Contributions to statistical learning in sparse models

Pierre Alquier

UNIVERSITÉ PARIS 6

MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

Pierre ALQUIER

Contributions à l'Apprentissage Statistique dans les Modèles Parcimonieux

Contribution to Statistical Learning in Sparse Models

Soutenu le 06/12/2013 devant le jury composé de:

M.	OLIVIER CATONI	CNRS
M.	OLEG LEPSKI	UNIVERSITÉ D'AIX-MARSEILLE
M.	PASCAL MASSART	UNIVERSITÉ DE PARIS-SUD
Mme.	DOMINIQUE PICARD	UNIVERSITÉ PARIS DIDEROT
M.	EMMANUEL RIO	UNIVERSITÉ DE VERSAILLES
M.	ALEXANDRE TSYBAKOV	UNIVERSITÉ PARIS 6 / ENSAE

Au vu des rapports de:

M.	OLEG LEPSKI	UNIVERSITÉ D'AIX-MARSEILLE
M.	EMMANUEL RIO	UNIVERSITÉ DE VERSAILLES
M.	TONG ZHANG	RUTGERS UNIVERSITY

Résumé

Ce mémoire d'habilitation a pour objet diverses contributions à l'estimation et à l'apprentissage statistique dans les modeles en grande dimension, sous différentes hypothèses de parcimonie. Dans une première partie, on introduit la problématique de la statistique en grande dimension dans un modèle générique de régression linéaire. Après avoir passé en revue les différentes méthodes d'estimation populaires dans ce modèle, on présente de nouveaux résultats tirés de l'article [A5] pour des estimateurs agrégés. La seconde partie a essentiellement pour objet d'étendre les résultats de la première partie à l'estimation de divers modèles de séries temporelles [A6, A8, A19, A19]. Enfin, la troisième partie présente plusieurs extensions à des modèles non paramétriques ou à des applications plus spécifiques comme la statistique quantique [A11, A10, A12, A13, A16, A4]. Dans chaque section, des estimateurs sont proposés, et, aussi souvent que possible, des inégalités oracles optimales sont établies.

Mots-clefs : Théorie de l'apprentissage statistique, estimateurs agrégés, inégalités PAC-Bayésiennes, statistique en grande dimension, parcimonie, estimateur LASSO, estimateurs pénalisés, dépendance faible, statistique quantique, régression matricielle, méthodes de Monte-Carlo.

Contributions to Statistical Learning in Sparse Models

Abstract

The aim of this habilitation thesis is to give an overview of my works on high-dimensional statistics and statistical learning, under various sparsity assumptions. In a first part, I will describe the major challenges of high-dimensional statistics in the context of the generic linear regression model. After a brief review of existing results, I will present the theoretical study of aggregated estimators that was done in [A5]. The second part essentially aims at providing extensions of the various theories presented in the first part to the estimation of time series models [A6, A8, A19, A19]. Finally, the third part presents various extensions to nonparametric models, or to specific applications such as quantum statistics [A11, A10, A12, A13, A16, A4]. In each section, we provide explicitely the estimators used and, as much as possible, optimal oracle inequalities satisfied by these estimators.

Keywords : Statistical learning theory, aggregated estimators, PAC-Bayesian inequalities, high-dimensional statistics, sparsity, LASSO estimator, penalized estimators, weak dependence, quantum statistics, reduced-rank regression, Monte-Carlo statistical methods.

Remerciements

Je tiens tout d'abord à réitérer mes remerciements à mon directeur de thèse, Olivier Catoni. J'avais écrit dans ma thèse une phrase que je considère toujours comme vraie et me permet de commencer par une auto-citation [A17] : "pour m'avoir proposé un sujet de recherche qui m'a passioné, pour sa grande gentillesse et disponibilité, et pour les Mathématiques que j'ai pu apprendre auprès de lui au cours des trois années passées." Je dois ajouter que son encadrement à l'époque et les discussions que nous avons pu avoir ensuite m'ont donné un éclairage original et intéressant sur les problèmes d'apprentissage statistique auxquels je m'intèresse aujourd'hui.

Je tiens à remercier Oleg Lepski, Emmanuel Rio et Tong Zhang d'avoir accepté de raporter mon mémoire d'habilitation. Des commentaires positifs et constructifs de la part d'un spécialiste de la sélection de modèles, des séries temporelles et de l'aggrégation de modèles sont très précieux ! Merci également Oleg pour avoir fait le voyage depuis Marseille à un moment de l'année aussi chargé !

Je tiens aussi à remercier chaleureusement, Dominique Picard, Pascal Massart, et Alexandre Tsybakov d'avoir accepté de faire partie de mon jury. Je vous ai presque tous eu comme profs, puis comme collègues, vous savez donc ce que je vous dois ! Un grand merci à Lucien Birgé et Elisabeth Gassiat qui avaient aussi accepté de faire partie du jury mais avec qui nous n'avons pas pu trouver de date commune. Je vous garde une coupe de champagne au frais.

Je remercie tous ceux avec qui j'ai pu collaborer sur différents projets de recherche (par ordre alphabéto-chronologique de publication des articles) : Mohamed Hebiri, Karim Lounici, Olivier Wintenberger, Paul Doukhan, Xiaoyin Li, Benjamin Guedj, Gérard Biau, Eric Moulines, Katia Meziani, Cristina Butucea, Gabriel Peyré, Tomoyuki Morimae, et ceux avec qui je suis actuellement au boulot : Judith Rousseau, Nial Friel, Aidan Boland, Richard Everitt, Nicolas Chopin, James Ridgway et Liang Feng. J'ai appris énormément de vous sur les sujets que nous avons pu aborder ensemble : LASSO, series temporelles, statistique quantique, ... Egalement Gilles Stoltz et Eric Gautier, pour le travail ensemble sur Stats in the Chateau, et les collaborations a venir ! Même si nous n'avons pas (encore ?) eu l'occasion d'écrire d'article ensemble, j'ai également appris beaucoup de discussions avec Arnak Dalalyan, Guillaume Lécué, Joseph Salmon sur tous les problèmes de statistique en grande dimension, de Jean-Paul Feugeas sur les applications en génomique, de Sébastien Gerchinovitz sur les suites individuelles, de Stéphane Boucheron sur les inégalités de concentration, de Karine Tribouley, Ghislaine Gayraud et Mathilde Mougeot sur l'agrégation et les tests, de Ismael Castillo sur le bayésien paramétrique ou non, et de Taiji Suzuki sur le modèle additif. Enfin, un merci particulier à Matthieu Cornec pour toutes les discussions que nous avons pu avoir sur tant de sujets différents, ainsi que pour l'organisation courageuse du groupe de travail "prévisions" pendant plusieurs années !

Lors de mes passages dans différentes équipes (équipe de stats du LPMA à Paris 6 et Paris 7, Laboratoire de Statistique du CREST, puis dans l'équipe de statistique et actuariat du département de Sciences Mathématiques à l'UCD), j'ai toujours été accueilli dans d'excellentes conditions de travail et humaines, et appris beaucoup de tous mes collegues, je remercie donc : Julyan Arbel, Jean-Yves Audibert, Norma Bargary, Patrice Bertail, Alan Benson, Soumyajit Biswas, Eleanor Brodie, Victor-Emmanuel Bru-

nel, Niamh Cahill, Christophe Chesneau, Olivier Collier, Laetitia Comminges, Sophie Dabo-Niang, Sylvain Delattre, Stéphanie Dupoiron, Céline Duval, Romuald Elie, Aurélie Fischer, Stéphane Gaiffas, Marie Galligan, Emmanuelle Gautherat, Marc-Antoine Giuliani, Claire Gormley, Belinda Hernandez, Mary Hall, Hugo Harari-Kermadec, Marc Hoffmann, Adrian Iuga, Pierre Jacob, Cyrille Joutard, Gabrielle Kelly, Gérard Kerkyacharian, Frédéric Lavancier, Catherine Laredo, Damien McParland, Jean-Baptiste Monnier, Brendan Murphy, Patrick Murphy, Nancy Duong Nguyen, Adrian O'Hagan, Sarah O'Rourke, Andrew Parnell, Vianney Perchet, Adrian Raftery, Mavo Ralazamahaleo, Jin Renhao, Philippe Rigollet, Christian Robert, Marc Roger de Campagnolle, Etienne Roquain, Judith Rousseau, Mathieu Rosenbaum, Robin Ryder, Jean-Bernard Salomond, Christian Schäfer, Johaness Schmidt-Hieber, James Sweeney, Jessica Tressou, Lionel Truquet, Thomas Vareschi, Nicolas Vayatis, Fanny Villers, Shane Whelan, Thomas Willer, Arthur White, Jason Wyse.

Je remercie S. Ejaz Ahmed, Jean-Marc Bardet, Gilles Blanchard, Laurent Cavalier, Bernard Delyon, Thinh Doan, Christian Francq, Hannes Leeb, Sébastien Loustau, Christophe Pouet, Vincent Rivoirard, Samy Tindel et Jean-Michel Zakoian pour leurs très gentilles invitations à venir parler dans différents séminaires et conférences.

J'ai assisté à de nombreuses conférences qui m'ont été très utiles depuis que j'ai commencé à faire de la recherche, mais je voulais avoir un mot particulier pour les journées de statistique mathématique à Luminy et l'école d'été de probabilités de Saint-Flour auxquelles j'ai assisté plusieurs fois : que ceci soit l'occasion, d'une part d'en remercier les organisateurs, d'autre part de leur demander continuer ces conférences le plus longtemps possible ! Les cours de Saint-Flour de A. Nemirovki, O. Catoni, P. Massart et V. Koltchinskii sont parmi les livres que j'ai le plus lus et relus (en incluant les livres non mathématiques !), et j'attends avec impatience la publication de ceux de E. Candès et A. Tsybakov.

Enfin, je remercie mes amis, et ma famille : Marie, qui nous a quittés quelques jours avant cette soutenance et pour qui je voudrais avoir une pensée particulière ici, Gilles, Josiane, Aline, Vincent, Nicole, Hugues Etienne, et Jean-Loup pour leur soutien permanent.

Table of contents

Introduction

Most of my work is about what is usually refered as *model selection* and *variable selection* in statistical learning theory. In this introduction we first provide a general introduction to learning theory and then, an example of variable selection problem.

A - short - introduction to statistical learning theory

Let us assume that a scientist wants to attribute labels to objects. Let $\mathcal{X}$ denote the set of possible objects and $\mathcal{Y}$ the set of labels. The objective is to build a function $f : \mathcal{X} \to \mathcal{Y}$ such that $y = f(x)$ is a satisfactory label for the object x (we will give precise statements of what we mean by *satisfactory* below). For example :

1. the object x is an image and the corresponding label y is binary : $y = 0$ if this image contains no human face, $y = 1$ if the image contains at least one human face.
2. we consider a given disease and a possible cure, x is the medical record of an patient suffering from the disease and $y = 1$ if the cure is efficient for this patient, $y = 0$ otherwise.
3. x contains information about a chemical reaction : quantity of reactants, pressure and temperature, while $y \in \mathbb{R}_+$ is the quantity of a product of interest produced by this reaction.

When $\mathcal{Y}$ is finite, this problem is refered as a classification problem. Any function $f : \mathcal{X} \to \mathcal{Y}$ is called a classification function. When $\mathcal{Y}$ is $\mathbb{R}$ or an interval of real numbers, it is called a regression problem. A function $f : \mathcal{X} \to \mathcal{Y}$ is a regression function.

In order to quantify the quality of the prediction made by a function f, one usually introduce a *loss* function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. The idea is that $\ell(y, y')$ measures the distance between two objects y and y', howevever, we do not generally assume that ℓ satisfies the axioms of a metric on $\mathcal{Y}$. When given an object x, and a function $f : \mathcal{X} \to \mathcal{Y}$, we can use f to predict the associated label by $f(x)$. If the actual label is y, the error encountered when using f on the pair (x, y) is given by $\ell(f(x), y)$. In binary classification, the most famous example of loss function is given by $\ell(y, y') = \mathbb{1}(y \neq y')$. However, as it leads to computational difficulties, it is often replaced by loss functions that lead to convex minimization programs, such as $\ell(y, y') = \exp(-yy')$, see Zhang [232] and the references therein on this topic. In regression, the most studied example is the quadratic loss $\ell(y, y') = (y - y')^2$.

In order to build the function f, any information can be used - e.g. we can use knowledge coming from theoretical physics, economics, biology... that can help to understand how objects and labels are related. Sometimes, however, such an information is not available. Even when it is available, it is often hard to use in practice, and sometimes even not completely reliable. On the other hand, based on past observations, on survey pools or on a series of experiments designed for the occasion, a set of examples of pairs object-label $(X_1, Y_1), \ldots, (X_n, Y_n)$ might be available. The objective is therefore to *learn* or to *infer* the function f based on these examples.

Different type of assumptions on the examples are possible. In statistical learning theory, it is usually assumed that the set of examples $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is a

random vector with some probability distribution $\mathbb{P}$ (in most cases, it is assumed that the pairs (X_i, Y_i) are independent from each other under $\mathbb{P}$, however, this assumtion is not always realistic, e.g. when dealing with time series as in Part II).

An example of variable selection problem

We now consider the following example : $\mathcal{Y} = \mathbb{R}$, $\mathcal{X} = \mathbb{R}^p$ for some $p > 0$. We propose to attribute labels to object according to linear functions : for any $(\theta, x) \in \mathbb{R}^p$, we define $f_\theta(x) = \theta.x$. We would like to identify, in some sense, the "best" parameter θ on the basis of the examples $\mathcal{D}_n$ only : we will use estimators, namely, functions

$$\hat{\theta} : (\mathbb{R}^p \times \mathbb{R})^n \to \mathbb{R}^p$$

such that $\hat{\theta}(\mathcal{D}_n)$ is as close as possible to the "best" θ (it is usual to write $\hat{\theta}$ instead of $\hat{\theta}(\mathcal{D}_n)$, thus making the dependence on the sample $\mathcal{D}_n$ implicit). However, depending on the situations, what is meant by "best" and "close to the best" differ :

Prevision objective : in this case, the scientist primary objective is to be able to predict labels of new objects. We can model the situation as follows : nature draws a pair (X, Y) from a probability distribution P, the statistician is given X and must guess Y. In this case, the objective is to minimize

$$R(\theta) := \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X), Y)].$$

This is called the *prevision risk* of the parameter θ (or simply *risk*). In this case, the "best" θ is not necessarily unique, it is the set of minimizers of R

$$\arg\min_{\theta\in\mathbb{R}^p} R(\theta),$$

we will use the notation $\overline{\theta}$ for a member of this set. We are satisfied with an estimator $\hat{\theta}$ when $R(\hat{\theta}) - R(\overline{\theta})$ is small, in expectation or with large probability. The usual assumptions in this settint is that P is unknown, but that under $\mathbb{P}$, the examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with common distribution P.

Transduction : close to the previous objective is the so-called *transduction objective* defended by Vapnik [213]. In this case, we assume that in addition to the sample $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, nature draws m pairs $(X'_1, Y'_1), \ldots, (X'_m, Y'_m)$ and the scientist is given $\mathcal{D}_n$ together with $X'_1, \ldots, X'_m$. The objective is to minimize

$$R'(\theta) := \frac{1}{m}\sum_{i=1}^{m} \ell(f_\theta(X'_i), Y'_i).$$

So we don't expect to be able to attribute a label to *any* object, but only to $X'_1, \ldots, X'_m$. Here again, the objective is to make $R'(\hat{\theta}) - \inf_{\mathbb{R}^p} R'$ as small as possible.

Reconstruction objective in situations where the X_i are deterministic, the prevision risk is often replaced with a reconstruction risk :

$$R_n(\theta) := \mathbb{E}_{((X'_1,Y'_1),\ldots,(X'_n,Y'_n))\sim\mathbb{P}} \left[\frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X'_i), Y'_i)\right].$$

While it makes sense to consider that the X_i are deterministic when they correspond to an experimental design chosen by the scientist, this criterion is less pertinent than the previous ones for prevision purposes.

Parameter estimation here, we assume that the objects are actually related to the labels by a linear relation :

$$\mathbb{E}(Y_i|X_i) = \theta_0 \cdot X_i.$$

The scientist is only interested in the estimation of θ_0 and not in prediction, so $\hat{\theta}$ must be such that $d(\hat{\theta}, \theta_0)$ is small for some distance d. For example, $d(\hat{\theta}, \theta_0) = \|\hat{\theta} - \theta_0\|$.

Variable selection : under the same setting, let us assume that only a few coordinates in X_i are relevant to predict Y_i. In other words, we assume that most coordinates in θ_0 are equal to 0. To identify these coordinates might be an objective in itself to the scientist. For any $\theta \in \mathbb{R}^p$, let $\text{supp}(\theta)$ be the set $I \subset \{1, \dots, p\}$ such that $\theta_i \neq 0 \Leftrightarrow i \in I$. Let Δ denote the symmetric difference for sets. Then in this case we look for an estimator $\hat{\theta}$ such that

$$R^{\Delta}(\hat{\theta}) = \text{card}\left(\text{supp}(\hat{\theta})\Delta\text{supp}(\theta_0)\right)$$

is small - if possible, $R^{\Delta}(\hat{\theta}) = 0$ with large probability.

Quite surprinsingly, these different objectives are sometimes not compatible. For example, Yang [226] proved in a quite general setting that an estimator that is consistent for the selection criterion (in the sense that $\mathbb{P}(R^{\Delta}(\hat{\theta}) = 0) \to 1$ when the sample size $n \to \infty$) is suboptimal for the parameter estimation and the prevision criterion. For results more specific to the linear case, see the nice results by Leeb and Pötscher [139], and Zhao and Yu [236] for the LASSO estimator (introduced in Part I below). In most of my work I focused on prevision and transduction objectives. For the sake of simplicity, in this thesis, I will present most of my results for the prevision objective only. However, I will present some results for the reconstruction objective when the corresponding results for the prevision objective are not available.

In order to quantify how additional assumption (such as "$\overline{\theta}$ has only a few nonzero coordinates") can help to improve the prediction criterion, we will establish so-called oracle inequalities, following Donoho and Johstone's terminology [78], see page 13 below.

Remark 0.1. *One might ask the following question : why did we only consider linear functions in this introduction ? For example it makes sense to define, for any measurable $f : \mathbb{R}^p \to \mathbb{R}$, a prevision risk*

$$R(f) = \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)].$$

However, the quest of an estimator $\hat{f}$ that would make the quantity $R(\hat{f}) - \inf_f R(f)$ as small as possible for any distribution P is in some way hopeless, as shown by the so-called no free-lunch theorem : Theorem 7.2 page 114 in the monograph by Devroye, Györfi and Lugosi [77], we refer the reader to the whole Chapter 7 in [77] for more details on this topic. Their are two ways to circumvent this difficulty : first, to impose some assumptions on P. However, it is often hard to know whether these assumptions are satisfied in practice. Alternatively, instead of competing against the best function f,

we can compete agains the best function in a restricted class $\mathcal{F}$. An example of such a class is the set of linear functions $\{f_\theta, \theta \in \mathbb{R}^p\}$. Other classes $\mathcal{F}$ are possible, Part III of this thesis provides a wide range of examples. The choice of the class is motivated by theoretical and practical reasons. First, precise assumptions on the complexity of the class $\mathcal{F}$ exists to ensure the convergence of $R(\hat{f}) - \inf_{f\in\mathcal{F}} R(f)$ to 0, with a given rate, uniformly on the probability distribution P. This major breakthrough is due to Vapnik and Chervonenkis, see [214], or [213] for an English translation and a more comprehensive introduction. Then, we must also keep in mind that the estimator must be calculable in practice. This remark was one of the major ideas in Valiant's PAC theory [210].

Outline of this thesis

In Part I, we focus on regression in large dimension under the classical assumption that the examples are independent. In Section 1 we introduce the notations and some classical estimators that satisfy oracle inequalities. However, for computational reasons, these estimators cannot be used when the dimension p is large. This led to the introduction of ℓ_1-penalized methods : Tibshirani's LASSO [205], Chen and Donoho's basis pursuit [59] or Candès and Tao's Dantzig selector [48]. These methods lead to computationally feasible estimators. On the other hand, stringent assumptions are required on the examples to ensure that these estimators enjoy good statistical properties. We finally present in Section 3 PAC-Bayesian aggregation methods, that lead to a good compromise between both situations [A5].

In Part II, we relax the assumption that the examples are independent in order to deal with time series. Technical definitions on time series are given in Section 4. In Section 5 we present the results of our paper [A6] : we extend the oracle inequalities known for the LASSO to the case of time series. In Section 6 we present the results of our papers [A8, A15, A19] : we extend PAC-Bayesian inequalities for time series.

In Part III, we consider more general models, where the idea of model or variable selection is still relevant : single-index models [A11] in Section 7, additive models [A10] in Section 8, quantum tomography [A13, A12] in Section 9, reduced-rank regression and matrix completion [A16] in Section 10 and finally we give a general result from [A4] in Section 11.

Part I

Estimation in sparse linear models

Summary

In this first part, we focus on estimation in high dimensional linear models.

