MEMOIRE D'HABILITATION A DIRIGER DES RECHERCHES

présenté par

Matthieu LERASLE

"Sélection d'estimateurs, validation croisée et relations entre tests agrégés et tests multiples"

Soutenue publiquement le 29 mars 2016 devant le jury composé de

Francis Bach	INRIA	Examinateur
Yannick Baraud	Université de Côte d'Azur	Examinateur
Béatrice Laurent	INSA de Toulouse	Examinateur
Pascal Massart	Université d'Orsay	Président
Eric Moulines	Ecole Polytechnique	Examinateur
Alexandre Tsybakov	ENSAE	Examinateur

et au vue des rapports également écrits par

Stéphane Boucheron	Université Paris VII	Rapporteur
Richard Nickl	Cambridge University	Rapporteur
Vladimir Koltchinskii	Georgia Tech University	Rapporteur

Remerciements

Je souhaite tout d'abord remercier les rapporteurs Stéphane Boucheron, Richard Nickl et Vladimir Koltchinskii pour leur travail minutieux et leurs commentaires enthousiastes. Merci beaucoup Stéphane pour ton rapport extrêmement fouillé et précis. Tes connaissances encyclopédiques en statistique théorique, mais également dans les quelques domaines connexes ou j'ai puisé avec mes collaborateurs quelques idées font de ton rapport un document précieux et inspirant. Thank you very much Richard for your precise and motivating report. It's been some years now that I am amazed by your extraordinary achievements, since your Post-Doc with Evarist. Your encouragements are particularly appreciated for these reasons also. Finally, thank you very much Vladimir for your deep comments on my work. You have been an inspiring researcher from the beginning of my career and it was an honor to have your opinion on my research.

Je voudrais également remercier les membres de mon jury aujourd'hui. Francis, après t'avoir croisé plusieurs fois place d'Italie en venant travailler avec Sylvain, ça a été un vrai plaisir d'écouter ton cours sur les liens étroits entre statistiques et optimisation. Merci Yannick pour tes conseils prodigués ces dernières années, tes points de vue tranchants et extrêmement perspicaces en statistiques, tes solutions simples et élégantes tant en mathématiques que dans les labyrinthes administratifs. Merci aussi de ton soutien constant dans la préparation de cette habilitation, à la fois pour ta relecture complète du manuscrit et tes innombrables conseils qui ont tant permis d'améliorer la qualité de ce manuscrit que pour ta disponibilité et tes conseils sûrs à l'approche des dates butoirs qui m'ont évité bien des moments de panique. Merci Béatrice pour ton soutien constant depuis le début de ma carrière. Ton calme imperturbable a été mon refuge chaque fois que j'ai pu douter, merci de m'avoir permis de faire ce métier par la qualité de ton encadrement en thèse et pour tes conseils extrêmement pertinents de problèmes à aborder, je sais que je te dois d'avoir travaillé sur l'estimation de densité par les moindres carrés et sur les tests, je pense que ce mémoire rend compte de l'immense influence de ces thématiques sur ma recherche. Merci Pascal, tu es avec Béatrice l'autre constante de mes premières années de recherche, tu as toujours su en quelques phrases m'inviter à réfléchir à un problème très profond, du bootstrap pendant ma thèse à la méthode de Goldenschluger et Lepski plus récemment, ou prodigué des conseils pertinents pour prendre les décisions importantes de ma carrière. Merci Eric pour ta fantastique énergie, tes encouragements répétés et les opportunités que tu m'as poussées à saisir depuis que nous nous sommes rencontrés aux abords de São Paulo. C'est extrêmement stimulant de te côtoyer plus régulièrement depuis quelques mois et il me tarde que l'on commence à travailler sur des problèmes de recherche ensemble. Enfin, merci Sacha pour le rôle immense que tu joues depuis le début de ma carrière, par tes travaux de recherche si variés, précis et tout le temps en évolution et tes livres de référence.

