,\dots,P} \in \mathbb{R}^P$ such that

$$P_{\mathcal{D}} f_0 = \sum_{p=1}^{P} \lambda_{0,p} \varphi_p$$

where $P_{\mathcal{D}}$ is the projection on the space spanned by $\mathcal{D}$. We have

$$(G\lambda_0)_p = \int (P_{\mathcal{D}} s_0)\varphi_p = \int s_0 \varphi_p = \beta_{0,p}.$$

So, Theorem 13 proves that $\lambda_0$ satisfies the adaptive Dantzig constraint (4.1) with probability larger than $1 - C_1(\varepsilon, \delta, \gamma) P^{1 - \frac{\gamma}{1+\varepsilon}}$ for any $\varepsilon > 0$. Actually, we force the set of parameters $\lambda$ satisfying the adaptive Dantzig constraint to contain $\lambda_0$ with large probability and to be as small as possible. Therefore, $\hat{s}^D = s_{\hat{\lambda}^{D,\gamma}}$ is a good candidate among sparse estimates linearly decomposed on $\mathcal{D}$ for estimating $s_o$.

We mention that Assumption (4.4) can be relaxed and we can take $P < N$ provided the definition of $\eta_{\gamma,p}$ is modified.

It turns out that a similar bound, with slightly better constant, could also be obtained using some results due to Maurer and Pontil [MP09]. More precisely, if we let now

$$\eta_{\gamma,p} = \sqrt{\frac{2\hat{\sigma}_p^2 \gamma \log P}{N}} + \frac{7\|\varphi_p\|_\infty \gamma \log P}{3(N-1)},$$

where we use directly $\hat{\sigma}_p^2$ instead of $\tilde{\sigma}_p^2$ but have replaced a 2 with a 7 in the second term and a $N$ by $N-1$ then

$$\mathbb{P}\left\{\exists p \in \{1, \dots, P\}, \quad |\beta_{0,p} - \hat{\beta}_p| \geq \eta_{\gamma,p}\right\} \leq 2P^{1-\gamma}$$

and

$$\mathbb{P}\left\{\exists p \in \{1, \dots, P\}, \quad |\hat{\sigma}_p - \sigma_p| > 2u\frac{\|\phi_p\|_\infty}{N}\right\} \leq 2P^{1-\gamma}.$$

So that our results holds also for this modified thresholds up to some straightforward modifications.

### 4.2.3 Results for the Dantzig estimators

In the sequel, we will denote $\hat{\lambda}^D = \hat{\lambda}^{D,\gamma}$ to simplify the notations, but the Dantzig estimator $\hat{f}^D$ still depends on $\gamma$. Moreover, we assume that (4.4) is true and we denote the vector $\eta_\gamma = (\eta_{\gamma,p})_{p=1,\dots,P}$ considered with the Dantzig constant $\gamma > 1$.

**The main result under local assumptions**

Let us state the main result. For any $J \subset \{1,\dots,P\}$, we set $J^C = \{1,\dots,P\} \setminus J$ and define $\lambda_J$ the vector which has the same coordinates as $\lambda$ on $J$ and zero coordinates on $J^C$. We introduce a local assumption indexed by a subset $J_0$.

- **Local Assumption** Given $J_0 \subset \{1,\dots,M\}$, for some constants $\kappa_{J_0} > 0$ and $\mu_{J_0} \geq 0$ depending on $J_0$, we have for any $\lambda$,

$$\|s_\lambda\|_2 \geq \kappa_{J_0} \frac{\|\lambda_{J_0}\|_{\ell_1}}{\sqrt{|J_0|}} - \frac{\mu_{J_0}}{\sqrt{|J_0|}} \left( \|\lambda_{J_0^C}\|_{\ell_1} - \|\lambda_{J_0}\|_{\ell_1} \right)_+ . \qquad (LA(J_0, \kappa_{J_0}, \mu_{J_0}))$$

Note that this Assumption is a slight generalization of the one published in [*Art-BLPR11*]. We obtain the following oracle type inequality without any assumption on $s_0$.

**Theorem 14.** *With probability at least $1 - C_1(\varepsilon,\delta,\gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, for all $J_0 \subset \{1,\dots,P\}$ such that there exist $\kappa_{J_0} > 0$ and $\mu_{J_0} \geq 0$ for which $(LA(J_0,\kappa_{J_0},\mu_{J_0}))$ holds, we have, for any $\alpha > 0$,*

$$\|\hat{s}^D - s_0\|_2^2 \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \|s_\lambda - s_0\|_2^2 + \alpha \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\Lambda(\lambda, J_0^c)^2}{|J_0|} + 16|J_0| \left(\frac{1}{\alpha} + \frac{1}{\kappa_{J_0}^2}\right) \|\eta_\gamma\|_{\ell_\infty}^2 \right\}, \qquad (4.5)$$

*with*

$$\Lambda(\lambda, J_0^c) = \|\lambda_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+}{2}.$$

Let us comment each term of the right hand side of (4.5). The first term is an approximation term which measures the closeness between $s_0$ and $s_\lambda$. This term can vanish if $s_0$ can be decomposed on the dictionary. The second term, a bias term, is a price to pay when either $\lambda$ is not supported by the subset $J_0$ considered or it does not satisfy the condition $\|\hat{\lambda}^D\|_{\ell_1} \leq \|\lambda\|_{\ell_1}$ which holds as soon as $\lambda$ satisfies the adaptive Dantzig constraint. Finally, the last term, which does not depend on $\lambda$, can be viewed as a variance term corresponding to the estimation on the subset $J_0$. The parameter $\alpha$ calibrates the weights given for the bias and variance terms in the oracle inequality. Concerning the last term, remember that $\eta_{\gamma,p}$ relies on an estimate of the variance of $\hat{\beta}_p$. Furthermore, we have with high probability:

$$\|\eta_\gamma\|_{\ell_\infty}^2 \leq 2 \sup_p \left( \frac{16\sigma_{0,p}^2 \gamma \log P}{N} + \left( \frac{10\|\varphi_p\|_\infty \gamma \log P}{N} \right)^2 \right).$$

So, if $s_0$ is bounded then, $\sigma_{0,m}^2 \leq \|s_0\|_\infty$ and if there exists a constant $c_1$ such that for any $m$,

$$\|\varphi_p\|_\infty^2 \leq c_1 \left( \frac{N}{\log P} \right) \|s_0\|_\infty, \qquad (4.6)$$

(which is true for instance for a bounded dictionary), then

$$\|\eta_\gamma\|_{\ell_\infty}^2 \leq C\|s_0\|_\infty \frac{\log P}{N},$$

(where $C$ is a constant depending on $\gamma$ and $c_1$) and tends to 0 when $N$ goes to $\infty$. We obtain thus the following result.

**Corollary 2.** *With probability at least $1 - C_1(\varepsilon,\delta,\gamma)P^{1-\frac{\gamma}{1+\varepsilon}}$, if (4.6) is satisfied, then, for all $J_0 \subset \{1,\dots,P\}$ such that there exist $\kappa_{J_0} > 0$ and $\mu_{J_0} \geq 0$ for which $(LA(J_0,\kappa_{J_0},\mu_{J_0}))$ holds, we have, for any $\alpha > 0$ and for any $\lambda$ that satisfies the adaptive Dantzig constraint,*

$$\|\hat{s}^D - s_0\|_2^2 \leq \|s_\lambda - s_0\|_2^2 + \alpha c_2 (1 + \kappa_{J_0}^{-2}\mu_{J_0}^2) \frac{\|\lambda_{J_0^C}\|_{\ell_1}^2}{|J_0|} + c_3(\alpha^{-1} + \kappa_{J_0}^{-2})|J_0|\|f_0\|_\infty \frac{\log M}{n}, \qquad (4.7)$$

*where $c_2$ is an absolute constant and $c_3$ depends on $c_1$ and $\gamma$.*

*If $s_0 = s_{\lambda_0}$ and if $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ holds with $J_0$ the support of $\lambda_0$ then, under (4.6), with probability at least $1 - C_1(\varepsilon, \delta, \gamma) M^{1-\frac{\gamma}{1+\varepsilon}}$, we have*

$$\|\hat{s}^D - s_0\|_2^2 \leq C' |J_0| \|f_0\|_\infty \frac{\log P}{N},$$

*where $C' = c_3 \kappa_{J_0}^{-2}$.*

Note that the second part of Corollary 2 is, strictly speaking, not a consequence of Theorem 14 but only of its proof.

Assumption $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ is local, in the sense that the constants $\kappa_{J_0}$ and $\mu_{J_0}$ (or their mere existence) may highly depend on the subset $J_0$. For a given $\lambda$, the best choice for $J_0$ in Inequalities (4.5) and (4.7) depends thus on the interaction between these constants and the value of $\lambda$ itself. Note that the assumptions of Theorem 14 are reasonable as the next section gives conditions for which Assumption $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ holds simultaneously with the same constant $\kappa$ and $\mu$ for all subsets $J_0$ of the same size.

**Results under global assumptions**

As usual, when $P > N$, properties of the Dantzig estimate can be derived from assumptions on the structure of the dictionary $\mathcal{D}$. For $l \in \mathbb{N}$, we denote

$$\phi_{\min}(l) = \min_{|J| \leq l} \min_{\substack{\lambda \in \mathbb{R}^P \\ \lambda_J \neq 0}} \frac{\|f_{\lambda_J}\|_2^2}{\|\lambda_J\|_{\ell_2}^2} \qquad \text{and} \qquad \phi_{\max}(l) = \max_{|J| \leq l} \max_{\substack{\lambda \in \mathbb{R}^P \\ \lambda_J \neq 0}} \frac{\|f_{\lambda_J}\|_2^2}{\|\lambda_J\|_{\ell_2}^2}.$$

These quantities correspond to the *restricted* eigenvalues of the Gram matrix $G$. Assuming that $\phi_{\min}(l)$ and $\phi_{\max}(l)$ are close to 1 means that every set of columns of $G$ with cardinality less than $l$ behaves like an orthonormal system. We also consider the restricted correlations

$$\theta_{l,l'} = \max_{\substack{|J| \leq l \\ |J'| \leq l' \\ J \cap J' = \emptyset}} \max_{\substack{\lambda, \lambda' \in \mathbb{R}^P \\ \lambda_J \neq 0, \lambda'_{J'} \neq 0}} \frac{\langle f_{\lambda_J}, f_{\lambda'_{J'}} \rangle}{\|\lambda_J\|_{\ell_2} \|\lambda'_{J'}\|_{\ell_2}}.$$

Small values of $\theta_{l,l'}$ mean that two disjoint sets of columns of $G$ with cardinality less than $l$ and $l'$ span nearly orthogonal spaces. We will use one of the following assumptions considered by Bickel, Ritov, and Tsybakov [BRT09].

- **Assumption 1** For some integer $1 \leq s \leq P/2$, we have

$$\phi_{\min}(2s) > \theta_{s,2s}. \tag{A1(s)}$$

  Oracle inequalities of the Dantzig selector were established under this assumption in the parametric linear model by Candès and Tao [CT07]. It was also considered by Bickel, Ritov, and Tsybakov [BRT09] for non-parametric regression and for the Lasso estimate. The next assumption, proposed by Bickel, Ritov, and Tsybakov [BRT09], constitutes an alternative to Assumption 1.

- **Assumption 2** For some integers $s$ and $l$ such that

$$1 \leq s \leq \frac{P}{2}, \quad l \geq s \quad \text{and} \quad s + l \leq P, \tag{4.8}$$

  we have

$$l \phi_{\min}(s + l) > s \phi_{\max}(l). \tag{A2(s,l)}$$

  If Assumption 2 holds for $s$ and $l$ such that $l \gg s$, then Assumption 2 means that $\phi_{\min}(l)$ cannot decrease at a rate faster than $l^{-1}$ and this condition is related to the *incoherent designs* condition stated by Meinhausen and Yu [MY09].

In the sequel, we set, under Assumption 1,

$$\kappa_{1,s} = \sqrt{\phi_{\min}(2s)}\left(1 - \frac{\theta_{s,2s}}{\phi_{\min}(2s)}\right) > 0, \quad \mu_{1,s} = \frac{\theta_{s,2s}}{\sqrt{\phi_{\min}(2s)}}$$

and under Assumption 2,

$$\kappa_{2,s,l} = \sqrt{\phi_{\min}(s+l)}\left(1 - \sqrt{\frac{\phi_{\max}(l)}{\phi_{\min}(s+l)}}\sqrt{\frac{s}{l}}\right) > 0, \quad \mu_{2,s,l} = \sqrt{\phi_{\max}(l)}\sqrt{\frac{s}{l}}.$$

Now, to apply Theorem 14, we need to check $(LA(J_0, \kappa_{J_0}, \mu_{J_0}))$ for some subset $J_0$ of $\{1, \dots, P\}$. Either Assumption 1 or Assumption 2 implies this assumption. Indeed, we have the following result.

**Proposition 9.** *Let $s$ and $l$ two integers satisfying (4.8). We suppose that (A1(s)) or (A2(s,l)) holds. Let $J_0 \subset \{1, \dots, P\}$ of size $|J_0| = s$ and $\lambda \in \mathbb{R}^P$, then Assumption $LA(J_0, \kappa_{s,l}, \mu_{s,l})$, namely,*

$$\|f_\lambda\|_2 \geq \kappa_{s,l}\frac{\|\lambda_{J_0}\|_{\ell_1}}{\sqrt{s}} - \frac{\mu_{s,l}}{\sqrt{s}}\left(\|\lambda_{J_0^C}\|_{\ell_1} - \|\lambda_{J_0}\|_{\ell_1}\right)_+,$$

*holds with $\kappa_{s,l} = \kappa_{1,s}$ and $\mu_{s,l} = \mu_{1,s}$ under (A1(s)) (respectively $\kappa_{s,l} = \kappa_{2,s,l}$ and $\mu_{s,l} = \mu_{2,s,l}$ under (A2(s,l)). If (A1(s)) and (A2(s,l)) are both satisfied, $\kappa_{s,l} = \max(\kappa_{1,s}, \kappa_{2,s,l})$ and $\mu_{s,l} = \min(\mu_{1,s}, \mu_{2,s,l})$.*

Proposition 9 proves that Theorem 14 can be applied under Assumptions 1 or 2. In addition, the constants $\kappa_{s,l}$ and $\mu_{s,l}$ are the same for all subset $J_0$ of size $|J_0| = s$. From Theorem 14, we deduce the following result.

**Theorem 15.** *With probability at least $1 - C_1(\varepsilon, \delta, \gamma)P^{1-\frac{\gamma}{1+\varepsilon}}$, for any two integers $s$ and $l$ satisfying (4.8) such that (A1(s)) or (A2(s,l)) holds, we have for any $\alpha > 0$,*

$$\|\hat{s}^D - s_0\|_2^2 \leq \inf_{\lambda \in \mathbb{R}^P} \inf_{\substack{J_0 \subset \{1,\dots,P\} \\ |J_0| = s}} \left\{ \|s_\lambda - s_o\|_2^2 + \alpha\left(1 + \frac{2\mu_{s,l}}{\kappa_{s,l}}\right)^2 \frac{\Lambda(\lambda, J_0^c)^2}{s} + 16s\left(\frac{1}{\alpha} + \frac{1}{\kappa_{s,l}^2}\right)\|\eta_\gamma\|_{\ell_\infty}^2 \right\}$$

*where*

$$\Lambda(\lambda, J_0^c) = \|\lambda_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+}{2},$$

*and $\kappa_{s,l}$ and $\mu_{s,l}$ are defined as in Proposition 9.*

Remark that the best subset $J_0$ of cardinal $s$ in Theorem 15 can be easily chosen for a given $\lambda$: it is given by the set of the $s$ largest coordinates of $\lambda$. This was not necessarily the case in Theorem 14 for which a different subset may give a better local condition and then may provide a smaller bound. If we further assume the mild assumption (4.6) on the sup norm of the dictionary introduced in the previous section, we deduce the following result.

**Corollary 3.** *With probability at least $1 - C_1(\varepsilon, \delta, \gamma)P^{1-\frac{\gamma}{1+\varepsilon}}$, if (4.6) is satisfied, for any integers $s$ and $l$ satisfying (4.8) such that (A1(s)) or (A2(s,l)) holds, we have for any $\alpha > 0$, any $\lambda$ that satisfies the adaptive Dantzig constraint, and for the best subset $J_0$ of cardinal $s$ (that corresponds to the $s$ largest coordinates of $\lambda$ in absolute value),*

$$\|\hat{s}^D - s_0\|_2^2 \leq \|s_\lambda - s_0\|_2^2 + \alpha c_2(1 + \kappa_{s,l}^{-2}\mu_{s,l}^2)\frac{\|\lambda_{J_0^C}\|_{\ell_1}^2}{s} + c_3(\alpha^{-1} + \kappa_{s,l}^{-2})s\|f_0\|_\infty^2\frac{\log M}{n}, \tag{4.9}$$

*where $c_2$ is an absolute constant, $c_3$ depends on $c_1$ and $\gamma$, and $\kappa_{s,l}$ and $\mu_{s,l}$ are defined as in Proposition 9.*

Note that, when $\lambda$ is $s$-sparse so that $\lambda_{J_0^C} = 0$, the oracle inequality (4.9) corresponds to the classical oracle inequality obtained in parametric frameworks (see Candès and Plan [CP09]; Candès and Tao [CT07] for instance) or in non-parametric settings. See, for instance Bunea [Bu09]; Bunea, Tsybakov, and Wegkamp [BTW06; BTW07a; BTW07b; BTW07c] or Geer [Ge08] but in these works, the functions of the dictionary are assumed to be bounded by a constant independent of $M$ and $n$. So, the adaptive Dantzig estimate requires weaker conditions since under (4.6), $\|\varphi_p\|_\infty$ can go to $\infty$ when $n$ grows. This point is capital for practical purposes, in particular when wavelet bases are considered.

### 4.2.4 Connections between the Dantzig and Lasso estimates

We show in this section the strong connections between Lasso and Dantzig estimates, which has already been illustrated by Bickel, Ritov, and Tsybakov [BRT09] for non-parametric regression models. By choosing convenient random weights depending on $\eta_\gamma$ for $\ell_1$-minimization, the Lasso estimate satisfies the adaptive Dantzig constraint. More precisely, we consider the Lasso estimator given by the solution of the following minimization problem

$$\hat{\lambda}^{L,\gamma} = \underset{\lambda \in \mathbb{R}^P}{\operatorname{argmin}} \left\{ \frac{1}{2} R(\lambda) + \sum_{p=1}^{P} \eta_{\gamma,p} |\lambda_p| \right\}, \tag{4.10}$$

where

$$R(\lambda) = \|s_\lambda\|_2^2 - \frac{2}{N} \sum_{i=1}^{n} s_\lambda(X_i).$$

Note that $R(\cdot)$ is the quantity minimized in unbiased estimation of the risk. For simplifications, we write $\hat{\lambda}^L = \hat{\lambda}^{L,\gamma}$. We denote $\hat{f}^L = f_{\hat{\lambda}^L}$. As said in Introduction, classical Lasso estimates are defined as the minimizer of expressions of the form

$$\left\{ \frac{1}{2} R(\lambda) + \eta \sum_{p=1}^{P} |\lambda_p| \right\},$$

where $\eta$ is proportional to $\sqrt{\frac{\log P}{N}}$. So, $\hat{\lambda}^L$ appears as a data-driven version of classical Lasso estimates.

The first order condition for the minimization of the expression given in (4.10) corresponds exactly to the adaptive Dantzig constraint and thus Theorem 15 always applies to $\hat{\lambda}^L$. Working along the lines of the proof of Theorem 15, one can prove a slightly stronger result.

**Theorem 16.** *With probability at least $1 - C_1(\varepsilon, \delta, \gamma) P^{1 - \frac{\gamma}{1+\varepsilon}}$, for any integers $s$ and $l$ satisfying (4.8) such that (A1(s)) or (A2(s,l)) holds, we have, for any $J_0$ of size $s$ and for any $\alpha > 0$,*

$$\left| \|\hat{s}^D - s_0\|_2^2 - \|\hat{s}^L - s_0\|_2^2 \right| \leq \alpha \left( 1 + \frac{2\mu_{s,l}}{\kappa_{s,l}} \right)^2 \frac{\|\hat{\lambda}^L_{J_0^C}\|_{\ell_1}^2}{s} + 16 s \left( \frac{1}{\alpha} + \frac{1}{\kappa_{s,l}^2} \right) \|\eta_\gamma\|_{\ell_\infty}^2$$

*where $\kappa_{s,l}$ and $\mu_{s,l}$ are defined as in Proposition 9.*

To extend this theoretical result, numerical performances of the Dantzig and Lasso estimates have been performed in [*Art-BLPR11*].

### 4.2.5 Calibration

We present here only the results concerning the calibration of the previous estimate. We show that the sufficient condition $\gamma > 1$ is *almost* a necessary condition since we derive a special and very simple framework in which Lasso and Dantzig estimates cannot achieve the optimal rate if $\gamma < 1$ (*almost* means that the case $\gamma = 1$ remains an open question). Let us describe this simple framework. The dictionary $\mathcal{D}$ considered in this section is the orthonormal Haar system:

$$\mathcal{D} = \left\{ \phi_{jk} : \quad -1 \leq j \leq j_0,\ 0 \leq k < 2^j \right\},$$

with $\phi_{-10} = \mathbf{1}_{[0,1]}$, $2^{j_0+1} = n$, and for $0 \leq j \leq j_0$, $0 \leq k \leq 2^j - 1$,

$$\phi_{jk} = 2^{j/2} \left( 1_{[k/2^j, (k+0.5)/2^j]} - 1_{[(k+0.5)/2^j, (k+1)/2^j]} \right).$$

In this case, $P = N$ and, since functions of $\mathcal{D}$ are orthonormal, the Gram matrix $G$ is the identity. Thus, the Lasso and Dantzig estimates both correspond to the soft thresholding rule:

$$\hat{s}^D = \hat{s}^L = \sum_{p=1}^{p} \operatorname{sign}(\hat{\beta}_p) \left( |\hat{\beta}_p| - \eta_{\gamma,p} \right) 1_{\{|\hat{\beta}_p| > \eta_{\gamma,p}\}} \varphi_p.$$

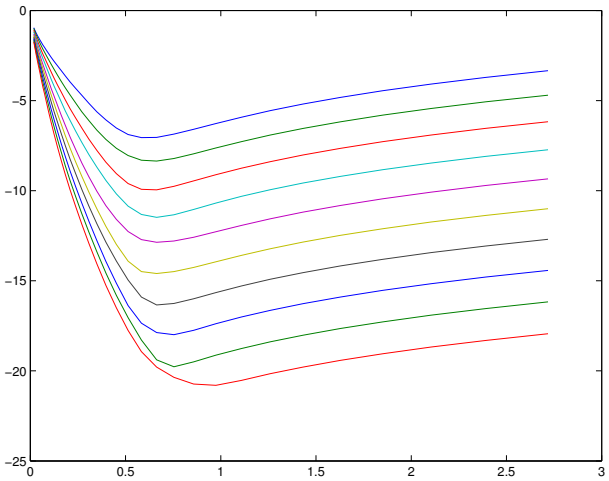



Figure 4.1: Graphs of $\gamma \mapsto \log_2(\overline{R_n}(\gamma))$ for $n = 2^J$ with, from top to bottom, $J = 4, 5, 6, \ldots, 13$

Now, our goal is to estimate $f_0 = \phi_{-10} = \mathbf{1}_{[0,1]}$ by using $\hat{f}^D$ depending on $\gamma$ and to show the influence of this constant. Unlike previous results stated in probability, we consider the expectation of the $\mathbb{L}_2$-risk:

**Theorem 17.** *On the one hand, if $\gamma > 1$, there exists a constant $C$ such that*

$$\mathbb{E}\|\hat{s}^D - s_0\|_2^2 \leq \frac{C \log n}{n}.$$

*On the other hand, if $\gamma < 1$, there exist a constant $c$ and $\delta < 1$ such that*

$$\mathbb{E}\|\hat{s}^D - s_0\|_2^2 \geq \frac{c}{n^\delta}.$$

This result shows that choosing $\gamma < 1$ is a bad choice in our setting. Indeed, in this case, the Lasso and Dantzig estimates cannot estimate a very simple signal ($s_0 = \mathbf{1}_{[0,1]}$) at a convenient rate of convergence.