We assume that we observe, on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, n random pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ that are independent under $\mathbb{P}$, with $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. We assume that

$$Y_i = f(X_i) + W_i$$

for some measurable function f and some random variables W_i with $\mathbb{E}(W_i|X_i) = 0$ and $\mathbb{E}(W_i^2|X_i) \leq \sigma^2$ for some constant σ^2. We define, for any $\theta \in \mathbb{R}^p$, and any $x \in \mathbb{R}^p$, $f_\theta(x) = \theta \cdot x$. We define the empirical risk

$$r(\theta) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(X_i))^2$$

and the risk

$$R(\theta) := \mathbb{E}\left[r(\theta)\right].$$

Note that when the (X_i, Y_i) are actually i.i.d. with common distribution P under $\mathbb{P}$, this is actually equal to the prevision risk

$$R(\theta) = \mathbb{E}\left[(Y - f_\theta(X))^2\right]$$

where (X, Y) is another pair also distributed according to P. On the other hand, when the X_i's are deterministic, this is equal to the reconstruction criterion defined in the introduction. Let us chose a $\overline{\theta}$ in the set of minimizers of $R : R(\overline{\theta}) = \inf_{\theta \in \mathbb{R}^p} R(\theta)$. When the X_i's are deterministic, Pythagorean theorem leads to

$$R(\theta) - R(\overline{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[(\theta - \overline{\theta}) \cdot X_i\right]^2 =: \|\theta - \overline{\theta}\|_n^2$$

for short (this norm, $\|\cdot\|_n$, is often refered as the *empirical norm*). In this case, we let M denote the design matrix $M = (X_1|\dots|X_n)$, and it is convenient to assume that the variables are renormalized in such a way that the diagonal coefficients of $M^T M/n$ are equal to 1. The most classical estimator in this setting is the so-called least square estimator $\hat{\theta}^{LSE}$, it is actually any value that minimizes the empirical risk r. For any $\theta \in \mathbb{R}^p$ we remind that $\text{supp}(\theta) = \{i \in \{1, \dots, p\}, \theta_i \neq 0\}$ and define $\|\theta\|_0 := \text{card}(\text{supp}(\theta))$.

Sections 1 and 2 are short reviews of the literature, we introduce ℓ_0 and ℓ_1-penalized estimators. In Section 3 we present our results on aggregated estimators.

1 Setting of the problem and first oracle inequalities

In Sections 1 and 2, we assume that the X_i's are deterministic and that the W_i's are Gaussian random variables, $W_i \sim \mathcal{N}(0, \sigma^2)$. It is quite straightforward to prove that

$$\mathbb{E}\left(\|\hat{\theta}^{LSE} - \overline{\theta}\|_n^2\right) = \frac{\sigma^2 \text{rank}(M)}{n}.$$

So, when $p \ll n$, we have

$$\mathbb{E}\left(\|\hat{\theta}^{LSE} - \overline{\theta}\|_n^2\right) \leq \frac{\sigma^2 p}{n} \ll 1.$$

On the other hand, when p is large, this result is not satisfying.

Remark 1.1. *In contrast with the deterministic case, when the pairs (X_i, Y_i) are i.i.d., one have to impose strong assumptions on their probability distribution in order to prove such a bound. For example, we have to assume that the X_i's are bounded, see various bounds in the paper by Birgé and Massart [35], in Section 5 in Catoni's monograph [54] or Sections 10, 11 and 12 in Györfi, Kohler, Krzyzak and Walk's book [107]. At the price of substituting a complicated estimator to $\hat{\theta}_{LSE}$, Audibert and Catoni [13] considerably improved on this condition, but some assumptions are still required. A discussion on the different results available for regression with random design can be found in Subsection 4.2 p. 2749 of the paper by Maillard and Munos [151].*

As mentioned in the introduction, in some situations, it makes sense to assume that a large portion of the coordinates in X_i are not actually related to Y_i. In other words, we expect that most coordinates in $\overline{\theta}$ are zero (or non significantly different from zero). Let us fix a set $I \subset \{1, \dots, p\}$. We put

$$\overline{\theta}_I \in \arg \min_{\mathrm{supp}(\theta) \subset I} R(\theta) \text{ and } \hat{\theta}_I^{LSE} \in \arg \min_{\mathrm{supp}(\theta) \subset I} r(\theta).$$

We have

$$\mathbb{E}\left(\|\hat{\theta}_I^{LSE} - \overline{\theta}\|_n^2\right) \leq \|\overline{\theta}_I - \overline{\theta}\|_n^2 + \frac{\sigma^2 \mathrm{card}(I)}{n}.$$

For example, the assumption that only the coordinates in X_i corresponding to a given set I are related to Y_i can be expressed as $\overline{\theta}_I = \overline{\theta}$. This leads to

$$\mathbb{E}\left(\|\hat{\theta}_I^{LSE} - \overline{\theta}\|_n^2\right) \leq \frac{\sigma^2 \mathrm{card}(I)}{n}$$

and it is a considerable improvement on the rate $\sigma^2 p/n$ when $\mathrm{card}(I) \ll p$. Even better, we can consider the set I^* given by

$$I^* = \arg \min_{I \subset \{1,\dots,p\}} \mathbb{E}\left(\|\hat{\theta}_I^{LSE} - \overline{\theta}\|_n^2\right).$$

The corresponding $\hat{\theta}_{I^*}^{LSE}$ satisfies obviously

$$\mathbb{E}\left(\|\hat{\theta}_{I^*}^{LSE} - \overline{\theta}\|_n^2\right) \leq \inf_{I \subset \{1,\dots,p\}} \left\{\|\overline{\theta}_I - \overline{\theta}\|_n^2 + \frac{\sigma^2 \mathrm{card}(I)}{n}\right\}. \tag{1}$$

This set I^* is unfortunately unknown in practice as it depends on the unknown $\overline{\theta}$. For this reason, $\hat{\theta}_{I^*}^{LSE}$ is not an estimator, it is often refered as an *oracle*. The question is now : is it possible to build an estimator $\hat{\theta}$ for which inequality (1) would hold ? It is actually possible to build an estimator satisfying an inequality close to (1). Such a result is usually refered as an *oracle inequality*, this terminology is due to Donoho and Johnstone [78], the first examples of oracle inequalities (in different contexts) can be found in papers by Donoho, Johstone, Kerkyacharian and Picard [79, 80]. Let us introduce a first example of estimator that satisfies an oracle inequality.

Definition 1.2 (ℓ_0-penalized estimator)**.** *For a nondecreasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ we define*

$$\hat{\theta}_g^0 = \arg \min_{\theta \in \mathbb{R}^p} \left\{r(\theta) + g(\|\theta\|_0)\right\}.$$

When $g(\|\theta\|_0) = \lambda\|\theta\|_0$ for some constant $\lambda > 0$, depending on the value of λ, this estimator can actually be seen as Akaike's AIC [5] or Schwarz' BIC [191] (see also Mallows [152]). Theoretical properties of penalization by $\|\cdot\|_0$ were studied in a wide variety of settings by different authors, see Barron, Birgé and Massart [21], Massart [154], Bunea, Tsybakov and Wegkamp [43], Golubev [101] among others.

Theorem 1.3 (Theorem 3.1 page 1679 in Bunea, Tsybakov and Wegkamp [43])**.** *Under the assumption that $\|f\|_\infty \leq L$ and that almost surely, $\|X_i\|_\infty \leq L$ for some constant $L > 0$, for any $a > 0$, there is a known function g (that does not depend on L but depends on a) and a constant C_a such that*

$$\mathbb{E}\left(\|\hat{\theta}_g^0 - \overline{\theta}\|_n^2\right) \leq (1+a) \inf_{I \subset \{1,\dots,p\}} \left\{ \|\overline{\theta}_I - \overline{\theta}\|_n^2 + \frac{10\sigma^2 \mathrm{card}(I)\left[\log\left(\frac{p}{\max(\mathrm{card}(I),1)}\right) + C_a\right]}{n} \right\}.$$

When compared to the upper bound on the oracle (1), we remark that we lose constants and a logarithmic term $\log(p/\mathrm{card}(I))$. This term is in some way the price to pay for the estimation of a set I. Theorem 5.1 page 1684 in [43] states that no estimator can satisfy an oracle inequality without this term uniformly over all possible $\overline{\theta}$.

So, from a theoretical perspective, the estimator $\hat{\theta}_g^0$ is optimal. The trouble is that, in practice, the minimization program in Definition 1.2 is not tractable when p is large : the best known method consists essentially in computing the constrained least square estimator $\hat{\theta}_I^{LSE}$ for any $I \subset \{1,\dots,p\}$. As there is 2^p such possible subsets, this is not feasible as soon as p is large, say $p > 50$. Actually, Theorem 1 p. 228 in Natarajan [167] states that this problem is NP-hard.

2 Convex relaxation

To this regard, a major breakthrough was the idea of a convex relaxation of the program in Definition 1.2 : this led to Tibshirani's LASSO estimator [205] and simultaneously to Chen and Donoho's basis pursuit [59, 58] (related estimators appeared earlier, such as Frank and Friedman's bridge regression [90] and Breiman non-negative garrote [39] but it seems that the full power of this idea was not identified at that time). We provide the definition of the LASSO estimator.

Definition 2.1 (ℓ_1-penalized estimator a.k.a. LASSO [205])**.** *For a constant $\lambda > 0$ we put*

$$\hat{\theta}_\lambda^1 = \arg\min_{\theta \in \mathbb{R}^p} \left\{ r(\theta) + \lambda\|\theta\|_1 \right\}.$$

The "LASSO" acronym stands for Least Angle Selection and Shrinkage Operator. This estimator became highly popular since its introduction. One of the main reason is that on the contrary to the ℓ_0-penalized estimator, there exists efficient algorithms to compute $\hat{\theta}_\lambda^1$. We mention, among others, the popular LARS algorithm by Efron, Hastie, Johnstone and Tibshirani [87] and the coordinate descent algorithm studied by Friedman, Hastie, Höfling and Tisbshirani [91]. But a natural question arises : does an oracle

inequality hold for the LASSO ? That question was studied by many authors : Candès and Tao provided inequalities for an estimator that they called Dantzig selector [48] that is closely related to the LASSO, Bunea, Tsybakov and Wegkamp [44] proved oracle inequalities for the LASSO, and Bickel, Ritov and Tsybakov [33] established stronger results, they also proved a kind of equivalence between the LASSO and Dantzig estimators. Koltchinskii [129, 130] proved oracle inequalities for the LASSO in the random design case. Hebiri and Lederer [113] analyzed the effect of the design matrix M on the quality of the prediction when using the LASSO. Zhao and Yu [236] and Lounici [150] provided conditions for the variable selection consistency of the LASSO. The conclusion of all these works is that it is necessary to introduce some assumption on the matrix $M = (X_1|\dots|X_n)$ in order to prove some oracle inequality for ℓ_1-penalized estimators. These conditions are studied and compared in the paper by Bühlmann and van de Geer [40]. We refer the reader to Part III page 121 in [A18] and to the monograph by Bühlmann and van de Geer [41] for a general introduction.

For the sake of completeness, we state Bickel, Ritov and Tsybakov's oracle inequality.

Definition 2.2. *We say that the matrix M satisfies the Restricted Eigenvalue Property* $\mathrm{REP}(s,c)$ *for $s \in \{1,\dots,p\}$ and $c > 0$ if $\kappa(s) > 0$ where*

$$\kappa(s) = \inf\left\{ \frac{v'M^TMv}{n\sum_{j\in J} v_j^2} \;\middle|\; \begin{array}{l} v \in \mathbb{R}^p, \quad J \subset \{1,\dots,p\}, \quad |J| < s \\ \sum_{j\notin J}|v_j| \le c\sum_{j\in J}|v_j| \end{array} \right\}.$$

Theorem 2.3 (Theorem 6.1 page 1716 in Bickel, Ritov and Tsybakov [33])**.** *Assume that* $\mathrm{REC}(s, 3+4/a)$ *holds for some $s \in \{1,\dots,p\}$ and $a > 0$. Fix $\varepsilon > 0$ and put $\lambda = 2\sigma\sqrt{2\log(p/\varepsilon)/n}$. Then*

$$\mathbb{P}\left(\|\hat\theta^1_\lambda - \overline\theta\|_n^2 \le (1+a) \inf_{\mathrm{card}(I)\le s} \left\{ \|\overline\theta_I - \overline\theta\|_n^2 + \frac{8C_a\sigma^2}{\kappa^2\left(s, 3+\frac{4}{a}\right)} \frac{\mathrm{card}(I)\log\left(\frac{p}{\varepsilon}\right)}{n} \right\} \right) \ge 1-\varepsilon$$

for some constant $C_a > 0$.

In practice, we encounter several difficulties :

- Assumption $\mathrm{REP}(\|\overline\theta\|_0, 3+4/a)$ might not hold on the matrix M. It is impossible to check this assumption on the data. Even to check $\mathrm{REC}(s, 3+4/a)$ for some known s is computationally demanding.
- The choice $\lambda = 2\sigma\sqrt{2\log(p/\varepsilon)/n}$ dependend on σ that is usually not known to the statistician. Some variants that would not depend on σ were proposed by Belloni and Chernozhukov [27] and Huet and Verzelen [100]. In simulations, however, even when σ is known, $\lambda = 2\sigma\sqrt{2\log(p/\varepsilon)/n}$ does not perform well in practice. Some propositions of λ based on resampling procedures can be found in papers by Bach [16] under the name BOLASSO and Meinshausen and Bühlmann [160] and Shah and Samworth [194] under the name "stability selection".
- Finally, the LASSO performs poorly in many practical situations, it is rather unstable and highly biased. Some variants of the ℓ_1-penalty were proposed that led to notable improvements in practice : Fan and Li's SCAD [88], Zou and Hastie's elastic net [240], Zou's adaptive LASSO [239]. One of the simplest way to attenuate the bias of the LASSO is to use the least square estimator constrained to a set of variables selected by the LASSO. This procedure is very popular in practice and its statistical properties were studied by Belloni and Chernozhukov [26]. All these procedures require an assumption on M, like $\mathrm{REC}(s,c)$ or more stringent ones.

Remark 2.4. *Due to its nice algorithmic properties, the LASSO was extended to a lot of different situations. In the case where the parameter $\overline{\theta}$ is sparse and smooth, Hebiri and van de Geer proposed the smooth LASSO [114]. In the case where it is sparse by blocs, Yuan and Lin introduced the group LASSO [230]. Tibshirani, Saunders, Rosser, Zhu and Knight proposed the fused LASSO [206] when the parameter is sparse and constant by blocs, a situation that makes sense in genomics. Econometric models with instrumental variables were considered Belloni, Chen, Chernozhukov and Hansen [25] and Gautier and Tsybakov [93]. We proposed a transductive version of the LASSO with Hebiri [A7, A9]. In order not to loose the guideline of this thesis we won't describe all these variants here.*

Remark 2.5. *Other ideas of efficient algorithms for variable selection were proposed. Screening methods aim at removing irrelevant variables based on a correlation criterion : see Fan and Lv's SIS method [89], our papers on Iterative Feature Selection [A1, A3, A14], Mougeot, Picard and Tribouley's LOL [166], see also Comminges and Dalalyan [63], Kolar and Liu [128]. These methods rely on* REC *type assummption, or stronger assumptions like coherence assumption or even orthogonality of the design. Greedy algorithms are another alternative, see Barron, Cohen, Dahmen and DeVore [22], Zhang [234, 235]. Huang, Cheang and Barron [121] established a link between greedy algorithms and ℓ_1-penalization. Finally, when the design matrix M is diagonal, the LASSO is equivalent to the soft-thresholding procedure studied by Donoho and Johstone [78, 79], with Kerkyacharian and Picard [80], see also the monograph by Härdle, Kerkyacharian, Picard and Tsybakov [110].*

3 Agregated estimators and PAC-Bayesian bounds

In [70], Dalalyan and Tsybakov studied an aggregated estimator that they called EWA (Exponentially Weighted Aggregation). This estimator enjoys nice theoretical properties : an oracle inequality withouth any REC-type assumption. On the other hand, Monte Carlo methods allow to compute this estimator for reasonably large values of p. Note, however, that the theoretical results were given in the fixed design setting only, and required $\|\overline{\theta}\|_1$ to be bounded. In our joint paper with Lounici [A5] we improved on [70] in two directions :

- We proposed an estimator that enjoys the same properties as the one in [70] without the boundedness assumption.
- We proposed another estimator which satisfies an oracle inequality for the prevision risk (allowing random design).

We now introduce both estimators. We put $\mathcal{P}_n(\{1,\dots,p\})$ the set of all subsets I of $\{1,\dots,p\}$ with $\mathrm{card}(I) \leq n$. We put $\Theta = \mathbb{R}^p$, for $I \subset \{1,\dots,p\}$ we put $\Theta_I = \{\theta \in \mathbb{R}^p : \mathrm{supp}(\theta) = I\}$. We also put $\Theta(K) = \{\theta \in \mathbb{R}, \|\overline{\theta}\|_1 \leq K\}$ and $\Theta_I(K) = \Theta_I \cap \Theta(K)$. We fix real number $\alpha, K > 0$ and put

$$\pi_I = \frac{\alpha^{\mathrm{card}(I)}}{\binom{p}{\mathrm{card}(I)} \sum_{j=0}^{n} \alpha^j}.$$

We also define u_I as the uniform probability measure on $\Theta_I(K+1)$ and the probability measure π by

$$\pi(\mathrm{d}\theta) = \sum_{I \in \mathcal{P}_n(\{1,\dots,p\})} \pi_I \mathrm{u}_i(\mathrm{d}\theta).$$

Definition 3.1. *For any $\lambda > 0$ we put*

$$\hat{\theta}_\lambda = \frac{\sum_{I\in\mathcal{P}_n(\{1,\dots,p\})} \pi_I \exp\left[-\lambda\left(r(\hat{\theta}_I^{LSE}) + \frac{2\sigma^2\mathrm{card}(I)}{n}\right)\right]\hat{\theta}_I^{LSE}}{\sum_{I\in\mathcal{P}_n(\{1,\dots,p\})} \pi_I \exp\left[-\lambda\left(r(\hat{\theta}_I^{LSE}) + \frac{2\sigma^2\mathrm{card}(I)}{n}\right)\right]}.$$

Definition 3.2. *For any $\lambda > 0$ we define the probability distribution $\tilde{\rho}_\lambda$ by*

$$\frac{\mathrm{d}\tilde{\rho}_\lambda}{\mathrm{d}\pi}(\theta) = \frac{\exp(-\lambda r(\theta))}{\int_{\Theta_K}\exp(-\lambda r(\theta'))\pi(\mathrm{d}\theta')} \text{ and } \tilde{\theta}_\lambda = \int_{\Theta_K}\theta\tilde{\rho}_\lambda(\mathrm{d}\theta).$$

Theorem 3.3 (Theorem 2.1 page 132 in [A5])**.** *As in Sections 1 and 2, let us assume that the X_i's are deterministic. Assume that the W_i are $\mathcal{N}(0,\sigma^2)$. Put $\lambda = 1/(4\sigma^2)$, then*

$$\mathbb{E}\left(\|\hat{\theta}_\lambda - \overline{\theta}\|_n^2\right) \le \inf_{I\in\mathcal{P}_n(\{1,\dots,p\})}\left\{\|\overline{\theta}_I - \overline{\theta}\|_n^2 + \frac{\sigma^2\mathrm{card}(I)}{n}\left(4\log\left(\frac{pe}{\alpha\mathrm{card}(I)}\right)+1\right) + \frac{4\sigma^2\log\left(\frac{1}{1-\alpha}\right)}{n}\right\}.$$

Definition 3.4. *We remind that the random variable W_1 is said to be sub-exponential with parameters $(\sigma,\xi)\in(\mathbb{R}_+^*)^2$ if $\mathbb{E}(W_1^2)\le\sigma^2$ and for any $k\ge 3$, $\mathbb{E}(|W_1|^k)\le\sigma^2 k!\xi^{k-2}$.*

Examples of sub-exponential random variables are bounded random variables : then $\xi = \sigma = \|W_1\|_\infty$, or Gaussian random variables $\mathcal{N}(0,s^2)$, then $\sigma = \xi = s$. We insist on the fact that the following theorem is valid both when the X_i's are deterministic or random with the pairs (X_i, Y_i) being i.i.d.

Theorem 3.5 (Theorem 3.1 page 133 in [A5])**.** *We assume that the W_i are sub-exponentials with parameters (σ,ξ). We assume that $\|f\|_\infty < L$ for some $L > 0$. We assume that a.s., $\|X_i\|_\infty \le L'$ for some $L' > 0$. Finally, assume that $\|\overline{\theta}\|_1 \le K$. For some (known) constant $\mathcal{C}$ that depends on σ, ξ, K, L, L', put $\lambda = n/\mathcal{C}$, we have for any $\varepsilon > 0$,*

$$\mathbb{P}\left\{R(\tilde{\theta}_\lambda) - R(\overline{\theta}) \le \frac{\mathcal{C}}{n}\left[\mathrm{card}(I)\log\left(\frac{enp(K+1)}{\alpha\mathrm{card}(I)}\right) + \log\left(\frac{2}{\varepsilon(1-\alpha)}\right) + \frac{3L'^2}{n}\right]\right\} \ge 1-\varepsilon.$$

Basically, these two results ensure that the estimators $\hat{\theta}_\lambda$ and $\tilde{\theta}_\lambda$ enjoys the same theoretical property as the LASSO, without any stringent assumption on the design like REP. On the other hand, when compared to the ℓ_0 penalized estimator, it is possible to compute $\hat{\theta}_\lambda$ and $\tilde{\theta}_\lambda$ for reasonably large p thanks to some Monte Carlo algorithm, see the simulation study below.

Remark 3.6. *The probability measure π can be seen as a prior probability distribution and it is possible to interpretate $\tilde{\rho}_\lambda$ as a posterior distribution in a Bayesian framework. There is a huge Bayesian litterature on the variable selection problem, various type of priors were considered (including some related to π above). We refer the reader to the surveys by George and McCulloch [95], George [94], and more recently to the papers by West [220], Cui and George [65], Liang, Paulo, Molina, Clyde and Berger [147], Scott and Berger [190] among others for the linear regression case and Nott and Leonte [169], Jiang [123] for more general models.*

The technique used to prove Theorem 3.5 relies on so-called PAC-Bayesian inequalities. As these inequalities will be heavily used in Parts II and III too, we now give more details on these.

PAC-Bayesian inequalities were initially introduced McAllester [155, 156, 157] based on earlier remarks by Shawe-Taylor and Williamson [195] ; more PAC-Bayesian bounds are proved by Langford, Seeger and Mediggo [133], Herbrich and Graepel [115] and Meir and Zhang [162] among others. In all these papers, the objective was to produce "PAC" type bounds for Bayesian estimators. "PAC" is an acronym meaning Probably Approximately Correct, it was introduced by Valiant [210] to refer to any bound valid with large probability (as the one in Theorem 3.5) together with the constraint that the estimator must be calculable in polynomial time w.r.t. n and $1/\varepsilon$, however, most authors refer now to PAC inequalities for any bound on the risk valid with large probability.