Je voudrai également remercier ici mes collègues au laboratoire de Nice en général et de l'équipe de probabilité et statistiques en particulier. Je souhaite bien entendu remercier en premier lieu Patricia, depuis 4 ans que je suis ici, tu as été une source constante d'inspiration, par ta curiosité, ton expertise, ta générosité mathématiques, tu as été évidemment ma première collaboratrice ici et tu m'as appris tellement pendant les heures que nous avons passées à noircir des cahiers, blanchir des tableaux en défrichant de nouveaux domaines que je n'aurais pas assez d'un autre manuscrit pour te remercier assez. Merci ensuite à toi Roland, mon collaborateur et ami. Je suis vraiment heureux qu'on ait entamé cette collaboration sur les statistiques des processus en milieux aléatoire à laquelle je promets un bel avenir. Merci Raphael d'avoir insisté pour qu'on travaille sur le modèle de Bradley-Terry, ça a été une expérience très riche de se pencher avec nos cultures différentes sur ce joli problème. Merci François pour ta patience et ta motivation infatigable dans la gestion des recrutements en statistiques, ta gestion apaisante et exemplaire restera un modèle pour moi. Merci aussi pour les nombreuses références que tu sors toujours instantanément. Merci enfin Thomas, Christine, Cédric, Dieter, Sylvain pour tous les cafés, repas, discussions et débats qui ont animés mes quelques années ici et font du laboratoire un endroit si agréable à vivre.

Je remercie aussi toutes les personnes qui m'ont accueillies dans les laboratoires que j'ai fréquentés depuis ma thèse avant de venir à Nice. Muito obrigado Roberto por este ano no Rio, foi uma experiença unica de trabalhar ao seu lado, voce e realmente um pesquisador excepcional, curioso de tantas coisas. Merci à toi Antonio de m'avoir accueilli tout frais sorti de la thèse dans ton paradis pour chercheurs au NUMEC. C'est là que j'ai pu être initié à la mécanique statistique, à la théorie de l'information, aux techniques de simulation parfaite, aux neurosciences. Merci aussi à mes amis et collaborateurs au Brésil, Sandro, Christophe pour votre disponibilité, votre humour et les cervejinhas que nous avons partagés pendant ces années! Finalmente, obrigadão a você Daniel para sua generosidade, seul calma, sua ciência, para simplesmente ser você cada dia. Você cuidou de tudo quando a gente chegou no Brasil, dos papeis, dos apartamentos, do meu trabalho tambem. Merci enfin à toi Clémentine pour ces années à Toulouse ou tu m'as appris les coefficients de dépendance, pour m'avoir envoyé au Brésil et aider à rentrer en France dans les meilleures conditions. Tu es aussi là depuis le début de ma carrière, par tes encouragements constants, tes conseils avisés et tes questions pertinentes.

Merci aussi à tous mes collaborateurs que je n'ai pas encore salué : Sylvain pour ta gentillesse, ta puissance et ta vision mathématiques sans égale, ton calme à toute épreuve, Nelo pour ton enthousiasme, ta motivation ton courage dans les calculs interminables et ta ténacité pour corriger les détails de mes preuves toujours un peu fausses, Luc et Gabor for being so kind and generous with someone like me when you have done so much for statistics and mathematics in general, Magalie pour ta capacité à faire d'un tas d'idées en désordre une merveille mathématique, Nicolas pour ta connaissance universelle et ta vitesse supersonique et Aurélien pour ta disponibilité, tes conseils, et ton goût des solutions élégantes.

Merci enfin à toi Aurélie et à vous mes filles Charlotte et Clémence, vous êtes les amours de ma vie, mes rayons de soleil.

Contents

Foreword

These notes present an overview of my research in statistics, with some incursions in applied probability, since my PhD thesis in 2009 [14]. In this PhD, I studied least-squares density estimation by model selection, I got particularly interested in resampling penalization to build data-driven procedures when the data are not independent but only weakly dependent [10, 11, 13]. I also used resampling estimators to build confidence balls for the density in [12].