A small simulation study is carried out to strengthen this theoretical asymptotic result. Performing our estimation procedure 100 times, we compute the average risk $\overline{R_n}(\gamma)$ for several values of the Dantzig constant $\gamma$ and several values of $n$. This computation is summarized in Figure 4.1 which displays the logarithm of $\overline{R_n}(\gamma)$ for $n = 2^J$ with, from top to bottom, $J = 4, 5, 6, \ldots, 13$ on a grid of $\gamma$'s around 1. To discuss our results, we denote by $\gamma_{\min}(n)$ the best $\gamma$: $\gamma_{\min}(n) = \operatorname{argmin}_{\gamma>0} \overline{R_n}(\gamma)$. We note that $1/2 \leq \gamma_{\min}(n) \leq 1$ for all values of $n$, with $\gamma_{\min}(n)$ getting closer to 1 as $n$ increases. Taking $\gamma$ too small strongly deteriorates the performance while a value close to 1 ensures a risk withing a factor 2 of the optimal risk. The assumption $\gamma > 1$ giving a theoretical control on the quadratic error is thus not too conservative. Following these results, we set $\gamma = 1.01$ in our numerical experiments in the next subsection.

## 4.3 Copula estimation by wavelet thresholding

We continue this chapter on density estimation by the study of a related, but different, problem: copula estimations. In risk management, in the areas of finance, insurance and climatology, for example, this new tool has been developed to model the dependence structure of data. It comes from the seminal results of Sklar [Sk59]

**Theorem 18** (Sklar (1959))**.** *Let $d \geq 2$ and $H$ be a $d-$variate distribution function. If each margin $S_m$, $m = 1, \ldots d$, of $H$ is continuous, a unique $d-$variate function $C$ with uniform margins $\mathcal{U}_{[0,1]}$ exists, so that*

$$\forall (x_1, \ldots, x_d) \in \mathbb{R}^d,\ H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

The distribution function $C$ is called **the copula** associated with the distribution $H$. Sklar's Theorem allows us to separately study the laws of the coordinates $X^m, m = 1, \dots d$, of any vector $X$, and the dependence between the coordinates.

The *copula model* has been extensively studied within a parametric framework. Numerous classes of parametric copulas, parametric distribution functions $C$, have been proposed. For instance there is the elliptic family, which contains the Gaussian copulas and the Student copulas, and the Archimedian family, which contains the Gumbel copulas, the Clayton copulas and the Frank copulas. The first step of such a parametric approach is to select the parametric family of the copula being considered. This is a *modeling* task that may require finding new copula and methodologies to simulate the corresponding data. Usual *statistical inference* (estimation of the parameters, goodness-of-fit test, etc) can only take place in a second step. Both tasks have been extensively studied.

With F. Autin and K. Tribouley [*Art-ALPT10*], we propose to study the copula model within a non-parametric framework. Our aim is to make very mild assumption about the copula. Thus, contrary to the parametric setting, no *a priori* model of the phenomenon is needed. For practitioners, non-parametric estimators could be seen as a *benchmark* that makes it possible to select the right parametric family by comparing them to an *agnostic* estimate. In fact, most of the time, practitioners observe the scatter plot of $\{(X_i^1, X_i^2), i = 1, \dots, N\}$, or $\{(R_i^1, R_i^2), i = 1, \dots, N\}$ where $R^1$ and $R^2S$ are the rank statistics of respectively $(X_i^1)$ and $(X_i^2)$, and then attempt, on the basis of these observations, only to guess the family of parametric copulas the target copula belongs to. Providing good non-parametric estimators of the copula makes this task easier and provides a more rigorous way to describe the copula.

In our study, we propose non-parametric procedures to estimate the *copula density* $c$ associated with the copula $C$. More precisely, we consider the following model. We assume that we are observing an $N$-sample $(X_1^1, \dots, X_1^d), \dots, (X_N^1, \dots, X_N^d)$ of independent data with the same distribution $S_0$ (and the same density $s_0$) as $(X^1, \dots, X^d)$. Referring to the margins of the coordinates of the vector $(X^1, \dots, X^d)$ as $F_1, \dots, F_d$, we are interested in estimating the copula density $c_0$ defined as the derivative (if it exists) of the copula distribution

$$c_0(u_1, \dots, u_d) = \frac{s_0(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \dots f_d(F_d^{-1}(u_d))}$$

where $F_p^{-1}(u_p) = \inf\{x \in R : F_p(x) \geq u_p\}$, $1 \leq p \leq d$ and $u = (u_1, \dots, u_d) \in [0,1]^d$. This would be a classical density model if the margins, and thus the direct observations, $(U_i^1 = F_1(X_i^1), \dots, U_i^d = F_d(X_i^d))$ for $i = 1, \dots, n$ , were known. Unfortunately, this is not the case. We can observe that this model is somewhat similar to the non-parametric regression model with unknown random design studied by Kerkyacharian and Picard [KP04] with their *warped wavelet families*.

Instead of using the $\ell_1$ strategy of the previous section, we use a projection based approach. More precisely, we propose two wavelet-based methods, a *Local Thresholding Method* and a *Global Thresholding Method*, that are both extensions of the methods studied by Donoho and Johnstone [DJ95]; Donoho, Johnstone, Kerkyacharian, and Picard [Do+96]; Kerkyacharian, Picard, and Tribouley [KPT96] in the classical density estimation framework. In the first, unbiased estimate of the wavelet coefficients are thresholded independently while in the second one they are thresholded levelwise. We first measure the performance for both estimators on all copula densities that are bounded and that belong to a very large class of regularity. The good behavior of our procedures is due to the approximation properties of the wavelet basis. A regular copula can be approximated by few non zero-wavelet coefficients leading to estimators with both a small bias and small variance. The wavelet representation is connected to well-known regularity spaces: Besov spaces, in particular, that contain Sobolev spaces or Holder spaces, can be defined through the wavelet coefficients. Our first result in this setting is that the rate of convergence of our estimators are:

1. optimal in the minimax sense (up to a logarithmic factor),
2. the same as in the standard density model. Using pseudo data instead of direct observations does not damage the quality of the procedures.

It should be observed that the same behavior also arises for linear wavelet procedures (see Genest, Masiello, and Tribouley [GMT09]). However, the linear procedure is not adaptive in the sense that we

need to know the regularity index of the copula density to obtain optimal procedures. We provide here a solution to this drawback.

Following the maxiset approach, we then characterize the precise set of copula densities estimated at a given polynomial rate for our procedures. We verify that the local one outperforms the others, in the sense that this is the procedure for which the set of copula densities estimated at a given rate is the largest.

One of the main difficulties of copula density estimation lies in the fact that most of the pertinent information is located near the boundaries of $[0,1]^d$ (at least for the most common copulas like the Gumbel copula or the Clayton copula). In the theoretical construction, we use a family of wavelets especially designed for this case: they extend only within the compact set $[0,1]^d$, do not thus cross the boundary and are optimal in terms of the approximation. In the practical construction, boundaries remain an issue. Note that the theoretically optimal wavelets are rarely implemented and when they are, they are not as efficient as in the theory. We propose an appropriate symmetrization/periodization process of the original data here in order to deal with this problem and also enhance the scheme by adding some translation invariance. We numerically verify the good behavior of the proposed scheme for simulated data with the usual parametric copula families. In [*Art-ALPT10*], we illustrate those resylts by an application on financial data by proposing a method to choose the parametric family and the parameters based on a preliminary non-parametric estimator used as a benchmark. Our contribution is thus also to propose an implementation that is very easy to use and that provides good estimators.

### 4.3.1 Estimation procedures

For a copula density $c_0$ belonging to $L_2([0,1]^d)$, estimation of $c_0$ is equivalent to estimation of its wavelet coefficients. It turns out that this can be easily done. Observe that, for any $d-$variate function $\Phi$

$$E_{c_0}(\Phi(U_1,\dots,U_d)) = E_{s_0}\left(\Phi(F_1(X^1),\dots,F_d(X^d))\right)$$

or equivalently

$$\int_{[0,1]^d} \Phi(u)c_0(u)du = \int_{\mathbb{R}^d} \Phi(F_1(x_1),\dots,F_d(x_d))s_0(x_1,\dots,x_d)dx_1\dots dx_d \quad .$$

This means that the wavelet coefficients of the copula density $c_0$ on the wavelet basis are equal to the coefficients of the joint density $s_0$ on the warped wavelet family

$$\{\phi_{j_0,k}(F_1(\cdot),\dots,F_d(\cdot)), \psi^{\epsilon}_{j,\ell}(F_1(\cdot),\dots,F_d(\cdot)) \\ |j \geq j_0, k \in \{0,\dots,2^{j_0}\}^d, \ell \in \{0,\dots,2^j\}^d, \epsilon \in S_d\}.$$

The corresponding empirical coefficients are

$$\widehat{c_{j_0,k}} = \frac{1}{n}\sum_{i=1}^{n} \phi_{j_0,k}(F_1(X_i^1),\dots,F_d(X_i^d))$$

and

$$\widehat{c^{\epsilon}_{j,k}} = \frac{1}{n}\sum_{i=1}^{n} \psi^{\epsilon}_{j,k}(F_1(X_i^1),\dots,F_d(X_i^d)). \tag{4.11}$$

These coefficients cannot be evaluated since the distributions functions associated to the margins $F_1,\dots,F_d$ are unknown. We propose to replace these unknown distributions functions by their corresponding empirical distributions functions $\widehat{F_1},\dots\widehat{F_d}$. The modified empirical coefficients are

$$\widetilde{c_{j_0,k}} = \frac{1}{n}\sum_{i=1}^{n} \phi_{j_0,k}(\widehat{F_1}(X_i^1),\dots,\widehat{F_d}(X_i^d)) = \frac{1}{n}\sum_{i=1}^{n} \phi_{j_0,k}\left(\frac{R_i^1-1}{n},\dots,\frac{R_i^d-1}{n}\right)$$

and

$$\widetilde{c_{j,k}^{\epsilon}} = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}^{\epsilon}(\widehat{F_1}(X_i^1), \ldots, \widehat{F_d}(X_i^d)) = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}^{\epsilon}\left(\frac{R_i^1 - 1}{n}, \ldots, \frac{R_i^d - 1}{n}\right)$$

where $R_i^p$ denotes the rank of $X_i^p$ for $p = 1, \ldots, d$

$$R_i^p = \sum_{l=1}^{n} \mathbf{1}\{X_l^p \leq X_i^p\}.$$

The most natural way to estimate the density $c_0$ is to reconstruct a function from the modified empirical coefficients. We consider here the very general family of **truncated estimators** of $c_0$ defined by

$$\widetilde{c_T} := \widetilde{c_T}(j_n, J_n) = \sum_k \widetilde{c_{j_n,k}}\phi_{j_n,k} + \sum_{j=j_n}^{J_n}\sum_{k,\epsilon} \omega_{j,k}^{\epsilon}\widetilde{c_{j,k}^{\epsilon}}\psi_{j,k}^{\epsilon}, \tag{4.12}$$

where the indices $(j_n, J_n)$ are such that $j_n \leq J_n$ and where, for any $(j, k, \epsilon)$, $\omega_{j,k}^{\epsilon}$ belongs to $\{0, 1\}$. Notice that $\omega_{j,k}^{\epsilon}$ may or may not depend on the observations.

The later case has been considered by Genest, Masiello, and Tribouley [GMT09] who proposed to use a *linear procedure*

$$\widetilde{c_L} := \widetilde{c_L}(j_n) = \sum_k \widetilde{c_{j_n,k}}\phi_{j_n,k} \tag{4.13}$$

for a suitable choice of $j_n$. The accuracy of this linear procedure relies on the fast uniform decay of the wavelets coefficients across the scale as soon as the function is uniformly regular. The trend at the chosen level $j_n$ becomes a sufficient approximation. The optimal choice of $j_n$ depends on the regularity of the unknown function to be estimated and thus the procedure is not data-driven.

We propose here to use some *non linear procedures* based on hard thresholding (see for instance Kerkyacharian and Picard [KP92], Kerkyacharian and Picard [KP00], and Donoho and Johnstone [DJ95])) that overcome this issue. In hard thresholding procedures, the *small* coefficients are killed by setting the corresponding $\omega_{j,k}^{\epsilon}$ to 0. They differ by the definition of *small*. We study here two strategies: a local one, where each coefficient is considered individually, and a global one, where all the coefficients at the same scale are considered globally.

For a given threshold level $\lambda_n > 0$ and a set of indices $(j_n, J_n)$, the local hard threshold weights $\omega_{j,k}^{\epsilon,L}$ and the global hard threshold weights $\omega_{j,k}^{\epsilon,G}$ are defined respectively by

$$\omega_{j,k}^{\epsilon,HL} = \mathbf{1}\{|\widetilde{c_{j,k}^{\epsilon}}| > \lambda_n\}. \qquad \text{and} \qquad \omega_{j,k}^{\epsilon,HG} = \mathbf{1}\{\sum_k |\widetilde{c_{j,k}^{\epsilon}}|^2 > 2^{jd}\lambda_n^2\} \quad .$$

Let us put $\widetilde{c_{j,k}^{\epsilon,HL}} = \omega_{j,k}^{\epsilon,HL}\widetilde{c_{j,k}^{\epsilon}}$ and $\widetilde{c_{j,k}^{\epsilon,HG}} = \omega_{j,k}^{\epsilon,HG}\widetilde{c_{j,k}^{\epsilon}}$. The corresponding local hard thresholding estimators $\widetilde{c_{HL}}$ and global hard thresholding estimators $\widetilde{c_{HG}}$ are defined respectively by

$$\widetilde{c_{HL}} := \widetilde{c_{HL}}(j_n, J_n, \lambda_n) = \sum_k \widetilde{c_{j_n,k}}\phi_{j_n,k} + \sum_{j=j_n}^{J_n}\sum_{k,\epsilon} \widetilde{c_{j,k}^{\epsilon,HL}}\psi_{j,k}^{\epsilon}. \tag{4.14}$$

and

$$\widetilde{c_{HG}} := \widetilde{c_{HG}}(j_n, J_n, \lambda_n) = \sum_k \widetilde{c_{j_n,k}}\phi_{j_n,k} + \sum_{j=j_n}^{J_n}\sum_{k,\epsilon} \widetilde{c_{j,k}^{\epsilon,HG}}\psi_{j,k}^{\epsilon}. \tag{4.15}$$

The non linear procedures given in (4.14) and (4.15) depend on the level indices $(j_n, J_n)$ and on the threshold value $\lambda_n$. In the next section, we define a criterion to measure the performance of our procedures and explain how to choose those parameters to achieve optimal performance.

### 4.3.2 Minimax Results

Provided the wavelet used is regular enough, Genest, Masiello, and Tribouley [GMT09] prove that the linear procedure $\widetilde{c_L} = \widetilde{c_L}(j_n^*)$ defined in (4.13) is minimax optimal on the Besov body $B^s_{2\infty}$ for all $s > 0$ provided $j_n^*$ is chosen so that:

$$2^{j_n^*-1} < n^{\frac{1}{2s+d}} \leq 2^{j_n^*}.$$

As hinted in a previous section, this result is not fully satisfactory because the optimal procedure depends on the regularity $s$ of the density which is generally unknown.

The thresholding procedures described in (4.14) and (4.15) do not suffer from this drawback: the same choice of parameters $j_n, J_N$ and $\lambda_n$ yields an almost minimax optimal estimator simultaneously for any $B^s_{2,\infty}$. The following theorem (which is a direct consequence of Theorem 20 established in the following section) ensure indeed that

**Theorem 19.** *Assume that the wavelet is continuously differentiable and let $s > 0$. For any choice of level $j_n$ and $J_n$ and threshold $\lambda_n$ such that*

$$2^{j_n-1} < (\log(n))^{1/d} \leq 2^{j_n}, \quad 2^{J_n-1} < \left(\frac{n}{\log n}\right)^{1/d} \leq 2^{J_n}, \quad \lambda_n = \sqrt{\frac{\kappa \log(n)}{n}}$$

*for some $\kappa$ large enough,*

$$\forall s > 0, \quad c \in \mathcal{B}^s_{2\infty} \cap L_\infty([0,1]^d) \implies \sup_n \left(\frac{n}{\log(n)}\right)^{\frac{2s}{2s+d}} E\|\widetilde{c} - c_0\|_2^2 < \infty$$

*where $\widetilde{c}$ stands either for the hard local thresholding procedure $\widetilde{c_{HL}}(j_n, J_n, \lambda_n)$ or for the hard global thresholding procedure $\widetilde{c_{HG}}(j_n, J_n, \lambda_n)$.*

Observe that, when $s > d/2$, one has the embedding $\mathcal{B}^s_{2\infty} \subsetneq L_\infty([0,1]^d)$, and thus the assumption $c_0 \in \mathcal{B}^s_{2\infty} \cap L_\infty([0,1]^d)$ in Theorem 19 could be replaced simply by $c_0 \in \mathcal{B}^s_{2\infty}$.

We immediately deduce

**Corollary 4.** *The hard local thresholding procedure $\widetilde{c_{HL}}$ and the hard global thresholding procedure $\widetilde{c_{HG}}$ are adaptive minimax optimal up to a logarithmic factor on the Besov bodies $\mathcal{B}^s_{2\infty}$ for the quadratic loss function.*

Notice that this logarithmic factor is nothing but the classical *price* of adaptivity.

As always, the minimax theory requires the choice of the functional space $\mathcal{F}$ or of a sequence of functional spaces $\mathcal{F}_s$. The arbitrariness of this choice is the main drawback of the minimax approach. Indeed, Corollary 4 establishes that no other procedures could be uniformly better on the spaces $\mathcal{B}^s_{2\infty}$ but it does not address two important questions. What about a different choice of spaces? Both of our thresholding estimators achieve the minimax rate on the spaces $\mathcal{B}^s_{2,\infty}$ but is there a way to distinguish their performance? To answer to these questions, we propose to explore the **maxiset approach**.

### 4.3.3 Maxiset Results

We define here the local weak Besov spaces $\mathcal{W}_L(r)$ and the global weak Besov spaces $\mathcal{W}_G(r)$ by

**Definition 4 (Local weak Besov spaces).** *For any $0 < r < 2$, a function $c \in L_2([0,1]^d)$ belongs to the local weak Besov space $\mathcal{W}_L(r)$ if and only if its sequence of wavelet coefficients $c^\epsilon_{j,k}$ satisfies the following equivalent properties:*

- $\sup_{0<\lambda\leq 1} \lambda^{r-2} \sum_{j\geq 0} \sum_{k,\epsilon} (c^\epsilon_{j,k})^2 \mathbf{1}\{|c^\epsilon_{j,k}| \leq \lambda\} < \infty,$
- $\sup_{0<\lambda\leq 1} \lambda^{r} \sum_{j\geq 0} \sum_{k,\epsilon} \mathbf{1}\{|c^\epsilon_{j,k}| > \lambda\} < \infty.$

and

**Definition 5** (Global weak Besov spaces)**.** *For any $0 < r < 2$, a function $c \in L_2([0,1]^d)$ belongs to the global weak Besov space $\mathcal{W}_G(r)$ if and only if its sequence of wavelet coefficients $c^\epsilon_{j,k}$ satisfies the following equivalent properties:*

- $\sup_{0<\lambda\leq 1} \lambda^{r-2} \sum_{j\geq 0} \sum_{k,\epsilon} (c^\epsilon_{j,k})^2 \mathbf{1}\{\sum_k (c^\epsilon_{j,k})^2 \leq 2^{dj}\lambda^2\} < \infty,$
- $\sup_{0<\lambda\leq 1} \lambda^{r} \sum_{j\geq 0} 2^{dj} \sum_{\epsilon} \mathbf{1}\{\sum_k (c^\epsilon_{j,k})^2 > 2^{dj}\lambda^2\} < \infty.$

As in the definition of the Besov bodies, the definition depends on the wavelet basis. However, as established by Meyer [Me90] and Cohen, De Vore, Kerkyacharian, and Picard [Co+01], this dependency is quite weak. Note that the equivalences between the properties used in the definitions of the weak Besov spaces can be proved as in Cohen, De Vore, Kerkyacharian, and Picard [Co+01].

These spaces are clearly related to the Besov bodies $\mathcal{B}^s_{2,\infty}$. Indeed some computation proves that $\mathcal{B}^s_{2,\infty} \subset \mathcal{W}_G\left(\frac{2d}{2s+d}\right)$ and $\mathcal{B}^s_{2,\infty} \subset \mathcal{W}_L\left(\frac{2d}{2s+d}\right)$. We have obtained the following strict inclusion property

**Proposition 10.** *For any $0 < r < 2$, $\quad \mathcal{W}_G(r) \subsetneq \mathcal{W}_L(r)$.*

We study the maxisets of the linear procedure and of the thresholding procedures. We focus on the near minimax optimal procedures that is we use the only following choices of parameters:

$$2^{j_n-1} < (\log(n))^{1/d} \leq 2^{j_n}, \qquad 2^{J_n-1} < \left(\frac{n}{\log(n)}\right)^{1/d} \leq 2^{J_n}$$

$$2^{j^*_n-1} < \left(\frac{n}{\log(n)}\right)^{\frac{1}{2s+d}} \leq 2^{j^*_n}, \qquad \lambda_n = \sqrt{\frac{\kappa \log(n)}{n}}$$

for some $\kappa > 0$ and we study the linear estimator $\widetilde{c_L} = \widetilde{c_L}(j^*_n)$, the local thresholding estimator $\widetilde{c_{HL}} = \widetilde{c_{HL}}(j_n, J_n, \lambda_n)$ and the global thresholding estimator $\widetilde{c_{HG}} = \widetilde{c_{HG}}(j_n, J_n, \lambda_n)$.

Let us fix $s > 0$. We focus on the rate $r_n = \left(n^{-1}\log(n)\right)^{2s/(2s+d)}$ which is the (near) minimax rate achieved on the space $B^s_{2\infty}$. The following theorem exhibits the maxisets of the three procedures with this target rate $r_n$.

**Theorem 20.** *Let $s > 0$, and assume that $c \in L_\infty([0,1]^d)$. For a large enough $\kappa$, we get*

$$\sup_n \left(\frac{n}{\log(n)}\right)^{\frac{2s}{2s+d}} E\|\widetilde{c_L} - c\|_2^2 < \infty \iff c \in \mathcal{B}^s_{2\infty}, \tag{4.16}$$

$$\sup_n \left(\frac{n}{\log(n)}\right)^{\frac{2s}{2s+d}} E\|\widetilde{c_{HL}} - c\|_2^2 < \infty \iff c \in \mathcal{B}^{\frac{ds}{2s+d}}_{2\infty} \cap \mathcal{W}_L(\frac{2d}{2s+d}), \tag{4.17}$$

$$\sup_n \left(\frac{n}{\log(n)}\right)^{\frac{2s}{2s+d}} E\|\widetilde{c_{HG}} - c\|_2^2 < \infty \iff c \in \mathcal{B}^{\frac{ds}{2s+d}}_{2\infty} \cap \mathcal{W}_G(\frac{2d}{2s+d}). \tag{4.18}$$

Note that the same spaces arise if we assume that the marginal distributions are known (see Autin, Le Pennec, and Tribouley [*Unpub-ALPT08*]). This is also a nice result to prove that the lack of direct observations does not make the problem harder.