McAllester's type PAC-Bayesian inequalities are empirical bounds, in the sense that the upper bound on the risk depends on the observations, and not on unknown quantities such as $\mathbb{P}$ or $\overline{\theta}$. Catoni [54, 55] extended PAC-Bayesian bounds to prove oracle-type inequalities for aggregated estimators, see also earlier works on aggregation and oracle inequalities : Chapters 5 (page 183) and 6 (page 207) in Nemirovksi [168], papers by Juditsky and Nemirovki [125], Vert [216], Catoni [52], Yang [224, 225], and the lower bounds by Tsybakov [208]. In Catoni's works on PAC-Bayesian bounds, the aggregated estimators are refered as "Gibbs estimators". Catoni's technique relies on two main ingredients :

- First, a deviation inequality is used to upper bound the distance between $r(\theta)$ and $R(\theta)$ for a fixed $\theta \in \Theta$, for example, the so called Hoeffding's or Hoeffding-Azuma's inequality [117, 14], Bernstein's inequality [31] or Bennett's one [28]. We refer the reader to the paper by Bercu, Gassiat and Rio [30], Chapter 14 page 481 in Buhlmann and van de Geer's book [41] or to Chapter 2 page 18 in the comprehensive monograph by Boucheron, Lugosi and Massart [38] for more details on these inequalities.
- The second step is to make this bound valid for any $\theta \in \Theta$ simultaneously. In the study of penalized empirical risk minimizers, this step is done thanks to concentration of measure inequalities, see e.g. Ledoux and Talagrand [138] or the aforementioned book [38] for concentration inequalities, and Massart [154] for the study of these estimators. In PAC-Bayesian theory, the approach is sligthly different. Instead of all possible parameters θ, Catoni considers the set of all probability distributions on Θ equiped with some suitable σ-algebra, and make the deviation inequality uniform on this set thanks to Donsker and Varadhan's variational inequality [81].

This last step requires to fix a reference distribution m on the space Θ (in Theorem 3.5 above, $\mathrm{m} = \pi$). By analogy with Bayesian statistics, this distribution is called a prior. However, it is used to control the complexity of the parameter space rather than to include some prior belief. Catoni [52, 54] also makes links with information theory and Rissanen's MDL principle [179] (minimum description length, see also Barron, Rissanen and Yu [23]). This is explored further by Zhang [233] who adapted the method to prove lower bounds too. Audibert and Bousquet [12] studied the link with generic chaining. Fore more recent advances on PAC-Bayesian theory see also Jiang and Tanner [124], Audibert [11], Audibert and Catoni [13].

Following a method initiated by Leung and Barron [145], Dalalyan and Tsybakov [70] replaced the first step (deviation inequality) by Stein's formula [197]. As a pro, this makes life easier when dealing with unbounded parameter spaces, so we also used this technique to prove Theorem 3.3. On the other hand, this produces results valid in expectation only. This paper was further investigated and improved in a series of papers, Dalalyan and Tsybakov [71, 72], Dalalyan and Salmon [69]. See Rigollet and Tsybakov [176], Dai, Rigollet, Xia and Zhang [67, 66], Lecué and Rigollet [137] for more recent advances on aggregation. Gerchinovitz [97] extended the technique of [70] to the online case (i.e. where data is given sequentially). Also, recently, Arias-Castro and Lounici [8] studied the variable selection abilities of a thresholded version of $\tilde{\theta}_\lambda$. Percival [172] studied aggregated estimators under priors inducing group sparsity.

Remark 3.7. *We also want to mention that PAC-Bayesian and related aggregation methods were extended to other problems than regression and classification, such as density estimation (Catoni [52, 54]) and variants like taylored density estimation (Higgs and Shawe-Taylor [116]), clustering (Seldin and Tishby [193]), ranking (Li, Jiang and Tanner [146] and Robbiano [181]), multiple testing (Blanchard and Fleuret [36]). Salmon and Le Pennec also interpretated the popular NL-means image denoising technique in this framework [187].*

A short simulation study

We implemented a Monte Carlo algorithm to compare the performances of $\hat{\theta}_\lambda$ and $\tilde{\theta}_\lambda$ to the ones of Tibshirani's LASSO $\hat{\theta}^1_\lambda$ [205]. Namely, we used different versions of the popular Metropolis-Hastings algorithm (see the monograph by Robert and Casella [182] for an introduction to Monte Carlo methods) :

- For $\hat{\theta}_\lambda$, we just have to approximate a mean of a finite (but huge) numbers of estimators $\hat{\theta}^{LSE}_I$ for $I \in \mathcal{P}_n$, so we used a standard version of Metropolis-Hastings on the finite set $I \in \mathcal{P}_n$.
- The situation is a bit more intricate for $\tilde{\theta}_I$, we used Green's version of Metropolis-Hastings [102] called Reversible Jump Monte Carlo Markov Chain (RJMCMC). This strategy turned out to be successful in many Bayesian model selection problems, see the examples in [102] and in Green and Richardson [103].

The complete description of the algorithm can be found in a preliminary version of [A5] :

`http://arxiv.org/pdf/1009.2707v1.pdf`

For the sake of completeness, we provide here some simulation results.

Description of the experiments : we consider variants of the toy example in Tibshirani's paper [205] :

$$\forall i \in \{1, \dots, n\}, \quad Y_i = \overline{\theta} \cdot X_i + W_i$$

with $X_i \in \mathcal{X} = \mathbb{R}^p$, $\overline{\theta} \in \mathbb{R}^p$ and the W_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, $n = 20$. The X_i's are i.i.d. (and independent from the W_i) and drawn from the gaussian distribution with zero mean and Toeplitz variance matrix :

$$\Sigma(\rho) = \left(\rho^{|i-j|}\right)_{(i,j)\in\{1,\dots,p\}^2}$$

TABLE 1 – Results for $\theta = 5 \times (e^{-1}, e^{-2}, ..., e^{-p})$.

σ^2	p		$\hat{\theta}^1_\lambda$	$\hat{\theta}_\lambda$	$\tilde{\theta}_\lambda$
1	8	median	0.138	**0.089**	0.121
		mean	0.178	**0.137**	**0.137**
		s.d.	0.145	0.121	0.116
3	8	median	0.397	0.437	**0.364**
		mean	0.434	0.430	**0.400**
		s.d.	0.178	0.271	0.282
1	30	median	0.262	**0.203**	0.205
		mean	0.277	0.247	**0.240**
		s.d.	0.147	0.149	0.149
3	30	median	0.593	0.519	**0.423**
		mean	0.630	0.665	**0.534**
		s.d.	0.409	0.684	0.383
1	100	median	0.276	**0.256**	0.261
		mean	0.375	0.353	**0.342**
		s.d.	0.256	0.200	0.199
3	100	median	1.045	0.687	**0.680**
		mean	1.023	0.809	**0.760**
		s.d.	0.364	0.476	0.464
1	1000	median	0.486	**0.390**	0.407
		mean	0.464	**0.373**	0.386
		s.d.	0.207	0.108	0.103
3	1000	median	1.549	**1.199**	1.288
		mean	1.483	1.268	**1.245**
		s.d.	0.460	0.702	0.692

for some $\rho \in [0, 1)$. Note that Tibshirani's toy example is set with $p = 8$ whereas we will consider here $p \in \{8, 30, 100, 1000\}$. We use two regression vectors :

$$\overline{\theta} = (5e^{-1}, 5e^{-2}, 5e^{-3}, 5e^{-4}, ...)$$
$$\text{and then } \overline{\theta} = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, ...)$$

respectively an *approximately sparse* parameter, and a *sparse* parameter. We will take σ^2 respectively equal to 1, *low noise* situation, and 3, *noisy* case ; the value of ρ is fixed to 0.5. We fix $\alpha = 1/10$.

The LASSO parameters are optimized on a grid Λ, as well as the λ parameter for our aggregates $\hat{\theta}_\lambda$ and $\tilde{\theta}_\lambda$. We report the compare the oracle results here, in practice, one would optimize these parameters through cross validation or a related method. For example, for $\hat{\theta}_\lambda$ we report $\min_{\lambda \in \Lambda} \|\hat{\theta}_\lambda - \overline{\theta}\|_n^2$.

We perform each experiment 20 times and report the mean, median and standard deviation of the results for the sparse situation in Table 1 and for the approximately sparse situation in Table 2.

We can see on these experiments that the aggregated estimators outperforms the LASSO in the low noise case $\sigma = 1$. When σ grows, the performances of our estimators

TABLE 2 – Numerical results for the estimation of $\theta = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, ...)$.

σ^2	p		$\hat{\theta}^1_\lambda$	$\hat{\theta}_\lambda$	$\tilde{\theta}_\lambda$
1	8	median	0.302	0.176	**0.172**
		mean	0.291	0.215	**0.209**
		s.d.	0.211	0.190	0.185
3	8	median	0.437	0.533	**0.370**
		mean	0.535	0.612	**0.527**
		s.d.	0.398	0.420	0.395
1	30	median	0.355	0.157	**0.143**
		mean	0.360	0.217	**0.209**
		s.d.	0.189	0.151	0.150
3	30	median	1.459	1.511	**1.267**
		mean	1.431	1.809	**1.333**
		s.d.	0.702	1.143	0.607
1	100	median	0.399	0.244	**0.204**
		mean	0.471	0.248	**0.212**
		s.d.	0.222	0.162	0.130
3	100	median	**1.378**	1.674	1.409
		mean	1.396	1.800	**1.365**
		s.d.	0.687	0.653	0.562

are still better, but the difference is less significative ; moreover, $\hat{\theta}_\lambda$ seems to become less stable (in particular Table 2, $p = 30$ and $\sigma^2 = 3$).

This simulation study clearly shows the advantage to use the exponential weights estimators, in particular $\tilde{\theta}_\lambda$, especially in the situation of approximate sparsity. As we mentioned in the introduction, the main advantage of the LASSO and of related methods is the computational efficiency. When p becomes larger ($p > 1000$), the RJMCMC algorithm takes much time to converge and the computation time becomes prohibitive. The strength of ℓ_1-penalized estimators is that they can be computed say for $p \simeq 10^7$ in a reasonable amount of time.

Part II

Extension to time series

Summary

In this part, we focus on time series. The assumption that the observations are independent does not make sense any more : in practice, times series always exhibit some kind of temporal dependence.

First, in Section 4, we introduce various weak dependence and mixing assumptions required to prove oracle inequalities. As a first step, we prove some results on the LASSO in Section 5. However, this covers only a special case (namely, when the design is deterministic and only the noise exhibit some dependence). In Section 6, we extend full parts of the PAC-Bayesian theory to the time series setting.

4 Mixing, weak dependence and examples

In order to study the theoretical properties of statistical procedures for time series, several measures of the dependence of random variables were introduced - the idea being that when the dependence is not strong, we should obtain results close to the ones known in the independent setting. The first example example of such a measure was Rosenblatt's α-mixing coefficient [184]. Other mixing coefficients were introduced : β-mixing and φ-mixing. As we will use φ-mixing coefficients later in this part, we give their definition.

Definition 4.1 (Ibragimov φ-mixing coefficient [122])**.** *Let $(W_i)_{i\in\mathbb{Z}}$ be a sequence of random variables. We put*

$$\varphi^W(r) = \sup\Big\{ \left|\mathbb{P}(B|A) - \mathbb{P}(B)\right|, A \in \sigma(W_k, W_{k-1}, W_{k-2}, \dots), B \in \sigma(W_{k+r}), k \in \mathbb{Z}\Big\}.$$

Finally, we say that $(W_i)_{i\in\mathbb{Z}}$ satisfies Assumption PhiMix($\mathcal{C}$) *for $\mathcal{C} \in \mathbb{R}_+$ if*

$$1 + \sum_{j=1}^{\infty} \sqrt{\varphi^W(j)} \leq \mathcal{C}.$$

Intuitively, when W_{k+r} is "almost independent" of $W_k, W_{k-1}, W_{k-2}, \dots$ then $\varphi^W(r)$ must be very small. Conditions on various mixing coefficients, like convergence to 0 when $r \to \infty$ or summability, allow to prove law of large numbers and central limit theorem for the process (W_i). We refer the reader to the monographs by Doukhan [82] and more recently Rio [177] for more details.

The trouble with mixing coefficients is that they provide a very restrictive notion of dependence. For example, let us consider the following definition.

Definition 4.2 (Causal Bernoulli shifts with bounded innovations)**.** *Let $k \in \mathbb{N}$. Let $(\xi_i)_{i\in\mathbb{Z}}$ be a sequence of i.i.d. bounded $\mathbb{R}^k$-valued random variables : a.s., $\|\xi_i\| \leq c_\xi$. Let $H : (\mathbb{R}^k)^{\mathbb{N}} \to \mathbb{R}^k$ be a measurable function, and $a = (a_i)_{i\in\mathbb{N}}$ be a sequence of non-negative numbers such that*

$$\|H(v) - H(v')\| \leq \sum_{j=0}^{\infty} a_j \|v_j - v'_j\|$$

and

$$\sum_{j=0}^{\infty} a_j < +\infty.$$

In this case, the process $W_i = H(\xi_i, \xi_{i-1}, \xi_{i-2} \dots)$ *exists, is stationary and is said to be a CBS (causal Bernoulli shift) with bounded innovations. For short we will say that* (W_i) *satisfies Assumption* $\mathrm{CBS}(c_\xi, a)$.

Intuitively, when the coefficients a_j decay to 0 fast enough, W_{k+r} is "almost independent" of $W_k, W_{k-1}, W_{k-2}, \dots$ However, there are known examples where the a_j decay exponentially fast, but α, β and φ-mixing coefficients are all (non zero) constant. See, for example, the third remark page 325 in Doukhan and Louhichi [83]. Remark that the class of CBS with bounded innovations is large, it includes among others all $\mathrm{ARMA}(p, q)$ time series with bounded innovations. More generally, it includes a wide set of chains with infinite memory $W_t = F(W_{t-1}, W_{t-2}, \dots; \xi_t)$ (the conditions on the function F can be found in the paper by Doukhan and Wintenberger [85]).

For this reason (among others), another type of dependence conditions was developed : weak dependence. There are a lot of various type of weak dependence coefficients, we won't mention them all here. We refer the reader to Chapter 2 in the monograph by Dedecker, Doukhan, Lang, Léon, Louhichi and Prieur [73] for a comprehensive survey. We give here the following θ-dependence coefficient that we will use later in this part.

Definition 4.3 (Dedecker, Doukhan, Lang, Léon, Louhichi and Prieur [73])**.** *For* $k \in \mathbb{N}$, *let* $(W_i)_{i\in\mathbb{Z}}$ *be a sequence of* $\mathbb{R}^k$*-valued random variables. For any* $q \in \mathbb{N}$, *for any* $\mathbb{R}^k$*-valued random variable* $Z_1, \dots, Z_q$ *defined on* $(\Omega, \mathcal{A}, \mathbb{P})$, *we define*

$$\theta_\infty(\mathfrak{S}, (Z_1, \dots, Z_q)) = \sup_{f\in\Lambda_1^q} \Big\| \mathbb{E}\left[f(Z_1, \dots, Z_q)|\mathfrak{S}\right] - \mathbb{E}\left[f(Z_1, \dots, Z_q)\right] \Big\|_\infty$$

where

$$\Lambda_1^q = \left\{ f : (\mathbb{R}^p)^q \to \mathbb{R}, \quad \frac{|f(z_1, \dots, z_q) - f(z_1', \dots, z_q')|}{\sum_{j=1}^q \|z_j - z_j'\|} \le 1 \right\},$$

and

$$\theta_{\infty,h}^W(1) := \sup\left\{\theta_\infty(\sigma(W_t, t \le p), (W_{j_1}, \dots, W_{j_\ell})), \quad p < j_1 < \dots < j_\ell, 1 \le \ell \le h\right\}.$$

Finally, we say that Assumption $\mathrm{ThetaDep}(\mathcal{C})$ *is satisfied for* $\mathcal{C} \in \mathbb{R}_+$ *if, for any* h, $\theta_{\infty,h}^W(1) \le \mathcal{C}$.

We relate the previous definitions of dependence in the following proposition.

Proposition 4.4. *Let* $k \ge 1$ *and* $(W_i)_{i\in\mathbb{Z}}$ *be a sequence of* $\mathbb{R}^k$*-valued random variables. Then :*

1. $\mathrm{PhiMix}(\mathcal{C})$ *and* $\|W_i\| \le c_W$ *a.s. for any* $i \Rightarrow \mathrm{ThetaDep}(c_W\mathcal{C})$.
2. $\mathrm{CBS}(c_\xi, a)$ *and* $\sum_{j=0}^\infty j a_j < \infty \Rightarrow \mathrm{ThetaDep}(2c_\xi \sum_{j=0}^\infty j a_j)$.

So θ-dependence is a quite general notion of dependence. The proof of the first point is actually a byproduct of the proof of Corollaire 1 page 907 in a paper by Rio [178] that will be discussed below. The second point is Proposition 4.2 page 891 in our joint paper with Wintenberger [A8].

As we mentioned in Part I, the main tool we use to control the risk of various estimation procedures is a deviation inequality like Bernstein and Hoeffding's inequalities

in the i.i.d. case. Since the 80's, such inequalities were proved under various mixing and weak dependence assumptions, see Statuljavichus and Yackimavicius [196], Doukhan and Louhichi [83], Rio [178], Samson [188], Dedecker and Prieur [74], Doukhan and Neumann [84], Merlevede, Peligrad and Rio [163, 164] and Wintenberger [221] and in the aforementioned monographs [82, 177, 73]. Deviation inequalities were also proved for more specific type of time series, like dynamical systems, by Collet, Martinez and Schmitt [62] or Markov Chains, by Clémençon [61], Catoni [53], Adamczak [2], Bertail and Clémençon [32].

5 LASSO in the dependent setting

In this section, we consider the same model as in Part I Sections 1 and 2 : we assume that there are n random variables W_i on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and n deterministic numbers $X_i \in \mathbb{R}^p$ and

$$Y_i = f(X_i) + W_i$$

for $1 \leq i \leq n$ for some measurable function f, that $\mathbb{E}(W_i|X_i) = 0$ and $\mathbb{E}(W_i^2|X_i) \leq \sigma^2$. We keep the same notations $f_\theta(x) = \theta \cdot x$, $r(\theta) := \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2$, $R(\theta) := \mathbb{E}[r(\theta)]$, $\overline{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} R(\theta)$, $\|\theta - \overline{\theta}\|_n^2 = R(\theta) - R(\overline{\theta})$ and M is the design matrix $M = (X_1|\dots|X_n)$. We still assume that the X_i's are normalized in such a way that $M^T M/n$ has only 1 on its diagonal. We put

$$c_X = \sup_{1 \leq i \leq n} \|X_i\|_\infty.$$

But this time, we do not assume any longer that the W_i are independent.

We have the following result on the LASSO estimator (we remind that the Definition is given in Definition 2.1 page 14).

Theorem 5.1 (Theorem 2.1 page 755 in Alquier and Doukhan [A6])**.** *We assume that M satisfies* $\mathrm{REC}(\|\overline{\theta}\|, 3)$ *(Definition 2.2 page 15). We assume that there is a constant $\alpha \in [0, 1/2]$ and a decreasing continuous function ψ such that for any $j \in \{1, \dots, p\}$, for any $t > 0$,*

$$\mathbb{P}\left(\left|\frac{c_X}{n}\sum_{i=1}^n W_i\right| \geq n^{\alpha - \frac{1}{2}} t\right) \leq \psi(t). \tag{2}$$

Let us choose $\varepsilon > 0$ and put $\lambda \geq \lambda^* := 4n^{\alpha-\frac{1}{2}}\psi^{-1}(\varepsilon/p)$. Then

$$\mathbb{P}\left(\|\hat{\theta}_\lambda^1 - \overline{\theta}\|_n^2 \leq \frac{4\lambda^2 \|\overline{\theta}\|_0}{\kappa(\|\theta\|_0)}\right) \geq 1 - \varepsilon.$$

Remember that the definition of $\kappa(\cdot)$ is given in Definition 2.2 page 15. Note that, in particular, we have

$$\mathbb{P}\left(\|\hat{\theta}_{\lambda^*}^1 - \overline{\theta}\|_n^2 \leq \frac{64\|\overline{\theta}\|_0 \left[\psi^{-1}\left(\frac{\varepsilon}{p}\right)\right]^2}{\kappa(\|\theta\|_0) n^{1-\frac{2}{\alpha}}}\right) \geq 1 - \varepsilon.$$

Also, note that our paper [A6] also contains results on the estimation of the density based on a dependent sample using a LASSO type estimator, studied under the name SPADES (SPArse Density EStimator) in the i.i.d. setting by Bunea, Tsybakov and Wegkamp [45].

The proof follows the main steps of the one of Theorem 6.1 in Bickel, Ritov and Tsybakov [33]. Actually, if we assume that the W_i are Gaussian or bounded, we can use a Hoeffding type inequality to check that (2) is satisfied with $\alpha = 0$ and $\psi(t)$ in $\exp(-t^2)$, and we obtain a result close to the one in [33]. In the case were the W_i are not independent, we have to use a deviation inequality suited for dependent random variables in order to check that (2) is satisfied.

In [A6], we gave some weak-dependence conditions under which such a deviation inequality is satisfied : one of them is base on a version of Bernstein's inequality by Doukhan and Neumann [84] and the other one on Doukhan and Louhichi's Marcinkiewicz-Zygmund type inequality [83]. Here, we provide a slightly different (and simpler) bound based on Rio's version of Hoeffding's inequality [178].