I studied with S. Arlot cross-validation schemes for model selection. Even if we started working on this project right after my PhD, we only finished our first paper recently [3]. The original project evolved a lot during this period and we were joined by N. Magalhães to extend the first results on model selection to linear estimator selection, this work is still in progress [17]. With N. Magalhães and P. Reynaud-Bouret, we already studied the optimal selection and minimal penalties for linear estimators [4].

Model selection was also central in my research during my Post-doct in São-Paulo where I studied discrete random fields with D. Y. Takahashi, a work that led to the articles [6, 9]. Our main motivation came from neuroscience where discrete random fields are used to represent brain activity. I continued working on problems related to neuroscience in Nice, particularly with A. Muzy and F. Grammont with whom I wrote a paper recently [16].

In São-Paulo, I also met A. Garivier and started to work with him on context tree estimation [18], which was natural since it was an important topic of research of my advisor there, A. Galves. In particular, I also worked on context trees, but on a probabilistic problem of perfect simulation with A. Galves' students S. Gallo and D. Y. Takahashi [7]. The idea of the paper [18] was to adapt the prediction approach from model selection theory to context tree estimation. Unfortunately, we did not publish this preprint and I won't present it in these notes.

Right after I got contracted by the CNRS in 2011 and before I even came to Nice, I was invited in IMPA in Rio de Janeiro by R. I. Oliveira. We started to work on subgaussian estimators and wrote a first paper [19] that wasn't published. I will only briefly discuss the material of this paper, mostly to motivate the new article [1]. This new paper deals with the problem of subgaussian estimation, it presents some estimators that strongly outperform the empirical mean when the distribution has heavy tails. It is interesting to notice that good subgaussian estimators of the expectation of a real-valued random variable are sufficient to build very general estimator-selection procedures. The problem of estimator selection in its general form was introduced by Y. Baraud [Bar11], I had the opportunity to write a short review on estimator selection for the Journées MAS 2014 in Toulouse, [5].

In Nice, and even already from Rio, I started to work with M. Fromont and P.

Reynaud-Bouret on aggregated tests. We discussed the problem with my former PhD advisor B. Laurent with whom we wrote a first paper [8] extending some results they recently had on two sample tests in a Poisson framework [FLRB11, FLRB13] to density estimation and some regression frameworks. This original paper led us to interesting questions on the links between aggregated tests and multiple testing. We started to investigate these links in the article [2] and we will hopefully continue in the following years.

Finally, still in Nice, I also had the opportunity to work with two colleagues R. Chetrite and R. Diel on the Bradley-Terry model. Our original motivations came from statistical questions as the estimation of the strength of the players, but we finally came up with nice probabilistic results on this models in random environment, when the strength are i.i.d. distributed [15]. It turns out that the probabilistic tools required are common with those I used to solve many model selection models, like concentration inequalities and control of the expectation of suprema of empirical processes.

These notes don't give the details of all the results I had, but hopefully provide some perspectives on my main topics of research. To keep the presentation as clear as possible, many mathematical details have been left appart and most results are presented in an informal way in simple examples rather than in their full generality. Rigorous results with many discussions and examples can be found in the articles, precise references are given in the notes. I also changed several notations of the original papers to keep a coherent presentation in the manuscript.

Bibliography

Published and accepted

[1] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. *Subgaussian mean estimators under heavy-tails.* To appear in Ann. Statist., (2016), preprint available in arXiv:1509.05845.

[2] M. Fromont, M. Lerasle, P. Reynaud-Bouret *Family Wise Separation Rates for multiple testing*, to appear in Ann. Statist. (2016), preprint available in https://hal.archives-ouvertes.fr/hal-01107321.

[3] S. Arlot and M. Lerasle. *Choice of V for V-fold cross-validation in least-squares density estimation,* to appear in J. Mach. Learn. Res., (2016), preprint available in arXiv:1210.5830.