The following strict embedding,

$$\mathcal{B}^s_{2\infty} \quad \subsetneq \quad \mathcal{B}^{\frac{ds}{2s+d}}_{2\infty} \cap \mathcal{W}_G(\frac{2d}{2s+d})$$

implies

**Corollary 5.** *Let $s > 0$ and let us consider the target rate*

$$r_n = \left(\frac{\log(n)}{n}\right)^{\frac{2s}{2s+d}} \quad . \tag{4.19}$$

*Then we get*

$$\mathcal{MS}(\widetilde{c_L}, r_n) \subsetneq \mathcal{MS}(\widetilde{c_{HG}}, r_n) \subsetneq \mathcal{MS}(\widetilde{c_{HL}}, r_n).$$

In other words, in the maxiset point of view and when the quadratic loss is considered, the thresholding rules outperform the linear procedure. Moreover, the hard local thresholding estimator $\widetilde{c_{HL}}$ appears to be the best estimator among the considered procedures since it strictly outperforms the hard global thresholding estimator $\widetilde{c_{HG}}$.

Numerical aspects of the thresholding estimation have been considered. We refer to the article for those aspects.

To summarize, when the unknown copula density is uniformly regular (in the sense that it is not too peaky on the corners), the thresholding wavelet procedures associated with the symmetrization extension produce good non parametric estimation. If the copula presents strong peaks at the corner (for instance the Clayton copula which has a large Kendall tau), our method is much less efficient. We think that improvements will come from a new family of wavelet adapted to singularity on the corners.

As shown in our numerical experiments, those procedures can be used in the popular two steps decision procedure: first use a nonparametric estimator to decide which copula family to consider and second estimate the parameters within this family. We do not claim that the plug-in method used with our estimate as a benchmark is optimal (it is slightly biased), but it provides a simple single framework. We did not study here the properties of such an estimator or of the corresponding goodness-of-fit test problem and refer to Gayraud and Tribouley [GT11] for this issue.

# Chapter 5

# Conditional density estimation

## 5.1 Conditional density, maximum likelihood and model selection

This is a model I have studied the most during the last three years at SELECT. In this framework, we observe $N$ pairs $((X_i, Y_i))_{1\leq i\leq N}$ of random variables, we are interested in estimating the law of the second one $Y_i \in \mathcal{Y}$, called variable, conditionally to the first one $X_i \in \mathcal{X}$, called covariate. In this study, we assume that the pairs $(X_i, Y_i)$ are independent while $Y_i$ depends on $X_i$ through its law. More precisely, we assume that the covariates $X_i$'s are independent but not necessarily identically distributed. The assumptions on the $Y_i$s are stronger: we assume that, conditionally to the $X_i$'s, they are independents and each variable $Y_i$ follows a law with density $s_0(\cdot|X_i)$ with respect to a common known measure $\mathrm{d}\lambda$. Our goal is to estimate this two-variable conditional density function $s_0(\cdot|\cdot)$ from the observations.

This problem has been introduced by Rosenblatt [Ro69] in the late 60's. He considered a stationary framework in which $s_0(y|x)$ is linked to the supposed existing densities $s_{0'}(x)$ and $s_{0''}(x, y)$ of respectively $X_i$ and $(X_i, Y_i)$ by

$$s_0(y|x) = \frac{s_{0''}(x, y)}{s_{0'}(x)},$$

and proposed a plugin estimate based on kernel estimation of both $s_{0'}(x)$ and $s_{0''}(x, y)$. Few other references on this subject seem to exist before the mid 90's with a study of a spline tensor based maximum likelihood estimator proposed by Stone [St94] and a bias correction of Rosenblatt's estimator due to Hyndman, Bashtannyk, and Grunwald [HBG96].

Kernel based method have been much studied since. For instance, Fan, Yao, and Tong [FYT96] and Gooijer and Zerom [GZ03] consider local polynomial estimator, Hall, Wolff, and Yao [HWY99] study a locally logistic estimator that is later extended by Hyndman and Yao [HY02]. In this setting, pointwise convergence properties are considered, and extensions to dependent data are often obtained. The results depend however on a critical bandwidth that should be chosen according to the regularity of the unknown conditional density. Its practical choice is rarely discussed with the notable exceptions of Bashtannyk and Hyndman [BH01], Fan and Yim [FY04] and Hall, Racine, and Li [HRL04]. Extensions to censored cases have also been discussed for instance by Keilegom and Veraverbeke [KV02]. See for instance Li and Racine [LR07] for a comprehensive review of this topic.

In the approach of Stone [St94], the conditional density is estimated through a parametrized modelization. This idea has been reused since by Györfi and Kohler [GK07] with a histogram based approach, by Efromovich [Ef07; Ef10] with a Fourier basis, and by Brunel, Comte, and Lacour [BCL07] and Akakpo and Lacour [AL11] with piecewise polynomial representation. Those authors are able to control an integrated estimation error: with an integrated total variation loss for the first one and a quadratic distance loss for the others. Furthermore, in the quadratic framework, they manage to construct adaptive estimators, estimators that do not require the knowledge of the regularity to be minimax optimal (up to a logarithmic factor), using respectively a blockwise attenuation principle and a model selection by penalization approach. Note that Brunel, Comte, and Lacour [BCL07] extend their result to censored cases while Akakpo and Lacour [AL11] are able to consider weakly dependent data.

The very frequent use of conditional density estimation in econometrics, see Li and Racine [LR07] for instance, could have provided a sufficient motivation for this study. However it turns out that this work stems from a completely different subject: unsupervised hyperspectral image segmentation. Using the synchrotron beam of Soleil, the IPANEMA platform [*Art-Be+11*], in which S. Cohen is working, is able to acquire high quality hyperspectral images, high resolution images for which a spectrum is measured at each pixel location. This provides rapidly a huge amount of data for which an automatic processing is almost necessary. One of these processings is the segmentation of these images into homogeneous zones, so that the spectral analysis can be performed on fewer places and the geometrical structures can be exhibited. The most classical unsupervised classification method relies on the density estimation of Gaussian mixture by a maximum likelihood principle. Components of the estimated mixtures correspond to classes. In the spirit of Kolaczyk, Ju, and Gopal [KJG05] and Antoniadis, Bigot, and Sachs [ABS08], with S. Cohen, I have extended this method by taking into account the localization of the pixel in the mixing proportions, going thus from density estimation to conditional density estimation. As stressed by Maugis and Michel [MM12a; MM12b], understanding finely the density estimator is crucial to be able to select the right number of classes. This work has been motivated by a similar issue for the conditional density estimation case.

Measuring losses in a conditional density framework can be performed in various way. We focus here on averaged density losses. Namely, let $\ell$ be a density loss and a design on the $X_i$'s, we define a corresponding *tensorized* loss $\ell^{\otimes n}$ by

$$\ell^{\otimes n}(s,t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\ell\left(s(\cdot|X_i), t(\cdot|X_i)\right)\right].$$

Although this loss may seems, at first, artificial, it is the most natural one. Furthermore, it reduces to classical one in several cases:

- If the law of $Y_i$ is independent of $X_i$, that is $s(\cdot|X_i) = s(\cdot)$ and $t(\cdot|X_i) = t(\cdot)$ do not depend on $X_i$, this loss reduces to the classical $\ell(s,t)$.
- If the $X_i$'s are not random but fixed, that is we consider a fixed design case, this loss is the classical fixed design type loss in which there is no expectation.
- If the $X_i$'s are i.i.d., this divergence is nothing but $\mathbb{E}\left[\ell(s(\cdot|X_1), t(\cdot|X_1))\right]$.
- If the density of the law of an $X_i$ is lower and upper bounded on its support then $\mathbb{E}\left[\ell(s(\cdot|X_i), t(\cdot|X_i))\right]$ can be replaced by $\int_{\mathrm{Supp} X_i} \ell(s(\cdot|X_i), t(\cdot|X_i))$, up to a multiplicative constant.

We stress that these types of loss is similar to the one used in the machine-learning community (see for instance Catoni [Ca07] that has inspired our notations). Such kind of losses appears also, but less often, in regression with random design (see for instance Birgé [Bi04]) or in other conditional density estimation studies (see for instance Brunel, Comte, and Lacour [BCL07] and Akakpo and Lacour [AL11]). When $\hat{s}$ is an estimator, or any function that depends on the observation, $KL_\lambda^{\otimes n}(s,\hat{s})$ measures this (random) integrated divergence between $s$ and $\hat{s}$ conditionally to the observation while $\mathbb{E}\left[KL_\lambda^{\otimes n}(s,\hat{s})\right]$ is the average of this random quantity with respect to the observations.

When the losses considered are the quadratic loss or the Kullback-Leibler divergence, two natural contrasts appear: for the quadratic loss,

$$\gamma(s) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\|s(\cdot|X_i)\|_2^2 - s(Y_i|X_i) \tag{5.1}$$

and for the Kullback-Leibler divergence

$$\gamma(s) = \frac{1}{N}\sum_{i=1}^{N} -\log\left(s(Y_i|X_i)\right). \tag{5.2}$$

We refer to Brunel, Comte, and Lacour [BCL07] and Akakpo and Lacour [AL11] for the first case and focus only to the second case.

More precisely, we consider a direct estimation of the conditional density function through this maximum likelihood approach. Although natural, this approach has been considered so far only by Stone [St94] as mentioned before and by Blanchard, Schäfer, Rozenholc, and Müller [Bl+07] in a classification setting with histogram type estimators. Assume we have a set $S_m$ of candidate conditional densities, our estimate $\widehat{s}_m$ is simply the maximum likelihood estimate

$$\widehat{s}_m = \underset{s_m \in S_m}{\operatorname{argmin}} \left( -\sum_{i=1}^{N} \log s_m(Y_i|X_i) \right).$$

Although this estimator may look like a maximum likelihood estimator of the joint density of $(X_i, Y_i)$, it does not generally coincide, even when the $X_i$'s are assumed to be i.i.d., with such an estimator as every function of $S_m$ is assumed to be a conditional density and not a density. The only exceptions are when the $X_i$'s are assumed to be i.i.d. uniform or non random and equal. Our aim is then to analyze the finite sample performance of such an estimator in term of Kullback-Leibler type loss. As often, a trade-off between a bias term measuring the closeness of $s_0$ to the set $S_m$ and a variance term depending on the complexity of the set $S_m$ and on the sample size appears. A good set $S_m$ is thus one for which this trade-off leads to a small risk bound.

For any model $S_m$, a set comprising some candidate conditional densities, we estimate $s_0$ by the conditional density $\widehat{s}_m$ that maximizes the likelihood (conditionally to $(X_i)_{1\leq i\leq N}$) or equivalently that minimizes the opposite of the log-likelihood, denoted -log-likelihood from now on:

$$\widehat{s}_m = \underset{s_m \in S_m}{\operatorname{argmin}} \left( \sum_{i=1}^{N} -\log(s_m(Y_i|X_i)) \right).$$

To avoid existence issue, we should work with almost minimizer of this quantity and define a $\eta$ -log-likelihood minimizer as any $\widehat{s}_m$ that satisfies

$$\sum_{i=1}^{N} -\log(\widehat{s}_m(Y_i|X_i)) \leq \inf_{s_m \in S_m} \left( \sum_{i=1}^{N} -\log(s_m(Y_i|X_i)) \right) + \eta.$$

We are working with a maximum likelihood approach and thus the Kullback-Leibler divergence $KL$. As we consider law with densities with respect to the known measure $\mathrm{d}\lambda$, we use the following notation

$$KL_\lambda(s,t) = KL(s\mathrm{d}\lambda, t\mathrm{d}\lambda) = \begin{cases} -\int_\Omega \log\left(\frac{t}{s}\right) s \,\mathrm{d}\lambda & \text{if } s\mathrm{d}\lambda \ll t\mathrm{d}\lambda \\ +\infty & \text{otherwise} \end{cases}$$

where $s\mathrm{d}\lambda \ll t\mathrm{d}\lambda$ means $\Leftrightarrow \forall \Omega' \subset \Omega, \int_{\Omega'} t\mathrm{d}\lambda = 0 \implies \int_{\Omega'} s\mathrm{d}\lambda = 0$. As explained before, as we deal with conditional densities and not classical densities, the previous divergence should be adapted so that we use the following *tensorized* divergence:

$$KL_\lambda^{\otimes n}(s,t) = \mathbb{E}\left[ \frac{1}{N} \sum_{i=1}^{N} KL_\lambda(s(\cdot|X_i), t(\cdot|X_i)) \right].$$

## 5.2 Single model maximum likelihood estimate

### 5.2.1 Asymptotic analysis of a parametric model

Assume that $S_m$ is a parametric model of conditional densities,

$$S_m = \left\{ s_{\theta_m}(y|x) \middle| \theta_m \in \Theta_m \subset \mathbb{R}^{\mathcal{D}_m} \right\},$$

to which the true conditional density $s_0$ does not necessarily belongs. In this case, if we let

$$\widehat{\theta}_m = \operatorname*{argmin}_{\theta_m \in \Theta_m} \left( \sum_{i=1}^{N} -\log(s_{\theta_m}(Y_i|X_i)) \right)$$

then $\widehat{s}_m = s_{\widehat{\theta}_m}$. White [Wh92] has studied this *misspecified model* setting for density estimation but its results can easily been extended to the conditional density case.

If the model is identifiable and under some (strong) regularity assumptions on $\theta_m \mapsto s_{\theta_m}$, provided the $\mathcal{D}_m \times \mathcal{D}_m$ matrices $A(\theta_m)$ and $B(\theta_m)$ defined by

$$A(\theta_m)_{k,l} = \mathbb{E}\left[ \frac{1}{N} \sum_{i=1}^{N} \int \frac{-\partial^2 \log s_{\theta_m}}{\partial \theta_{m,k} \partial \theta_{m,l}}(y|X_i)\, s_0(y|X_i) \mathrm{d}\lambda \right]$$

$$B(\theta_m)_{k,l} = \mathbb{E}\left[ \frac{1}{N} \sum_{i=1}^{N} \int \frac{\partial \log s_{\theta_m}}{\partial \theta_{m,k}}(y|X_i)\, \frac{\partial \log s_{\theta_m}}{\partial \theta_{m,l}}(y|X_i)\, s_0(y|X_i) \mathrm{d}\lambda \right]$$

exists, the analysis of White [Wh92] implies that, if we let

$$\theta_m^\star = \operatorname*{argmin}_{\theta_m \in \Theta_m} KL_\lambda^{\otimes n}(s_0, s_{\theta_m}),$$

$\mathbb{E}\left[ KL_\lambda^{\otimes n}(s_0, \widehat{s}_m) \right]$ is asymptotically equivalent to

$$KL_\lambda^{\otimes n}(s_0, s_{\theta_m^\star}) + \frac{1}{2N} \operatorname{Tr}(B(\theta_m^\star) A(\theta_m^\star)^{-1}).$$

When $s_0$ belongs to the model, i.e. $s_0 = s_{\theta_m^\star}$, $B(\theta_m^\star) = A(\theta_m^\star)$ and thus the previous asymptotic equivalent of $\mathbb{E}\left[ KL_\lambda^{\otimes n}(s_0, \widehat{s}_m) \right]$ is the classical parametric one

$$\min_{\theta_m} KL_\lambda^{\otimes n}(s_0, s_{\theta_m}) + \frac{1}{2N} \mathcal{D}_m.$$

This simple expression does not hold when $s_0$ does not belong to the parametric model as $\operatorname{Tr}(B(\theta_m^\star) A(\theta_m^\star)^{-1})$ cannot generally be simplified.

A short glimpse on the proof of the previous result shows that it depends heavily on the asymptotic normality of $\sqrt{N}(\widehat{\theta}_m - \theta_m^\star)$. One may wonder if extension of this result, often called the Wilk's phenomenon [Wi38], exists when this normality does not hold, for instance in non parametric case or when the model is not identifiable. Along these lines, Fan, Zhang, and Zhang [FZZ01] propose a generalization of the corresponding Chi-Square goodness-of-fit test in several settings and Boucheron and Massart [BM11] study the finite sample deviation of the corresponding empirical quantity in a bounded loss setting.

Our aim is to derive a non asymptotic upper bound of type

$$\mathbb{E}\left[ KL_\lambda^{\otimes n}(s_0, \widehat{s}_m) \right] \leq \left( \min_{s_m \in S_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{1}{2N} \mathcal{D}_m \right) + C_2 \frac{1}{N}$$

with as few assumptions on the conditional density set $S_m$ as possible. Note that we only aim at having an upper bound and do not focus on the (important) question of the existence of a corresponding lower bound.

Our answer is far from definitive, the upper bound we obtained is the following weaker one

$$\mathbb{E}\left[ JKL_{\rho,\lambda}^{\otimes n}(s_0, \widehat{s}_m) \right] \leq (1+\epsilon) \left( \inf_{s_m \in S_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{\kappa_0}{N} \mathfrak{D}_m \right) + C_2 \frac{1}{N}$$

in which the left-hand $KL_\lambda^{\otimes n}(s_0, \widehat{s}_m)$ has been replaced by a smaller divergence $JKL_{\rho,\lambda}^{\otimes n}(s_0, \widehat{s}_m)$ described below, $\epsilon$ can be chosen arbitrary small, $\mathfrak{D}_m$ is a model complexity term playing the role of the dimension $\mathcal{D}_m$ and $\kappa_0$ is a constant that depends on $\epsilon$. This result has nevertheless the right bias/variance trade-off flavor and can be used to recover usual minimax properties of specific estimators.

### 5.2.2 Jensen-Kullback-Leibler divergence and bracketing entropy

The main visible loss is the use of a divergence smaller than the Kullback-Leibler one (but larger than the squared Hellinger distance and the squared $L_1$ loss whose definitions are recalled later). Namely, we use the Jensen-Kullback-Leibler divergence $JKL_\rho$ with $\rho \in (0,1)$ defined by

$$JKL_\rho(s\mathrm{d}\lambda, t\mathrm{d}\lambda) = JKL_{\rho,\lambda}(s,t) = \frac{1}{\rho} KL_\lambda\left(s, (1-\rho)s + \rho t\right).$$

Note that this divergence appears explicitly with $\rho = \frac{1}{2}$ in Massart [Ma07], but can also be found implicitly in Birgé and Massart [BM98] and Geer [Ge95]. We use the name Jensen-Kullback-Leibler divergence in the same way Lin [Li91] uses the name Jensen-Shannon divergence for a sibling in his information theory work. The main tools in the proof of the previous inequality are deviation inequalities for sums of random variables and their suprema. Those tools require a boundness assumption on the controlled functions that is not satisfied by the -log-likelihood differences $-\log\frac{s_m}{s_0}$. When considering the Jensen-Kullback-Leibler divergence, those ratios are implicitly replaced by ratios $-\frac{1}{\rho}\log\frac{(1-\rho)s_0+\rho s_m}{s_0}$ that are close to the -log-likelihood differences when the $s_m$ are close to $s_0$ and always upper bounded by $-\frac{\log(1-\rho)}{\rho}$. This divergence is smaller than the Kullback-Leibler one but larger, up to a constant factor, than the squared Hellinger one, $d^2_\lambda(s,t) = \int_\Omega |\sqrt{s} - \sqrt{t}|^2 \mathrm{d}\lambda$, and the squared $L_1$ distance, $\|s-t\|^2_{\lambda,1} = \left(\int_\Omega |s-t|\mathrm{d}\lambda\right)^2$, as proved in our technical report [*Unpub-CLP11a*]

**Proposition 11.** *For any probability measures* $s\mathrm{d}\lambda$ *and* $t\mathrm{d}\lambda$ *and any* $\rho \in (0,1)$

$$C_\rho\, d^2_\lambda(s,t) \le JKL_{\rho,\lambda}(s,t) \le KL_\lambda(s,t).$$

*with* $C_\rho = \frac{1}{\rho}\min\left(\frac{1-\rho}{\rho}, 1\right)\left(\log\left(1 + \frac{\rho}{1-\rho}\right) - \rho\right)$ *while*

$$\max(C_\rho/4, \rho/2)\|s-t\|^2_{\lambda,1} \le JKL_{\rho,\lambda}(s,t) \le KL_\lambda(s,t).$$

*Furthermore, if* $s\mathrm{d}\lambda \ll t\mathrm{d}\lambda$ *then*

$$d^2_\lambda(s,t) \le KL_\lambda(s,t) \le \left(2 + \log\left\|\frac{s}{t}\right\|_\infty\right) d^2_\lambda(s,t)$$

*while*

$$\frac{1}{2}\|s-t\|^2_{\lambda,1} \le KL_\lambda(s,t) \le \left\|\frac{1}{t}\right\|_\infty \|s-t\|^2_{\lambda,2}.$$

More precisely, as we are in a conditional density setting, we use their *tensorized* versions

$$d^{2\otimes n}_\lambda(s,t) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N d^2_\lambda(s(\cdot|X_i), t(\cdot|X_i))\right] \quad \text{and} \quad JKL^{\otimes n}_{\rho,\lambda}(s,t) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N JKL_{\rho,\lambda}(s(\cdot|X_i), t(\cdot|X_i))\right].$$

We focus now on the definition of the model complexity $\mathfrak{D}_m$. It involves a bracketing entropy condition on the model $S_m$ with respect to the Hellinger type divergence $d^{\otimes n}_\lambda(s,t) = \sqrt{d^{2\otimes n}_\lambda(s,t)}$. A bracket $[t^-, t^+]$ is a pair of functions such that $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, t^-(y|x) \le t^+(y|x)$. A conditional density function $s$ is said to belong to the bracket $[t^-,t^+]$ if $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, t^-(y|x) \le s(y|x) \le t^+(y|x)$. The bracketing entropy $H_{[\cdot],d^{\otimes n}_\lambda}(\delta, S)$ of a set $S$ is defined as the logarithm of the minimum number of brackets $[t^-,t^+]$ of width $d^{\otimes n}_\lambda(t^-,t^+)$ smaller than $\delta$ such that every function of $S$ belongs to one of these brackets. $\mathfrak{D}_m$ depends on the bracketing entropies not of the global models $S_m$ but of the ones of smaller localized sets $S_m(\widetilde{s},\sigma) = \left\{s_m \in S_m \middle| d^{\otimes n}_\lambda(\widetilde{s}, s_m) \le \sigma\right\}$. Indeed, we impose a structural assumption:

**Assumption ($\mathbf{H}_m$).** *There is a non-decreasing function* $\phi_m(\delta)$ *such that* $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ *is non-increasing on* $(0,+\infty)$ *and for every* $\sigma \in \mathbb{R}^+$ *and every* $s_m \in S_m$

$$\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}_\lambda}\left(\delta, S_m(s_m,\sigma)\right)}\,\mathrm{d}\delta \le \phi_m(\sigma).$$

Note that the function $\sigma \mapsto \int_0^\sigma \sqrt{H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m)}\,\mathrm{d}\delta$ does always satisfy this assumption. $\mathfrak{D}_m$ is then defined as $n\sigma_m^2$ with $\sigma_m^2$ the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{N}\sigma$. A good choice of $\phi_m$ is one which leads to a small upper bound of $\mathfrak{D}_m$. This bracketing entropy integral, often call Dudley integral, plays an important role in empirical processes theory, as stressed for instance in Vaart and Wellner [VW96] and in Kosorok [Ko08]. The equation defining $\sigma_m$ corresponds to a crude optimization of a supremum bound as shown explicitly in the proof. This definition is obviously far from being very explicit but it turns out that it can be related to an entropic dimension of the model. Recall that the classical entropy dimension of a compact set $S$ with respect to a metric $d$ can be defined as the smallest non negative real $\mathcal{D}$ such that there is a non negative $\mathcal{V}$ such that

$$\forall \delta > 0, H_d(\delta, S) \leq \mathcal{V} + \mathcal{D}\log\left(\frac{1}{\delta}\right)$$

where $H_d$ is the classical entropy with respect to metric $d$. The parameter $\mathcal{V}$ can be interpreted as the logarithm of the volume of the set. Replacing the classical entropy by a bracketing one, we define the bracketing dimension $\mathcal{D}_m$ of a compact set as the smallest real $\mathcal{D}$ such that there is a $\mathcal{V}$ such

$$\forall \delta > 0, H_{[\cdot],d}(\delta, S) \leq \mathcal{V} + \mathcal{D}\log\left(\frac{1}{\delta}\right).$$

As hinted by the notation, for parametric model, under mild assumption on the parametrization, this bracketing dimension coincides with the usual one. Under such assumption, one can prove that $\mathfrak{D}_m$ is proportional to $\mathcal{D}_m$. More precisely, working with the localized set $S_m(s,\sigma)$ instead of $S_m$, we obtain, in our technical report [*Unpub-CLP11a*],

**Proposition 12.**

- *if* $\exists \mathcal{D}_m \geq 0, \exists \mathcal{C}_m \geq 0, \forall \delta \in (0, \sqrt{2}], H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m) \leq \mathcal{V}_m + \mathcal{D}_m \log\frac{1}{\delta}$ *then*
  - *if* $\mathcal{D}_m > 0$, *(*$H_m$*) holds with* $\mathfrak{D}_m \leq \left(2C_{\star,m} + 1 + \left(\log\frac{N}{eC_{\star,m}\mathcal{D}_m}\right)_+\right)\mathcal{D}_m$ *with* $C_{\star,m} = \left(\sqrt{\frac{\mathcal{V}_m}{\mathcal{D}_m}} + \sqrt{\pi}\right)^2$,
  - *if* $\mathcal{D}_m = 0$, *(*$H_m$*) holds with* $\phi_m(\sigma) = \sigma\sqrt{\mathcal{V}_m}$ *such that* $\mathfrak{D}_m = \mathcal{V}_m$,
- *if* $\exists \mathcal{D}_m \geq 0, \exists \mathcal{V}_m \geq 0, \forall \sigma \in (0, \sqrt{2}], \forall \delta \in (0, \sigma], H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m(s_m,\sigma)) \leq \mathcal{V}_m + \mathcal{D}_m \log\frac{\sigma}{\delta}$ *then*
  - *if* $\mathcal{D}_m > 0$, *(*$H_m$*) holds with* $\phi_m$ *such that* $\mathfrak{D}_m = C_{\star,m}\mathcal{D}_m$ *with* $C_{\star,m} = \left(\sqrt{\frac{\mathcal{V}_m}{\mathcal{D}_m}} + \sqrt{\pi}\right)^2$,
  - *if* $\mathcal{D}_m = 0$, *(*$H_m$*) holds with* $\phi_m(\sigma) = \sigma\sqrt{\mathcal{V}_m}$ *such that* $\mathfrak{D}_m = \mathcal{V}_m$.