Theorem 5.2 (Variant of Corollary 4.3 page 462 in Alquier and Doukhan [A6])**.** *We assume that M satisfies* $\mathrm{REC}(\|\overline{\theta}\|, 3)$ *and that the W_i satisfy assumption* $\mathrm{ThetaDep}(\mathcal{C})$ *and that $|W_i| < c_W$ a.s. for any i, then for $\lambda = 4c_X(c_W + \mathcal{C})\sqrt{2\log(2p/\varepsilon)/n}$ we have*

$$\mathbb{P}\left\{ \|\hat{\theta}^2_\lambda - \overline{\theta}\|^2_n \leq \frac{128 c_X^2 (c_W + \mathcal{C})^2}{\kappa(\|\overline{\theta}\|_0)} \frac{\|\overline{\theta}\|_0 \log\left(\frac{2p}{\varepsilon}\right)}{n} \right\} \geq 1 - \varepsilon.$$

Note that, if the W_i are actually independent, then we can take $\mathcal{C} = 0$ and in this case, we obtain exactly what we would have obtained with the original Hoeffding's inequality [117], this is due to the fact that the constants are tight in Rio's inequality.

Yoon, Park and Lee [227] also studied the LASSO under the same setting (deterministic design, dependent noise). Instead of weak dependence assumptions, they use a parametric assumption on the noise : the noise is autoregressive. They don't provide non asymptotic results such as Theorem 5.1, but on the other hand, they provide the exact asymptotic distribution of the LASSO in this setting.

6 PAC-Bayesian bounds for stationary time series

In many applications, the assumption that the X_i's are deterministic is too restrictive. In order to predict time series using autoregressive type models, we need to relax this assumption. The LASSO was used to select variables in VAR (vectorial autoregressive) models by Hsu, Hung and Chang [120], without theoretical justification. Wang, Li and Tsai [217] proved asymptotic results for the LASSO in the AR (autoregressive) case. However, to the best of our knowledge, there is no non-asymptotic theory for the LASSO in this context.

Other estimation methods were studied in the context of time series. While there is a long history of nonparametric statistics for time series (we refer the reader to the aforementioned books [82, 73] and the references therein and a recent different approach by Delattre and Gaïffas [75]), there were only a few attempts to generalize the statistical learning approach to this context. We already mentioned in Part I Massart's approach

for penalized estimators based on concentration inequalities [154], Baraud, Comte and Viennet [18] extended this approach for regression and autoregression with β-mixing time series. However, the risk function they used does not exactly correspond to a prevision risk, it is actually the empirical norm. In order to study risks more suited to prediction, several authors established Vapnik's type bound (see [213]) under various mixing assumtions : Modha and Masry [165], Meir [161], Xu and Chen [223], Zou, Li and Xu [238], Steinward and Christmann [198], Steinwart, Hush and Scovel [199], Hang and Steinwart [108]. Up to our knowledge, none of these authors considered weak dependence assumptions. We also want to mention another approach used to establish bounds on the prevision risk for time series : to use bounds coming from the theory of individual sequences prediction, see Cesa-Bianchi and Lugosi [57] or Stoltz [200] for an introduction to this theory. With this approach, the observations are usually considered as deterministic, but it is possible to obtain bounds for previsions of random variables thanks to a trick very well described, for example, in the introduction of the Ph.D. thesis of Gerchinovitz [96] in the i.i.d. case. This approach was used by Duchi, Agarwal, Johansson and Jordan [86, 3] to predict φ-mixing time series.

In our papers with Li and Wintenberger [A8, A15, A19], as an alternative, we extended PAC-Bayesian inequalities to the context of weakly dependent time series forecasting. We introduce these results in this section.

Let $(X_t)_{t\in\mathbb{Z}}$ be a $\mathbb{R}^p$-valued, stationary, bounded time series : $\|X_t\| \leq c_X$ almost surely. In a first time, we will consider a rather general family of predictors : we fix an integer k and a set of predictors $\left\{f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta\right\}$: for any θ and any t, f_θ applied to the last past values $(X_{t-1}, \ldots, X_{t-k})$ is a possible prediction of X_t. For short we put, for any $t \in \mathbb{Z}$ and any $\theta \in \Theta$,

$$\hat{X}_t^\theta := f_\theta(X_{t-1}, \ldots, X_{t-k}).$$

However, due to the fact that some weak dependence properties are only hereditary through Lipshitz functions, we have to assume some structure on (f_θ). Namely, there is an $L > 0$ such that for any $\theta \in \Theta$, there are coefficients $a_j(\theta)$ for $1 \leq j \leq k$ satisfying, for any x_1, ..., x_k and y_1, ..., y_k in $\mathbb{R}^p$,

$$\|f_\theta(x_1, \ldots, x_k) - f_\theta(y_1, \ldots, y_k)\| \leq \sum_{j=1}^k a_j(\theta) \|x_j - y_j\|, \text{ with } \sum_{j=1}^k a_j(\theta) \leq L. \quad (3)$$

Example 6.1 (Linear Auto-Regressive (AR) predictors)**.** *When $p = 1$ (real valued time series) we put $\theta = (\theta_0, \theta_1, \ldots, \theta_k) \in \Theta = \{\theta \in \mathbb{R}^{k+1} : |\theta_1| + \cdots + |\theta_k| \leq L\}$ and*

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \theta_0 + \sum_{j=1}^k \theta_j X_{t-j}.$$

We also fix a loss function ℓ is given by : $\ell(x, x') = g(x - x')$ for some convex function g with : $g \geq 0$, $g(0) = 0$ and g is K-Lipschitz.

Definition 6.2 (Risk function)**.** *We put, for any $\theta \in \Theta$, $R(\theta) = \mathbb{E}\left[\ell\left(\hat{X}_t^\theta, X_t\right)\right]$, and as usual we fix $\overline{\theta}$ a minimizer of R.*

Note that because of the stationarity, $R(\theta)$ does not depend on t.

Definition 6.3 (Empirical risk). *For any* $\theta \in \Theta$, $r(\theta) = \frac{1}{n-k}\sum_{i=k+1}^{n} \ell\left(\hat{X}_i^{\theta}, X_i\right)$.

As in Section 3, we fix a suitable σ-algebra (say $\mathcal{T}$) on Θ. Let $\mathcal{M}_+^1(\Theta)$ be the set of all probability distributions on $(\Theta, \mathcal{T})$. We fix a prior probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$.

Definition 6.4 (Gibbs estimator). *We put, for any* $\lambda > 0$,

$$\tilde{\theta}_\lambda = \int_\Theta \theta \tilde{\rho}_\lambda(\mathrm{d}\theta), \text{ where } \tilde{\rho}_\lambda(\mathrm{d}\theta) = \frac{e^{-\lambda r(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda r(\theta')}\pi(\mathrm{d}\theta')}.$$

The first PAC-Bayesian bound in this context is in our paper with Wintenberger [A8], we state here a slightly more general version that can be found in our more recent papers with Li [A15] and Wintenberger and Li [A19].

Theorem 6.5 (Theorem 1 page 26 in [A15]). *Assume that the time series* (X_t) *satisfies assumption* ThetaDep($\mathcal{C}$). *Let us put* $\kappa = \kappa(K, L, c_X, \mathcal{C}) := K(1+L)(c_X + \mathcal{C})/\sqrt{2}$. *Then, for any* $\lambda > 0$, *for any* $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\tilde{\theta}_\lambda\right) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{2\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right]\right) \geq 1-\varepsilon$$

where $\mathcal{K}$ *stands for the Kullback divergence.*

We remind that $\mathcal{K}$ is given by $\mathcal{K}(\rho,\pi) = \int \log\left[\mathrm{d}\rho/\mathrm{d}\pi(\theta)\right]\rho(\mathrm{d}\theta)$ if $\rho \ll \pi$ and $+\infty$ otherwise. The proof of this theorem follows the general guidelines described in Section 3, however, we replace Hoeffding's inequality by its extention for dependent observations by Rio [178].

Depending on the parameter set Θ, we can use different priors. For example, when Θ is finite, we have the following corollary.

Corollary 6.6 (Theorem 1 page 8 in [A19]). *Assume that* $\mathrm{card}(\Theta) < \infty$ *and that all the assumptions of Theorem 6.5 are satisfied. Let* π *be the uniform probability distribution on* Θ. *Then for any* $\lambda > 0$, $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\tilde{\theta}_\lambda\right) \leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{2\log\left(\frac{2M}{\varepsilon}\right)}{\lambda}\right) \geq 1-\varepsilon.$$

Note that if we take $\lambda = \sqrt{n\log(M)}(1-k/n)/\kappa$, we obtain

$$R\left(\tilde{\theta}_\lambda\right) \leq R(\overline{\theta}) + \frac{4\kappa}{\left(1-\frac{k}{n}\right)}\sqrt{\frac{\log(M)}{n}} + \frac{2\kappa\log\left(\frac{2}{\varepsilon}\right)}{\left(1-\frac{k}{n}\right)\sqrt{n\log(M)}}.$$

When ℓ is the absolute loss, the rate $\sqrt{\log(M)/n}$ is known to be optimal in the i.i.d. case : Theorem 8.3 page 1618 in the paper by Audibert [11]. So this bound cannot be improved. In practice, however, this value of λ is not known to the statistician : it depends on $\kappa = \kappa(K, L, c_X, \mathcal{C})$ and the weak dependence constant $\mathcal{C}$ is not observable. We overcome this difficulty in the most recent version of this work, where we prove that the ERM estimator (Empirical Risk Minimizer) satisfies the same result than $\tilde{\theta}_\lambda$

and does not depend on $\mathcal{C}$, we refer the reader to Theorem 2 page 9 in [A19], where a procedure to estimate L is also described. We also want to mention that in this result, k (the number of delays in the predictors) is fixed but a procedure to chose k on the basis of the observations is described in [A8].

As shown on this example, Theorem 6.5 cannot lead to better rates in n than $1/\sqrt{n}$. So, when we apply this result to an autoregressive linear model with a quadratic loss, it will lead to rates that are generally suboptimal - these bounds are however provided in [A15, A19]. Optimal rates of convergence in this case are obtained at the price of more restrictive assumptions on the time series. We provide here the analogous of Theorem 3.5 page 17 for time series. From now, ℓ is the quadratic loss $\ell(x, x') = (x - x')^2$, $p = 1$ (real-valued time series), we fix a dictionary of $\mathbb{R}^k \to \mathbb{R}$ functions $(\phi_j)_{j=1}^h$. For $\theta \in \mathbb{R}^h$,

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \sum_{j=1}^{h} \theta_j \phi_j(X_{t-1}, \ldots, X_{t-k})$$

and as in Section 3 we put $\Theta = \mathbb{R}^h$, for $I \subset \{1, \ldots, h\}$ we put $\Theta_I = \{\theta \in \mathbb{R}^h : \mathrm{supp}(\theta) = I\}$. We also put $\Theta(L) = \{\theta \in \mathbb{R}^h, \|\theta\|_1 \leq L\}$ and $\Theta_I(L) = \Theta_I \cap \Theta(L)$. We assume that the ϕ_j satisfy a Lipshitz condition so that for any $\theta \in \Theta(L)$, the function f_θ satisfies the condition given by (3) page 28. For the sake of shorteness we don't describe explicitely the prior π here, it is exactly similar to the one used in Section 3. Finally we fix $\overline{\theta}_I$ as a minimizer of R over $\Theta_I(L)$ for each $I \subset \{1, \ldots, h\}$.

Theorem 6.7 (Corollary 1 page 16 in [A19])**.** *Assume that the time series (X_t) satisfies assumption* PhiMix($\mathcal{C}$)*. Let us assume that $\overline{\theta}$, the minimizer of R over Θ, actually belongs to $\Theta(L)$. Then there is a known constant $\mathcal{C}' = \mathcal{C}'(\mathcal{C}, L, c_X)$ such that, for $\lambda = \mathcal{C}'/n$, for any $\varepsilon > 0$,*

$$\mathbb{P}\left(R(\tilde{\theta}_\lambda) - R(\overline{\theta}) \leq 4 \inf_J \left\{ R(\overline{\theta}_J) - R(\overline{\theta}) + \mathcal{C}' \frac{\mathrm{card}(J) \log\left(\frac{(n-k)h}{\mathrm{card}(J)}\right) + \log\left(\frac{2}{\varepsilon}\right)}{n-k} \right\} \right) \geq 1 - \varepsilon.$$

The proof rely on Samson's version of Bernstein inequality for φ-mixing time series [188]. Note that Assumption PhiMix($\mathcal{C}$) only requires the summability of the $[\varphi^X(s)]^{1/2}$ while the assumptions in the aforementioned paper [3] require that $\varphi^X(s)$ is exponentially decreasing in s, that is a stronger assumption.

Since this work, Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer [192] also extended the PAC-Bayesian approach to the study of martingales. Moreover, we want to mention Wintenberger's recent work [222] that aims to generalize Samson's inequality for a wider class of time series using weak transportation inequalities. This opens the path for interesting generalizations of Theorem 6.7.

Part III

Extensions : beyond linear models

Summary

In this part, the objective is to provide oracle inequalities in the spirit of Theorem 3.5 (page 17) for models that go beyond the linear model. In Sections 7 and 8, we extend Theorem 3.5 to non-parametric models : the single-index model and the additive non-parametric model, thanks to PAC-Bayesian bounds. We then turn to rather different models. In Section 9, we prove an oracle inequality in a matrix estimation problem motivated by quantum physics. We use penalized estimators in this case. We focus on matrix estimation in a general setting in Section 10 and establish the first PAC-Bayesian bound in this context. Finally, in Section 11 we provide a PAC-Bayesian bound in a general model selection problem.

7 Sparse single-index

In many applications, linear predictors as studied in Part I do not lead to good predictions, and non-parametric families of predictors are prefered. However, if we consider as a set of predictors all functions $\mathbb{R}^p \to \mathbb{R}$ with a given regularity s, the rate of convergence for the quadratic loss is usually given by $n^{-2s/(2s+p)}$. When p is large, the convergence is very slow, this phenomenon is usually refered as *the curse of dimensionality*. This motiated the introduction of the so-called single index model in econometrics. This corresponds to predictors under the form $f_\theta(x) = f_{(g,\beta)}(x) = g(\beta \cdot x)$ where $\beta \in \mathbb{R}^p$ and $g : \mathbb{R} \to \mathbb{R}$. In this case, the rate for the estimation of an s-smooth function g is $n^{-2s/(2s+1)}$, the estimation of β is meant to be "easier" and we avoid the curse of dimensionality. We refer the reader to the monographs by McCullagh and Nedler [158] and Horowitz [119] and the paper by Härdle, Hall and Ichimura [109] for an introduction. See also the papers by Delecroix, Hristache and Patilea [76], Dalalyan, Judistki and Spokoiny [68] or Lopez [149] for more recent advances. Gaiffas and Lecué [92] also studied the single-index model through PAC-Bayesian methods.

The motivation of this section is the following : when p is very large (say $p > n$), the estimation of β itself is a problem. In this case, it does not make sense to assume that the estimation of β is easier than the estimation of g - nor that the leading term in the rate of convergence is the one corresponding to the estimation of g. As argued in Part I, we cannot estimate β properly without any additional assumption, such as sparsity or approximate sparsity. So we introduceded with Biau in [A11] the *sparse single-index model*.

We assume that we observe n pairs (X_i, Y_i) on some space $(\Omega, \mathcal{A}, \mathbb{P})$. Under $\mathbb{P}$, the pairs are i.i.d. with common distribution P, $X_i \in [-1, 1]^p$,

$$Y_i = \overline{g}(\overline{\beta} \cdot X_i) + W_i$$

with $\mathbb{E}(W_i|X_i) = 0$, $\overline{g}$ is a measurable bounded function $\mathbb{R} \to [-C, C]$ for some $C \geq 1$ and $\overline{\beta} \in \mathbb{R}^p$. Note that the model is not identifiable. While this is not a major problem for prediction purpose, it is more convenient to deal with identifiable models in order to define a prior on the parameters in this context. So we impose the following restriction : $\overline{\beta} \in \mathcal{S}^p_{1,+} = \{\beta \in \mathbb{R}^p, \|\beta\|_1 = 1$ and the first nonzero coordinate in β is positive$\}$. Let us also define, for any $I \subset \{1, \dots, p\}$, $\mathcal{S}^p_{1,+}(I) = \{\beta \in \mathcal{S}^p_{1,+}, \mathrm{supp}(\beta) = I\}$.

We put $\overline{\theta} = (\overline{g}, \overline{\beta})$ and for any $\theta = (g, \beta)$ where g is some measurable function $\mathbb{R} \to \mathbb{R}$ and $\beta \in \mathcal{S}^p_{1,+}$ we put $R(\theta) = R(g, \beta) = \mathbb{E}_{(X,Y)\sim P}[(Y - g(\beta \cdot X))^2]$ and $r(\theta) = r(g, \beta) =$

$(1/n)\sum_{i=1}^{n}[Y_i - g(\beta \cdot X_i)]^2$.

The definition of our estimator $\check{\theta}_\lambda$ is quite cumbersome. The reader may want to skip in a first time the details of the construction and proceed directly to Theorem 7.5. We define a prior $\nu(\mathrm{d}g)$ for the function g, a prior $\mu(\mathrm{d}\beta)$ for the parameter β and finally define the prior on the pair $\theta = (g, \beta)$ as $\pi(\mathrm{d}\theta) = \pi(\mathrm{d}(f, \beta)) = \nu(\mathrm{d}g)\mu(\mathrm{d}\beta)$.

Definition 7.1. *Let (ϕ_j) be the non-normalized Fourier basis $\phi_{2j}(t) = \cos(\pi j t)$ and $\phi_{2j+1}(t) = \sin(\pi j t)$. For any $M \in \mathbb{N}$ and $\Lambda > 0$ let $\mathcal{B}_M(\Lambda)$ be the set*

$$\mathcal{B}_M(\Lambda) = \left\{ (\beta_1, \dots, \beta_M) \in \mathbb{R}^M : \sum_{j=1}^{M} j|\beta_j| \leq \Lambda \text{ and } \beta_M \neq 0 \right\}.$$

Let $\mathcal{F}_M(\Lambda)$ be a set of functions defined as the image of $\mathcal{B}_M(\Lambda)$ through the map

$$\Phi_M : (\beta_1, \dots, \beta_M) \mapsto \sum_{j=1}^{M} \beta_j \varphi_j.$$

Finally, we define $\nu_M(\mathrm{d}g)$ on the set $\mathcal{F}_M(C+1)$ as the image of the uniform probability measure on $\mathcal{B}_M(C+1)$ induced by the map Φ_M, and take

$$\nu(\mathrm{d}g) = \frac{\sum_{M=1}^{n} 10^{-M} \nu_M(\mathrm{d}g)}{1 - (\frac{1}{10})^n}.$$

Note that the idea beyond the prior ν is that Sobolev spaces are well approximated by $\mathcal{F}_M(\Lambda)$ as M grows.

Definition 7.2. *We put*

$$\mu(\mathrm{d}\beta) = \frac{\sum_{i=1}^{p} 10^{-i} \binom{p}{i}^{-1} \sum_{I \subset \{1,\dots,p\}, \mathrm{card}(I)=i} \mu_I(\mathrm{d}\beta)}{1 - (\frac{1}{10})^p},$$

where μ_I is the uniform probability measure on the set $\mathcal{S}^p_{1,+}(I)$.

Note the similarity with the sparsity inducing prior in Section 3.

Definition 7.3. *We put $\pi(\mathrm{d}\theta) = \pi(\mathrm{d}(f, \beta)) = \nu(\mathrm{d}g)\mu(\mathrm{d}\beta)$, and we define as previously, for any $\lambda > 0$, the probability distribution $\tilde{\rho}_\lambda(\mathrm{d}\theta) \propto \pi(\mathrm{d}\theta)\exp[-\lambda r(\theta)]$. This time, the estimator $\check{\theta}_\lambda$ is simply drawn randomly from $\hat{\rho}_\lambda$.*

Actually, it would be possible to define an aggregated predictor in this case :

$$\hat{f}_\lambda(x) = \int g(\beta.x)\hat{\rho}_\lambda(\mathrm{d}(g, \beta))$$

but note that it would not generally hold that $\hat{f}_\lambda(x) = \hat{g}(\hat{\beta} \cdot x)$ for some estimators $\hat{g}$ and $\hat{\beta}$. The predictor $\hat{f}_\lambda$ would satisfy a result exactly to Theorem 7.5 below but, for the sake of simplicity, we only state this result for the randomized estimator $\check{\theta}_\lambda$.

Definition 7.4. *We put, for any $I \subset \{1,\dots,p\}$ and $M \in \{1,\dots,n\}$,*

$$\overline{\theta}_{I,M} = \arg \min_{\theta \in \mathcal{F}_M(C)\times \mathcal{S}^p_{1,+}(I)} R(\theta).$$

Theorem 7.5 (Theorem 2 page 249 in [A11])**.** *Let us assume that the W_i are sub-exponentials with parameters (σ,ξ) (Definition 3.4 page 17). Set $w = 8(2C+1)\max[\xi, 2C+1]$ and $\lambda = n/\{w + 2\left[(2C+1)^2 + 4\sigma^2\right]\}$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\Bigg\{ R(\check{\theta}_\lambda) - R(\overline{\theta}) \leq \Xi \inf_{I\subset\{1,\dots,p\}} \inf_{1\leq M\leq n} \Bigg\{ R(\overline{\theta}_{I,M}) - R(\overline{\theta}) + \frac{M\log(Cn) + \mathrm{card}(I)\log(pn) + \log\left(\frac{2}{\varepsilon}\right)}{n} \Bigg\} \Bigg\} \geq 1-\varepsilon,$$

where $\Xi = \Xi(\xi, C, \sigma) > 0$.

Note that in this theorem, the result holds with large probability on the drawing of the sample $((X_i,Y_i))_{i=1}^n$ *and on the drawing of the estimator* $\check{\theta}_\lambda \sim \tilde{\rho}_\lambda$. We explicit the consequences of this oracle inequality when $\overline{g}$ actually belongs so a Sobolev ellipsoid.