[4] M. Lerasle, N. Magalhães, and P. Reynaud-Bouret. *Optimal kernel selection for density estimation,* to appear in High Dimensional Probability VII : The Cargese Volume (2016), preprint available in arXiv:1511.02112.

[5] M. Lerasle *Estimator Selection*, ESAIM Proc, Vol. 51 (2015), 232-245.

[6] M. Lerasle and D. Y. Takahashi. *Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields,* Bernoulli Vol. 22, No.1 (2016), 325-344.

[7] S. Gallo, M. Lerasle, D.Y. Takahashi *Markov approximation of chains of infinite order in the $\bar{d}$-metric*, Markov Process. Related Fields, Vol 19 (2013), 51-82.

[8] M. Fromont, B. Laurent, M Lerasle, and P. Reynaud-Bouret. *Kernel based tests with non-asymptotic bootstrap approaches for two-sample problems.* J. Mach. Learn. Res. W & CP, (2012), 23:23.1–23.23.

[9] M. Lerasle and D. Y. Takahashi. *An oracle approach for interaction neighborhood estimation in random fields.* Electron. J. Stat., 5 (2011), 534–571.

[10] M. Lerasle. *Optimal model selection in density estimation.* Ann. Inst. Henri Poincaré P & S, 48 (2011), 884–908.

[11] M. Lerasle *Optimal Model Selection for Stationary Data under various Mixing Conditions*, Ann. Statist., Vol. 39, No. 4 (2011), 1852-1877.

[12] M. Lerasle *Adaptive, non asymptotic Confidence Balls in density Estimation*, ESAIM P&S Vol. 16 (2010), 61-84.

[13] M. Lerasle *Adaptive density estimation of stationary β-mixing and τ-mixing processes.* Math. Meth. Statist., Vol. 18, No. 1 (2009), 59-83.

[14] M. Lerasle. *Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité.* PhD thesis, INSA Toulouse, (2009).

Submitted

[15] R. Chetrite, R. Diel, and M. Lerasle. *The number of potential winners in Bradley-Terry Models in random environment* Submitted, (2015), preprint available in arXiv:1509.07265.

[16] A. Muzy, M. Lerasle, F. Grammont, V.T. Doan and D.R.C. Hill *Discrete event system specification with parallel and stochastic simulation in the context of biological neural networks models.* Sumitted (2016).

In preparation

[17] S. Arlot, M. Lerasle, and N. Magalhães. *Selection of kernel estimators by cross-validation.* (2015), preprint available in Chapter 3 of N. Magalhães PhD dissertation <tel-01164581>.

Unpublished

[18] A. Garivier, M. Lerasle *Oracle properties of BIC estimators and slope heuristic in context tree estimation.* Preprint available in ArXiv:1111.2191, (2011).

[19] M. Lerasle and R. I. Oliveira. *Robust empirical mean estimators.* Preprint available in Arxiv:1112.3914, (2011).

Chapter 1

Estimation by selection of estimators

This first chapter starts with a presentation of model selection theory of Barron, Birgé and Massart [BBM99, BM97, BM01] in the least-squares density estimation framework. We also refer to Massart's book [Mas07] for an overview.
Section 1.1.1 fixes notation, Section 1.1.2 presents the basic ideas to derive asymptotically optimal inequalities and Section 1.1.3 a simple strategy to perform selection among large collections of models where oracle inequalities cannot be achieved. My goal is to provide some guidelines in simple examples that may be useful to understand the following sections. I also stress the central role of concentration of measure in model selection and show that the most elementary concentration tools are sufficient to perform model selection in rich collections of models. The reader familiar with model selection can have a quick look at Section 1.1.1 and jump to Section 1.2 where I start to present my own results.
I present in Section 1.2 my results on cross-validation [3, 17] and in Section 1.3 the results on linear estimator selection [4, 17]. I conclude in Section 1.4 by some applications of model selection in statistical physics models for neuroscience [6, 9].