Note that we assume bounds on the entropy only for $\delta$ and $\sigma$ smaller than $\sqrt{2}$, but, as for any conditional densities pair $(s,t)$ $d_\lambda^{\otimes n}(s,t) \leq \sqrt{2}$,

$$H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m(s_m,\sigma)) = H_{[\cdot],d_\lambda^{\otimes n}}(\delta \wedge \sqrt{2}, S_m(s_m, \sigma \wedge \sqrt{2}))$$

which implies that those bounds are still useful when $\delta$ and $\sigma$ are large. Assume now that all models are such that $\frac{\mathcal{V}_m}{\mathcal{D}_m} \leq \mathcal{C}$, i.e. their log-volumes $\mathcal{V}_m$ grow at most linearly with the dimension (as it is the case for instance for hypercubes with the same width). One deduces that Assumptions ($\mathrm{H}_m$) hold simultaneously for every model with a common constant $C_\star = \left(\sqrt{\mathcal{C}} + \sqrt{\pi}\right)^2$. The model complexity $\mathfrak{D}_m$ can thus be chosen roughly proportional to the dimension in this case, this justifies the notation as well as our claim at the end of the previous section.

### 5.2.3 Single model maximum likelihood estimation

For technical reason, we also need a separability assumption on our model:

**Assumption ($\mathbf{Sep}_m$).** *There exist a countable subset $S'_m$ of $S_m$ and a set $\mathcal{Y}'_m$ with $\lambda(\mathcal{Y} \setminus \mathcal{Y}'_m) = 0$ such that for every $t \in S_m$, there exists a sequence $(t_k)_{k\geq 1}$ of elements of $S'_m$ such that for every $x$ and for every $y \in \mathcal{Y}'_m$, $\log(t_k(y|x))$ goes to $\log(t(y|x))$ as $k$ goes to infinity.*

We are now ready to state our risk bound theorem:

**Theorem 21.** *Assume we observe $(X_i, Y_i)$ with unknown conditional density $s_0$. Assume $S_m$ is a set of conditional densities for which Assumptions ($H_m$) and ($Sep_m$) hold and let $\widehat{s}_m$ be a $\eta$ -log-likelihood minimizer in $S_m$*

$$\sum_{i=1}^{N} -\log(\widehat{s}_m(Y_i|X_i)) \leq \inf_{s_m \in S_m} \left( \sum_{i=1}^{N} -\log(s_m(Y_i|X_i)) \right) + \eta$$

*Then for any $\rho \in (0,1)$ and any $C_1 > 1$, there are two constants $\kappa_0$ and $C_2$ depending only on $\rho$ and $C_1$ such that, for $\mathfrak{D}_m = n\sigma_m^2$ with $\sigma_m$ the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{N}\sigma$, the likelihood estimate $\widehat{s}_m$ satisfies*

$$\mathbb{E}\left[ JKL_{\rho,\lambda}^{\otimes_n}(s_0, \widehat{s}_m) \right] \leq C_1 \left( \inf_{s_m \in S_m} KL_{\lambda}^{\otimes_n}(s_0, s_m) + \frac{\kappa_0}{N}\mathfrak{D}_m \right) + C_2 \frac{1}{N} + \frac{\eta}{N}.$$

This theorem holds without any assumption on the design $X_i$, in particular we do not assume that the covariates admit upper or lower bounded densities. The law of the design appears however in the divergence $JKL_{\lambda}^{\otimes_n}$ and $KL_{\lambda}^{\otimes_n}$ used to assess the quality of the estimate as well as in the definition of the divergence $d_{\lambda}^{\otimes_n}$ used to measure the bracketing entropy. By construction, those quantities however do not involve the values of the conditional densities outside the support of the $X_i$s and put more focus on the regions of high density of covariates than the other. Note that Assumption $\mathrm{H}_m$ could be further localized: it suffices to impose that the condition on the Dudley integral holds for a sequence of minimizer of $d_{\lambda}^{2\otimes_n}(s_0, s_m)$.

We obtain thus a bound on the expected loss similar to the one obtained in the parametric case that holds for finite sample and that do not require the strong regularity assumptions of White [Wh92]. In particular, we do not even require an identifiability condition in the parametric case. As often in empirical processes theory, the constant $\kappa_0$ appearing in the bound is pessimistic. Even in a very simple parametric model, the current best estimates are such that $\kappa_0 \mathfrak{D}_m$ is still much larger than the variance of Section 5.2.1. Numerical experiments show there is a hope that this is only a technical issue. The obtained bound quantifies however the expected classical bias-variance trade-off: a good model should be large enough so that the true conditional density is close from it but, at the same time, it should also be small so that the $\mathfrak{D}_m$ term does not dominate.

It should be stressed that a result of similar flavor could have been obtained by the information theory technique of Barron, Huang, Li, and Luo [Ba+08] and Kolaczyk, Ju, and Gopal [KJG05]. Indeed, if we replace the set $S_m$ by a *discretized* version $\mathfrak{S}_m$ so that

$$\inf_{\mathfrak{s}_m \in \mathfrak{S}_m} KL_{\lambda}^{\otimes_n}(s_0, \mathfrak{s}_m) \leq \inf_{s_m \in S_m} KL_{\lambda}^{\otimes_n}(s_0, s_m) + \frac{1}{N},$$

then, if we let $\widehat{\mathfrak{s}}_m$ be a -log-likelihood minimizer in $\mathfrak{S}_m$,

$$\mathbb{E}\left[ \mathcal{D}_{\lambda}^{2\otimes_n}(s_0, \widehat{\mathfrak{s}}_m) \right] \leq \inf_{s_m \in S_m} KL_{\lambda}^{\otimes_n}(s_0, s_m) + \frac{1}{N}\log|\mathfrak{S}_m| + \frac{1}{N}$$

where $\mathcal{D}_{\lambda}^{2\otimes_n}$ is the tensorized Bhattacharyya-Renyi divergence, another divergence smaller than $KL^{\otimes_n}$, $|\mathfrak{S}_m|$ is the cardinality of $\mathfrak{S}_m$ and expectation is taken conditionally to the covariates $(X_i)_{1\leq i\leq N}$. As verified by Barron, Huang, Li, and Luo [Ba+08] and Kolaczyk, Ju, and Gopal [KJG05], $\mathfrak{S}_m$ can be

chosen of cardinality of order $\log n\, \mathcal{D}_m$ when the model is parametric. We obtain thus also a bound of type

$$\mathbb{E}\left[\mathcal{D}_\lambda^{2\otimes_n}(s_0, \widehat{\mathfrak{s}}_m)\right] \leq \inf_{s_m \in S_m} KL_\lambda^{\otimes_n}(s_0, s_m) + \frac{C_1}{N}\log n\, \mathcal{D}_m + \frac{1}{N}.$$

with better constants but with a different divergence. The bound holds however only conditionally to the design, which can be an issue as soon as this design is random, and requires to compute an adapted discretization of the models.

## 5.3 Model selection and penalized maximum likelihood

### 5.3.1 Framework

A natural question is then the choice of the model. In the model selection framework, instead of a single model $S_m$, we assume we have at hand a collection of models $\mathcal{S} = \{S_m\}_{m\in\mathcal{M}}$. If we assume that Assumptions ($\mathrm{H}_m$) and ($\mathrm{Sep}_m$) hold for all models, then for every model $S_m$

$$\mathbb{E}\left[JKL_{\rho,\lambda}^{\otimes_n}(s_0, \widehat{s}_m)\right] \leq C_1 \left( \inf_{s_m \in S_m} KL_\lambda^{\otimes_n}(s_0, s_m) + \frac{\kappa_0}{N}\mathfrak{D}_m \right) + C_2 \frac{1}{N} + \frac{\eta}{N}.$$

Obviously, one of the models minimizes the right hand side. Unfortunately, there is no way to know which one without knowing $s_0$, i.e. without an oracle. Hence, this *oracle* model can not be used to estimate $s_0$. We nevertheless propose a data-driven strategy to select an estimate among the collection of estimates $\{\widehat{s}_m\}_{m\in\mathcal{M}}$ according to a selection rule that performs almost as well as if we had known this *oracle*.

As always, using simply the -log-likelihood of the estimate in each model

$$\sum_{i=1}^{N} -\log(\widehat{s}_m(Y_i|X_i))$$

as a criterion is not sufficient. It is an underestimation of the true risk of the estimate and this leads to choose models that are too complex. By adding an adapted penalty $\mathrm{pen}(m)$, one hopes to compensate for both the *variance* term and the bias between $\frac{1}{N}\sum_{i=1}^{N} -\log \frac{\widehat{s}_{\widehat{m}}(Y_i|X_i)}{s_0(Y_i|X_i)}$ and $\inf_{s_m\in S_m} KL_\lambda^{\otimes_n}(s_0, s_m)$. For a given choice of $\mathrm{pen}(m)$, the *best* model $S_{\widehat{m}}$ is chosen as the one whose index is an almost minimizer of the penalized $\eta$ -log-likelihood :

$$\sum_{i=1}^{N} -\log(\widehat{s}_{\widehat{m}}(Y_i|X_i)) + \mathrm{pen}(\widehat{m}) \leq \inf_{m\in\mathcal{M}} \left( \sum_{i=1}^{N} -\log(\widehat{s}_m(Y_i|X_i)) + \mathrm{pen}(m) \right) + \eta'.$$

The analysis of the previous section turns out to be crucial as the intrinsic complexity $\mathfrak{D}_m$ appears in the assumption on the penalty. It is no surprise that the complexity of the model collection itself also appears. We need an information theory type assumption on our collection; we assume thus the existence of a Kraft type inequality for the collection:

**Assumption (K).** *There is a family $(x_m)_{m\in\mathcal{M}}$ of non-negative number such that*

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < +\infty$$

It can be interpreted as a coding condition as stressed by Barron, Huang, Li, and Luo [Ba+08] where a similar assumption is used. Remark that if this assumption holds, it also holds for any permutation of the coding term $x_m$. We should try to mitigate this arbitrariness by favoring choice of $x_m$ for which the ratio with the intrinsic entropy term $\mathfrak{D}_m$ is as small as possible. Indeed, as the condition on the penalty is of the form

$$\mathrm{pen}(m) \geq \kappa\left(\mathfrak{D}_m + x_m\right),$$

this ensures that this lower bound is dominated by the intrinsic quantity $\mathfrak{D}_m$.

### 5.3.2 A general theorem for penalized maximum likelihood conditional density estimation

Our main theorem is then:

**Theorem 22.** *Assume we observe $(X_i, Y_i)$ with unknown conditional density $s_0$. Let $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ be at most countable collection of conditional density sets. Assume Assumption (K) holds while Assumptions $(H_m)$ and $(Sep_m)$ hold for every model $S_m \in \mathcal{S}$. Let $\widehat{s}_m$ be a $\eta$ -log-likelihood minimizer in $S_m$*

$$\sum_{i=1}^{N} -\log(\widehat{s}_m(Y_i|X_i)) \leq \inf_{s_m\in S_m}\left(\sum_{i=1}^{N} -\log(s_m(Y_i|X_i))\right) + \eta$$

*Then for any $\rho \in (0,1)$ and any $C_1 > 1$, there are two constants $\kappa_0$ and $C_2$ depending only on $\rho$ and $C_1$ such that, as soon as for every index $m \in \mathcal{M}$*

$$\mathrm{pen}(m) \geq \kappa\,(\mathfrak{D}_m + x_m) \quad \text{with } \kappa > \kappa_0$$

*where $\mathfrak{D}_m = n\sigma_m^2$ with $\sigma_m$ the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{N}\sigma$, the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that*

$$\sum_{i=1}^{N} -\log(\widehat{s}_{\widehat{m}}(Y_i|X_i)) + \mathrm{pen}(\widehat{m}) \leq \inf_{m\in\mathcal{M}}\left(\sum_{i=1}^{N} -\log(\widehat{s}_m(Y_i|X_i)) + \mathrm{pen}(m)\right) + \eta'$$

*satisfies*

$$\mathbb{E}\left[JKL_{\rho,\lambda}^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{N}\right) + C_2\frac{\Sigma}{N} + \frac{\eta+\eta'}{N}.$$

Note that, as in 5.2.3, the approach of of Barron, Huang, Li, and Luo [Ba+08] and Kolaczyk, Ju, and Gopal [KJG05] could have been used to obtain a similar result with the help of discretization.

This theorem extends Theorem 7.11 Massart [Ma07] which handles only density estimation. As in this theorem, the cost of model selection with respect to the choice of the best single model is proved to be very mild. Indeed, let $\mathrm{pen}(m) = \kappa(\mathfrak{D}_m + x_m)$ then one obtains

$$\begin{aligned}
&\mathbb{E}\left[JKL_{\rho,\lambda}^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \\
&\quad\leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{\kappa}{N}(\mathfrak{D}_m + x_m)\right) + C_2\frac{\Sigma}{N} + \frac{\eta+\eta'}{N} \\
&\quad\leq C_1\frac{\kappa}{\kappa_0}\left(\max_{m\in\mathcal{M}}\frac{\mathfrak{D}_m + x_m}{\mathfrak{D}_m}\right)\inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{\kappa_0}{N}\mathfrak{D}_m\right) + C_2\frac{\Sigma}{N} + \frac{\eta+\eta'}{N}.
\end{aligned}$$

As soon as the term $x_m$ is always small relatively to $\mathfrak{D}_m$, we obtain thus an oracle inequality that show that the penalized estimate satisfies, up to a small factor, the bound of Theorem 21 for the estimate in the best model. The price to pay for the use of a collection of model is thus small. The gain is on the contrary very important: we do not have to know the best model within a collection to almost achieve its performance.

So far we do not have discussed the choice of the model collection, it is however critical to obtain a *good* estimator. There is unfortunately no universal choice and it should be adapted to the specific setting considered. Typically, if we consider conditional density of *regularity* indexed by a parameter $\alpha$, a good collection is one such that for every parameter $\alpha$ there is a model which achieves a quasi optimal bias/variance trade-off. Efromovich [Ef07; Ef10] considers Sobolev type regularity and use thus models generated by the first elements of Fourier basis. Brunel, Comte, and Lacour [BCL07] and Akakpo and Lacour [AL11] considers anisotropic regularity spaces for which they show that a collection of piecewise polynomial models is adapted. Although those choices are justified, in these papers, in a quadratic loss approach, they remain good choices in our maximum likelihood approach with a Kullback-Leibler type loss. Estimator associated to those collections are thus *adaptive* to the regularity: without knowing the

*regularity* of the true conditional density, they select a model in which the estimate performs almost as well as in the *oracle* model, the best choice if the regularity was known. In both cases, one could prove that those estimators achieve the minimax rate for the considered classes, up to a logarithmic factor.

As in Section 5.2.3, the known estimate of constant $\kappa_0$ and even of $\mathfrak{D}_m$ can be pessimistic. This leads to a theoretical penalty which can be too large in practice. A natural question is thus whether the constant appearing in the penalty can be estimated from the data without loosing a theoretical guaranty on the performance? No definitive answer exists so far, but numerical experiment in specific case shows that the *slope heuristic* proposed by Birgé and Massart [BM07] may yield a solution.

The assumptions of the previous theorem are as general as possible. It is thus natural to question the existence of interesting model collections that satisfy its assumptions. We have mention so far the Fourier based collection proposed by Efromovich [Ef07; Ef10] and the piecewise polynomial collection of Brunel, Comte, and Lacour [BCL07] and Akakpo and Lacour [AL11] considers anisotropic regularity. We focus on a variation of this last strategy. Motivated by an application to unsupervised image segmentation, we consider model collection in which, in each model, the conditional densities depend on the covariate only in a piecewise constant manner. After a general introduction to these partition-based strategies, we study two cases: a classical one in which the conditional density depends in a piecewise polynomial manner of the variables and a newer one, which correspond to the unsupervised segmentation application, in which the conditional densities are Gaussian mixture with common Gaussian components but mixing proportions depending on the covariate.

## 5.4 Partition-based conditional density models

### 5.4.1 Covariate partitioning and conditional density estimation

Following an idea developed by Kolaczyk, Ju, and Gopal [KJG05], we partition the covariate domain and consider candidate conditional density estimates that depend on the covariate only through the region it belongs. We are thus interested in conditional densities that can be written as

$$s(y|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} s(y|\mathcal{R}_l)\mathbf{1}_{\{x \in \mathcal{R}_l\}}$$

where $\mathcal{P}$ is partition of $\mathcal{X}$, $\mathcal{R}_l$ denotes a generic region in this partition, $\mathbf{1}$ denotes the characteristic function of a set and $s(y|\mathcal{R}_l)$ is a density for any $\mathcal{R}_l \in \mathcal{P}$. Note that this strategy, called as in Willet and Nowak [WN07] partition-based, shares a lot with the CART-type strategy proposed by Donoho [Do97] in an image processing setting.

Denoting $\|\mathcal{P}\|$ the number of regions in this partition, the model we consider are thus specified by a partition $\mathcal{P}$ and a set $\mathcal{F}$ of $\|\mathcal{P}\|$-tuples of densities into which $(s(\cdot|\mathcal{R}_l))_{\mathcal{R}_l \in \mathcal{P}}$ is chosen. This set $\mathcal{F}$ can be a product of density sets, yielding an independent choice on each region of the partition, or have a more complex structure. We study two examples: in the first one, $\mathcal{F}$ is indeed a product of piecewise polynomial density sets, while in the second one $\mathcal{F}$ is a set of $\|\mathcal{P}\|$-tuples of Gaussian mixtures sharing the same mixture components. Nevertheless, denoting with a slight abuse of notation $S_{\mathcal{P},\mathcal{F}}$ such a model, our $\eta$-log-likelihood estimate in this model is any conditional density $\widehat{s}_{\mathcal{P},\mathcal{F}}$ such that

$$\left(\sum_{i=1}^{N} -\log(\widehat{s}_{\mathcal{P},\mathcal{F}}(Y_i|X_i))\right) \leq \min_{s_{\mathcal{P},\mathcal{F}} \in S_{\mathcal{P},\mathcal{F}}} \left(\sum_{i=1}^{N} -\log(s_{\mathcal{P},\mathcal{F}}(Y_i|X_i))\right) + \eta.$$

We first specify the partition collection we consider. For the sake of simplicity, our description is restricted to the case where the covariate space $\mathcal{X}$ is simply $[0,1]^{d_X}$. We stress that the proposed strategy can easily be adapted to more general settings including discrete variable ordered or not. We impose a strong structural assumption on the partition collection considered that allows to control their *complexity* and only consider five specific hyperrectangle based collections of partitions of $[0,1]^{d_X}$:

- Two are recursive dyadic partition collections.

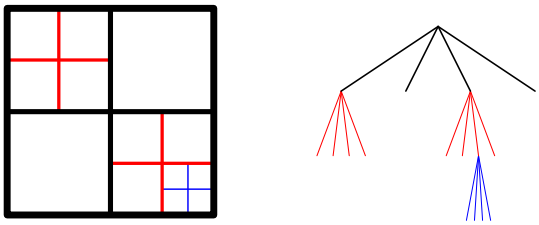


Figure 5.1: Example of a recursive dyadic partition with its associated dyadic tree.

  - The uniform dyadic partition collection (UDP($\mathcal{X}$)) in which all hypercubes are subdivided in $2^{d_X}$ hypercubes of equal size at each step. In this collection, in the partition obtained after $J$ step, all the $2^{d_X J}$ hyperrectangles $\{\mathcal{R}_l\}_{1 \leq l \leq \|\mathcal{P}\|}$ are thus hypercubes whose measure $|\mathcal{R}_l|$ satisfies $|\mathcal{R}_l| = 2^{-d_X J}$. We stop the recursion as soon as the number of steps $J$ satisfies $\frac{2^{d_X}}{N} \geq |\mathcal{R}_l| \geq \frac{1}{N}$.
  - The recursive dyadic partition collection (RDP($\mathcal{X}$)) in which at each step a hypercube of measure $|\mathcal{R}_l| \geq \frac{2^{d_X}}{N}$ is subdivided in $2^{d_X}$ hypercubes of equal size.

- Two are recursive split partition collections.