Definition 7.6 (Sobolev ellipsoid)**.** *For $s > 0$ and $\mathcal{D} > 0$,*

$$\mathcal{W}(s,\mathcal{D}) = \left\{ f \in L_2([-1,1]) : f = \sum_{j=1}^\infty \beta_j \varphi_j \text{ and } \sum_{j=1}^\infty j^{2s}\beta_j^2 \leq \mathcal{D} \right\}.$$

Theorem 7.7 (Corollary 4 page 250 in [A11])**.** *Under the assumptions of Theorem 7.5, and under the following additional assumptions :*

1. *the random variable $\overline{\theta}.X_1$ has a probability density on $[-1,1]$ bounded above by a positive constant B,*
2. *$\overline{g} \in \mathcal{W}\left(s, \frac{6C^2}{\pi^2}\right)$ for some $s \geq 2$, for some unknown s,*

we have, for any $\varepsilon > 0$,

$$\mathbb{P}\left(R(\check{\theta}_\lambda) - R(\overline{\theta}) \leq \Xi' \left\{ \left(\frac{\log(Cn)}{n}\right)^{\frac{2s}{2s+1}} + \frac{\|\overline{\beta}\|_0 \log(pn)}{n} + \frac{\log\left(\frac{2}{\varepsilon}\right)}{n} \right\} \right) \geq 1-\varepsilon$$

where $\Xi' = \Xi'(\sigma,\xi,C,B,s) > 0$.

The first term in the right-hand side is the, up to a $\log(n)$ term, the minimax rate of estimation in $W(s, 6C^2/\pi^2)$ (we refer the reader to the monograph by Tsybakov [209] and the references therein for the lower bounds). Note that this result is adaptive in s in the sense that the estimator (including the parameter λ) does not depend on s. The second term is, here again up to a $\log(n)$ term, the optimal rate of estimation of $\overline{\beta}$. So, in general this upper bound is optimal. See [A11] for a more detailed discussion.

Also, in this paper, we proposed a Monte-Carlo method to draw $\check{\theta}_\lambda$ from $\tilde{\rho}_\lambda$. Here again, our algorithm is based on Green's RJMCMC algorithm [102]. We refer the reader to the paper for an extensive simulation study as well as tests on real data. A related

algorithm was used by Wang [218] to select variables in the single-index model (without theoretical study).

Since the publication of this work, Lepski and Serdyukova [144] proposed an alternative method for adaptation in the single-index model, they reach the minimax rate $n^{2s/(2s+1)}$ (without the $\log(n)$ term). However, their method is not designed for sparse $\overline{\beta}$, it means that they pay the price p/n instead of $\|\overline{\beta}\|_0 \log(pn)/n$ for the estimation of $\overline{\beta}$. Finally, Wang, Xu and Zhu [219] and Zhang, Wang, Yu and Gai [231] proposed alternative methods for sparse single-index model based on penalization.

8 Additive model

Another popular model useful to avoid the curse of dimensionality is the additive non-parametric model. Here, for $x \in \mathbb{R}^p$, the predictor is given by $f_1(x_1) + \cdots + f_p(x_p)$. So, we only have to estimate p functions $\mathbb{R} \to \mathbb{R}$. Classical references on additive models are the paper by Stone [201] and Hastie and Tibshirani [111] and the monograph by the same authors [112].

Here again, for relatively small p, it is easier to estimate p functions $\mathbb{R} \to \mathbb{R}$ than one function $\mathbb{R}^p \to \mathbb{R}$. However, when p is too large, both tasks become impossible. In this case, it makes sense to impose sparsity on the model : namely, most of the functions f_i are (close to) zero. A natural approach is the following : fix a dictionary of functions $(\phi_j)_{j\in S}$, e.g. splines, Fourier basis, wavelets... and expand each of the f_i in this dictionary. This leads to a predictor

$$\sum_{i=1}^{p} \sum_{j\in S} \theta_{i,j} \phi_j(x_i)$$

that is linear in $\theta = (\theta_{i,j})_{i\in\{1,\dots,p\},j\in S}$. So, it makes sense to consider the penalties mentioned in Part I. For example, Yuan and Lin's group LASSO penalty [230] enforces some groups of coordinates $(\theta_{i,j})_{j\in\mathcal{S}}$ for fixed i to be null, it means that we estimate some of the functions f_i by 0. Bach [17] provided hypothesis under which this approach leads to a consistent estimation. Later, oracle inequalities were established by Meier, van de Geer and Bühlmann [159], Ravikumar, Lafferty, Liu and Wasserman [173], Koltchinskii and Yuan [132], Suzuki and Sugiyama [203]. However, an assumption on the design (like REP, Definition 2.2 page 15 for the LASSO) is required to prove these inequalities.

Following the ideas introduced in Section 3, PAC-Bayesian methods seems to be a nice alternative to establish oracle inequalities in this context without any REP type assumption. This idea was used simultaneously in two papers : one by Suzuki [202], the other being our joint work with Guedj [A10]. Suzuki's result applies in the fixed design case while our result is valid in the random design case and hence holds on the prevision risk R. On the other hand, Suzuki considers unbounded spaces of functions, while our method requires the set of predictors to be bounded. Suzuki considered a Gaussian prior on a reproducing kernel Hilbert space (RKHS) and used results from van der Vaart and van Zanten [212] to control the Kullback-Leibler divergence between the posterior and the prior. We used an approach closer to the one developped in Section 7 with the Fourier basis.

We assume that we observe n pairs (X_i, Y_i) on some space $(\Omega, \mathcal{A}, \mathbb{P})$. Under $\mathbb{P}$, the pairs are i.i.d. with common distribution P, $X_i \in [-1,1]^p$,

$$Y_i = \psi(X_i) + W_i$$

with $\mathbb{E}(W_i|X_i) = 0$, ψ is a measurable function $\mathbb{R}^p \to [-C, C]$ for some $C \geq 1$. We let (ϕ_j) denote the Fourier basis (Definition 7.1 page 34). For any $m = (m_1, \ldots, m_p) \in \mathbb{N}^p$, and $\mathcal{D} > 0$, we define $\mathcal{B}_m^1(0, \mathcal{D})$ as

$$\mathcal{B}_m^1(0,\mathcal{D}) = \left\{ \theta = (\theta_{j,k})_{j \in \mathrm{supp}(m), k \in \{1,\ldots,m_j\}} \text{ with, for any } j, \sum_{k=1}^{m_j} |\theta_{j,k}| \leq \mathcal{D} \right\}$$

and put, for any $x \in \mathbb{R}^p$, m and $\theta \in \mathcal{B}_m^1(0,\mathcal{D})$,

$$f_\theta(x) = \sum_{j \in \mathrm{supp}(m)} \sum_{k=1}^{m_j} \theta_{j,k} \phi_k(x_j).$$

Finally, as usual, $R(\theta) = \mathbb{E}_{(X,Y)\sim P}[(Y - f_\theta(X))^2]$ and $r(\theta) = (1/n)\sum_{i=1}^n [Y_i - f_\theta(X_i)]^2$, and $\overline{\theta}$ is a minimizer of R over the reunion of all $\mathcal{B}_m^1(0,\mathcal{D})$ for $m \in \mathbb{N}^p$ and $\mathcal{D} > 0$. We considered the following prior.

Definition 8.1 (Priors)**.** *We put, for any $\alpha \in (0, 1/2)$,*

$$\eta_\alpha(m) = \frac{1 - \frac{\alpha}{1-\alpha}}{1 - \left(\frac{\alpha}{1-\alpha}\right)^{p+1}} \frac{\alpha^{\sum_{j=1}^p m_j}}{\binom{p}{S(m)}}.$$

We fix α once and for all, and define π_m as the uniform measure on $\mathcal{B}_m^1(0, C)$. Finally, we put

$$\pi(\mathrm{d}\theta) = \sum_{m \in \mathbb{N}^p} \eta_\alpha(m) \pi_m(\mathrm{d}\theta).$$

Then, as in the previous section, we define $\tilde{\rho}_\lambda(\mathrm{d}\theta) \propto \pi(\mathrm{d}\theta)\exp(-\lambda r(\theta))$ and the estimator $\check{\theta}_\lambda$ drawn from $\tilde{\rho}_\lambda$. Note that an aggregated version $\tilde{\theta}_\lambda$ is also studied in [A10] but, for the sake of simplicity, we only report the results for the randomized estimator.

Theorem 8.2 (Theorem 2.1 page 269 in [A10])**.** *Let us assume that the W_i are sub-exponentials with parameters (σ, ξ). Set $w = 8C\max(\xi, C)$ and $\lambda = n/[2w + 4(\sigma^2 + C^2)]$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\Bigg\{ R(\check{\theta}_\lambda) - R(\overline{\theta}) \leq \Xi \inf_{m \in \mathbb{N}^p} \inf_{\theta \in \mathcal{B}_m^1(0,C)} \Bigg[R(\theta) - R(\overline{\theta}) + \frac{\|m\|_0 \log\left(\frac{p}{\|m\|_0}\right)}{n} + \frac{\log(n)}{n} \sum_{j \in \mathrm{supp}(m)} m_j + \frac{\log\left(\frac{2}{\varepsilon}\right)}{n} \Bigg] \Bigg\} \geq 1 - \varepsilon,$$

where $\Xi = \Xi(\sigma, \xi, C, \alpha > 0)$.

Theorem 8.3 (Theorem 2.2 page 271 in [A10]). *Assume that the assumptions of Theorem 8.2 are satisfied. Assume that the regression function ψ actually satisfies*

$$\psi(x) = \sum_{j \in S^*} \psi_j(x_j)$$

for some $S_* \subset \{1, \dots, p\}$ and that for each j, $\psi_j \in \mathcal{W}(s_j, D_j)$. Finally, assume that when $(X, Y) \sim P$, P has a density with respect to the Lebesgue measure bounded from above by $B > 0$. Then,

$$\mathbb{P}\left\{ R(\check{\theta}_\lambda) - R(\overline{\theta}) \leq \Xi' \left[\sum_{j \in S^*} D_j^{\frac{1}{2s_j+1}} \left(\frac{\log(n)}{2ns_j} \right)^{\frac{2s_j}{2s_j+1}} + \frac{\operatorname{card}(S^*) \log\left(\frac{p}{\operatorname{card}(S^*)} \right)}{n} + \frac{\log\left(\frac{2}{\varepsilon}\right)}{n} \right] \right\} \geq 1 - \varepsilon$$

for some $\Xi' = \Xi'(\sigma, \xi, C, \alpha, B) > 0$.

We still obtain the minimax rate of convergence (up to a $\log(n)$ term). In [A10], we also present a Monte-Carlo algorithm to draw $\check{\theta}_\lambda$, based on Carlin and Chib [50] algorithm. We also provide a detailed simulation study.

Finally, we want to mention a very recent preprint by Abramovich and Lahav [1] in which the authors study the theoretical properties of a Bayesian MAP estimator (Maximum A Posteriori) in the additive nonparametric model with fixed design. The MAP achieves the minimax rate in this setting.

9 Statistical models in quantum physics

Quantum physics recently provided a wide range of new statistical estimation problems. It is not the purpose of this thesis to provide a complete introduction to the field of quantum physics or quantum statistics, we rever the reader to Holevo's monograph for the probabilistic and statistical aspects of quantum physics [118] or to the more recent introductory paper on quantum statistics by Barndorff-Nielsen, Gill and Jupp [19]. In quantum physics, the state of a system is represented by a (complex) matrix ρ with

- $\rho^* = \rho$ (ρ is Hermitian),
- $\operatorname{tr}(\rho) = 1$ (the trace of the matrix is one),
- for any column vector v, $v^* \rho v \geq 0$ (ρ is non-negative).

This matrix ρ is called the density matrix of the state. In many problems of quantum statistics, the objective is to estimate the density matrix ρ of a system on the basis experimental observations.

Depending on the system, the dimension of ρ can be finite or infinite, and can satisfy various additional assumptions. We studied two cases of interest for physicists. First, in a joint paper with Meziani and Peyré [A12] we studied a model of quantum homodyne tomography. In this case ρ is an infinite matrix with a regularity assumption : the coefficients $\rho_{i,j}$ of ρ decay exponentially fast in $i + j$. A review on quantum homodyne tomography can be found in the paper by Artiles, Gill and Guţă [9]. Some rates of

convergence were provided in a paper by Aubry, Butucea and Meziani [10], but the corresponding estimators depended on the rate of decay of the coefficients $\rho_{i,j}$. In [A12], we obtained the same rate of convergence for an adaptive estimator based on a soft-thresholding procedure (see Remark 2.5 page 16 above).

In quantum computing, the system of interest is called a n-qubit and the corresponding density matrix ρ is a $2^n \times 2^n$ matrix. Moreover, physicists are interested in generating systems in *pure states*, corresponding to ρ with $\text{rank}(\rho) = 1$. In our joint work with Butucea, Hebiri, Meziani and Morimae [A13], we developped a penalized estimator for this problem. In this section, we present the results of this paper. First, we introduce the basic notations.

On a n-qubit system, there are 3^n possible experimental measurements and each possible returns a vector in $\{-1, 1\}^n$ (this number, 3^n, might seem a bit arbitrary, but it comes from quantum theory, we refer the reader to [A13] and the references therein). The probability to obtain a given vector $v \in \{-1, +1\}^n$ as an outcome is a function of the measurement and of the density matrix ρ of the system. We use the notation $M = (M_{i,v})_{i\in\{1,\dots,3^n\},v\in\{-1,+1\}^n}$, $M_{i,v}$ is the probability to obtain the outcome v when we perform experiment i. Quantum theory provides a linear function F such that $M = F(\rho)$.

Note that, one of the striking facts in quantum theory is that every measurement changes the state of the observed system. So, once a measurement is performed on a system in state ρ, the system is no longer in state ρ after the measurement. As the physicists want to test the ability of a device to produce systems in a given state of interest ρ_0, they proceed as follow. For each measurement type $i \in \{1, \dots, 3^n\}$, they repeat a given number of times (say m) :

1. use the device to produce a system of n-qubit,
2. perform measurement i it on the system.

From these observations, we can build a natural estimator of M, $\hat{M}$, where each probability $M_{i,v}$ is estimated by the corresponding empirical frequency $\hat{M}_{i,v}$ (note that, once the device exists, it is usually possible to repeat a large number of experiments, $m \geq 1000$ is possible; moreover, we can assume that the different repetitions of the experiment are independent). In order to make things more clear, we provide a toy example.

Example 9.1. *In this example, $n = 2$, so there are 9 possible measurement on the system. Each measurement is performed $m = 1000$ times. We stick to the following notation : ρ_0 is the state of interest, ρ is the actual state produced by the device. So, in theory :*

	probability of outcome			
measurement	$(-1,-1)$	$(-1,+1)$	$(+1,-1)$	$(+1,+1)$
1	0.00	0.43	0.43	0.14
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
9	0.50	0.12	0.22	0.16

$$\Rightarrow F(\rho) = M = \begin{pmatrix} 0.00 & 0.43 & 0.43 & 0.14 \\ \vdots & \vdots & \vdots & \vdots \\ 0.50 & 0.12 & 0.22 & 0.16 \end{pmatrix}$$

but ρ, and as a consequence M, is unknow. However, we observe

	number of outcomes observed			
measurement	$(-1,-1)$	$(-1,+1)$	$(+1,-1)$	$(+1,+1)$
1	0	437	440	123
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
9	584	125	227	164

$$\Rightarrow \hat{M} = \begin{pmatrix} 0.000 & 0.437 & 0.440 & 0.123 \\ \vdots & \vdots & \vdots & \vdots \\ 0.584 & 0.125 & 0.227 & 0.164 \end{pmatrix}.$$

A way to test whether $\rho = \rho_0$ would be to compare $F(\rho_0)$ with $\hat{M}$. However, physicists are not interested in the matrix M, the object that makes sense in physics is ρ, not M. In particular, we already mentioned that physicists are also interested in the question : do we have $\text{rank}(\rho) = 1$?

Guţă, Kypraios and Dryden [106] wrote the likelihood of ρ in this context and proposed as an estimator a maximizer of a version of the likelihood penalized by the rank of ρ. While this works in theory, the maximization is computationally intensive, so we proposed in our paper [A13] an alternative penalized moment method. Our estimator $\hat{\rho}_t$ is given by

$$\hat{\rho}_t = \arg\min_p \left\{ \|\hat{M} - F(p)\|_F^2 + t.\text{rank}(p) \right\}$$

for some parameter $t > 0$, where $\|\cdot\|_F$ is the Frobenius norm on matrices given by $\|A\|_F^2 = \text{tr}(A^*A)$.

Theorem 9.2 (Corollary 4.3 page 8 in [A13]). *There is some constant $C > 0$ such that, when $t \geq C(4/3)^n(n - \log\varepsilon)/m$,*

$$\mathbb{P}\Big(\|\hat{\rho}_t - \rho\|_F^2 \leq Ct\text{rank}(\rho) \Big) \geq 1 - \varepsilon.$$

So, the smaller the rank of ρ is, the easier the estimation task. Moreover, under additional assumptions on the eigenvalues of ρ, it is possible to prove that the probability that $\text{rank}(\hat{\rho}_t) = \text{rank}(\rho)$ is large too (Corollary 4.4 page 8 in [A13]). This is coherent with the results by Gross, Liu, Flammia, Becker and Eisert [105] who proved that, when we actually know than $\text{rank}(\rho) = s < 2^n$, it is not necessary to go through all the 3^n possible type of measurements. In [A13] we also provide an explicit procedure to compute $\hat{\rho}_t$ as well as some simulations and we test our procedure on the experimental data coming from the paper by Barreiro, Müller, Schindler, Nigg, Monz, Chwalla, Hennrich, Roos, Zoller and Blatt [20]. Note that this is the first nonasymptotic bound proved for any estimator in this context, it gives an idea of how many measurements m are necessary in order to ensure an accurate recovery of ρ. However, we don't know yet whether this rate is optimal, the lower bounds will be the object of a future work.

The proof of Theorem 9.2 relies essentially on Corollary 6 page 1290 in Bunea, She and Wegkamp [42], who proposed a general procedure for the estimation of low-rank matrices in a general context. However, in order to prove that the assumptions of this

corollary are satisfied, a deviation inequality for random matrices is required. Such inequalities are usually refered as *non-commutative* deviation inequalities. A first generalization of Bernstein's inequality for random matrices was proved by Ahlswede and Winter [4], improvements can be found in Oliveira [170], Gross [104], Recht [174]. An Hoeffding's inequality for matrices was first proved by Christofides and Markström [60]. We also refer the reader to the very nice and comprehensive survey papers (that also contain new results) by Tropp [207] and Vershynin [215] and to Section 3.2 page 252 in Tao's monograph [204]. Here, we actually used the version of non-commutative Hoeffding's inequality of [207].

10 Bayesian low-rank matrix estimation

Several statistical problems involve the estimation of large but potentially low-rank matrices. Beyond the classical PCA, there was a recent interest in matrix completion and in the the reduced-rank regression model. Penalized estimators appeared as a computationally efficient ways to perform optimal matrix completion, we refer the reader to the striking papers by by Candès and Tao [48], Candés and Plan [46], Candés and Recht [47] and Gross [104]. Regarding reduced-rank regression, we refer the reader to the monograph by Reinsel and Velu [175] for an introduction and to the paper by Bunea, She and Wegkamp [42] for recent results based on penalized estimators. More recently, the so-called trace regression model was introduced as a general model that would include linear regression, reduced-rank regression and matrix completion as special cases. Penalized estimators for this general model are studied in Koltchinskii [130], Rohde and Tsybakov [183], Klopp [127], Koltchinskii, Lounici and Tsybakov [131].

However, little has been done on Bayesian estimators in this context. Special cases of reduced-rank regression are used in econometrics and have been estimated by Bayesian estimators, we mention the nice survey by Geweke [98] and the references therein. Bayesian model selection in order to estimate the rank of the involved matrix was done by Kleibergen and Paap [126], Corander and Villani [64] proved the model selection consistency of these procedures. However, the recent applications of matrix completion imposed the additional constraint of computational efficiency : for example, the Netflix challenge proposed a database containing 100,480,507 ratings that 480,189 users gave to 17,770 movies, the objective being to reconstruct a matrix with 8,532,958,530 entries, see Bennett and Lanning [29] for a more complete description. A few Bayesian methods taking this constraint into account were proposed : Yu, Tresp and Schwaighofer [229], Lim and Teh [148], Salakhutdinov and Mnih [185, 186], Lawrence and Urtasun [134], Yu, Lafferty, Zhu and Gong [228], Zhou, Wang, Chen, Paisley and Carin [237], Babacan, Luessi, Molina and Katsaggelos [15]. In all these papers, the authors used Monte-Carlo methods or Variational Bayes (VB) methods to compute their estimators (see e.g. Beal [24] for an introduction to VB). A reasonable computational efficiency was reached : in some of these papers, the authors obtained good performances on well known large datasets such as Netflix and MovieLens. Finally, we mention that Aoyagi and Watanabe provided general conditions for consistency of low-rank matrix estimation [6, 7]. However, their method is only valid for priors on bounded spaces, while in all the papers mentioned previously, computational efficiency is reached thanks to Gaussian priors.

One of my current research projects is to investigate the properties of Bayesian es-

timators for the estimation of low-rank matrices in these various problems. As a first step, in my paper [A16], I proved that the prior in [15] lead to an estimator that satisfies an oracle inequality with the optimal rate of convergence, up to log terms, in the reduced-rank regression model with fixed design (note that the priors in the other papers mentioned above are very similar).

We assume that the latent probability space is $(\Omega, \mathcal{A}, \mathbb{P})$ and that we observe an $\ell \times m$ random matrix Y and an $\ell \times p$ deterministic matrix X with

$$Y = XB + \mathcal{E}$$

where B is some unknown $p \times m$ deterministic matrix and $\mathcal{E}$ is some unknown $\ell \times p$ random matrix. We will make one of the following assumptions on the noise :

- **Assumption (A1) :** the entries $\mathcal{E}_{i,j}$ of the matrix $\mathcal{E}$ are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$, and we know an upper bound s^2 for σ^2.
- **Assumption (A2) :** the entries of $\mathcal{E}$ are i.i.d. according to any distribution supported by the compact interval $[-\zeta, \zeta]$ with a density f w.r.t. the Lebesgue measure and $f(x) \geq f_{\min} > 0$, and we know an upper bound $s^2 \geq \mathbb{E}(|\mathcal{E}_{1,1}|)/(2f_{\min})$.