1.1 Model selection for least-squares density estimation

1.1.1 Position of the problem

Let X be a random variable with distribution P on a measurable space $(\mathbb{X}, \mathcal{X}, \mu)$. Assume P is absolutely continuous with respect to (w.r.t.) μ and denote by f its density. We want to estimate f based on the observation of independent and identically distributed (i.i.d.) copies $X_1, \ldots, X_n$ of X. The performance of any estimator $\widehat{f}$ is measured by the quadratic risk $R_f(\widehat{f}) = \mathbb{E}\|f - \widehat{f}\|^2$, where $\|.\|$ denotes the usual $L^2(\mu)$-norm. For any probability measure Q and any real-valued function $g \in L^1(Q)$, let $Qg = \int g dQ$ and let P_n be the empirical measure based on the observations, that is $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. The loss $\|f - \widehat{f}\|^2$ of $\widehat{f}$ satisfies

$$\|f - \widehat{f}\|^2 - \|f\|^2 = \|\widehat{f}\|^2 - 2P\widehat{f} = P\gamma(\widehat{f}) \ ,$$

where, for any $g \in \mathcal{L}^2(\mu)$ and any $x \in \mathbb{X}$, $\gamma(g, x) = \|g\|^2 - 2g(x)$ denotes the least-squares contrast. We are given a (finite, but possibly growing with n) collection of finite-dimensional linear subspaces $(S_m)_{m\in\mathcal{M}_n}$ of $L^2(\mu)$ (the models), which are used to produce a collection of estimators

$$\forall m \in \mathcal{M}_n, \qquad \widehat{f}_m = \arg\min_{g\in S_m} P_n\gamma(g) \ .$$

These estimators are easy to compute, given an orthonormal basis $(\varphi_i)_{i\in\mathcal{I}_m}$ of S_m, we have

$$\widehat{f}_m = \sum_{i\in\mathcal{I}_m} (P_n\varphi_i)\,\varphi_i \ . \tag{1.1}$$

They are usually called *projection estimators* to emphasize that they actually estimate the orthogonal projection of f onto S_m

$$f_m = \sum_{i\in\mathcal{I}_m} (P\varphi_i)\,\varphi_i \ .$$

In this presentation, we shall focus on spaces of histograms: let $\mathcal{A}_m$ be a finite partition of measurable subsets of $\mathbb{X}$. The histogram space S_m associated to $\mathcal{A}_m$ is the linear span of the functions $(\mathbf{1}_A)_{A\in\mathcal{A}_m}$. The projection estimator on S_m is

$$\widehat{f}_m = \sum_{A\in\mathcal{A}_m} \frac{P_n\mathbf{1}_A}{\mu(A)}\mathbf{1}_A \ .$$

To illustrate our result, we specify three families of histograms on $\mathbb{X} = [0,1]$ endowed with its Borel σ-algebra and the Lebesgue measure μ.
The collection of regular histograms $(S_d)_{d\in\mathcal{M}_n^r}$ is the collection of histograms based on the regular partitions $(\mathcal{A}_d^r)_{d\in\mathcal{M}_n^r}$, where $\mathcal{M}_n^r = \{1,\dots,n\}$ and, for all $d\in\mathcal{M}_n^r$,

$$\mathcal{A}_d^r = \left\{ \left[\frac{k}{d},\frac{k+1}{d}\right)_{k=0,\dots,d-2}, \left[1-\frac{1}{d},1\right] \right\} \ .$$

For $d\in\mathcal{M}_n^r$, the histogram associated to the partition $\mathcal{A}_d^r$ is therefore

$$\widehat{f}_d = d\sum_{A\in\mathcal{A}_d^r} (P_n\mathbf{1}_A)\mathbf{1}_A \ .$$