  - The recursive dyadic split partition (RDSP($\mathcal{X}$)) in which at each step a hyperrectangle of measure $|\mathcal{R}_l| \geq \frac{2}{N}$ can be subdivided in 2 hyperrectangles of equal size by an even split along one of the $d_X$ possible directions.
  - The recursive split partition (RSP($\mathcal{X}$)) in which at each step a hyperrectangle of measure $|\mathcal{R}_l| \geq \frac{2}{N}$ can be subdivided in 2 hyperrectangles of measure larger than $\frac{1}{N}$ by a split along one a point of the grid $\frac{1}{N}\mathbb{Z}$ in one the $d_X$ possible directions.

- The last one does not possess a hierarchical structure. The hyperrectangle partition collection (HRP($\mathcal{X}$)) is the full collection of all partitions into hyperrectangles whose corners are located on the grid $\frac{1}{N}\mathbb{Z}^{d_X}$ and whose volume is larger than $\frac{1}{N}$.

We denote by $\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$ the corresponding partition collection where $\star(\mathcal{X})$ is either UDP($\mathcal{X}$), RDP($\mathcal{X}$), RDSP($\mathcal{X}$), RSP($\mathcal{X}$) or HRP($\mathcal{X}$).

As noticed by Kolaczyk and Nowak [KN05], Huang, Pollak, Do, and Bouman [Hu+06] or Willet and Nowak [WN07], the first four partition collections, ($\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDSP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RSP}(\mathcal{X})}$), have a tree structure. Figure 5.1 illustrates this structure for a RDP($\mathcal{X}$) partition. This specific structure is mainly used to obtain an efficient numerical algorithm performing the model selection. For sake of completeness, we have also added the much more complex to deal with collection $\mathcal{S}_{\mathcal{P}}^{\mathrm{HRP}(\mathcal{X})}$, for which only exhaustive search algorithms exist.

As proved in our technical report [*Unpub-CLP11a*], those partition collections satisfy Kraft type inequalities with weights constant for the UDP($\mathcal{X}$) partition collection and proportional to the number $\|\mathcal{P}\|$ of hyperrectangles for the other collections. Indeed,

**Proposition 13.** *For any of the five described partition collections $\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$, $\exists A_0^\star, B_0^\star, c_0^\star$ and $\Sigma_0$ such that for all $c \geq c_0^{\star(\mathcal{X})}$:*

$$\sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}} e^{-c\left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} \|\mathcal{P}\|\right)} \leq \Sigma_0^{\star(\mathcal{X})} e^{-c \max\left(A_0^{\star(\mathcal{X})}, B_0^{\star(\mathcal{X})}\right)}.$$

This will prove useful to verify Assumption ($K$) for the model collections of the next sections.

In those sections, we study the two different choices proposed above for the set $\mathcal{F}$. We first consider a piecewise polynomial strategy similar to the one proposed by Willet and Nowak [WN07] defined for

$\mathcal{Y} = [0,1]^{d_Y}$ in which the set $\mathcal{F}$ is a product of sets. We then consider a Gaussian mixture strategy with varying mixing proportion but common mixture components that extends the work of Maugis and Michel [MM12a] and has been the original motivation of this work. In both cases, we prove that the penalty can be chosen roughly proportional to the dimension.

### 5.4.2 Piecewise polynomial conditional density estimation

In this section, we let $\mathcal{X} = [0,1]^{d_X}$, $\mathcal{Y} = [0,1]^{d_Y}$ and $\lambda$ be the Lebesgue measure $\mathrm{d}y$. Note that, in this case, $\lambda$ is a probability measure on $\mathcal{Y}$. Our candidate density $s(y|x \in \mathcal{R}_l)$ is then chosen among piecewise polynomial densities. More precisely, we reuse a hyperrectangle partitioning strategy this time for $\mathcal{Y} = [0,1]^{d_Y}$ and impose that our candidate conditional density $s(y|x \in \mathcal{R}_l)$ is a square of polynomial on each hyperrectangle $\mathcal{R}^{\mathcal{Y}}_{l,k}$ of the partition $\mathcal{Q}_l$. This differs from the choice of Willet and Nowak [WN07] in which the candidate density is simply a polynomial. The two choices coincide however when the polynomial is chosen among the constant ones. Although our choice of using squares of polynomial is less natural, it already ensures the positiveness of the candidates so that we only have to impose that the integrals of the piecewise polynomials are equal to 1 to obtain conditional densities. It turns out to be also crucial to obtain a control of the local bracketing entropy of our models. Note that this setting differs from the one of Blanchard, Schäfer, Rozenholc, and Müller [Bl+07] in which $\mathcal{Y}$ is a finite discrete set.

We should now define the sets $\mathcal{F}$ we consider for a given partition $\mathcal{P} = \{\mathcal{R}_l\}_{1 \le l \le \|\mathcal{P}\|}$ of $\mathcal{X} = [0,1]^{d_X}$. Let $\mathbf{D} = (\mathbf{D}_1, \dots, \mathbf{D}_{d_Y})$, we first define for any partition $\mathcal{Q} = \{\mathcal{R}^{\mathcal{Y}}_k\}_{1 \le k \le \|\mathcal{Q}\|}$ of $\mathcal{Y} = [0,1]^{d_Y}$ the set $\mathcal{F}_{\mathcal{Q},\mathbf{D}}$ of squares of piecewise polynomial densities of maximum degree $\mathbf{D}$ defined in the partition $\mathcal{Q}$:

$$\mathcal{F}_{\mathcal{Q},\mathbf{D}} = \left\{ s(y) = \sum_{\mathcal{R}^{\mathcal{Y}}_k \in \mathcal{Q}} P^2_{\mathcal{R}^{\mathcal{Y}}_k}(y) \mathbf{1}_{\left\{y \in \mathcal{R}^{\mathcal{Y}}_k\right\}} \;\middle|\; \begin{array}{l} \forall \mathcal{R}^{\mathcal{Y}}_k \in \mathcal{Q}, P_{\mathcal{R}^{\mathcal{Y}}_k} \text{ polynomial of degree at most } \mathbf{D}, \\ \sum_{\mathcal{R}^{\mathcal{Y}}_k \in \mathcal{Q}} \int_{\mathcal{R}^{\mathcal{Y}}_k} P^2_{\mathcal{R}^{\mathcal{Y}}_k}(y) = 1 \end{array} \right\}$$

For any partition collection $\mathcal{Q}^{\mathcal{P}} = (\mathcal{Q}_l)_{1 \le l \le \|\mathcal{P}\|} = \left(\{\mathcal{R}^{\mathcal{Y}}_{l,k}\}_{1 \le k \le \|\mathcal{Q}_l\|}\right)_{1 \le l \le \|\mathcal{P}\|}$ of $\mathcal{Y} = [0,1]^{d_Y}$, we can thus defined the set $\mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ of $\|\mathcal{P}\|$-tuples of piecewise polynomial densities as

$$\mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \left\{ (s(\cdot|\mathcal{R}_l))_{\mathcal{R}_l \in \mathcal{P}} \middle| \forall \mathcal{R}_l \in \mathcal{P}, s(\cdot|\mathcal{R}_l) \in \mathcal{F}_{\mathcal{Q}_l,\mathbf{D}} \right\}.$$

The model $S_{\mathcal{P},\mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}}$, that is denoted $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ with a slight abuse of notation, is thus the set

$$\begin{aligned} S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} &= \left\{ s(y|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} s(y|\mathcal{R}_l) \mathbf{1}_{\{x \in \mathcal{R}_l\}} \middle| (s(y|\mathcal{R}_l)_{\mathcal{R}_l \in \mathcal{P}} \in \mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \right\} \\ &= \left\{ s(y|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} \sum_{\mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l} P^2_{\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}}(y) \mathbf{1}_{\left\{y \in \mathcal{R}^{\mathcal{Y}}_{l,k}\right\}} \mathbf{1}_{\{x \in \mathcal{R}_l\}} \middle| \begin{array}{l} \forall \mathcal{R}_l \in \mathcal{P}, \forall \mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l, \\ \quad P_{\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}} \text{ polynomial of degree at most } \mathbf{D}, \\ \forall \mathcal{R}_l \in \mathcal{P}, \sum_{\mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l} \int_{\mathcal{R}^{\mathcal{Y}}_{l,k}} P^2_{\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}}(y) = 1 \end{array} \right\} \end{aligned}$$

Denoting $\mathcal{R}^{\times}_{l,k}$ the product $\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}$, the conditional densities of the previous set can be advantageously rewritten as

$$s(y|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} \sum_{\mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l} P^2_{\mathcal{R}^{\times}_{l,k}}(y) \mathbf{1}_{\left\{(x,y) \in \mathcal{R}^{\times}_{l,k}\right\}}$$

As shown by Willet and Nowak [WN07], the maximum likelihood estimate in this model can be obtained by an independent computation on each subset $\mathcal{R}^{\times}_{l,k}$:

$$\widehat{P}_{\mathcal{R}^{\times}_{l,k}} = \frac{\sum_{i=1}^N \mathbf{1}_{\left\{(X_i,Y_i) \in \mathcal{R}^{\times}_{l,k}\right\}}}{\sum_{i=1}^N \mathbf{1}_{\{X_i \in \mathcal{R}_l\}}} \underset{P, \deg(P) \le \mathbf{D}, \int_{\mathcal{R}^{\mathcal{Y}}_{l,k}} P^2(y)\mathrm{d}y = 1}{\operatorname{argmin}} \sum_{i=1}^N \mathbf{1}_{\left\{(X_i,Y_i) \in \mathcal{R}^{\times}_{l,k}\right\}} \log\left(P^2(Y_i)\right).$$

This property is important to be able to use the efficient optimization algorithms of Willet and Nowak [WN07] and Huang, Pollak, Do, and Bouman [Hu+06].

Our model collection is obtained by considering all partitions $\mathcal{P}$ within one of the UDP($\mathcal{X}$), RDP($\mathcal{X}$), RDSP($\mathcal{X}$), RSP($\mathcal{X}$) or HRP($\mathcal{X}$) partition collections with respect to $[0,1]^{d_X}$ and, for a fixed $\mathcal{P}$, all partitions $\mathcal{Q}_l$ within one of the UDP($\mathcal{Y}$), RDP($\mathcal{Y}$), RDSP($\mathcal{Y}$), RSP($\mathcal{Y}$) or HRP($\mathcal{Y}$) partition collections with respect to $[0,1]^{d_Y}$. By construction, in any cases,

$$\dim\left(()\, S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}\right) = \sum_{\mathcal{R}_l \in \mathcal{P}} \left( \|\mathcal{Q}_l\| \prod_{d=1}^{d_Y} (\mathbf{D}_d + 1) - 1 \right).$$

To define the penalty, we use a slight upper bound of this dimension

$$\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \sum_{\mathcal{R}_l \in \mathcal{P}} \|\mathcal{Q}_l\| \prod_{d=1}^{d_Y} (\mathbf{D}_d + 1) = \|\mathcal{Q}^{\mathcal{P}}\| \prod_{d=1}^{d_Y} (\mathbf{D}_d + 1)$$

where $\|\mathcal{Q}^{\mathcal{P}}\| = \sum_{\mathcal{R}_l \in \mathcal{P}} \|\mathcal{Q}_l\|$. is the total number of hyperrectangles in all the partitions:

**Theorem 23.** *Fix a collection $\star(\mathcal{X})$ among* UDP($\mathcal{X}$), RDP($\mathcal{X}$), RDSP($\mathcal{X}$), RSP($\mathcal{X}$) *or* HRP($\mathcal{X}$) *for $\mathcal{X} = [0,1]^{d_X}$, a collection $\star(\mathcal{Y})$ among* UDP($\mathcal{Y}$), RDP($\mathcal{Y}$), RDSP($\mathcal{Y}$), RSP($\mathcal{Y}$) *or* HRP($\mathcal{Y}$) *and a maximal degree for the polynomials $\mathbf{D} \in \mathcal{N}^{\lceil \mathcal{Y}}$.*

*Let*

$$\mathcal{S} = \left\{ S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \middle| \mathcal{P} = \{\mathcal{R}_l\} \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})} \text{ and } \forall \mathcal{R}_l \in \mathcal{P}, \mathcal{Q}_l \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{Y})} \right\}.$$

*Then there exist a $C_\star > 0$ and a $c_\star > 0$ independent of $n$, such that for any $\rho$ and for any $C_1 > 1$, the penalized estimator of Theorem 22 satisfies*

$$\mathbb{E}\left[ JKL_{\rho,\lambda}^{\otimes n}(s_0, \widehat{s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}}) \right] \leq C_1 \inf_{S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in \mathcal{S}} \left( \inf_{s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}} KL_{\lambda}^{\otimes n}(s_0, s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}) + \frac{\mathrm{pen}(\mathcal{Q}^{\mathcal{P}}, \mathbf{D})}{N} \right) + C_2 \frac{1}{N} + \frac{\eta + \eta'}{N}$$

*as soon as*

$$\mathrm{pen}(\mathcal{Q}^{\mathcal{P}}, \mathbf{D}) \geq \widetilde{\kappa}\, \mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$$

*for*

$$\widetilde{\kappa} > \kappa_0 \left( C_\star + c_\star \left( A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})} + B_0^{\star(\mathcal{Y})} \right) + 2\log n \right).$$

*where $\kappa_0$ and $C_2$ are the constants of Theorem 22 that depend only on $\rho$ and $C_1$. Furthermore $C_\star \leq \frac{1}{2}\log(8\pi e) + \sum_{d=1}^{d_Y} \log\left(\sqrt{2}(\mathbf{D}_d + 1)\right)$ and $c_\star \leq 2\log 2$.*

A penalty chosen proportional to the dimension of the model, the multiplicative factor $\widetilde{\kappa}$ being constant over $n$ up to a logarithmic factor, is thus sufficient to guaranty the estimator performance. Furthermore, one can use a penalty which is a sum of penalties for each hyperrectangle of the partition:

$$\mathrm{pen}(\mathcal{Q}^{\mathcal{P}}, \mathbf{D}) = \sum_{\mathcal{R}_{l,k}^{\times} \in \mathcal{Q}^{\mathcal{P}}} \widetilde{\kappa} \left( \prod_{d=1}^{d_Y} (\mathbf{D}_d + 1) \right).$$

This additive structure of the penalty allows to use the fast partition optimization algorithm of Donoho [Do97] and Huang, Pollak, Do, and Bouman [Hu+06] as soon as the partition collection is tree structured.

The requirement on the penalty can be weakened to

$$\mathrm{pen}(\mathcal{Q}^{\mathcal{P}},\mathbf{D}) \geq \kappa\Bigg(\left(C_\star + 2\log\frac{N}{\sqrt{\|\mathcal{Q}^{\mathcal{P}}\|}}\right)\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} + c_\star\left(A_0^{\star(\mathcal{X})} + \left(B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})}\right)\|\mathcal{P}\| + B_0^{\star(\mathcal{Y})}\sum_{\mathcal{R}_l\in\mathcal{P}}\|\mathcal{Q}_l\|\right)\Bigg)$$

in which the complexity part and the coding part appear more explicitly. This smaller penalty is no longer proportional to the dimension but still sufficient to guaranty the estimator performance. Using the crude bound $\|\mathcal{Q}^{\mathcal{P}}\| \geq 1$, one sees that such a penalty can still be upper bounded by a sum of penalties over each hyperrectangle. The loss with respect to the original penalty is of order $\kappa \log\|\mathcal{Q}^{\mathcal{P}}\|\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$, which is negligible as long as the number of hyperrectangle remains small with respect to $n^2$.

Some variations around this Theorem can be obtained through simple modifications of its proof. For example, the term $2\log(n/\sqrt{\|\mathcal{Q}^{\mathcal{P}}\|})$ disappears if $\mathcal{P}$ belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$ while $\mathcal{Q}_l$ is independent of $\mathcal{R}_l$ and belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$. Choosing the degrees $\mathbf{D}$ of the polynomial among a family $\mathcal{D}^M$ either globally or locally as proposed by Willet and Nowak [WN07] is also possible. The constant $C_\star$ is replaced by its maximum over the family considered, while the coding part is modified by replacing respectively $A_0^{\star(\mathcal{X})}$ by $A_0^{\star(\mathcal{X})} + \log|\mathcal{D}^M|$ for a global optimization and $B_0^{\star(\mathcal{Y})}$ by $B_0^{\star(\mathcal{Y})} + \log|\mathcal{D}^M|$ for a local optimization. Such a penalty can be further modified into an additive one with only minor loss. Note that even if the family and its maximal degree grows with $n$, the *constant* $C_\star$ grows at a logaritmic rate in $n$ as long as the maximal degree grows at most polynomially with $n$.

Finally, if we assume that the true conditional density is lower bounded, then

$$KL_\lambda^{\otimes n}(s,t) \leq \left\|\frac{1}{t}\right\|_\infty \|s-t\|_{\lambda,2}^{\otimes_n,2}$$

as shown by Kolaczyk and Nowak [KN05]. We can thus reuse ideas from Willet and Nowak [WN07], Akakpo [Ak12] or Akakpo and Lacour [AL11] to infer the quasi optimal minimaxity of this estimator for anisotropic Besov spaces (see for instance in Karaivanov and Petrushev [KP03] for a definition) whose regularity indices are smaller than 1 along the axes of $\mathcal{X}$ and smaller than $\mathbf{D}+1$ along the axes of $\mathcal{Y}$.

### 5.4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties

In this section, we consider an extension of Gaussian mixture that takes account into the covariate into the mixing proportion. This model has been motivated by the unsupervised hyperspectral image segmentation problem mentioned in the introduction. We recall first some basic facts about Gaussian mixtures and their uses in unsupervised classification.

In a classical Gaussian mixture model, the observations are assuming to be drawn from several different classes, each class having a Gaussian law. Let $K$ be the number of different Gaussians, often call the number of clusters, the density $s_0$ of $Y_i$ with respect to the Lebesgue measure is thus modeled as

$$s_{K,\theta,\pi}(\cdot) = \sum_{k=1}^{K}\pi_k\Phi_{\theta_k}(\cdot)$$

where

$$\Phi_{\theta_k}(y) = \frac{1}{(2\pi\det\Sigma_k)^{p/2}}\,e^{-\frac{1}{2}(y-\mu_k)'\Sigma_k^{-1}(y-\mu_k)}$$

with $\mu_k$ the mean of the $k$th component, $\Sigma_k$ its covariance matrix, $\theta_k = (\mu_k,\Sigma_k)$ and $\pi_k$ its mixing proportion. A model $S_{K,\mathcal{G}}$ is obtained by specifying the number of component $K$ as well as a set $\mathcal{G}$ to which should belong the $K$-tuple of Gaussian $(\Phi_{\theta_1},\ldots,\Phi_{\theta_K})$. Those Gaussians can share for instance the same shape, the same volume or the same diagonalization basis. The classical choices are described for

instance in Biernacki, Celeux, Govaert, and Langrognet [Bi+06]. Using the EM algorithm, or one of its extension, one can efficiently obtain the proportions $\widehat{\pi}_k$ and the Gaussian parameters $\widehat{\theta}_k$ of the maximum likelihood estimate within such a model. Using tools also derived from Massart [Ma07], Maugis and Michel [MM12a] show how to choose the number of classes by a penalized maximum likelihood principle. These Gaussian mixture models are often used in unsupervised classification application: one observes a collection of $Y_i$ and tries to split them into homogeneous classes. Those classes are chosen as the Gaussian components of an estimated Gaussian mixture close to the density of the observations. Each observation can then be assigned to a class by a simple maximum likelihood principle:

$$\widehat{k}(y) = \underset{1\leq k\leq \widehat{K}}{\operatorname{argmax}}\, \widehat{\pi}_k \Phi_{\widehat{\theta}_k}(y).$$

This methodology can be applied directly to an hyperspectral image and yields a segmentation method, often called spectral method in the image processing community. This method however fails to exploit the spatial organization of the pixels.

To overcome this issue, Kolaczyk, Ju, and Gopal [KJG05] and Antoniadis, Bigot, and Sachs [ABS08] propose to use mixture model in which the mixing proportions depend on the covariate $X_i$ while the mixture components remain constant. We propose to estimate simultaneously those mixing proportions and the mixture components with our partition-based strategy. In a semantic analysis context, in which documents replace pixels, a similar Gaussian mixture with varying weight, but without the partition structure, has been proposed by Si and Jin [SJ05] as an extension of a general mixture based semantic analysis model introduced by Hofmann [Ho99] under the name *Probabilistic Latent Semantic Analysis*. A similar model has also been considered in the work of Young and Hunter [YH10]. In our approach, for a given partition $\mathcal{P}$, the conditional density $s(\cdot|x)$ are modeled as

$$s_{\mathcal{P},K,\theta,\pi}(\cdot|x) = \sum_{\mathcal{R}_l\in\mathcal{P}} \left( \sum_{k=1}^{K} \pi_k[\mathcal{R}_l] \Phi_{\theta_k}(\cdot) \right) \mathbf{1}_{\{x\in\mathcal{R}_l\}}$$

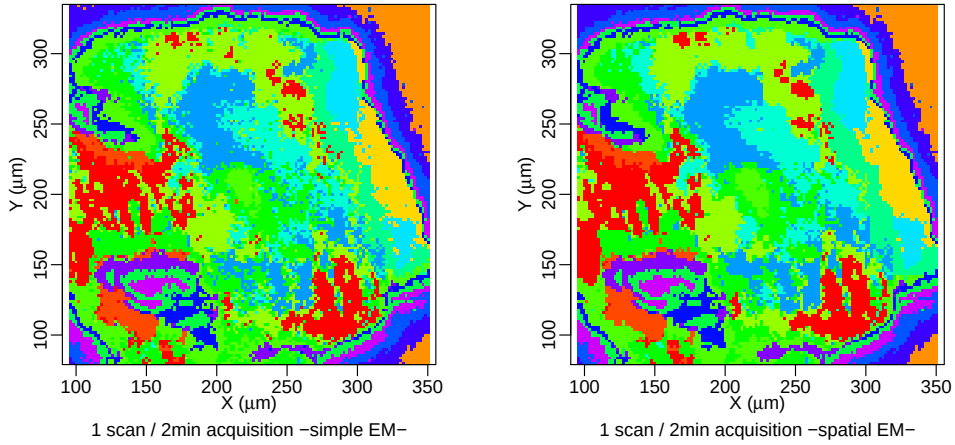
which, denoting $\pi[\mathcal{R}(x)] = \sum_{\mathcal{R}_l\in\mathcal{P}} \pi[\mathcal{R}_l]\, \mathbf{1}_{\{x\in\mathcal{R}_l\}}$, can advantageously be rewritten

$$= \sum_{k=1}^{K} \pi_k[\mathcal{R}(x)] \Phi_{\theta_k}(\cdot)\,.$$

The $K$-tuples of Gaussian can be chosen is the same way as in the classical Gaussian mixture case. Using a penalized maximum likelihood strategy, a partition $\widehat{\mathcal{P}}$, a number of Gaussian components $\widehat{K}$, their parameters $\widehat{\theta}_k$ and all the mixing proportions $\widehat{\pi}[\widehat{\mathcal{R}_l}]$ can be estimated. Each pair of pixel position and spectrum $(x, y)$ can then be assigned to one of the estimated mixture components by a maximum likelihood principle:

$$\widehat{k}(x,y) = \underset{1\leq k\leq \widehat{K}}{\operatorname{argmax}}\, \widehat{\pi}_k[\widehat{\mathcal{R}_l}(x)] \Phi_{\widehat{\theta}_k}(y).$$

This is the strategy we have used at IPANEMA [*Art-Be+11*] to segment, in an unsupervised manner, hyperspectral images. In these images, a spectrum $Y_i$, with around 1000 frequency bands, is measured at each pixel location $X_i$ and our aim was to derive a partition in *homogeneous* regions without any human intervention. This is a precious help for users of this imaging technique as this allows to focus the study on a few representative spectrums. Combining the classical EM strategy for the Gaussian parameter estimation (see for instance Biernacki, Celeux, Govaert, and Langrognet [Bi+06]) and dynamic programming strategies for the partition, as described for instance by Kolaczyk, Ju, and Gopal [KJG05], we have been able to implement this penalized estimator and to test it on real datasets. Figure 5.2 illustrates this methodology. The studied sample is a thin cross-section of maple with a single layer of hide glue on top of it, prepared recently using materials and processes from the Cité de la Musique, using materials of the same type and quality that is used for lutherie. We present here the result for a




(a) (b)

Figure 5.2: Unsupervised segmentation result: a) with constant mixing proportions b) with piecewise constant mixing proportions.

low signal to noise ratio acquisition requiring only two minutes of scan. Using piecewise constant mixing proportions instead of constant mixing proportions leads to a better geometry of the segmentation, with less isolated points and more structured boundaries. As described in a more applied study [*Unpub-CLP12c*], this methodology permits to work with a much lower signal to noise ratio and thus allows to reduce significantly the acquisition time.