Note that **(A1)** and **(A2)** are special case of Assumption **A** used by Dalalyan and Tsybakov (page 99 in in [70]), our result would actually hold under this more general condition.

We now describe the prior. First, we replace the matrix parameter B by two matrices M and N, with $B = MN^T$. Here $k \leq \min(p, m)$ is a fixed integer, M is $p \times k$, N is $m \times k$, and then

$$\pi(\mathrm{d}(M, N)|\Gamma) \propto \exp\left[-\frac{1}{2}\left(\mathrm{Tr}(M^T\Gamma^{-1}M) + \mathrm{Tr}(N^T\Gamma^{-1}N)\right)\right]\mathrm{d}(M, N)$$

where $\mathrm{d}(M, N)$ stands for the product of the Lebesgue measure on each component of M and N, and Γ is some random matrix

$$\Gamma = \begin{pmatrix} \gamma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \gamma_k \end{pmatrix},$$

the γ_j being i.i.d. and $1/\gamma_j \sim \mathrm{Gamma}(a, b)$. So we have :

$$\pi(\mathrm{d}(M, N)) = \int_\Gamma \pi(\mathrm{d}(M, N)|\Gamma)\pi_\Gamma(\mathrm{d}\Gamma)$$

where

$$\pi_\Gamma(\mathrm{d}\Gamma) = \frac{b^{ka}}{\Gamma(a)^k}\prod_{j=1}^{k}\left\{\gamma_j^{-a-1}\exp\left(-\frac{b}{\gamma_j}\right)\right\}\mathrm{d}\gamma_1 \dots \mathrm{d}\gamma_k.$$

We define the Gibbs estimator by

$$\tilde{B}_\lambda = \int MN^T\tilde{\rho}_\lambda(\mathrm{d}(M, N))$$

where $\tilde{\rho}_\lambda$ is the probability distribution given by

$$\tilde{\rho}_\lambda(\mathrm{d}(M, N)) \propto \exp\left(-\lambda\|Y - XMN^T\|_F^2\right)\pi(\mathrm{d}(M, N)).$$

For $J \subset \{1,\dots,k\}$ let $\mathcal{M}_J$ denote the set of $p \times k$ matrices M such that every column M_j of M corresponding to an index $j \notin J$ is null. Similarly, $\mathcal{N}_J$ is the set of $m \times k$ matrices N such that every column N_j is null when $j \notin J$.

Theorem 10.1 (Theorem 1 page 304 in [A16]). *Assume that either* **(A1)** *or* **(A2)** *is satisfied. Let us put* $a = 1$ *and* $b = \frac{s^2}{2\ell p k^2(m^2+p^2)}$ *in the prior* π_Γ. *For* $\lambda = 1/(4s^2)$,

$$\mathbb{E}\left(\|X\tilde{B}_\lambda - XB\|_F^2\right) \leq \inf_{J\subset\{1,\dots,k\}} \inf_{M\in\mathcal{M}_J} \inf_{N\in\mathcal{N}_J} \left\{ \|X(MN^T - B)\|_F^2 + 6s^2(m+p)|J| \log\left(\frac{1.34\ell p}{s^2}\right) + 8s^2 k \log\left(\frac{22.17\ell p k^2(m^2+p^2)}{s^2}\right) + \frac{2s^2\|X\|_F^2}{\ell p}\left\{\|N\|_F^2 + \|M\|_F^2 + \frac{2s^2}{\ell p} + 16s^2\right\} + 8s^2\left(\|N\|_F^2 + \|M\|_F^2 + \log(2)\right)\right\}.$$

Assume that all the entries of X satisfy $|X_{i,j}| \leq C$ for some $C > 0$, then $\|X\|_F^2/(\ell p) \leq C^2$. Also assume that $\mathrm{rank}(B) = k_0$ and that $B = MN^T$ with $M_{k_0+1} = \cdots = M_k = 0$ and $N_{k_0+1} = \cdots = N_k = 0$ and $|N_{i,j}|, |M_{i,j}| \leq c$. We get

$$\mathbb{E}\left(\|X\tilde{B}_\lambda - XB\|_F^2\right) \leq 50s^2(m+p)k_0\left\{\log(\ell\max(p,m)) + \log\left[\max\left(\frac{1}{s^2},1\right)\right] + 1 + C^2(1+c^2+s^2)\right\}.$$

When $\mathrm{rank}(X) = p$, we recover the same upper bound as in Bunea, She and Wegkamp [42], up to a $\log(\ell\max(p,m))$ term. This rate (without the log) is known to be optimal, see remark (ii) page 1293 in [42] and Rohde and Tsybakov [183]. However, the terms $\|M\|_F^2$ and $\|N\|_F^2$ can lead to suboptimal rates in less classical asymptotics where $\|B\|_F$ would grow with the sample size ℓ. But, up to our knowledge, these terms cannot be avoided when using a Gaussian prior.

In the conclusion of this thesis, we will present some of our works in project in order to get rif of these terms. To find a prior that would lead simultaneously to a computationally feasible estimator and to an oracle inequality without the terms $\|M\|_F^2$ and $\|N\|_F^2$ is one of my obectives.

11 Models with controled complexity

This last section describes the results of the paper [A4], in which a very general model selection procedure for regression is proposed. This procedure is largely inspired by the procedure proposed in Subsection 2.2 page 68 in Catoni's monograph [55]. It relies on two steps :

1. Gibbs estimators are defined in each submodel,

2. one of them is selected through a selection procedure in the spirit of the celebrated Lepki's method, see Lepski [141, 142, 143] and Birgé [34].

The strength of this method is that it is simultaneously adaptive with respect to the dimension of the models and to the parameter in the so-called margin assumption introduced by Mammen and Tsbakov [153]. On the other hand, this method suffers the same drawback as ℓ_0-penalized methods introduced in Part I : it is not computationally feasible when the number of models is too large.

On a space $(\Omega, \mathcal{A}, \mathbb{P})$, we observe n i.i.d. pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ with $(X_i, Y_i) \sim P$, where X_i belongs to a set $\mathcal{X}$, Y_i to a set $\mathcal{Y}$. We consider as in the other sections a family of prediction functions $\{f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$ and a loss function $\ell : \mathcal{Y}^2 \to \mathcal{R}_+$. As mentioned in the introduction, ℓ can be the 0-1 loss if $\mathcal{Y}$ is a finite set, or a convex loss function such as the least square loss. For the sake of simplicity, we assume that $\mathbb{P}(\ell(f_\theta(X), Y) \leq C) = 1$ for some $C > 0$. For more general situations, a risk truncation procedure was proposed in [A4], at the cost of a loss in the rates of convergence. This procedure was recently improved by Audibert and Catoni [13, 56], we will give more comments on this below. We assume that we have a partition of Θ : $(\Theta_i)_{i \in I}$ where I is finite. We fix a prior π_i on each submodel Θ_i (equiped with a suitable σ-algebra) and weights $\mu_i > 0$ such that $\sum_{i \in I} \mu_i = 1$. As in the previous sections, we put

$$R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X), Y)], \, \overline{\theta} \in \arg\min_{\theta \in \Theta} R(\theta) \text{ and } r(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i).$$

Similarly, we define $\overline{\theta}_i$ as a minimizer of R over Θ_i. We define, for any $\lambda > 0$, the measure $\tilde{\rho}_{i,\lambda}$ by

$$\tilde{\rho}_{i,\lambda}(\mathrm{d}\theta) \propto \exp(-\lambda r(\theta)) \pi_i(\mathrm{d}\theta)$$

for $\theta \in \Theta_i$. Let us put $\Lambda = \{2^0, 2^1, \dots, 2^{\left\lfloor \frac{\log(n)}{\log(2)} \right\rfloor}\}$. The estimation procedure is as follows : for each $(i, \lambda) \in I \times \Lambda$, we draw

$$\check{\theta}_{i,\lambda} \sim \tilde{\rho}_{i,\lambda}.$$

The next step is to choose a pair $(\hat{\imath}, \hat{\lambda})$. This choice relies on two main facts (for a fixed confidence level $\varepsilon > 0$) :

– the existence of empirical bounds $\tilde{B}_\varepsilon((i,\lambda),(i',\lambda'))$, i.e. bounds that depend on the observations through empirical risk r, and not on the unknown P, such that

$$\mathbb{P}\left(\forall ((i,\lambda),(i',\lambda')) \in (I \times \Lambda)^2, R(\check{\theta}_{i,\lambda}) - R(\check{\theta}_{i',\lambda'}) \leq \tilde{B}_\varepsilon((i,\lambda),(i',\lambda'))\right) \geq 1 - \varepsilon$$

with a sub-additivity structure :

$$\tilde{B}_\varepsilon((i,\lambda),(i',\lambda')) \leq \tilde{B}_\varepsilon((i,\lambda),(i'',\lambda'')) + \tilde{B}_\varepsilon((i'',\lambda''),(i',\lambda'))$$

(the existence of these bounds is given by Theorem 2.5 page 288 and Definition 3.2 page 289 in [A4]),

– the existence of an empirical complexity measure $\mathcal{C}_\varepsilon((i,\lambda))$ (Definition 3.1 page 289 in [A4]).

The definition of these quantities will be provided below for the sake of completeness. We then follow the idea of Lepski's procedure [141, 142, 143]. We arrange the pairs $(i,\lambda) \in I \times \Lambda$ by increasing complexity, i.e. we define $M = \mathrm{card}(I)\mathrm{card}(\Lambda)$, and objects

t_i for $i \in \{1, \dots, M\}$ such that $\{t_i, i \in \{1, \dots, M\}\} = I \times \Lambda$ and $\mathcal{C}_\varepsilon(t_1) \leq \mathcal{C}_\varepsilon(t_2) \leq \cdots \leq \mathcal{C}_\varepsilon(t_M)$. For any $k \in \{1, \dots, M\}$ we put

$$s(k) = \inf\{j \in \{1, \dots, M\},\ \tilde{B}_\varepsilon(t_k, t_j) > 0\},$$

by convention $s(k) = 0$ if for any j, $\tilde{B}_\varepsilon(t_k, t_j) \leq 0$. We then put

$$\hat{k} = \min(\arg\max s)$$

and choose $(\hat{i}, \hat{\lambda})$ as the pair such that $(\hat{i}, \hat{\lambda}) = t_{\hat{k}}$, our estimator is finally given by

$$\check{\theta} = \check{\theta}_{\hat{i},\hat{\lambda}}.$$

We now state the oracle inequality satisfied by $\check{\theta}$ under two assumptions : Catoni's complexity assumption [55] and Mammen and Tsybakov's margin assumption [153].

Definition 11.1 (Margin assumption [153])**.** *Let us put, for any $(\theta, \theta') \in \Theta^2$,*

$$V(\theta, \theta') = \mathbb{E}_{(X,Y)\sim P}\left\{\left[\ell(f_\theta(X), Y) - \ell(f_{\theta'}(X), Y)\right]^2\right\}.$$

We say that Mammen and Tsybakov's margin assumption is satisfied with constants $(\kappa, c) \in [1, +\infty[\times\mathbb{R}_+^*$ if, for any $\theta \in \Theta$,

$$V(\theta, \overline{\theta}) \leq c\left[R(\theta) - R(\overline{\theta})\right]^\kappa.$$

In the case of the 0-1 loss, this assumption can be geometrically interpreted as the existence of a margin between classes, however, it is also meaningfull in other contexts, see the discussions in the paper by Lecué [135] among others.

Definition 11.2 (Complexity assumption [48])**.** *We say that Catoni's complexity assumption is satisfied for the partition $(\Theta_i)_{i\in I}$ with positive complexities $(d_i)_{i\in I}$ if, for any $i \in I$,*

$$\sup_{\xi\in\mathbb{R}}\left\{\xi\left[\int_{\Theta_i} R(\theta)\pi^i_{\exp(-\xi R)}(\mathrm{d}\theta) - R\left(\overline{\theta}_i\right)\right]\right\} \leq d_i$$

where $\pi^i_{\exp(-\xi R)}$ is the probability distribution defined by

$$\pi^i_{\exp(-\xi R)}(\mathrm{d}\theta) \propto \exp[-\xi R(\theta)]\pi_i(\mathrm{d}\theta)$$

for $\theta \in \Theta_i$.

This assumption is discussed in [55] and in our Ph.D. thesis [A17]. It is shown that, in many cases, when the Θ_i's are compact sets in finite dimensional space with respective dimension $\dim(\Theta_i)$, then we can take $d_i = \mathcal{C}\dim(\Theta_i)$ for some constant $\mathcal{C}$.

Theorem 11.3 (Theorem 3.2 page 290 in [A4])**.** *Let us assume the assumptions given by Definitions 11.1 and 11.2 are satisfied. Then there is a constant $\mathcal{C} = \mathcal{C}(\kappa, c, C)$ such that*

$$\mathbb{P}\Bigg\{R(\check{\theta}) - R(\overline{\theta}) \leq \inf_{i\in I}\Bigg[R\big(\overline{\theta}_i\big) - R(\overline{\theta}) + \mathcal{C}\max\Bigg\{\left(\frac{\left[R\left(\overline{\theta}_i\right) - R\left(\overline{\theta}\right)\right]^{\frac{1}{\kappa}}\left(d_i + \log\frac{1+\log_2(n)}{\varepsilon\mu_i}\right)}{n}\right)^{\frac{1}{2}}, \left(\frac{d_i + \log\frac{1+\log_2(n)}{\varepsilon\mu(i)}}{n}\right)^{\frac{\kappa}{2\kappa-1}}\Bigg\}\Bigg]\Bigg\} \geq 1 - \varepsilon.$$

This rate of convergence is known to be optimal (up to the $\log\log_2(n)$ term !), we refer the reader to Theorem 3.1 and 3.2 page 44 in Lecué's Ph.D. thesis [136], see also his paper [135].

For the sake of completeness and aesthetic considerations, we provide explicit formulas for $\tilde{B}_\varepsilon((i,\lambda),(i',\lambda'))$ and $\mathcal{C}_\varepsilon((i,\lambda))$. First, we define a natural estimator of $V(\theta,\theta')$:

$$v(\theta,\theta') = \frac{1}{n}\sum_{i=1}^{n}\left[\ell(f_\theta(X_i),Y_i) - \ell(f_{\theta'}(X_i),Y_i)\right]^2.$$

Fix some parameter $\zeta > 0$. We can now give the definition of the complexity measure

$$\mathcal{C}_\varepsilon(i,\lambda) = \inf_{\gamma\in[\zeta\lambda,\infty[}\left\{\frac{1}{1-\frac{\lambda}{\gamma}}\log\int_{\Theta_i}\exp\left[\frac{\lambda\gamma}{2n}v(\check{\theta}_{i,\lambda},\theta)\right]\tilde{\rho}_{\lambda,i}(\mathrm{d}\theta) + \left(1+\frac{1}{\zeta-1}+\frac{\lambda}{\gamma-\lambda}\right)\log\left(\frac{3}{\varepsilon\mu_i\mathrm{card}(\Lambda)^2}\right)\right\}.$$

We define, for any parameter $\alpha > 0$,

$$\Phi_\alpha(t) = \frac{\log(1-\alpha t)}{\alpha}, \tag{4}$$

note that this function is invertible and that for any $u \in \mathbb{R}$,

$$\Phi_\alpha^{-1}(u) = \frac{1-\exp(-\alpha u)}{\alpha}.$$

Actually, we will only use Φ_α^{-1}, but the role of the function Φ_α is important in the unbounded case, see [A4]. Then we put :

$$\begin{aligned}B_\varepsilon((i,\lambda),(i',\lambda')) = \inf_{\xi>0}\inf_{\gamma>\lambda}\inf_{\gamma'>\lambda'}\Phi_{\frac{\xi}{n}}^{-1}\Bigg\{& r(\check{\theta}_{i,\lambda}) - r(\check{\theta}_{i',\lambda'}) + \frac{\xi}{2n}v(\check{\theta}_{i,\lambda},\check{\theta}_{i',\lambda'}) \\ &+ \frac{1}{\xi}\Bigg[\frac{1}{1-\frac{\lambda}{\gamma}}\log\int_{\Theta_i}\exp\left[\frac{\lambda\gamma}{2n}v(\check{\theta}_{i,\lambda},\theta)\right]\tilde{\rho}_{\lambda,i}(\mathrm{d}\theta) \\ &+ \frac{1}{1-\frac{\lambda'}{\gamma'}}\log\int_{\Theta_{i'}}\exp\left[\frac{\lambda\gamma}{2n}v(\check{\theta}_{i',\lambda'},\theta)\right]\tilde{\rho}_{\lambda',i'}(\mathrm{d}\theta) \\ &+ \left(1+\frac{\lambda}{\gamma-\lambda}+\frac{\lambda'}{\gamma'-\lambda'}\right)\log\left(\frac{3}{\varepsilon\mu_i\mu_{i'}\mathrm{card}(\Lambda)^4}\right)\Bigg]\Bigg\}.\end{aligned}$$

It is possible to prove that this quantity satisfies almost all the required properties (Theorem 2.5 page 288 in [A4]), the only trouble being that it is not subadditive. This is why we define :

$$\tilde{B}_\varepsilon((i,\lambda),(i',\lambda')) = \inf\Bigg\{\sum_{k=1}^{h}B_\varepsilon((i_{k-1},\lambda_{k-1}),(i_k,\lambda_k)), \quad h\geq 1,(i_0,\ldots,i_h)\in I^{h+1},(\lambda_0,\ldots,\lambda_h)\in\Lambda^{h+1}\Bigg\}.$$

Conclusion and future works

We hope that the reader found these explanations useful and that he/she is convinced 1. that PAC-Bayesian bounds are useful tools to prove powerful oracle inequalities and 2. that aggregated or Bayesian-type estimators are a valuable alternative to penalized estimators, depending on the context and on the objective of the statistician.

I will conclude this thesis by an overview of a few open issues.

1. Regardind Part II, it would be nice to see to what extent statistical learning can be extended to a wider set of time series. A recent preprint by Wintenberger [222] extends Samson's version of Bernstein inequality [188] to a more general context thanks to weak transportation inequalities. Moreover, in some situations, the rates obtained are different to the ones in the i.i.d. case, so it seems necessary to study lower bounds in this case too.
2. From the algorithmic perspective, to implement the Gibbs estimator through Green's RJMCMC method [102] revealed a successful strategy for reasonably large dimension p, but there is room for improvement for larger p. Carlin and Chib's algorithm [50] used in [A10] was an improvement, it would be interesting to see if more sophisticated Monte-Carlo algorithms could help, e.g. Pandolfi, Bartolucci and Friel's [171] version of the multiple choice Metropolis-Hastings algorithm could help. In order to deal with very large datasets, such as NetFlix or MovieLens, computational efficiency is a crucial issue. The ℓ_1-penalized methods studied by Candès and different coauthors in the aforementioned papers [49, 46, 47] lead to very efficient algorithms. In order to challenge these methods, it is likely that a naive Monte-Carlo method is not enough. Another interesting alternative to compute Bayesian estimators is the family of Variational Bayes methods (e.g. [24]). I'm currently working on different variant of these methods to be able to process very large datasets. Also, it would be necessary to understand the rate of convergence of these Monte-Carlo algorithms, and probably to include the computational cost in the model selection procedures. To this regard, I would like to point out the very interesting recent preprint by Sanchez-Perez [189].
3. Still regarding matrices estimation, Theorem 10.1 page 43 shows that Bayesian estimators are (almost) optimal in the reduced-rank regression model, but the case of matrix completion is still an open issue. The situation is actually the following : both models are special cases of the trace regression model. But in order to prove results for matrix completion, one must study the trace regression model with random design, while we were only able to prove results with a fixed design until now. In order to use PAC-Bayesian bounds in the random design context, almost all known techniques require bounded parameter space, while - up to our knowledge - the only way to produce computationally feasible estimators is to use Gaussian priors. Finally, in order to get rid of the terms $\|M\|_F^2$ and $\|N\|_F^2$ is the bound, we probably have to use a heavy-tailed prior as in Dalalyan and Tsybakov's papers [70, 71, 72], but this would lead to serious computational and theoretical problems ! Audibert and Catoni [13, 56] proposed an interesting method to deal with unbounded parameter sets in PAC-Bayesian bounds with random design (it relies on many ideas, one of them is an improvement of the change of variables $\Phi_\alpha(\cdot)$ given by (4) page 46). But this does not solve the computational issue. All of this is to be investigated in depth.

4. PAC-Bayesian bounds study Bayesian estimators from a non asymptotic perspective. An alternative is to study them from an asymptotic perspective. In the parametric case, this is explained in details in van der Vaart's monograph [211] : one can prove consistency, and then exhibit rates of convergence thanks to Bernstein-von Mises theorem. Ghosal, Ghosh and van der Vaart [99] established general conditions for such a result to hold in the non-parametric case. More recently, we refer the reader to Boucheron and Gassiat [37], Rivoirard and Rousseau [180], Castillo and Nickl [51] for Bernstein-von Mises theorems in various contexts. When using these tools, boundedness of the parameter space is not an issue. So, to study connections between PAC-Bayes approach and Bernstein-von Mises theorems might lead to other ways of solving this problem, and would be of high interest in itself anyway.

Publication list

All the papers are available on my website :

`http://alquier.ensae.net/`

Articles in peer-reviewed journals

[A1] Alquier, P. (2008) Iterative Feature Selection in Least Square Regression Estimation. *Annales de l'Institut Henri Poincarré, Probabilités et Statistiques* **44**(1), pp. 47–88.

[A2] Alquier, P. (2008) Density Estimation with Quadratic Loss, a Confidence Intervals Method. *ESAIM Probability and Statistics* **12**, pp. 438–463.

[A3] Alquier, P. (2008) LASSO, Iterative Feature Selection and the Correlation Selector : Oracle Inequalities and Numerical Performances. *Electronic Journal of Statistics* **2**, pp. 1129–1152.