We should now specify the models we consider. As we follow the construction of Section 5.4.1, for a given segmentation $\mathcal{P}$, this amounts to specify the set $\mathcal{F}$ to which belong the $\|\mathcal{P}\|$-tuples of densities $(s(y|\mathcal{R}_l))_{\mathcal{R}_l\in\mathcal{P}}$. As described above, we assume that $s(y|\mathcal{R}_l) = \sum_{k=1}^{K} \pi_k[\mathcal{R}_l]\Phi_{\theta_k}(y)$. The mixing proportions within the region $\mathcal{R}_l$, $\pi[\mathcal{R}_l]$, are chosen freely among all vectors of the $K-1$ dimensional simplex $\mathcal{S}_{K-1}$:

$$\mathcal{S}_{K-1} = \left\{ \pi = (\pi_1, \ldots, \pi_k) \middle| \forall k, 1 \leq k \leq K, \pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1 \right\}.$$

As we assume the mixture components are the same in each region, for a given number of components $K$, the set $\mathcal{F}$ is entirely specified by the set $\mathcal{G}$ of $K$-tuples of Gaussian $(\Phi_{\theta_1}, \ldots, \Phi_{\theta_K})$ (or equivalently by a set $\Theta$ for $\theta = (\theta_1, \ldots, \theta_K)$).

To allow variable selection, we follow Maugis and Michel [MM12a] and let $E$ be an arbitrary subspace of $\mathcal{Y} = \mathbb{R}^p$, that is expressed differently for the different classes, and let $E^\perp$ be its orthogonal, in which all classes behave similarly. We assume thus that

$$\Phi_{\theta_k}(y) = \Phi_{\theta_{E,k}}(y_E)\Phi_{\theta_{E^\perp}}(y_{E^\perp})$$

where $y_E$ and $y_{E^\perp}$ denote, respectively, the projection of $y$ on $E$ and $E^\perp$, $\Phi_{\theta_{E,k}}$ is a Gaussian whose parameters depend on $k$ while $\Phi_{\theta_{E^\perp}}$ is independent of $k$. A model is then specified by the choice of a set $\mathcal{G}_E^K$ for the $K$-tuples $(\Phi_{\theta_{E,1}}, \ldots, \Phi_{\theta_{E,K}})$ (or equivalently a set $\Theta_E^K$ for the $K$-tuples of parameters $(\theta_{E,1}, \ldots, \theta_{E,K})$) and a set $\mathcal{G}_{E^\perp}$ for the Gaussian $\Phi_{\theta_{E^\perp}}$ (or equivalently a set $\Theta_{E^\perp}$ for its parameter $\theta_{E^\perp}$). The resulting model is denoted $S_{\mathcal{P},K,\mathcal{G}}$

$$S_{\mathcal{P},K,\mathcal{G}} = \left\{ s_{\mathcal{P},K,\theta,\pi}(y|x) = \sum_{k=1}^{K} \pi_k[\mathcal{R}(x)]\, \Phi_{\theta_{E,k}}(y_E)\, \Phi_{\theta_{E^\perp}}(y_{E^\perp}) \middle| \begin{array}{l} (\Phi_{\theta_{E,1}}, \ldots, \Phi_{\theta_{E,K}}) \in \mathcal{G}_E^K, \\ \Phi_{\theta_{E^\perp}} \in \mathcal{G}_{E^\perp}, \\ \forall \mathcal{R}_l \in \mathcal{P}, \pi[\mathcal{R}_l] \in \mathcal{S}_{K-1} \end{array} \right\}.$$

The sets $\mathcal{G}_E^K$ and $\mathcal{G}_{E^\perp}$ are chosen among the *classical* Gaussian $K$-tuples, as described for instance in Biernacki, Celeux, Govaert, and Langrognet [Bi+06]. For a space $E$ of dimension $p_E$ and a fixed number $K$ of classes, we specify the set

$$\mathcal{G} = \left\{ (\Phi_{E,\theta_1}, \ldots, \Phi_{E,\theta_K}) \middle| \theta = (\theta_1, \ldots, \theta_K) \in \Theta_{[\cdot]_{p_E}^K} \right\}$$

through a parameter set $\Theta_{[\cdot]_{p_E}^K}$ defined by some (mild) constraints on the means $\mu_k$ and some (strong) constraints on the covariance matrices $\Sigma_k$.

The $K$-tuple of means $\mu = (\mu_1, \ldots, \mu_K)$ is either known or unknown without any restriction. A stronger structure is imposed on the $K$-tuple of covariance matrices $(\Sigma_1, \ldots, \Sigma_K)$. To define it, we need to introduce a decomposition of any covariance matrix $\Sigma$ into $LDAD'$ where, denoting $|\Sigma|$ the determinant of $\Sigma$, $L = |\Sigma|^{1/p_E}$ is a positive scalar corresponding to the volume, $D$ is the matrix of eigenvectors of $\Sigma$ and $A$ the diagonal matrix of renormalized eigenvalues of $\Sigma$ (the eigenvalues of $|\Sigma|^{-1/p_E}\Sigma$). Note that this decomposition is not unique as, for example, $D$ and $A$ are defined up to a permutation. We impose nevertheless a structure on the $K$-tuple $(\Sigma_1, \ldots, \Sigma_K)$ through structures on the corresponding $K$-tuples of $(L_1, \ldots, L_K)$, $(D_1, \ldots, D_K)$ and $(A_1, \ldots, A_K)$. They are either known, unknown but with a common value or unknown without any restriction. The corresponding set is indexed by $[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]_{p_E}^K$ where $\star = 0$ means that the quantity is known, $\star = K$ that the quantity is unknown without any restriction and possibly different for every class and its lack means that there is a common unknown value over all classes.

To have a set with finite bracketing entropy, we further restrict the values of the means $\mu_k$, the volumes $L_k$ and the renormalized eigenvalue matrix $A_k$. The means are assumed to satisfy $\forall 1 \leq k \leq K, |\mu_k| \leq a$ for a known $a$ while the volumes satisfy $\forall 1 \leq k \leq K, L_- \leq L_k \leq L_+$ for some known positive values $L_-$ and $L_+$. To describe the constraints on the renormalized eigenvalue matrix $A_k$, we define the set $\mathcal{A}(\lambda_-, \lambda_+, p_E)$ of diagonal matrices $A$ such that $|A| = 1$ and $\forall 1 \leq i \leq p_E, \lambda_- \leq A_{i,i} \leq \lambda_+$. Our assumption is that all the $A_k$ belong to $\mathcal{A}(\lambda_-, \lambda_+, p_E)$ for some known values $\lambda_-$ and $\lambda_+$.

Among the $3^4 = 81$ such possible sets, six of them have been already studied by Maugis and Michel [MM12a; MM12b] in their classical Gaussian mixture model analysis:

- $[\mu_0 \, \mathrm{L}_K \, \mathrm{D}_0 \, \mathrm{A}_0]_{p_E}^K$ in which only the volume of the variance of a class is unknown. They use this model with a single class to model the non discriminant variables in $E^\perp$.
- $[\mu_K \, \mathrm{L}_K \, \mathrm{D}_0 \, \mathrm{A}_K]_{p_E}^K$ in which one assumes that the unknown variances $\Sigma_k$ can be diagonalized in the same known basis $\mathrm{D}_0$.
- $[\mu_K \, \mathrm{L}_K \, \mathrm{D}_K \, \mathrm{A}_K]_{p_E}^K$ in which everything is free,
- $[\mu_K \, \mathrm{L} \, \mathrm{D}_0 \, \mathrm{A}]_{p_E}^K$ in which the variances $\Sigma_k$ are assumed to be equal and diagonalized in the known basis $\mathrm{D}_0$.
- $[\mu_K \, \mathrm{L} \, \mathrm{D}_0 \, \mathrm{A}_K]_{p_E}^K$ in which the volumes $\mathbb{L}_k$ are assumed to be equal and the variance can be diagonalized in the known basis $\mathrm{D}_0$
- $[\mu_K \, \mathrm{L} \, \mathrm{D} \, \mathrm{A}]_{p_E}^K$ in which the variances $\Sigma_k$ are only assumed to be equal

All these cases, as well as the others, are covered by our analysis with a single proof.

To summarize, our models $S_{\mathcal{P},K,\mathcal{G}}$ are parametrized by a partition $\mathcal{P}$, a number of components $K$, a set $\mathcal{G}$ of $K$-tuples of Gaussian specified by a space $E$ and two parameter sets, a set $\Theta_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]_{p_E}^K}$ of $K$-tuples of Gaussian parameters for the differentiated space $E$ and a set $\Theta_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]_{p_{E^\perp}}}$ of Gaussian parameters for its orthogonal $E^\perp$. Those two sets are chosen among the ones described above with the same constants $a$, $L_-$, $L_+$, $\lambda_-$ and $\lambda_+$. One verifies that

$$\dim\left(()\, S_{\mathcal{P},K,\mathcal{G}}\right) = \|\mathcal{P}\|(K-1) + \dim\left(\left(\Theta_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]_{p_E}^K}\right)\right) + \dim\left(\left(\Theta_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]_{p_{E^\perp}}}\right)\right).$$

Before stating a model selection theorem, we should specify the collections $\mathcal{S}$ considered. We consider sets of model $S_{\mathcal{P},K,\mathcal{G}}$ with $\mathcal{P}$ chosen among one of the partition collections $\mathcal{S}_{\mathcal{P}}^\star$, $K$ smaller than $K_M$, which

can be theoretically chosen equal to $+\infty$, a space $E$ chosen as $\mathrm{Span}\{e_i\}_{i\in I}$ where $e_i$ is the canonical basis of $\mathbb{R}^p$ and $I$ a subset of $\{1,\dots,p\}$ is either known, equal to $\{1,\dots,p_E\}$ or free and the indices $[\mu_\star\, \mathrm{L}_\star\, \mathrm{D}_\star\, \mathrm{A}_\star]$ of $\Theta_E$ and $\Theta_{E^\perp}$ are chosen freely among a subset of the possible combinations.

Without any assumptions on the design, we obtain

**Theorem 24.** *Assume the collection $\mathcal{S}$ is one of the collections of the previous paragraph.*

*Then, there exist a $C_\star > \pi$ and a $c_\star > 0$, such that, for any $\rho$ and for any $C_1 > 1$, the penalized estimator of Theorem 22 satisfies*

$$\mathbb{E}\left[JKL^{\otimes n}_{\rho,\lambda}(s_0, \widehat{s}_{\widehat{\mathcal{P},K,\mathcal{G}}})\right] \leq C_1 \inf_{S_{\mathcal{P},K,\mathcal{G}}\in\mathcal{S}} \left( \inf_{s_{\mathcal{P},K,\mathcal{G}}\in S_{\mathcal{P},K,\mathcal{G}}} KL^{\otimes n}_{\lambda}(s_0, s_{\mathcal{P},K,\mathcal{G}}) + \frac{\mathrm{pen}(\mathcal{P},K,\mathcal{G})}{N} \right) + \frac{C_2}{N} + \frac{\eta+\eta'}{N}$$

*as soon as*

$$\mathrm{pen}(\mathcal{P},K,\mathcal{G}) \geq \widetilde{\kappa}_1 \dim\left(()\, S_{\mathcal{P},K,\mathcal{G}}\right) + \widetilde{\kappa}_2 \mathcal{D}_E$$

*for*

$$\widetilde{\kappa}_1 \geq \kappa\left(\left(2C_\star + 1 + \left(\log\frac{N}{eC_\star}\right)_+ + c_\star\left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} + 1\right)\right)\right) \qquad \text{and} \qquad \widetilde{\kappa}_2 \geq \kappa c_\star$$

*with $\kappa > \kappa_0$ where $\kappa_0$ and $C_2$ are the constants of Theorem 22 that depend only on $\rho$ and $C_1$ and*

$$\mathcal{D}_E = \begin{cases} 0 & \text{if } E \text{ is known,} \\ p_E & \text{if } E \text{ is chosen among spaces spanned by the first coordinates,} \\ (1+\log 2 + \log\frac{p}{p_E})p_E & \text{if } E \text{ is free.} \end{cases}$$

As in the previous section, the penalty term can thus be chosen, up to the variable selection term $\mathcal{D}_E$, proportional to the dimension of the model, with a proportionality factor constant up to a logarithmic term with $n$. A penalty proportional to the dimension of the model is thus sufficient to ensure that the model selected performs almost as well as the best possible model in term of conditional density estimation. As in the proof of Antoniadis, Bigot, and Sachs [ABS08], we can also obtain that our proposed estimator yields a minimax estimate for spatial Gaussian mixture with mixture proportions having a geometrical regularity even without knowing the number of classes.

Moreover, again as in the previous section, the penalty can have an additive structure, it can be chosen as a sum of penalties over each hyperrectangle plus one corresponding to $K$ and the set $\mathcal{G}$. Indeed

$$\mathrm{pen}(\mathcal{P},K,\mathcal{G}) = \sum_{\mathcal{R}_l\in\mathcal{P}} \widetilde{\kappa}_1(K-1) + \widetilde{\kappa}_1\left(\dim\left(\left(\Theta_{[\mu_\star\, \mathrm{L}_\star\, \mathrm{D}_\star\, \mathrm{A}_\star]^K_{p_E}}\right)\right) + \dim\left(\left(\Theta_{[\mu_\star\, \mathrm{L}_\star\, \mathrm{D}_\star\, \mathrm{A}_\star]_{p_{E^\perp}}}\right)\right)\right) + \widetilde{\kappa}_2\mathcal{D}_E$$

satisfies the requirement of Theorem 24. This structure is the key for our numerical minimization algorithm in which one optimizes alternately the Gaussian parameters with an EM algorithm and the partition with the same fast optimization strategy as in the previous section.

In Appendix, we obtain a weaker requirement

$$\mathrm{pen}(\mathcal{P},K,\mathcal{G}) \geq \kappa\left(\left(2C_\star + 1 + \left(\log\frac{N}{eC_\star \dim\left(()\, S_{\mathcal{P},K,\mathcal{G}}\right)}\right)_+\right)\dim\left(()\, S_{\mathcal{P},K,\mathcal{G}}\right) + c_\star\left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})}\|\mathcal{P}\| + (K-1) + \mathcal{D}_E\right)\right)$$

in which the complexity and the coding terms are more explicit. Again up to a logarithmic term in $\dim\left(()\, S_{\mathcal{P},K,\mathcal{G}}\right)$, this requirement can be satisfied by a penalty having the same additive structure as in the previous paragraph.

Our theoretical result on the conditional density estimation does not guaranty good segmentation performance. If data are generated according to a Gaussian mixture with varying mixing proportions,

one could nevertheless obtain the asymptotic convergence of our class estimator to the optimal Bayes one. We have nevertheless observed in our numerical experiments at IPANEMA that the proposed methodology allow to reduce the signal to noise ratio while keeping meaningful segmentations.

Two major questions remain nevertheless open. Can we calibrate the penalty (choosing the constants) in a datadriven way while guaranteeing the theoretical performance in this specific setting? Can we derive a non asymptotic classification result from this conditional density result? The *slope heuristic*, proposed by Birgé and Massart [BM07], we have used in our numerical experiments, seems a promising direction. Deriving a theoretical justification in this conditional estimation setting would be much better. Linking the non asymptotic estimation behavior to a non asymptotic classification behavior appears even more challenging.

With our PhD student L. Montuelle, we are considering extensions of this framework. We have first considered a similar model in which both the proportions and the means of the Gaussian depends on the covariate. Using a logistic model for the weights, she has been able to show that a penalty proportional to the dimension of the models also leads to oracle inequalities in this case [*Unpub-MCLP12*]. She has also obtained results when replacing the Gaussian mixtures by Poissonian Mixtures.

## 5.5 Binary choice model

I would like to conclude this chapter with a description of work in progress with E. Gautier [*Unpub-GLP11*]. The model we consider is a complex inverse model called the binary choice model in which the unknown is a density which is observed indirectly through a conditional density. In this model, one observe i.i.d. samples $(X_i, Y_i)$ where the $X_i$ are random vectors of norm 1 that has density $s_X$ on the hypersphere and $Y_i = 2\mathbf{1}_{\{\langle X_i, \beta_i\rangle > 0\}} - 1$ where the $\beta_i$ are random vectors of norm 1 of density $s_\beta$ on the hypersphere independents of all $X_i$'s. The goal is to estimate $s_\beta$ from the observation. As explained later, the conditional probability $\mathbb{P}\{Y = 1|X = x\}$ is related to the density $s_\beta$ by a simple known compact estimator. The key to estimate $s_\beta$ is thus to estimate first the conditional probability.

This model originates from econometrics in which Discrete choice models are important models in economics for the choice of agents between a number of exhaustive and mutually exclusive alternatives. They have applications in many areas ranging from empirical industrial organizations, labor economics, health economics, planning of public transportation, evaluation of public policies, etc. For a review, the interested reader can refer to the Nobel lectures of Mc Fadden [McFa01]. We consider here a binary choice model where individuals only have two options. In a random utility framework, an agent chooses the alternative that yields the higher utility. Assume that the utility for each alternative is linear in regressors which are observed by the statistician. The regressors are typically attributes of the alternative faced by the individuals, *e.g.* the cost or time to commute from home to one's office for each of the two transport alternatives. Because this linear structure is an ideal situation and because the statistician is missing some factors, the utilities are written as the linear combination of the regressors plus some random error term. When the utility difference is positive the agent chooses the first alternative, otherwise he chooses the second. The Logit, Probit or Mixed-Logit models are particular models of this type. We consider the case where the coefficients of the regressors are random. This accounts for heterogeneity or taste variation: each individual is allowed to have his own set of coefficients (the preferences or tastes). Like in the work of Gautier and Kitamura [GK12]; Ichimura and Thompson [IT98], we consider a nonparametric treatment of the joint distribution of the error term and vector of random coefficients. Nonparametric treatment of unobserved heterogeneity is very important in economics, references include the work of Beran and Hall [BH92]; Elbers and Ridder [ER82]; Gautier and Kitamura [GK12]; Heckman and Singer [HS84]; Hoderlein, Klemelä, and Mammen [HKM10]; Ichimura and Thompson [IT98]. It allows to be extremely flexible about the joint distribution of the preferences (as well as the error term). Random coefficients models can be viewed as mixture models. They also have a Bayesian interpretation, see for example the work of Healy and Kim [HK96] for a similar model on the sphere. Nonparametric estimation of the density of the vector of random coefficients corresponds to nonparametric estimation of a prior in the empirical Bayes setting.

In the nonparametric random coefficients binary choice model we assume that we have $n$ i.i.d. observations $(x_i, y_i)$ of $(X, Y)$ where $X$ is a random vector of Euclidean norm 1 in $\mathbb{R}^d$ and $Y$ is a discrete

random variable and $Y$ and $X$ are related through a non observed random vector $\beta$ of norm 1 by

$$Y = 2\mathbf{1}_{\{\langle X,\beta\rangle>0\}} - 1 = \begin{cases} 1 & \text{if } X \text{ and } \beta \text{ are in the same hemisphere} \\ -1 & \text{otherwise.} \end{cases} \tag{5.3}$$

In (5.3), $\langle \cdot, \star \rangle$ is the scalar product in $\mathbb{R}^d$. We make the assumption that $X$ and $\beta$ are independent. This assumption corresponds to the exogeneity of the regressors. It could be relaxed using instrumental variables (see Gautier and Kitamura [GK12]). $-1$ and 1 are labels for the two choices. They correspond to the sign of $\langle X, \beta \rangle$. $X$ and $\beta$ are assumed to be of norm 1 because only the sign of $\langle X, \beta \rangle$ matters in the choice mechanism. The regressors in the latent variable model are thus assumed to be properly rescaled. Model (5.3) allows for arbitrary dependence between the random unobservables. In this model, $X$ corresponds to a vector of regressors where, in an original scale, the first component is 1 and the remaining components are the regressors in the binary choice model. The 1 stands because in applications we always include a constant in the latent variable model for the binary choice model. The first element of $\beta$ in this formulation absorbs the usual error term as well as the constant in standard binary choice models with non-random coefficients. We assume that $X$ and $\beta$ have densities $s_X$ and $s_\beta$ with respect to the spherical measure $\sigma$ on the unit sphere $\mathbb{S}^{d-1}$ of the Euclidean space $\mathbb{R}^d$. Because in the original scale the first component of $X$ is 1, the support of $X$ is included in $H^+ = \{x \in \mathbb{S}^{d-1} : < x, (1, 0, \dots, 0) >\geq 0\}$. We assume, for simplicity, through this paper, that the support of $X$ satisfies supp $s_X = H^+$. In the work of Gautier and Kitamura [GK12], the case of regressors with limited support, including dummy variables is also studied but identification requires that these variables, as well as one continuously distributed regressor, are not multiplied by random coefficients.

The estimation of the density of the random coefficient can be rewritten as a kind of linear ill-posed inverse problem. We can write for $x \in H^+$,

$$\mathbb{E}[Y|X = x] = \int_{b\in\mathbb{S}^{d-1}} \operatorname{sign}\left(\langle x, b\rangle\right) s_\beta(b) d\sigma(b) \tag{5.4}$$

where sign denotes the sign. This can then be rewritten in terms of another operator from integral geometry:

$$\mathbb{P}(Y = 1|X = x) = \frac{\mathbb{E}[Y|X = x] + 1}{2} = \int_{b\in\mathbb{S}^{d-1}} \mathbf{1}_{\{\langle x,b\rangle>0\}} s_\beta(b) d\sigma(b) \triangleq \mathcal{H}\left(s_\beta\right)(x). \tag{5.5}$$

The operator $\mathcal{H}$ is called the hemispherical transform. $\mathcal{H}$ is a special case of the Pompeiu operator (see, *e.g.*, Zalcman [Za92]). The operator $\mathcal{H}$ arises when one wants to reconstruct a star-shaped body from its *half-volumes* (see Funk [Fu16]). Inversion of this operator was studied by Funk [Fu16]; Rubin [Ru99], it can be achieved in the spherical harmonic basis (also called the Fourier Laplace basis as the extension of the Fourier basis on $\mathbb{S}^1$ and the Laplace basis in $\mathbb{S}^2$), using polynomials in the Laplace-Beltrami operator for certain dimensions and using a continuous wavelet transform. Rubin [Ru99], and in a certain extent Groemer [Gr96], also discuss some of its properties. It is an operator which is diagonal in the spherical harmonic basis and which eigenvalues are known explicitly. Were the function $\mathbb{P}\{Y = 1|X = x\}$ exactly known, this would have been a deconvolution problem on the sphere. It is however not the case and this function can only be estimated: this can be seen as a regression problem with unknown random design or equivalently here as a conditional density estimation problem.

Deconvolution on the sphere has been studied by various authors among which Healy and Kim [HK96]; Kerkyacharian, Phan Ngoc, and Picard [KPNP11]; Kim and Koo [KK00]. Because of the indicator function, this is a type of boxcar deconvolution. Boxcar deconvolution has been studied in specific cases by Johnstone and Raimondo [JR04]; Kerkyacharian, Picard, and Raimondo [KPR07]. There are two important difficulties regarding identification: (1) because of the intercept in the latent variable model, the left hand side of (5.5) is not a function defined on the whole sphere, (2) $\mathcal{H}$ is not injective. Proper restrictions are imposed to identify $s_\beta$. Treatment of the random design (possibly inhomogeneous) with unknown distribution appearing in the regression function that has to be inverted is an important difficulty. Regression with random design is a difficult problem, see for example Kerkyacharian and Picard [KP04]; Kulik and Raimondo [KR09] for the case of wavelet thresholding estimation using warped

wavelet for a regression model on an interval, or Gaïffas [Ga09] in the case of inhomogeneous designs. For the conditional density estimation method, we refer at the beginning of this chapter.

Gautier and Kitamura [GK12] propose an estimator using smoothed projections on the finite dimensional spaces spanned by the first vectors of the spherical harmonics basis. It is straightforward to compute in every dimension $d$. Convergence rates for the $L^p$-losses for $p \in [1, \infty]$ and CLT are obtained by Gautier and Kitamura [GK12]. They depend on the degree of smoothing of the operator which is $\nu = d/2$ in the Sobolev spaces based on $L^2$, the smoothness of the unknown function, the smoothness of $s_X$ as well as its degeneracy (when it takes small values or is 0, in particular when $x$ is approaching the boundary of $H^+$). The treatment of the random design is a major difficulty that we have try to cope with.

The goal of our study was to provide an estimator of $s_\beta$ which is adaptive in the unknown smoothness of the function. As the eigenfunctions of the operator are spherical harmonic basis, we have used the corresponding needlets introduced by Narcowich, Petrushev, and Ward [NPW06a]. They were successfully used in statistics to provide adaptive estimation procedures of Baldi, Kerkyacharian, Marinucci, and Picard [Ba+09]; Kerkyacharian, Petrushev, Picard, and Willer [Ke+07]; Kerkyacharian, Phan Ngoc, and Picard [KPNP11] and Kerkyacharian, Kyriazis, Le Pennec, Petrushev, and Picard [*Art-Ke+10*]. As described in our preliminary report [*Unpub-GLP11*], we have considered a method based on the observation that

$$\int_{\mathbb{S}^{d-1}} \mathbb{P}(Y = 1|X = x)b(x)dx = \int_{\mathbb{S}^{d-1}} \frac{\mathbb{P}(Y = 1|X = x)}{s_X(x)} b(x)s(x)dx$$

which can be estimated without bias by

$$\sum_{i=1}^{n} \frac{\mathbf{1}_{\{Y_i=1\}}}{s_X(X_i)} b(X_i).$$

As $s_X$ is unknown, we rely on a preliminary estimate of this quantity $\hat{s}_X$ that we plug in this question to estimate the spherical harmonic coefficients. Those coefficients are recombined to obtain needlet coefficients that are then, in the spirit of [*Art-BLPR11*], thresholded using a data-drive threshold. In our study, we obtain a general oracle inequality which, unfortunately, depends heavily on the properties of the preliminary estimate $\hat{s}_X$. After proving lower bounds, we show that those bounds are close to be optimal if $s_X$ is lower and upper bounded and smooth enough. Better results could probably be obtained by estimation $\mathbb{P}\{Y = 1|X = x\}$, a conditional density, without resorting to the usual plugin estimate. Indeed, as shown in the beginning of this chapter, we do not really need to estimate $s_X$ to estimate $\mathbb{P}\{Y = 1|X = x\}$...

# Chapter 6

# Conclusion

This manuscript is an attempt to summarize my contributions of the last ten years. They share some properties: I have studied statistical problems and propose solutions with a strong flavor of approximation theory. The main principle is the one I have attributed to Ockham's: *A good solution is a tradeoff between fidelity and complexity.* In all my contribution, I have tried to give both theoretical and practical, if not numerical, answers. All the themes I have mentioned in the overview are not in the same states.

The NL-Means study (Theme 6) is still a preliminary study from my side but not from J. Salmon's side, which has further studied this model and related ones. The bandlet studies (Themes 0 and 1) as well as the maxiset study of model selection estimator can be considered as closed from my point of view. Most of my questions around those themes are answered, but one that relates precisely the two: Is there a simple characterization of the maxisets of bandlet estimators? The study of wavelet based copula estimation (Theme 5) is also quite comprehensive, but preliminary numerical studies show that better estimators can be obtained by using needlet based estimators. Concerning the Radon transform (Theme 4), the needlet approach has proved to be successful and the results on the axisymmetric case are still to be published. Conditional density estimation appears now as my main subject. On one side, I'm working on the maximum likelihood approach and its application to hyperspectral image analysis (Themes 9 and 8) which are the sources of numerous extension. One the other side, the study of the binary choice model (Theme 7) which falls in this setting still requires some polishing.

I'm currently working on two new projects: the combination of random projection and GMM to reduce the computational cost of unsupervised clustering methods and a project on anomaly detection in complex texture. In the first project with L. Montuelle and S. Cohen, we want to obtain theoretical guarantees of the robustness of the randomly projected GMM method. The last project takes place in an industrial setting as this is the subject of S. Thivin's beginning PhD thesis performed in collaboration with M. Prenat from Thales Optronics. On a lighter note, I'm confident I will eventually encounter the most famous absent statistical model of my previous work: the regression one. My feeling is that this absence is more probably due to technical reasons than practical ones...

# List of communications

## List of publications


[*Art-CLP12b*] S. Cohen and E. Le Pennec. "Partition-Based Conditional Density Estimation". *ESAIM Probab. Stat.* (2012). DOI: 10.1051/ps/2012017.

[*Art-KLPP12*] G. Kerkyacharian, E. Le Pennec, and D. Picard. "Radon needlet thresholding". *Bernoulli* 18.2 (2012), pp. 391–433. DOI: 10.3150/10-BEJ340.

[*Art-BLPR11*] K. Bertin, E. Le Pennec, and V. Rivoirard. "Adaptive Dantzig density estimation". *Ann. Inst. H. Poincaré Probab. Statist.* 47.1 (2011), pp. 43–74. DOI: 10.1214/09-AIHP351.

[*Art-Be+11*] L. Bertrand et al. "European research platform IPANEMA at the SOLEIL synchrotron for ancient and historical materials". *J. Synchrotron Radiat.* 18.5 (2011), pp. 765–772. DOI: 10.1107/S090904951102334X.

[*Art-DLPM11*] Ch. Dossal, E. Le Pennec, and S. Mallat. "Bandlets Image Estimation with Model Selection". *Sig. Process.* 91.12 (2011), pp. 2743–2753. DOI: 10.1016/j.sigpro.2011.01.013.

[*Art-Au+10*] F. Autin, E. Le Pennec, J.-M. Loubes, and V. Rivoirard. "Maxisets for Model Selection". *Constr. Approx.* 31.2 (2010), pp. 195–229. DOI: 10.1007/s00365-009-9062-2.

[*Art-ALPT10*] F. Autin, E. Le Pennec, and K. Tribouley. "Thresholding methods to estimate the copula density". *J. Multivariate Anal.* 101.1 (2010), pp. 200–222. DOI: 10.1016/j.jmva.2009.07.009.

[*Art-Ke+10*] G. Kerkyacharian, G. Kyriazis, E. Le Pennec, P. Petrushev, and D. Picard. "Inversion of noisy Radon transform by SVD based needlets". *Appl. Comput. Harmon. Anal.* 28.1 (2010), pp. 24–45. DOI: 10.1016/j.acha.2009.06.001.

[*Art-LPM05a*] E. Le Pennec and S. Mallat. "Bandelet Image Approximation and Compression". *Multiscale Model. Sim.* 4.3 (2005), pp. 992–1039. DOI: 10.1137/040619454.

[*Art-LPM05b*] E. Le Pennec and S. Mallat. "Sparse Geometrical Image Representation with Bandelets". *IEEE Trans. Image Process.* 14.4 (2005), pp. 423–438. DOI: 10.1109/TIP.2005.843753.


## List of proceedings


[*Proc-CLP11b*] S. Cohen and E. Le Pennec. "Segmentation non supervisée d'image hyperspectrale par mélange de gausiennes spatialisé". In: *GRETSI 11*. Bordeaux, 2011.

[*Proc-LPS09a*] E. Le Pennec and J. Salmon. "Agrégation d'estimateurs pour le débruitage d'image". In: *Journée de la SFdS*. Bordeaux, 2009.

[*Proc-LPS09b*] E. Le Pennec and J. Salmon. "An aggregator point of view on NL-Means". In: *SPIE Wavelet XIII 09*. San Diego, 2009. DOI: 10.1117/12.826881.

[*Proc-SLP09*] J. Salmon and E. Le Pennec. "NL-Means and aggregation procedures". In: *ICIP 09*. 2009, pp. 2977–2980. DOI: 10.1109/ICIP.2009.5414512.

[*Proc-LePe+07*] E. Le Pennec, Ch. Dossal, G. Peyré, and S. Mallat. "Débruitage géométrique d'image dans des bases orthonormées de bandelettes". In: *GRETSI 07*. Troyes, 2007.

[*Proc-Pe+07*] G. Peyré, E. Le Pennec, Ch. Dossal, and S. Mallat. "Geometrical Image Estimation with Orthogonal Bandlets Bases". In: *SPIE Wavelet XII 07*. San Diego, 2007. DOI: 10.1117/12.731227.

[*Proc-LPM03a*] E. Le Pennec and S. Mallat. "Bandelettes et représentation géométrique des images". In: *GRETSI 03*. Paris, 2003.

[*Proc-LPM03b*] E. Le Pennec and S. Mallat. "Geometrical Image Compression with Bandelets". In: *VCIP 03*. Special Session. Lugano, 2003. DOI: 10.1117/12.509904.

[*Proc-LPM01a*] E. Le Pennec and S. Mallat. "Bandelet Representations for Image Compression". In: *ICIP 01*. Special Session. Thessaloniki, 2001. DOI: 10.1109/ICIP.2001.958939.

[*Proc-LPM01b*] E. Le Pennec and S. Mallat. "Représentation d'image par bandelettes et application à la compression". In: *GRETSI 01*. Toulouse, 2001.

[*Proc-LPM00*] E. Le Pennec and S. Mallat. "Image Compression with Geometrical Wavelets". In: *ICIP 00*. Vancouver, 2000. DOI: 10.1109/ICIP.2000.901045.

## List of technical reports and preprints

[*Unpub-BLPT12*] M. Bergounioux, E. Le Pennec, and E. Trelat. "A needlet based inversion for Radon transform of axially symmetric objects". Work in progress. 2012.

[*Unpub-CLP12a*] S. Cohen and E. Le Pennec. "Conditional Density Estimation by Penalized Likelihood Model Selection". *Submitted* (2012).

[*Unpub-CLP12c*] S. Cohen and E. Le Pennec. "Unsupervised segmentation of hyperspectral images with spatialized Gaussian mixture model and model selection". Work in progress. 2012.

[*Unpub-MCLP12*] L. Montuelle, S. Cohen, and E. Le Pennec. "Conditional density estimation by penalized maximum likelihood and model selection". Work in progress. 2012.

[*Unpub-CLP11a*] S. Cohen and E. Le Pennec. *Conditional Density Estimation by Penalized Likelihood Model Selection and Applications*. Tech. rep. INRIA, 2011.

[*Unpub-GLP11*] E. Gautier and E. Le Pennec. "Adaptive Estimation in the Nonparametric Random Coefficients Binary Choice Model by Needlet Thresholding". *Submitted* (2011).

[*Unpub-ALPT08*] F. Autin, E. Le Pennec, and K. Tribouley. *Thresholding methods to estimate the copula density*. Tech. rep. Extended version arXiv:0802.2424. LPMA, 2008.

# Bibliography


[Ak73] H. Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Second International Symposium on Information Theory (Tsahkadsor,1971)*. 1973, pp. 267–281.

[Ak74] H. Akaike. "A new look at the statistical model identification". *IEEE Trans. Autom. Control* 19.6 (1974), pp. 716–723.

[Ak12] N. Akakpo. "Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection". *Mathematical Methods of Statistics* 21.1 (2012), pp. 1–28.

[AL11] N. Akakpo and C. Lacour. "Inhomogeneous and anisotropic conditional density estimation from dependent data". *Electon. J. Statist.* 5 (2011), pp. 1618–1653.

[Al92] B. Alpert. "Wavelets and Other Bases for Fast Numerical Linear Algebra". In: San Diego, CA, USA: C. K. Chui, editor, Academic Press, 1992, pp. 181–216.

[AK10] A. Ammari and A. Karoui. "Stable inversion of the Abel integral equation of the first kind by means of orthogonal polynomials". *Inverse Probl.* 26.10 (2010), p. 105005.

[AAR06] G. Andrew, R. Askey, and R. Roy. *Special Functions*. Cambridge University Press, 2006.

[ABS08] A. Antoniadis, J. Bigot, and R. von Sachs. "A multiscale approach for statistical characterization of functional images". *J. Comput. Graph. Statist.* 18.1 (2008), pp. 216–237.

[Ar+08] F. Arandiga, A. Cohen, R. Donat, N. Dyn, and B. Matei. "Approximation of piecewise smooth images by edge-adapted techniques". *Appl. Comput. Harmon. Anal.* 24 (2008), pp. 225–250.

[AM09] S. Arlot and P. Massart. "Data-driven calibration of penalties for least-squares regression". *J. Mach. Learn. Res.* 10 (2009), pp. 245–279.

[Au06] F. Autin. "Maxiset for density estimation on $\mathbb{R}$". *Math. Methods of Statist.* 15.2 (2006), pp. 123–145.

[Au08] F. Autin. "Maxisets for $\mu$-thresholding rules". *Test* 17.2 (2008), pp. 332–349.

[APR06] F. Autin, D. Picard, and V. Rivoirard. "Large variance Gaussian priors in Bayesian nonparametric estimation: a maxiset approach". *Math. Methods Statist.* 15.4 (2006), pp. 349–373.

[Ba+09] P. Baldi, G. Kerkyacharian, D. Marinucci, and D. Picard. "Adaptive density estimation for directional data using needlets". *Ann. Statist.* 37.6A (2009), pp. 3362–3395. ISSN: 0090-5364.

[Ba00] Y. Baraud. "Model selection for regression on a fixed design". *Probab. Theory Related Fields* 117.4 (2000), pp. 467–493.

[Ba02] Y. Baraud. "Model selection for regression on a random design". *ESAIM Probab. Stat.* 6 (2002), pp. 127–146.

[BBM99] A. Barron, L. Birgé, and P. Massart. "Risk bounds for model selection via penalization". *Probability Theory and Related Fields* 113 (3 1999). 10.1007/s004400050210, pp. 301–413. ISSN: 0178-8051. URL: http://dx.doi.org/10.1007/s004400050210.

[Ba+08] A. Barron, C. Huang, J. Li, and X. Luo. "MDL Principle, Penalized Likelihood, and Statistical Risk". In: *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*. Tampere University Press, 2008.

[BH01] D. Bashtannyk and R. Hyndman. "Bandwidth selection for kernel conditional density estimation". *Comput. Statist. Data Anal.* 36.3 (2001), pp. 279–298.

[BH92] R. Beran and P. Hall. "Estimating coefficient distributions in random coefficient regression". *Ann. Statist.* 20 (1992), pp. 1970–1984.

[BT10] M. Bergounioux and E. Trélat. "A variational method using fractional order Hilbert spaces for tomographic reconstruction of blurred and noised binary images". *J. Funct. Anal.* 259.10 (2010), pp. 2296–2332.

[BR09] K. Bertin and V. Rivoirard. "Maxiset in sup-norm for kernel estimators". *Test* 18.3 (2009), pp. 475–496.

[BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". *Ann. Statist.* 37 (2009), pp. 1705–1732.

[Bi+06] Ch. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. "Model-based cluster and discriminant analysis with the MIXMOD software". *Comput. Statist. Data Anal.* 51.2 (2006), pp. 587–600.

[Bi04] L. Birgé. "Model selection for Gaussian regression with random design". *Bernoulli* 10.6 (2004), pp. 1039–1051.

[Bi08] L. Birgé. "Model selection for density estimation with $\mathbb{L}_2$-loss". arXiv 0808.1416. 2008.

[BM97] L. Birgé and P. Massart. "From model selection to adaptive estimation". In: *Festschrift for Lucien Lecam: Research papers in Probability and Statistics*. Ed. by D. Pollard, E. Torgersen, and G. Yang. New-York: Springer-Verlag, 1997, pp. 55–87.

[BM98] L. Birgé and P. Massart. "Minimum contrast estimators on sieves: exponential bounds and rates of convergence". *Bernoulli* 4.3 (1998), pp. 329–375.

[BM00] L. Birgé and P. Massart. "An adaptive compression algorithm in Besov spaces". *Constr. Approx.* 16.1 (2000), pp. 1–36.

[BM01] L. Birgé and P. Massart. "Gaussian model selection". *J. Eur. Math. Soc.* 3.3 (2001), pp. 203–268.

[BM07] L. Birgé and P. Massart. "Minimal penalties for Gaussian model selection". *Probability theory and related fields* 138.1-2 (2007), pp. 33–73.

[Bl+07] G. Blanchard, C. Schäfer, Y. Rozenholc, and K. Müller. "Optimal dyadic decision trees". *Machine Learning* 66.2 (2007), pp. 209–241.

[BBL05] S. Boucheron, O. Bousquet, and G. Lugosi. "Theory of classification: a survey of some recent advances". *ESAIM Probab. Stat.* 9 (2005), pp. 323–375.

[BM11] S. Boucheron and P. Massart. "A high-dimensional Wilks phenomenon". *Probability Theory and Related Fields* 150.3-4 (2011), pp. 405–433.

[BCL07] E. Brunel, F. Comte, and C. Lacour. "Adaptive Estimation of the Conditional Density in Presence of Censoring". *Sankhyā* 69.4 (2007), pp. 734–763.

[BCM05] A. Buades, B. Coll, and J.-M. Morel. "A review of image denoising algorithms, with a new one". *Multiscale Model. Simul.* 4.2 (2005), 490–530 (electronic). ISSN: 1540-3459.

[Bu09] F. Bunea. "Consistent selection via the Lasso for high dimensional approximating regression models". In: *Pushing the Limits of Contemporary Statistics: Contributions in Honor of J. K. Ghosh*. Vol. 37. Inst. Math. Stat. Collect. Beachwood: IMS, 2009, pp. 2145–2177.

[BTW06] F. Bunea, A. Tsybakov, and M. Wegkamp. "Aggregation and sparsity via $\ell_1$ penalized least squares". In: *Learning Theory*. Vol. 4005. Lecture Notes in Comput. Sci. Berlin: Springer, 2006, pp. 379–391.

[BTW07a] F. Bunea, A. Tsybakov, and M. Wegkamp. "Aggregation for Gaussian regression". *Ann. Statist.* 35.4 (2007), pp. 1674–1697. ISSN: 0090-5364.

[BTW07b] F. Bunea, A. Tsybakov, and M. Wegkamp. "Sparse density estimation with $\ell_1$ penalties". In: *Learning Theory*. Vol. 4539. Lecture Notes in Comput. Sci. Berlin: Springer, 2007, pp. 530–543.

[BTW07c] F. Bunea, A. Tsybakov, and M. Wegkamp. "Sparsity Oracle Inequalities for the Lasso". *Electron. J. Statist.* 1 (2007), pp. 169–194.

[Bu+10] F. Bunea, A. Tsybakov, M. Wegkamp, and A. Barbu. "Spades and Mixture Models". *Ann. Statist.* 38.4 (2010), pp. 2525–2558.

[Ca06] E. Candès. "Modern statistical estimation via oracle inequalities". *Acta Numer.* 15 (2006), 257–325.

[CD99] E. Candès and D. Donoho. "A surprinsingly effective nonadaptive representation for objects with edges". In: *Curves and Surfaces*. 1999.

[CP09] E. Candès and Y. Plan. "Near-ideal model selection by $\ell_1$ minimization". *Ann. Statist.* 37 (2009), pp. 2145–2177.

[CT07] E. Candès and T. Tao. "The Dantzig selector: statistical estimation when $p$ is much larger than $n$". *Ann. Statist.* 35 (2007), pp. 2313–2351.

[Ca04] O. Catoni. *Statistical learning theory and stochastic optimization*. Vol. 1851. Lecture Notes in Mathematics. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. Berlin: Springer-Verlag, 2004, pp. viii+272. ISBN: 3-540-22572-2.

[Ca07] O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Vol. 56. Lecture Notes–Monograph Series. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2007.

[Ca11] L. Cavalier. "Inverse problems in statistics". In: *Inverse Problems and High-Dimensional Estimation*. Ed. by P. Alquier, E. Gautier, and G. Stoltz. Vol. 203. Lecture Notes in Statistics. Springer, 2011.

[Ca+02] L. Cavalier, G. Golubev, D. Picard, and A. Tsybakov. "Oracle inequalities for inverse problems". *Ann. Statist.* 30.3 (2002), pp. 843–874. ISSN: 0090-5364.

[CT02] L. Cavalier and A. Tsybakov. "Sharp adaptation for inverse problems with random noise". *Probab. Theory Related Fields* 123.3 (2002), pp. 323–354. ISSN: 0178-8051.

[CDS01] D. Chen, D. Donoho, and M. Saunders. "Atomic decomposition by basis pursuit". *SIAM Rev.* 43 (2001), pp. 129–159.

[Ch03] O. Christensen. *An introduction to frames and Riesz bases*. Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Boston Inc., 2003, pp. xxii+440. ISBN: 0-8176-4295-1.

[Co+01] A. Cohen, R. De Vore, G. Kerkyacharian, and D. Picard. "Maximal spaces with given rate of convergence for thresholding algorithms". *Appl. Comput. Harmon. Anal.* 11 (2001), pp. 167–191.

[CM91] R. Coifman and Y. Meyer. "Remarques sur l'analyse de Fourier à fenêtre". *C. R. Acad. Sci. Paris Sér. I Math.* 312.3 (1991), pp. 259–261. ISSN: 0764-4442.

[CW92] R. Coifman and M. Wickerhauser. "Entropy-Based Algorithms for Best Basis Selection". *IEEE Trans. Inform. Theory* 38.2 (1992), pp. 713–718.

[Co64] A. Cormack. "Representation of a function by its line integrals with some radiological applications II". *J. Appl. Phys.* 35 (1964), pp. 2908–2913.

[CKP12] T. Coulhon, G. Kerkyacharian, and P. Petrushev. "Heat kernel generated frames in the setting of Dirichlet spaces". arXiv:1206.0463. 2012.

[DT07] A. Dalalyan and A. Tsybakov. "Aggregation by Exponential Weighting, Sharp Oracle Inequalities and Sparsity". In: *COLT*. 2007, pp. 97–111.

[Da92] I. Daubechies. *Ten Lectures on Wavelets*. Philadelphia: SIAM, 1992.

[Da81] M. Davison. "A singular value decomposition for the Radon transform in $n$-dimensional Euclidean space". *Numer. Funct. Anal. Optim.* 3.3 (1981), pp. 321–340. ISSN: 0163-0563.