[A4] Alquier, P. (2008) PAC-Bayesian Bounds for Randomized Empirical Risk Minimizers. *Mathematical Methods of Statistics* **17**(4), pp. 279–304.

[A5] Alquier, P. and Lounici, K. (2011) PAC-Bayesian Theorems for Sparse Regression Estimation with Exponential Weights. *Electronic Journal of Statistics* **5**, pp. 127–145.

[A6] Alquier, P. and Doukhan, P. (2011) Sparsity Considerations for Dependent Observations. *Electronic Journal of Statistics* **5**, pp. 750–774.

[A7] Alquier, P. and Hebiri, M. (2011) Generalization of L1 Constraints for High Dimensional Regression Problems. *Statistics and Probability Letters* **81**(12), pp. 1760–1765.

[A8] Alquier, P. and Wintenberger, O. (2012) Model Selection for Weakly Dependent Time Series Forecasting. *Bernoulli* **18**(3), pp. 883–913.

[A9] Alquier, P. and Hebiri, M. (2012) Transductive Versions of the LASSO and the Dantzig Selector. *Journal of Statistical Planning and Inference* **142**(9), pp. 2485–2500.

[A10] Guedj, B. and Alquier, P. (2013) PAC-Bayesian Estimation and Prediction in Sparse Additive Models. *Electronic Journal of Statistics* **7**, pp. 264–291.

[A11] Alquier, P. and Biau, G. (2013) Sparse Single-Index Models. *Journal of Machine Learning Research* **14**, pp. 243–280.

[A12] Alquier, P., Meziani, K. and Peyré, G. (2013) Adaptive Estimation of the Density Matrix in Quantum Homodyne Tomography with Noisy Data. *Inverse Problems* **29**(7), 075017.

[A13] Alquier, P., Butucea, C., Hebiri, M., Meziani, K. and Morimae, T. (2013) Rank Penalized Estimation of a Quantum System. *Physical Review A* **88**(3), 032113.

Articles in peer-reviewed conferences

[A14] Alquier, P. (2010) An Algorithm for Iterative Selection of Blocks of Features. In *Proceedings of the 21th International Conference on Algorithmic Learning Theory (ALT'10)*, M. Hutter, F. Stephan, V. Vovk and T. Zeugmann Editors, Springer Lecture Notes in Artificial Intelligence 6331, pp. 35–49.

[A15] Alquier, P. and Li, X. (2012) Prediction of Quantiles by Statistical Learning and Application to GDP Forecasting. In *Proceedings of the 15th International Conference on Discovery Science (DS'12)*, J.-G. Ganascia, P. Lenca and J.-M. Petit Editors, Springer Lecture Notes in Artificial Intelligence 7569, pp. 22–36.

[A16] Alquier, P. (2013) Bayesian Methods for Low-rank Matrix Estimation : Short Survey and Theoretical Study. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory (ALT'13)*, S. Jain, R. Munos, F. Stephan and T. Zeugmann Editors, Springer Lecture Notes in Artificial Intelligence 8139, pp. 309–323.

Ph.D. Thesis

[A17] Alquier, P. (2006) Transductive and Inductive Adaptive Inference for Density and Regression Estimation. *Ph.D. Thesis*, Université Paris 6.

Book (as an editor)

[A18] Alquier, P., Gautier, E. and Stoltz, G. (Editors) (2011) *Inverse Problems and High-Dimensional Estimation.* Stats in the Château Summer School, August 31 - September 4, 2009. Springer Lecture Notes in Statistics 203.

Submitted preprints

[A19] Alquier, P., Li, X. and Wintenberger, O. (2012) Prediction of Time Series by Statistical Learning : General Losses and Fast Rates. *Preprint arXiv* :1202.4283v1. Currently in minor revision for *Dependence Modeling.*

References

[1] Abramovich, F. and Lahav, T. (2013) Sparse Additive Regression on a Regular Lattice. *Preprint arXiv* :1307.5992.

[2] Adamczak, R. (2008) A Tail Inequality for Suprema of Unbounded Empirical Processes with Applications to Markov Chains. *Electronic Journal of Probability* **13**, pp. 1000–1034.

[3] Agarwal, A. and Duchi, J. C. (2013) The Generalization Ability of Online Algorithms for Dependent Data. *IEEE Transactions of Information Theory* **59**(1), pp. 573–587.

[4] Ahlswede, R. and Winter, A. (2002) Strong Converse for Identification Quantum Channels. *IEEE Transactions of Information Theory* **48**(3), pp. 569–579.

[5] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, Akadémiai Kiadó, Budapest, pp. 267–281.

[6] Aoyagi, M. and Watanabe, S. (2004) The Generalization Error of Reduced Rank Regression in Bayesian Estimation. *International Symposium on Information Theory and its Applications (ISITA2004)*, pp. 1068–1073.

[7] Aoyagi, M. and Watanabe, S. (2005) Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation. *Neural Networks* **18**(7), pp. 924–933.

[8] Arias-Castro, E. and Lounici, K. (2012) Variable Selection with Exponential Weights and ℓ_0-Penalization. *Preprint arXiv* :1208 :2635.

[9] Artiles, L., Gill, R. and Guţă, M. (2005) An invitation to Quantum Tomography. *Journal of the Royal Statistical Society : Series B* **67**(1), pp. 109–134.

[10] Aubry, J.-M., Butucea, C. and Meziani, K. (2009) State Estimation in Quantum Homodyne Tomography with Noisy Data. *Inverse Problems* **25**(1), 015003.

[11] Audibert, J.-Y. (2009) Fast Learning Rates in Statistical Inference Through Aggregation. *The Annals of Statistics* **38**(4), pp. 1591–1646.

[12] Audibert, J.-Y. and Bousquet, O. (2007) Combining PAC-Bayesian and Generic Chaining Bounds. *Journal of Machine Learning Research* **8**, pp. 863–889.

[13] Audibert, J.-Y. and Catoni, O. (2011) Robust Linear Least Square Regression. *The Annals of Statistics* **39**(5), pp. 2766–2794.

[14] Azuma, K. (1967) Weighted Sums of Certain Dependent Random Variables. *Tohoku Mathematical Journal* **2**(19), pp. 357–367.

[15] Babacan, S. D., Luessi, M., Molina, R. and Katsaggelos, A. K. (2011) Low-rank Matrix Completion by Variational Sparse Bayesian Learning. *IEEE Internationan Conference on Audio, Speech and Signal Processing*, pp. 2188–2191.

[16] Bach, F. (2008) BOLASSO : Model Consistent LASSO Estimation Through the Bootstrap. *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, ACM, New York, pp. 33–40.

[17] Bach, F. (2008) Consistency of the Groupe LASSO and Multiple Kernel Learning. *Journal of Machine Learning Research* **9**, pp. 1019–1048.

[18] Baraud, Y., Comte, F. and Viennet, G. (2001) Model Selection for Autoregression with Dependent Data. *ESAIM Probability and Statistics* **5**(1), pp. 33–49.

[19] Barndorff-Nielsen, O. E., Gill, R. and Jupp, P. E. (2003) On Quantum Statistical Inference. *Journal of the Royal Statistical Society : Series B* **65**(4), pp. 775–816.

[20] Barreiro, J. T., Müller, M., Schindler, P., Nigg, D., Monz, T., Chwalla, M., Hennrich, M., Roos, C. F., Zoller, P. and Blatt, R. (2011) An open-system quantum simulator with trapped ions. *Nature* **470**, pp. 486–491.

[21] Barron, A. R., Birgé, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields* **113**(3), pp. 301–413.

[22] Barron, A. R., Cohen, A., Dahmen, W. and DeVore, R. (2008) Adaptive Approximation and Learning by Greedy Algorithms. *The Annals of Statistics* **36**(1), pp. 685–736.

[23] Barron, A. R., Rissanen, J. and Yu, B. (1998) The Minimum Description Length Principle in Coding and Modeling. *IEEE Transactions on Information Theory* **44**(6), pp. 2743–2760.

[24] Beal, M. J. (2003) Variational Algorithm for Approximate Bayesian Inference. *Ph.D. Thesis*, University College London.

[25] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012) Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica* **80**(6), pp. 2369–2430.

[26] Belloni, A. and Chernozhukov, V. (2013) Least Squares after Model Selection in High-dimensional Sparse Models. *Bernoulli* **19**(2), pp. 521–547.

[27] Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root LASSO : Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika* **98**(4), pp. 791–806.

[28] Bennett, G. (1962) Probability Inequalities for Sums of Independant Random Variables. *Journal of the American Statistical Association* **57**(297), pp. 33–54.

[29] Bennett, J. and Lanning, S. (2007) The Netflix Prize. *Proceedings of KDD Cup and Workshop 07*, pp. 3–6.

[30] Bercu, B., Gassiat, E. and Rio, E. (2002) Concentration Inequalities, Large and Moderate Deviations for Self-Normalized Empirical Processes. *Annals of Probability* **30**(4), pp. 1576–1604.

[31] Bernstein, S. N. (1946) *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow.

[32] Bertail, P. and Clémençon, S. Sharp Bounds for the Tails of Functionals of Harris Markov Chains. *Theory of Probability and its Applications* **54**(3), pp. 505–515.

[33] Bickel, P., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous Analysis of LASSO and Dantzig Selector. *The Annals of Statistics* **37**(4), pp. 1705–1732.

[34] Birgé, L. (1999) An Alternative Point of View on Lepki's Method. In *State of the Art in Probability and Statistics*, M. de Gunst, C. Klaassen and A. van der Vaart Editors, Institute of Mathematical Statistics (IMS) Lecture notes - monograph series 36, pp. 113–133.

[35] Birgé, L. and Massart, P. (1998) Minimum Contrast Estimators on Sieves. *Bernoulli* **4**(3), pp. 329–375.

[36] Blanchard, G. and Fleuret, F. (2007) Occam's Hammer. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, M. H. Bshouty and C. Gentile Editors, Springer Lecture Notes in Computer Science 4539, pp. 112-126.

[37] Boucheron, S. and Gassiat, E. (2009) A Bernstein-von Mises Theorem for Discrete Probability Distributions. *Electronic Journal of Statistics* **3**, pp. 114–148.

[38] Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities : A Nonasymptotic Theory of Independence*. Oxford University Press.

[39] Breiman, L. (1993) Better Subset Selection using Non-negative Garotte. *Technical Report*, University of California, Berkeley.

[40] Bühlmann, P. and van de Geer, S. (2009) On the Conditions Used to Prove Oracle Results for the LASSO. *Electronic Journal of Statistics* **3**, pp. 1360–1392.

[41] Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data*. Springer Series in Statistics.

[42] Bunea, F., She, Y. and Wegkamp, M. H. (2011) Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices. *The Annals of Statistics* **39**(2), pp. 1282–1309.

[43] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007) Aggregation for Gaussian Regression. *The Annals of Statistics* **35**(4), pp. 1674–1697.

[44] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007) Sparsity Oracle Inequalities for the LASSO. *Electronic Journal of Statistics* **1**, pp. 169–194.

[45] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007) Sparse Density Estimation with ℓ_1 Penalties. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, N. H. Bshouty and C. Gentile Editors, Springer Lecture Notes in Artificial Intelligence 4539, pp. 530–543.

[46] Candés, E. and Plan, Y. (2009) Matrix Completion with Noise. *Proceedings of the IEEE* **98**(6), pp. 925–936.

[47] Candés, E. and Recht, B. (2009) Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics* **9**(6), pp. 717–772.

[48] Candés, E. and Tao, T. (2007) The Dantzig Selector : Statistical Estimation when p is Much Larger Than n. *The Annals of Statistics* **35**(6), pp. 2313–2351.

[49] Candés, E. and Tao, T. (2009) The Power of Convex Relaxation : Near-Optimal Matrix Completion. *IEEE Transactions on Information Theory* **56**(5), pp. 2053–2080.

[50] Carlin, B. P. and Chib, S. (1995) Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society : Series B* **57**(3), pp. 473–484.

[51] Castillo, I. and Nickl, R. (2013) Nonparametric Bernstein-von Mises Theorems. *The Annals of Statistics* **41**(4), pp. 1999-2028.

[52] Catoni, O. (2002) Data Compression and Adaptive Histograms. In *Foundations of Computational Mathematics : Proceedings of the SMALEFEST 2000*, F. Cucker and J. Maurice Rojas Editors, World Scientific, pp. 35–60.

[53] Catoni, O. (2003) Laplace Transform Estimates and Deviation Inequalities. *Annales de l'Institut Henri Poincarré, Probabilités et Statistiques* **39**(1), pp. 1–26.

[54] Catoni, O. (2004) *Statistical Learning Theory and Stochastic Optimization*. Ecole d'Été de Probabilités de Saint-Flour XXXI - 2001. J. Picard Editor. Springer Lecture Notes in Mathematics 1851.

[55] Catoni, O. (2007) *PAC-Bayesian Supervised Classification : The Thermodnamics of Statistical Learning*. Institute of Mathematical Statistics (IMS) Lecture notes - monograph series 56.

[56] Catoni, O. (2012) Challenging the Empirical Mean and Empirical Variance : a Deviation Study. *Annales de l'Institut Henri Poincarré, Probabilités et Statistiques* **48**(4), pp. 1148–1185.

[57] Cesa-Bianchi, N. and Lugosi, G. (2006) *Prediction, Learning and Games*. Cambridge University Press, New York.

[58] Chen, S. (1995) Basis Pursuit. *Ph.D. Thesis*, Department of Statistics, Stanford University.

[59] Chen, S. and Donoho, D. (1994) Basis Pursuit. In *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*, pp. 41-44.

[60] Christofides, D. and Markström, K. (2008) Expansion Properties of the Asymptotic Efficiency for Tests of a Hypothesis based on the Sum of Observations. *Random Structures and Algorithms* **32**(8), pp. 88-100.

[61] Clémençon, S. (2001) Moment and Probability Inequalities for Sums of Bounded Additive Functionals of a Regular Markov Chains via the Nummelin Splitting Technique. *Statistics and Probability Letters* **3**(1), pp. 227–238.

[62] Collet, P., Martinez, S. and Schmitt, B. (2002) Exponential Inequalities for Dynamical Measures of Expanding Maps of the Interval. *Probability Theory and Related Fields* **123**(3), pp. 301–322.

[63] Comminges, L. and Dalalyan, A. (2012) Tight Conditions for Consistency of Variable Selection in the Context of High Dimensionality. *The Annals of Statistics* **40**(5), pp. 2667–2696.

[64] Corander, J. and Villani, M. (2004) Bayesian Assessment of Dimensionality in Reduced Rank Regression. *Statistica Neerlandica* **58**(3), pp. 255–270.

[65] Cui, W. and George, I. E. (2009) Empirical Bayes vs. Fully Bayes Variable Selection. *Journal of Statistical Planning and Inference* **138**(4), pp. 888–900.

[66] Dai, D., Rigollet, P., Xia, L. and Zhang, T. (2013) Aggregation of Affine Estimators. *Preprint* arXiv :1311.2799.

[67] Dai, D., Rigollet, P. and Zhang, T. (2012) Deviation Optimal Learning using Greedy Q-Aggregation. *The Annals of Statistics* **40**(3), pp. 1878–1905.

[68] Dalalyan, A., Judistki, A. and Spokoiny, V. (2008) A New Algorithm for Estimating the Effective Dimension-Reduction Subspace. *Journal of Machine Learning Research* **9**, pp. 1647–1678.

[69] Dalalyan, A. and Salmon, J. (2012) Sharp Oracle Inequalities for Aggregation of Affine Estimators. *The Annals of Statistics* **40**(4), pp. 2327–2355.

[70] Dalalyan, A. and Tsybakov, A. B. (2007) Aggregation by Exponential Weighting and Sharp Oracle Inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, M. H. Bshouty and C. Gentile Editors, Springer Lecture Notes in Computer Science 4539, pp. 97–111.

[71] Dalalyan, A. and Tsybakov, A. B. (2008) Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity. *Machine Learning* **72**(1-2), pp. 39–61.

[72] Dalalyan, A. and Tsybakov, A. B. (2012) Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences* **78**(5), pp. 1423–1443.

[73] Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S. and Prieur, C. (2007) *Weak Dependance : with Examples and Applications*. Springer Lecture Notes in Statistics 190.

[74] Dedecker, J. and Prieur, C. (2005) New Dependence Coefficients : Examples and Applications to Statistics. *Probability Theory and Related Fields* **132**(2), pp. 203–253.

[75] Delattre, S. and Gaïffas, S. (2011) Nonparametric Regression with Martingale Regression Errors. *Stochastic Processes and their Applications* **121**(12), pp. 2899–2924.

[76] Delecroix, M., Hristache, M. and Patilea, V. (2006) On Semi-parametric M-estimation in Single-index Regression. *Journal of Statistical Planning and Inference* **136**(3), pp. 730–769.

[77] Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. Springer Applications of Mathematics Series **31**.

[78] Donoho, D. and Johnstone, I. (1994) Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika* **81**(3), pp. 425–455.

[79] Donoho, D. and Johnstone, I. (1995) Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association* **90**(432), pp. 1200–1224.

[80] Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage : Asymptotia ? *Journal of the Royal Statistical Society : Series B* **57**(2), pp. 301–369.

[81] Donsker, M. D. and Varadhan, S. S. (1976) Asymptotic Evaluation of Certain Markov Process Expectation for Large Time. *Communications on Pure and Applied Mathematics I* **28**, pp. 389–461.

[82] Doukhan, P. (1994) *Mixing. Properties and Examples*. Springer Lecture Notes in Statistics **85**.

[83] Doukhan, P. and Louhichi, S. (1999) A New Weak Dependence Condition and Applications to Moment Inequalities. *Stochastic Processes and their Applications* **84**(2), pp. 313–342.

[84] Doukhan, P. and Neumann, M. H. (2007) Probability and Moment Inequalities for Sums of Weakly Dependent Random Variables, with Applications. *Stochastic Processes and their Applications* **117**(7), pp. 878–903.

[85] Doukhan, P. and Wintenberger, O. (2008) Weakly Dependent Chains with Infinite Memory. *Stochastic Processes and their Applications* **118**(11), pp. 1997–2013.

[86] Duchi, J. C., Agarwal, A., Johansson, M. and Jordan, M. I. (2012) Ergodic Mirror Descent. *SIAM Journal on Optimization* **22**(4), pp. 1549–1578.

[87] Efron, B., Hastie, T., Johnstone, I. and and Tibshirani, R. (2004) Least Angle Regression. *The Annals of Statistics* **32**(2), pp. 407–499.

[88] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* **96**(456), pp. 1348–1360.

[89] Fan, J. and Lv, J. (2006) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society : Series B* **70**(5), pp. 849–911.

[90] Frank, I. and Friedman, J. (1993) A Statistical View of some Chemometrics Regression Tools. *Technometrics* **35**(1), pp. 109-148.

[91] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise Coordinate Optimization. *Annals of Applied Statistics* **1**(2), pp. 302–332.

[92] Gaïffas, S. and Lecué, G. (2007) Optimal Rates and Adaptation in the Single-Index Model using Aggregation. *Electronic Journal of Statistics* **1**, pp. 538–573.

[93] Gautier, E. and Tsybakov, A. B. (2011) High-Dimensional Instrumental Variables Regression and Confidence Sets. *Preprint* arXiv :1105.2454.

[94] George, I. E. (2000) The Variable Selection Problem. *Journal of the American Statistical Association* **95**(452), pp. 1304–1308.

[95] George, I. E. and McCulloch, R. E. (1997) Approaches for Bayesian Model Selection. *Statistica Sinica* **7**(2), pp. 339–373.

[96] Gerchinovitz, S. (2011) Prediction of Individual Sequences and Prediction in the Statistical Framework : some Links around Sparse Regression and Aggregation Techniques. *Ph.D. Thesis*, Université Paris Sud.

[97] Gerchinovitz, S. (2013) Sparsity Regret Bounds for Individual Sequences in Online Linear Regression. *Journal of Machine Learning Research* **14**, pp. 729–769.

[98] Geweke, J. (1996) Bayesian Reduced-rank Regression in Econometrics. *Journal of Econometrics* **75**(1), pp. 121–146.

[99] Ghosal, S., Ghosh, J. and van der Vaart, A. W. (2000) Convergence Rates of Posterior Distributions. *The Annals of Statistics* **28**(2), pp. 500–531.

[100] Giraud, C., Huet, S. and Verzelen, N. (2012) High-dimensional Regression with Unknown Variance. *Statistical Science* **27**(4), pp. 508–518.

[101] Golubev, Y. (2010) On Universal Oracle Inequalities Related to High-dimensional Linear Models. *The Annals of Statistics* **38**(5), pp. 2751–2780.

[102] Green, P.-J. (1995) Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **82**(4), pp. 711–732.

[103] Green, P.-J. and Richardson, S. (1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society : Series B* **59**(4), pp. 731–792.

[104] Gross, D. (2011) Recovering Low-Rank Matrices from few Coefficients in any Basis. *IEEE Transactions on Information Theory* **57**(3), pp. 1548–1566.

[105] Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S. and Eisert, J. (2010) Quantum State Tomography via Compressed Sensing. *Physical Review Letters* **105**(15), 150401.

[106] Guţă, M., Kypraios, T. and Dryden, I. (2012) Rank-Based Model Selection for Multiple Ions Quantum Tomography. *New Journal of Physics* **14**, 105002.

[107] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2004) *A Distribution Free Theory of Nonparametric Regression.* Springer Series in Statistics.

[108] Hang, H. and Steinwart, I. (2012) Fast Learning from α-mixing Observations. *Technical report*, Fakultät für Mathematik und Physik, Universität Stutgart.

[109] Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal Smoothing in Single-index Models. *The Annals of Statistics* **21**(1), pp. 157–178.

[110] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. B., (1998) *Wavelets, Approximation, and Statistical Applications.* Springer Lecture Notes in Statistics 129.

[111] Hastie, T. and Tibshirani, R. (1986) Generalized Additive Models. *Statistical Science* **1**(3), pp. 297–318.

[112] Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models.* Chapman & Hall Monogaphs on Statistics and Applied Probability 43.