[DM96] V. Dicken and P. Maass. "Wavelet-Galerkin methods for ill-posed problems". *J. Inverse Ill-Posed Probl.* 4.3 (1996), pp. 203–221. ISSN: 0928-0219.

[Do93] D. Donoho. "Unconditional bases are optimal bases for data compression and for statistical estimation". *Appl. Comput. Harmon. Anal.* 1.1 (1993), pp. 100–115.

[Do97] D. Donoho. "CART and best-ortho-basis: a connection". *Ann. Statist.* 25.5 (1997), pp. 1870–1911.

[Do99] D. Donoho. "Wedgelets: Nearly-minimax Estimation of Edges". *Ann. Statist* 27 (1999), pp. 353–382.

[DET06] D. Donoho, M. Elad, and V. Temlyakov. "Stable recovery of sparse overcomplete representations in the presence of noise". *IEEE Trans. Inform. Theory* 52.1 (2006), pp. 6–18.

[DJ94a] D. Donoho and I. Johnstone. "Ideal denoising in an orthonormal basis chosen from a library of bases". *C. R. Acad. Sci. Paris Sér. I Math.* Serie 1.319 (1994), pp. 1317–1322.

[DJ94b] D. Donoho and I. Johnstone. "Ideal spatial adaptation by wavelet shrinkage". *Biometrika* 81.3 (1994), pp. 425–455.

[DJ95] D. Donoho and I. Johnstone. "Adapting to unknown smoothness via wavelet shrinkage." *J. Amer. Statist. Assoc* 90.432 (1995), pp. 1200–1224.

[Do+96] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. "Density estimation by wavelet thresholding". *Ann. Statist.* 24.2 (1996), pp. 508–539. ISSN: 0090-5364.

[DX01] Ch. Dunkl and Y. Xu. *Orthogonal polynomials of several variables*. Vol. 81. Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press, 2001, pp. xvi+390. ISBN: 0-521-80043-9.

[Ef07] S. Efromovich. "Conditional density estimation in a regression setting". *Ann. Statist.* 35.6 (2007), pp. 2504–2535.

[Ef10] S. Efromovich. "Oracle inequality for conditional density estimation and an actuarial example". *Ann. Inst. Statist. Math.* 62 (2 2010), pp. 249–275.

[EK01] S. Efromovich and V. Koltchinskii. "On inverse problems with unknown operators". *IEEE Trans. Inform. Theory* 47.7 (2001), pp. 2876–2894. ISSN: 0018-9448.

[Ef+04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. "Least angle regression". *Ann. Statist.* 32 (2004), pp. 407–499.

[ER82] C. Elbers and G. Ridder. "True and spurious duration dependence: the identifiability of the proportional hazard models". *Rev. Econom. Stud.* 49 (1982), pp. 403–410.

[Er+81] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi. *Higher transcendental functions. Vol. II*. Based on notes left by Harry Bateman, Reprint of the 1953 original. Melbourne, Fla.: Robert E. Krieger Publishing Co. Inc., 1981, pp. xviii+396. ISBN: 0-89874-069-X.

[FYT96] J. Fan, Q. Yao, and H. Tong. "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems". *Biometrika* 83.1 (1996), pp. 189–206.

[FY04] J. Fan and T. Yim. "A Crossvalidation Method for Estimating Conditional Densities". *Biometrika* 91.4 (2004), pp. 819–834.

[FZZ01] J. Fan, C. Zhang, and J. Zhang. "Generalized likelihood ratio statistics and Wilks phenomenon." *Ann. Statist.* 29.1 (2001), pp. 153–193.

[Fu16] P Funk. "Uber eine geometrische anwendung der abelschen integralgleichung". *Math. Ann.* 77 (1916), pp. 129–135.

[Ga09] S. Gaïffas. "Uniform estimation of a signal based on inhomogeneous data". *Statist. Sinica* 19 (2009), pp. 427–447.

[GK12] E. Gautier and Y. Kitamura. "Nonparametric estimation in random coefficients binary choice models". *Econometrica* To appear (2012).

[GT11] G. Gayraud and K. Tribouley. "A goodness of fit test for the copula density". *Test* 20 (2011), pp. 549–573.

[Ge95] S. van de Geer. "The method of sieves and minimum contrast estimators". *Math. Methods Statist.* 4 (1995), pp. 20–38.

[Ge08] S. van de Geer. "High dimensional generalized linear models and the Lasso". *Ann. Statist.* 36 (2008), pp. 614–645.

[GMT09] C. Genest, E. Masiello, and K. Tribouley. "Estimating copula densities through wavelets". *Insur. Math. Econ.* 44 (2009), pp. 170–181.

[GP03] A. Goldenshluger and S. Pereverzev. "On adaptive inverse estimation of linear functionals in Hilbert scales". *Bernoulli* 9.5 (2003), pp. 783–807. ISSN: 1350-7265.

[GZ03] J. de Gooijer and D. Zerom. "On conditional density estimation". *Stat. Neerlandica* 57.2 (2003), pp. 159–176.

[GS97] D. Gottlieb and C.-W. Shu. "On the Gibbs phenomenon and its resolution". *SIAM Rev.* 39.4 (1997), pp. 644–668.

[Gr96] H. Groemer. *Geometric Applications of Fourier Series and Spherical Harmonics*. Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press, 1996.

[GK07] L. Györfi and M. Kohler. "Nonparametric estimation of conditional distributions". *IEEE Trans. Inform. Theory* 53 (2007), pp. 1872–1879.

[HRL04] P. Hall, J. Racine, and Q. Li. "Cross-Validation and the Estimation of Conditional Probability Densities". *J. Amer. Statist. Assoc.* 99.468 (2004), pp. 1015–1026.

[HWY99] P. Hall, R. Wolff, and Q. Yao. "Methods for estimating a conditional distribution function". *J. Amer. Statist. Assoc.* 94 (1999), pp. 154–163.

[HK96] D. Healy and P. Kim. "An empirical Bayes approach to directional data and efficient computation on the sphere". *Ann. Statist.* 24 (1996), pp. 232–254.

[HS84] J. Heckman and B. Singer. "A method for minimizing the impact of distributional assumptions in econometric models for duration data". *Econometrica* 52 (1984), pp. 271–320.

[He99] S. Helgason. *The Radon transform*. Second. Vol. 5. Progress in Mathematics. Boston, MA: Birkhäuser Boston Inc., 1999, pp. xiv+188. ISBN: 0-8176-4109-2.

[HKM10] S. Hoderlein, J. Klemelä, and E. Mammen. "Reconsidering the random coefficient model." *Econometric Theory* 26 (2010), pp. 804–837.

[Ho99] T. Hofmann. "Probabilistic latent semantic analysis". In: *Proc. of Uncertainty in Artificial Intelligence*. 1999, pp. 289–296.

[Hu+06] Y. Huang, I. Pollak, M. Do, and C. Bouman. "Fast search for best representations in multitree dictionaries". *IEEE Trans. Image Process.* 15.7 (July 2006), pp. 1779–1793.

[HBG96] R. Hyndman, D. Bashtannyk, and G. Grunwald. "Estimating and visualizing conditional densities". *J. Comput. Graph. Statist.* 5 (1996), pp. 315–336.

[HY02] R. Hyndman and Q. Yao. "Nonparametric estimation and symmetry tests for conditional density functions". *Journal of nonparametric statistics* 14.3 (2002), pp. 259–278.

[IT98] H. Ichimura and T. Thompson. "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution". *J. Econometrics* 86 (1998), pp. 269–295.

[IPX12] K. Ivanov, P. Petrushev, and Y. Xu. "Decomposition of spaces of distributions induced by tensor products bases". *J. Funct. Anal.* 263 (2012), pp. 1147–1197.

[JS61] W. James and C. Stein. "Estimation with quadratic loss". *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* 1 (1961), pp. 361–379.

[JR04] I. Johnstone and M. Raimondo. "Periodic boxcar deconvolution and diophantine approximation". *Ann. Statist.* 32 (2004), pp. 1781–1804.

[JL04] A. Juditsky and S. Lambert-Lacroix. "On minimax density estimation on $\mathbb{R}$". *Bernoulli* 10 (2004), pp. 187–220.

[KP03] B. Karaivanov and P. Petrushev. "Nonlinear piecewise polynomial approximation beyond Besov spaces". *Appl. Comput. Harmon. Anal.* 15.3 (2003), pp. 177–223.

[KV02] I. van Keilegom and N. Veraverbeke. "Density and hazard estimation in censored regression models". *Bernoulli* 8.5 (2002), pp. 607–625.

[Ke+07] G. Kerkyacharian, P. Petrushev, D. Picard, and T. Willer. "Needlet algorithms for estimation in inverse problems". *Electron. J. Stat.* 1 (2007), 30–76 (electronic). ISSN: 1935-7524.

[KPNP11] G. Kerkyacharian, T. M. Phan Ngoc, and D. Picard. "Localized deconvolution on the sphere". *Ann. Statist.* 39 (2011), pp. 1042–1068.

[KP92] G. Kerkyacharian and D. Picard. "Density estimation in Besov spaces". *Stat. Probab. Lett.* 13 (1992), pp. 15–24.

[KP00] G. Kerkyacharian and D. Picard. "Thresholding algorithms, maxisets and well-concentrated bases". *Test* 9.2 (2000), pp. 283–344. ISSN: 1133-0686.

[KP04] G. Kerkyacharian and D. Picard. "Regression in random design and warped wavelets". *Bernoulli* 10 (2004), pp. 1053–1105.

[KPR07] G. Kerkyacharian, D. Picard, and M. Raimondo. "Adaptive boxcar deconvolution on full Lebesgue measure sets". *Statist. Sinica* 17 (2007), 317—340.

[KPT96] G. Kerkyacharian, D. Picard, and K. Tribouley. "$L^p$ adaptive density estimation". *Bernoulli* 2 (1996), pp. 229–247.

[KB06] C. Kervrann and J. Boulanger. "Optimal Spatial Adaptation for Patch-Based Image Denoising". *IEEE Trans. Image Process.* 15.10 (2006), pp. 2866–2878.

[KK00] P. Kim and J. Koo. "Directional mixture models and optimal esimation of the mixing density". *Canad. J. Statist.* 28 (2000), pp. 383–398.

[KF00] K. Knight and W. Fu. "Asymptotics for Lasso-type estimators". *Ann. Statist.* 28 (2000), pp. 1356–1378.

[KJG05] E. Kolaczyk, J. Ju, and S. Gopal. "Multiscale, multigranular statistical image segmentation". *J. Amer. Statist. Assoc.* 100.472 (2005), pp. 1358–1369.

[KN04] E. Kolaczyk and R. Nowak. "Multiscale likelihood analysis and complexity penalized estimation". *Ann. Statist.* 32 (2004), pp. 500–527.

[KN05] E. Kolaczyk and R. Nowak. "Multiscale generalised linear models for nonparametric function estimation". *Biometrika* 92.1 (2005), pp. 119–133.

[KT93] A. Korostelev and A. Tsybakov. *Minimax Theory of Image Reconstruction*. Vol. 82. Springer, 1993.

[Ko08] M. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer, 2008.

[KM56] H. Kramer and M. Mathews. "A linear coding for transmitting a set of correlated signals". *IRE Trans. Inform. Theory* 2.3 (1956), pp. 41–46.

[KR09] R. Kulik and R. Raimondo. "Wavelet regression in random design with heteroscedastic dependent errors". *Ann. Statist.* 37 (2009), pp. 3396–3430.

[KPX08] G. Kyriazis, P. Petrushev, and Y. Xu. "Decomposition of weighted Triebel-Lizorkin and Besov spaces on the ball". *Proc. Lond. Math. Soc. (3)* 97.2 (2008), pp. 477–513. ISSN: 0024-6115.

[La+05] D. Labate, W. Lim, G. Kutyniok, and G. Weiss. "Sparse multidimensional representation using shearlets". In: *Wavelets XI (San Diego, CA, 2005)*. SPIE. 2005, pp. 254–262.

[LB06] G. Leung and A. Barron. "Information theory and mixing least-squares regressions". *IEEE Trans. Inform. Theory* 52.8 (2006), pp. 3396–3410. ISSN: 0018-9448.

[LR07] Q. Li and J. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007.

[Li91] J. Lin. "Divergence measures based on the Shannon entropy". *IEEE Trans. Inform. Theory* 37.1 (Jan. 1991), pp. 145–151.

[LS75] B. Logan and L. Shepp. "Optimal reconstruction of a function from its projections". *Duke Math. J.* 42.4 (1975), pp. 645–659.

[LL08] J.-M. Loubes and C. Ludeña. "Adaptive complexity regularization for linear inverse problems". *Electron. J. Stat.* 2 (2008), pp. 661–677.

[LL10] J.-M. Loubes and C. Ludeña. "Penalized estimators for non linear inverse problems". *ESAIM Probab. Stat.* 14 (2010), pp. 173–191.

[LR09] J.-M. Loubes and V. Rivoirard. "Review of rates of convergence and regularity conditions for inverse problems". *Int. J. Tomogr. Stat.* 11.S09 (2009), pp. 61–82.

[Lo84] A. Louis. "Orthogonal function series expansions and the null space of the Radon transform". *SIAM J. Math. Anal.* 15.3 (1984), pp. 621–633. ISSN: 0036-1410.

[Lo08] K. Lounici. "Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators". *Electron. J. Statist.* 2 (2008), pp. 90–102.

[Ma08] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd ed., The Sparse Way*. 3rd. Academic Press, 2008.

[Ma73] C. Mallows. "Some Comments on Cp". *Technometrics* 15.4 (1973), pages. ISSN: 00401706.

[Ma07] P. Massart. *Concentration inequalities and model selection*. Vol. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Berlin: Springer, 2007, pp. xiv+337.

[Ma05] B. Matei. “Smoothness characterization and stability in nonlinear multiscale framework: theoretical results”. *Asymptot. Anal.* 41.3–4 (2005), pp. 277–309.

[MP03] P. Mathé and S. Pereverzev. “Geometry of linear ill-posed problems in variable Hilbert scales”. *Inverse Probl.* 19.3 (2003), pp. 789–803. ISSN: 0266-5611.

[MM12a] C. Maugis and B. Michel. “A non asymptotic penalized criterion for Gaussian mixture model selection”. *ESAIM Probab. Stat.* 15 (2012), pp. 41–68.

[MM12b] C. Maugis and B. Michel. “Data-driven penalty calibration: a case study for Gaussian mixture model selection”. *ESAIM Probab. Stat.* 15 (2012), pp. 320–339.

[MP09] A. Maurer and M. Pontil. “Empirical Bernstein bounds and sample variance penalization”. In: *COLT*. 2009.

[McFa01] D. Mc Fadden. “Economic choices - Nobel Lecture, December 2000”. *American Economic Review* 91 (2001), pp. 351–378.

[MY09] N. Meinhausen and B. Yu. “Lasso-type recovery of sparse representations for high-dimensional data”. *Ann. Statist.* 37 (2009), pp. 246–270.

[MB06] N. Meinshausen and P. Bühlmann. “High dimensional graphs and variable selection with the Lasso”. *Ann. Statist.* 34 (2006), pp. 1436–1462.

[Me90] Y. Meyer. *Ondelettes et opérateurs. I.* Actualités Mathématiques. [Current Mathematical Topics]. Paris: Hermann, 1990, pp. xii+215. ISBN: 2-7056-6125-0.

[NPW06a] F. Narcowich, P. Petrushev, and J. Ward. “Decomposition of Besov and Triebel-Lizorkin spaces on the sphere”. *J. Funct. Anal.* 238.2 (2006), pp. 530–564. ISSN: 0022-1236.

[NPW06b] F. Narcowich, P. Petrushev, and J. Ward. “Localized tight frames on spheres”. *SIAM J. Math. Anal.* 38.2 (2006), 574–594 (electronic). ISSN: 0036-1410.

[Na01] F. Natterer. *The mathematics of computerized tomography*. Vol. 32. Classics in Applied Mathematics. Reprint of the 1986 original. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2001, pp. xviii+222. ISBN: 0-89871-493-1.

[NCB04] R. Neelamani, H. Choi, and R. Baraniuk. “ForWaRD: Fourier-wavelet regularized deconvolution for ill-conditioned systems”. *IEEE Trans. Signal Process.* 52.2 (2004), pp. 418–433. ISSN: 1053-587X.

[Ne00] A. Nemirovski. “Topics in non-parametric statistics”. In: *Lectures on probability theory and statistics (Saint-Flour, 1998)*. Vol. 1738. Lecture Notes in Math. Berlin: Springer, 2000, pp. 85–277.

[OPT00a] M. Osborne, B. Presnell, and B. Turlach. “A new approach to variable selection in least squares problems”. *IMA J. Numer. Anal.* 20 (2000), pp. 389–404.

[OPT00b] M. Osborne, B. Presnell, and B. Turlach. “On the Lasso and its dual”. *J. Comput. Graph. Statist.* 9 (2000), pp. 319–337.

[PX05] P. Petrushev and Y. Xu. “Localized polynomial frames on the interval with Jacobi weights”. *J. Fourier Anal. Appl.* 11.5 (2005), pp. 557–575. ISSN: 1069-5869.

[PX08] P. Petrushev and Y. Xu. “Localized polynomial frames on the ball”. *Constr. Approx.* 27.2 (2008), pp. 121–148. ISSN: 0176-4276.

[Pe09] G. Peyré. “Manifold Models for Signals and Images”. *Comput. Vis. Image Und.* 113.2 (Feb. 2009), pp. 249–260.

[PM08] G. Peyré and S. Mallat. “Orthogonal Bandlets Bases for Geometric Images Approximation”. *Journal of Pure and Applied Mathematics* 61.9 (2008), pp. 1173–1212.

[RR10] P. Reynaud-Bouret and V. Rivoirard. “Near optimal thresholding estimation of a Poisson intensity on the real line”. *Electron. J. Statist.* 4 (2010), pp. 172–238.

[Ri04] V. Rivoirard. "Maxisets for linear procedures". *Stat. Probab. Lett.* 67.3 (2004), pp. 267–275.

[Ri05] V. Rivoirard. "Bayesian modeling of sparse sequences and maxisets for Bayes rules". *Math. Methods Statist.* 14.3 (2005), pp. 346–376.

[RT07] V. Rivoirard and K. Tribouley. "The maxiset point of view for estimating integrated quadratic functionals". *Statist. Sinica* 18.1 (2007), pp. 255–279.

[Ro69] M. Rosenblatt. "Conditional probability density and regression estimators". In: *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968*. New York: Academic Press, 1969, pp. 25–31.

[Ru99] B. Rubin. "Inversion and characterization of the hemispherical transform". *J. Anal. Math.* 77 (1999), pp. 105–128.

[Sa10] J. Salmon. "On two parameters for denoising with Non-Local Means". *Stat. Probab. Lett.* 17 (2010), pp. 269–272.

[SS10] J. Salmon and Y. Strozecki. "From Patches to Pixels in Non-Local methods: Weighted-Average Reprojection". In: *ICIP*. 2010, pp. 1929–1932.

[SS12] J. Salmon and Y. Strozecki. "Patch Reprojections for Non Local Methods". *Sig. Process.* 92.2 (2012), pp. 477–489.

[Sh+05] R. Shukla, P.L. Dragotti, M.. Do, and M. Vetterli. "Rate-distortion optimized tree structured compression algorithms for piecewise polynomial images". *IEEE Trans. Image Process.* 14.3 (Mar. 2005), pp. 343–359.

[SJ05] L. Si and R. Jin. "Adjusting Mixture Weights of Gaussian Mixture Model via Regularized Probabilistic Latent Semantic Analysis". In: *Advances in Knowledge Discovery and Data Mining*. 2005, pp. 218–252.

[Sk59] A. Sklar. "Fonctions de répartition à $n$ dimensions et leurs marges". *Publ. Inst. Statist. Univ. Paris* 8 (1959), pp. 229–231.

[St94] Ch. Stone. "The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation". *Ann. Statist.* 22.1 (1994), pp. 118–171.

[Sz75] G. Szegö. *Orthogonal polynomials*. Fourth. American Mathematical Society, Colloquium Publications, Vol. XXIII. Providence, R.I.: American Mathematical Society, 1975, pp. xiii+432.

[Ti96] R. Tibshirani. "Regression shrinkage and selection via the Lasso". *J. Roy. Statist. Soc. Ser. B* 58 (1996), pp. 267–288.

[Ts00] A. Tsybakov. "On the best rate of adaptive estimation in some inverse problems". *C. R. Acad. Sci. Paris Sér. I Math.* 330.9 (2000), pp. 835–840. ISSN: 0764-4442.

[Ts08] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008. ISBN: 0387790519, 9780387790510.

[VW96] A. van der Vaart and J. Wellner. *Weak Convergence*. Springer, 1996.

[Wh92] H. White. "Maximum Likelihood Estimation of Misspecified Models". *Econometrica* 50.1 (1992), pp. 1–25.

[Wi38] S. Wilks. "The large-sample distribution of the likelihood ratio for testing composite hypotheses". *Ann. Math. Statist.* 9 (1938), pp. 60–62.

[WN07] R. Willet and R. Nowak. "Multiscale Poisson Intensity and Density Estimation". *IEEE Trans. Inform. Theory* 53.9 (2007), pp. 3171–3187.

[Xu07] Y. Xu. "Reconstruction from Radon projections and orthogonal expansion on a ball". *J. Phys. A* 40.26 (2007), pp. 7239–7253. ISSN: 1751-8113.

[Ya00] Y. Yang. "Combining different procedures for adaptive regression". *J. Multivariate Anal.* 74.1 (2000), pp. 135–161. ISSN: 0047-259X.

[YH10] D. Young and D. Hunter. "Mixtures of regressions with predictor-dependent mixing proportions". *Comput. Statist. Data Anal.* 54.10 (2010), pp. 2253–2266.

[YZ06] B. Yu and P. Zhao. "On model selection consistency of Lasso estimators". *J. Mach. Learn. Res.* 7 (2006), pp. 2541–2567.

[Za92] L. Zalcman. "A bibliographic survey of the Pompeiu problem". In: *Approximation by Solutions of Partial Differential Equations*. Ed. by B. Fuglede et al. Amsterdam: Kluwer Academic Publ., 1992, pp. 185–194.

[ZH08] C. Zhang and J. Huang. "The sparsity and bias of the Lasso selection in high-dimensional linear regression". *Ann. Statist.* 36 (2008), pp. 1567–1594.

[Zo06] H. Zou. "The adaptive Lasso and its oracle properties". *J. Amer. Statist. Assoc.* 101 (2006), pp. 1418–1429.

# Appendix A

# Let It Wave

Working (and funding) *Let It Wave* with Ch. Bernard, J. Kalifa and S. Mallat has been such a tremendous scientific and human adventure that I feel that summarizing my contributions of the last ten years without mentionning the *Let It Wave* part would have been misleading... As I am not allowed describe accurately what I have done in *Let It Wave*/Zoran/CSR, I have decided to only make a list *à la Prévert* of some topics I have been working on along those years:

- face detection and compression,
- seismic flattening and inversion,
- video deinterlacing,
- video denoising and deblocking,
- video frame rate detection and conversion,
- video superresolution,
- video sharpening,
- video dynamic range compensation,
- LED backlight compensation,
- automatic 2D video to 3D video conversion,
- chroma upsampling,
- video scaling detection...

Working on those subjects has been a pleasure, but what I have probably enjoyed the most is the interaction with the employees, and the *Algo* team in particular. It is a pleasure to conclude this manuscript by a huge *Thank you* to them.