[113] Hebiri, M. and Lederer, J. (2013) How Correlation Influence LASSO Prediction. *IEEE Transactions on Information Theory* **59**(3), pp. 1846–1854.

[114] Hebiri, M. and van de Geer, S. (2011) The Smooth-LASSO and Other $\ell_1+\ell_2$-Penalized Methods. *Electronic Journal of Statistics* **5**, pp. 1184–1226.

[115] Herbrich, R. and Graepel, T. (2002) A PAC-Bayesian Margin Bound for Linear Classifiers. *IEEE Transactions on Information Theory* **48**(12), pp. 3140–3150.

[116] Higgs, M. and Shawe-Taylor, J. (2010) A PAC-Bayes Bound for Taylored Density Estimation. In *Proceedings of the 21th International Conference on Algorithmic Learning Theory (ALT'10)*, M. Hutter, F. Stephan, V. Vovk and T. Zeugmann Editors, Springer Lecture Notes in Artificial Intelligence 6331, pp. 148–162.

[117] Hoeffding, W. (1963) Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **58**(301), pp. 13–30.

[118] Holevo, A. S. (1982) *Probabilistic and Statistical Aspects of Quantum Theory.* North-Holland Series in Statistics and Probability.

[119] Horowitz, J. L. (1998) *Semiparametric Methods in Econometrics.* Springer Series in Statistics.

[120] Hsu, N.-J., Hung, H.-L. and Chang, Y.-M. (2008) Subset Selection for Vector Autoregressive Processes using LASSO. *Computational Statistics and Data Analysis* **52**(7), pp. 3645–3657.

[121] Huang, C., Cheang, G. H. L. and Barron, A. R. (2008) Risk of Penalized Least Squares, Greedy Selection and ℓ_1-penalization for Flexible Function Libraries. *Technical report*, Yale University.

[122] Ibragimov, I. A. (1962) Some Limit Theorems for Stationary Processes. *Theory of Probability and its Applications* **7**(4), pp. 349–382.

[123] Jiang, W. (2007) Bayesian Variable Selection for High Dimensional Generalized Linear Models : Convergence Rate of the Fitted Density. *The Annals of Statistics* **35**(4), pp. 1487–1511.

[124] Jiang, W. and Tanner, M. A. (2008) Gibbs Posterior for Variable Selection in High-dimensional Classification and Data Mining. *The Annals of Statistics* **36**(5), pp. 2270–2231.

[125] Juditsky, A. and Nemirovski, A. (2000) Functionnal Aggregation for Nonparametric Estimation. *The Annals of Statistics* **28**(3), pp. 681–712.

[126] Kleibergen, F. and Paap, R. (2002) Priors, Posteriors and Bayes Factors for a Bayesian Analysis of Cointegration. *Journal of Econometrics* **111**(2), pp. 223–249.

[127] Klopp, O. (2011) Rank-Penalized Estimators for High-Dimensional Matrices. *Electronic Journal of Statistics* **5**, pp. 1161–1183.

[128] Kolar, M. and Liu, H. (2012) Marginal Regression for Multitask Learning. *JMLR : Workshop and Conference Proceedings* **22** : Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS'12), pp. 647–655.

[129] Koltchinskii, V. (2009) The Dantzig Selector and Sparsity Oracle Inequalities. *Bernoulli* **15**(6), pp. 799–828.

[130] Koltchinskii, V. (2010) *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.* Ecole d'Été de Probabilités de Saint-Flour XXXVIII - 2008. J. Picard Editor. Springer Lecture Notes in Mathematics 2033.

[131] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011) Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion. *The Annals of Statistics* **39**(5), pp. 2302–2329.

[132] Koltchinskii, V. and Yuan, M. (2010) Sparsity in Multiple Kernel Learning. *The Annals of Statistics* **38**(6), pp. 3660–3695.

[133] Langford, J., Seeger, M. and Megiddo, N. (2001) An Improved Predictive Accuracy Bound for Averaging Classifiers. *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*, Morgan Kaufmann Publishers, pp. 290–297.

[134] Lawrence, N. D. and Urtasun, R. (2009) Non-linear Matrix Factorization with Gaussian Processes. *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, ACM, pp. 601–608.

[135] Lecué, G. (2007) Simultaneous Adaptation to the Margin and Complexity in Classification. *The Annals of Statistics* **35**(4), pp. 1698–1721.

[136] Lecué, G. (2007) Aggregation Procedures : Optimality and Fast Rates. *Ph.D. Thesis*, Université Paris 6.

[137] Lecué, G. and Rigollet, P. (2013) Optimal Learning with Q-aggregation. *Preprint* arXiv :1301.6080. To appear in *the Annals of Statistics*.

[138] Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces*. Springer - Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics 23.

[139] Leeb, H. and Pötscher, B. M. (2007) Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator. *Journal of Econometrics* **142**(1), pp. 201–211.

[140] Leng, C., Lin, Y. and Wahba, G. (2006) A Note on the LASSO and Related Procedures in Model Selection. *Statistica Sinica* **16**, pp. 1273–1284.

[141] Lepski, O. V. (1990) On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory of Probability and its Applications* **35**(3), pp. 454–466.

[142] Lepski, O. V. (1991) Asymptotically Minimax Adaptive Estimation I : Upper Bounds. *Theory of Probability and its Applications* **36**(4), pp. 682–697.

[143] Lepski, O. V. (1992) Asymptotically Minimax Adaptive Estimation II : Schemes without Optimal Adaptation : Adaptive Estimators. *Theory of Probability and its Applications* **37**(3), pp. 433-448.

[144] Lepski, O. and Serdyukova, N. (2013) Adaptive Estimation under Single-Index Constraint in a Regression Model. *Preprint* arXiv :1304 :7668.

[145] Leung, G. and Barron, A. R. (2006) Information Theory and Mixing Least-Square Regressions. *IEEE Transactions on Information Theory* **52**(8), pp. 3396–3410.

[146] Li, C., Jiang, W. and Tanner, M. A. (2013) General Oracle Inequalities for Gibbs Posterior with Application to Ranking. *JMLR : Workshop and Conference Proceedings* **30** : Conference on Learning Theory (COLT'13), pp. 512–521.

[147] Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. O. (2008) Mixture of g-Priors for Bayesian Variable Selection. *Journal of the American Statistical Association* **103**(481), pp. 410–423.

[148] Lim, Y. J. and Teh, Y. W. (2007) Variational Bayesian Approach to Movie Rating Prediction. *Proceedings of KDD Cup and Workshop 07*, pp. 15–21.

[149] Lopez, O. (2009) Single-Index Regression Models with Right-Censored Responses. *Journal of Statistical Planning and Inference* **139**(3), pp. 1082–1097.

[150] Lounici, K. (2008) Sup-norm Convergence Rate and Sign Concentration Property of LASSO and Dantzig Estimators. *Electronic Journal of Statistics* **2**, pp. 90–102.

[151] Maillard, O.-A. and Munos, R. (2012) Linear Regression with Random Projections. *Journal of Machine Learning Research* **13**, pp. 2735–2772.

[152] Mallows, C. L. (1973) Some Comments on C_P. *Technometrics* **15**(4), pp. 661-676.

[153] Mammen, E. and Tsybakov, A. B. (1999) Smooth Discrimination Analysis. *The Annals of Statistics* **27**(6), pp. 1808–1829.

[154] Massart, P. (2007) *Concentration Inequalities and Model Selection*. Ecole d'Été de Probabilités de Saint-Flour XXXIII - 2003. J. Picard Editor. Springer Lecture Notes in Mathematics 1896.

[155] McAllester, D. (1998) Some PAC-Bayesian Theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98)*, ACM, pp. 230–234.

[156] McAllester, D. (1999) PAC-Bayesian Model Averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT'99)*, ACM, pp. 164–170.

[157] McAllester, D. (2003) PAC-Bayesian Stochastic Model Selection. *Machine Learning* **51**(1), pp. 5–21.

[158] McCullagh, P. and Nedler, J. A. (1983) *Generalized Linear Models*. Chapman & Hall Monogaphs on Statistics and Applied Probability 37.

[159] Meier, L., van de Geer, S. and Bühlmann, P. (2009) High-dimensional Additive Modeling. *The Annals of Statistics* **37**(6B), pp. 3779–3821.

[160] Meinshausen, N. and Bühlmann, P. (2013) Stability Selection. *Journal of the Royal Statistical Society : Series B* **72**(4), pp. 417–473.

[161] Meir, R. (2000) Nonparametric Time Series Prediction through Adaptive Model Selection. *Machine Learning* **39**(1), pp. 5–34.

[162] Meir, R. and Zhang, T. (2003) Generalization Error Bounds for Bayesian Mixture Algorithms. *Journal of Machine Learning Research* **4**, pp. 839–860.

[163] Merlevede, F., Peligrad, M. and Rio, E. (2009) A Bernstein Type Inequality and Moderate Deviations under Strong Mixing Conditions. In *High Dimensional Probability V : The Luminy Volume*, C. Houdré, V. Koltchinskii, D. M. Mason and M. Peligrad Editors, Institute of Mathematical Statistics (IMS) Collections, pp. 273–292.

[164] Merlevede, F., Peligrad, M. and Rio, E. (2011) A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences. *Probability Theory and Related Fields* **151**(3-4), pp. 435–474.

[165] Modha, D. S. and Masry, E. (1998) Memory-Universal Prediction of Stationary Random Processes. *IEEE Transactions of Information Theory* **44**(1), pp. 117–133.

[166] Mougeot, M., Picard, D. and Tribouley, K. (2012) Learning Out of Leaders. *Journal of the Royal Statistical Society : Series B* **74**(3), pp. 475–513.

[167] Natarajan, B. K. (1995) Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing* **24**(2), pp. 227–234.

[168] Nemirovski, A. (2000) Topics in Non-Parametric Statistics. In *Ecole d'Été de Probabilités de Saint-Flour XXVIII - 1998*. P. Bernard Editor. Springer Lecture Notes in Mathematics 1738, pp. 85–277.

[169] Nott, D. J. and Leonte, D. (2004) Sampling Schemes for Bayesian Variable Selection in Generalized Linear Models. *Journal of Computational and Graphical Statistics* **13**(2), pp. 362–382.

[170] Oliveira, R. I. (2010) Sums of Random Hermitian Matrices and an Inequality by Rudelson. *Electronic Communications in Probability* **15**, pp. 203–212.

[171] Pandolfi, S., Bartolucci, F. and Friel, N. (2010) A Generalization of the Multiple-Try Metropolis Algorithm for Bayesian Estimation and Model Selection. *JMLR : Workshop and Conference Proceedings* **9** : Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS'10), pp. 41–48.

[172] Percival, D. (2013) Structured, Sparse Aggregation. *Journal of the American Statistical Association* **107**(498), pp. 814–823.

[173] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse Additive Models. *Journal of the Royal Statistical Society : Series B* **71**(5), pp. 1009–1030.

[174] Recht, B. (2011) A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research* **12**, pp. 3413–3430.

[175] Reinsel, G. C. and Velu, R. P. (1998) *Multivariate Reduced-Rank Regression : Theory and Applications*. Springer Lecture Notes in Statistics 136.

[176] Rigollet, P. and Tsybakov, A. B. (2011) Exponential Screening and Optimal Rates of Sparse Estimation. *The Annals of Statistics* **39**(2), pp. 731–771.

[177] Rio, E. (2000) *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants*. Springer "Mathématiques et Applications" 31.

[178] Rio, E. (2000) Inégalité de Hoeffding pour les Fonctions Lipshitziennes de Suites Dépendantes. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* **330**(10), pp. 905–908.

[179] Rissanen, J. (1978) Modelling by Shortest Data Description. *Automatica* **14**(5), pp. 465–471.

[180] Rivoirard, V. and Rousseau, J. (2012) Bernstein-von Mises Theorem for Linear Functionals of the Density. *The Annals of Statistics* **40**(3), pp. 1489–1523.

[181] Robbiano, S. (2013) Upper Bounds and Aggregation in Bipartite Ranking. *Electronic Journal of Statistics* **7**, pp. 1249–1271.

[182] Robert, C. and Casella, G. (2004) *Monte-Carlo Statistical Methods (2nd Edition)*. Springer Texts in Statistics.

[183] Rohde, A. and Tsybakov, A. B. (2011) Estimation of High-Dimensional Low-Rank Matrices. *The Annals of Statistics* **39**(2), pp. 887-930.

[184] Rosenblatt, M. (1956) A Central Limit Theorem and a Strong Mixing Condition. *Proceedings of the National Academy of Sciences of the United States of America* **42**(1), pp. 43–47.

[185] Salakhutdinov, R. and Mnih, A. (2008) Bayesian Probabilistic Matrix Factorization. *Advances in Neural Information Processing Systems (NIPS'07)*, J. C. Platt, D. Koller, Y Singer and S. Roweis Editors, Cambridge, MIT Press.

[186] Salakhutdinov, R. and Mnih, A. (2008) Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo. *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, ACM, New York, pp. 880–887.

[187] Salmon, J. and Le Pennec, E. (2009) An Aggregator Point of View on NL-Means. *Proceedings of SPIE* 7664, Wavelets XIII, 74461E.

[188] Samson, P.-M. (2000) Concentration of Measure Inequalities for Markov Chains and Φ-Mixing Processes. *Annals of Probability* **28**(1), pp. 416–461.

[189] Sanchez-Perez, A. (2013) Time Series Prediction via Aggregation : an Oracle Bound Including Numerical Cost. *Preprint arXiv* :1311.4500.

[190] Scott, J. G. and Berger, J. O. (2010) Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *The Annals of Statistics* **38**(5), pp. 2587–2619.

[191] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics* **6**(2), pp. 461–464.

[192] Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J. and Auer, P. (2012) PAC-Bayesian Inequalities for Martingales. *IEEE Transactions of Information Theory* **58**(12), pp. 7086–7093.

[193] Seldin, Y. and Tishby, N. (2010) PAC-Bayesian Analysis of Co-clustering and Beyond. *Journal of Machine Learning Research* **11**, pp. 3595–3646.

[194] Shah, R. D. and Samworth, R. J. (2013) Variable Selection with Error Control : Another Look at Stability Selection. *Journal of the Royal Statistical Society : Series B* **75**(1), pp. 55–80.

[195] Shawe-Taylor, J. and Williamson, R. (1997) A PAC Analysis of a Bayesian Estimator. In *Proceedings of the 10th Annual Conference on Computational Learning Theory (COLT'97)*, ACM, pp. 2–9.

[196] Statuljavichus, V. A. and Yackimavicius, D. A. (1989) Theorems on Large Deviations for Dependent Random Variables. *Soviet Mathematics Doklady* **38**(2), pp. 442-445.

[197] Stein, C. (1981) Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics* **9**(6), pp. 1135–1151.

[198] Steinwart, I. and Christmann, A. (2009) Fast-Learning from Non I.I.D. Observations. In *Advances in Neural Information Processing Systems (NIPS'09)*, Y. Bengio, D. Schurmans, J. Lafferty, C. K. I William and A. Culotta Editors, pp. 1768–1776.

[199] Steinwart, I. and Hush, D. and Scovel, C. (2009) Learning from Dependent Observations. *Journal of Multivariate Analysis* **100**(1), pp. 175–194.

[200] Stoltz, G. (2010) Agrégation Séquentielle de Prédicteurs : Méthodologie Générale et Applications à la Prévision de la Qualité de l'Air et à celle de la Consommation Electrique. *Journal de la SFDS* **151**(2), pp. 66–106.

[201] Stone, C. J. (1985) Additive Regression and Other Nonparametric Models. *The Annals of Statistics* **13**(2), pp. 689–705.

[202] Suzuki, T. (2012) PAC-Bayesian Bound for Gaussian Process and Multiple Kernel Additive Model. *JMLR : Workshop and Conference Proceedings* **23** : Conference on Learning Theory (COLT'12), pp. 8.1–20.

[203] Suzuki, T. and Sugiyama, M. (2013) Fast Learning Rate of Multiple Kernel Learning : Trade-Off between Sparsity and Smoothness. *The Annals of Statistics* **41**(3), pp. 1381–1405.

[204] Tao, T. (2012) *Topics in Random Matrix Theory*. American Mathematical Society, Graduate Studies in Mathematics 132.

[205] Tibshirani, R. (1996) Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society : Series B* **58**(1), pp. 267–288.

[206] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and Smoothness via the Fused LASSO. *Journal of the Royal Statistical Society : Series B* **67**(1), pp. 91–108.

[207] Tropp, J. A. (2011) User-Friendly Tail Bounds for Sums of Random Variables. *Foundation of Computational Mathematics* **12**(4), pp. 389–434.

[208] Tsybakov, A. B. (2003) Optimal Rates of Aggregation. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT'03)*, B. Schölkopf and M. K. Warmuth Editors, Springer Lecture Notes in Computer Science 2777, pp. 303–313.

[209] Tsybakov, A. B. (2009) *Introduction to Nonparametric Estimation*. Springer Series in Statistics.

[210] Valiant, L. (1984) A Theory of the Learnable. *Communications of the ACM* **27**(11), pp. 1134–1142.

[211] van der Vaart, A. W. (1998) *Asymptotics Statistics*. Cambridge University Press.

[212] van der Vaart, A. W. and van Zanten, J. H. (2008) Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors. *The Annals of Statistics* **36**(3), pp. 1435–1463.

[213] Vapnik, V. N. (1998) *Statistical Learning Theory*. Wiley-Blackwell.

[214] Vapnik, V. N. and Chervonenkis, A. Ya. (1971) On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications* **16**(2), pp. 264–280.

[215] Vershynin, R. (2012) Introduction to the Non-Asymptotic Analysis of Random Matrices. In *Compressed Sensing, Theory and Applications*, Y. Eldar and G. Kutyniok Editors, Cambridge University Press, pp. 303–313.

[216] Vert, J.-P. (2001) Adaptive Context Trees and Text Clustering. *IEEE Transactions on Information Theory* **47**(5), pp. 1884–1901.

[217] Wang, H., Li, G. and Tsai, C.-L. (2007) Regression Coefficient and Autoregressive Order Shrinkage via the LASSO. *Journal of the Royal Statistical Society : Series B* **69**(1), pp. 63–78.

[218] Wang, H.-B. (2009) Bayesian Estimation and Variable Selection for Single-Index Models. *Computational Statistics and Data Analysis* **53**(7), pp. 2617–2627.

[219] Wang, T., Xu, P.-R. and Zhu, L.-X. (2012) Non-Convex Penalized Estimation in High-Dimensional Models with Single-Index Structure. *Computational Statistics and Data Analysis* **53**(7), pp. 2617–2627.

[220] West, M. (2003) Bayesian Factors in the "Large p Small n" Paradigm. In *Bayesian Statistics* **7**, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. David, D. Heckerman, A. F. M. Smith and M. West Editors, Oxford University Press, pp. 723–732.

[221] Wintenberger, O. (2010) Deviation Inequalities for Sums of Weakly Dependent Time Series. *Electronic Communications in Probability* **15**, pp. 489–503.

[222] Wintenberger, O. (2012) Weak Transport Inequalities and Applications to Exponential and Oracle Inequalities. *Preprint arXiv* :1207.4951.

[223] Xu, Y.-L. and Chen, D.-R. (2008) Learning Rate of Regularized Regression for Exponentially Strongly Mixing Sequence. *Journal of Statistical Planning and Inference* **138**(7), pp. 2180–2189.

[224] Yang, Y. (2003) Regression with Multiple Candidate Models : Selecting or Mixing ? *Statistica Sinica* **13**(3), pp. 783–809.

[225] Yang, Y. (2004) Aggregating Regression Procedures to Improve Performance. *Bernoulli* **10**(1), pp. 25–47.

[226] Yang, Y. (2005) Can the Strengths of AIC and BIC be Shared ? A Conflict Between Model Indentification and Regression Estimation. *Biometrika* **92**(4), pp. 937–590.

[227] Yoon, Y. J., Park, C. and Lee, T. (2012) Penalized Regression Models with Autoregressive Error Terms. To appear in *Journal of Statisticsl Computation and Simulation*.

[228] Yu, K., Lafferty, J., Zhu, S. and Gong, Y. (2009) Large-Scale Collaborative Prediction using a Non-parametric Random Effects Model. *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, ACM, New York, pp. 1185–1192.

[229] Yu, K., Tresp, V. and Schwaighofer, A. (2005) Learning Gaussian Process for Multiple Tasks. *Proceedings of the 22th International Conference on Machine Learning (ICML'05)*, ACM, New York, pp. 1012–1019.

[230] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society : Series B* **68**(1), pp. 49–67.

[231] Zhang, J., Wang, X., Yu, Y. and Gai, Y. (2013) Estimation and Variable Selection in Partial Linear Single Index Models with Error-prone Linear Covariates. To appear in *Statistics*.

[232] Zhang, T. (2004) Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *The Annals of Statistics* **32**(1), pp. 56–85.

[233] Zhang, T. (2006) Information Theoretical Upper and Lower Bounds for Statistical Estimation. *IEEE Transaction on Information Theory* **52**(4), pp. 1307–1321.

[234] Zhang, T. (2011) Sparse Recovery With Orthogonal Matching Pursuit Under RIP. *IEEE Transaction on Information Theory* **57**(9), pp. 6215–6221.

[235] Zhang, T. (2011) Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. *IEEE Transaction on Information Theory* **57**(7), pp. 4689–4708.

[236] Zhao, R. and Yu, B. (2006) On Model Selection Consistency of LASSO. *Journal of Machine Learning Research* **7**, pp. 2541–2563.

[237] Zhou, M., Wang, C., Chen, M., Paisley, J. and Carin, L. (2010) Nonparametric Bayesian Matrix Completion. *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM2010)*, pp. 213–216.

[238] Zou, B., Li, L. and Xu, Z. (2009) The Generalization Performance of ERM Algorithm with Strongly Mixing Observations. *Machine Learning* **75**(3), pp. 275–295.

[239] Zou, H. (2006) The Adaptive LASSO and its Oracle Properties. *Journal of the American Statistical Association* **101**(476), pp. 1418–1429.

[240] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society : Series B* **67**(2), pp. 301–320.