The 1-breakpoint collection $(S_m)_{m\in\mathcal{M}_n^{1b}}$ is the collection of histograms associated to the partitions $(\mathcal{A}_m)_{m\in\mathcal{M}_n^{1b}}$, where $\mathcal{M}_n^{1b}$ is the collection of all triplets $m=(c,d_1,d_2)$ with $c\in\{1/n,\dots,1-1/n\}$ and $d_1\in[1,nc]\cap\mathbb{N}$, $d_2\in[1,n(1-c)]\cap\mathbb{N}$ and for any $m\in\mathcal{M}_n^{1b}$, $\mathcal{A}_m$ is the partition of $[0,1]$ with bin sizes equal to c/d_1 on $[0,c]$ and $(1-c)/d_2$ on $(c,1]$. If there exist integers $d\le n$ and $k\le d$ such that $c=k/d$, $d_1=k$ and $d_2=d-k$, $\mathcal{A}_m$ is the regular partition $\mathcal{A}_d^r$.
Finally, let $\phi=\mathbf{1}_{[0,1]}$ and $\psi=\mathbf{1}_{[0,1/2]}-\mathbf{1}_{(1/2,1]}$ and, for any $j\ge 0$ and $k\in\mathbb{Z}$, let $\psi_{j,k}=2^{j/2}\psi(2^j.-k)$. The collection $(\phi,(\psi_{j,k})_{j\in\mathbb{N},k\in\{0,\dots,2^j-1\}})$ defines the Haar basis and the last collection of histograms of interest is the collection $(S_m)_{m\in\mathcal{M}_n^H}$, where, denoting by

$$M_n = \left\{ (j,k),\ \text{s.t. } j\in\left\{0,\dots,\left\lfloor \log_2\left(\frac{n}{\log n}\right)\right\rfloor\right\},\ k\in\left\{0,\dots,2^j-1\right\} \right\} \ ,$$

$\mathcal{M}_n^H$ is the collection of all subsets $m\subset M_n$ and for any $m\in\mathcal{M}_n^H$, S_m is the linear span of ϕ and all $(\psi_{j,k})_{(j,k)\in m}$.

The main task in model selection is to select $\widehat{m} \in \mathcal{M}_n$ such that the risk of the final estimator $\widehat{f}_{\widehat{m}}$ is as small as possible. More precisely, the goal is to prove that, for some constant $C \geq 1$,

$$R_f(\widehat{f}_{\widehat{m}}) \leq C \inf_{m \in \mathcal{M}_n} R_f(\widehat{f}_m) \ . \tag{1.2}$$

This inequality is called an *oracle inequality* because the risk of the estimator $\widehat{f}_{\widehat{m}}$ is comparable to the one that an "oracle", who knows in advance the risks of all $(\widehat{f}_m)_{m \in \mathcal{M}_n}$, would have chosen to achieve $\inf_{m \in \mathcal{M}_n} R_f(\widehat{f}_m)$. We will also sometimes try to improve this inequality and prove that, for some $\epsilon_n \to 0$,

$$R_f(\widehat{f}_{\widehat{m}}) \leq (1+\epsilon_n) \inf_{m \in \mathcal{M}_n} R_f(\widehat{f}_m) + \Delta(\epsilon_n, \mathcal{M}_n) \ . \tag{1.3}$$

Inequality (1.2) is actually better than an oracle inequality when the remainder term $\Delta(\epsilon_n, \mathcal{M}_n) = o(\inf_{m \in \mathcal{M}_n} R_f(\widehat{f}_m))$, in this case, the oracle inequality is called *asymptotically optimal.* The term "oracle" is due to Donoho and Johnston [DJ94] and is now commonly accepted in various contexts including model selection [Mas07], selection of Parzen's estimators [GL11], aggregation [RT07] or thresholding [DJKP96, GN09].

1.1.2 Selection of regular histograms by deterministic penalization

To perform the selection of $\widehat{m}$, a natural idea is to estimate the loss $P\gamma(\widehat{f}_m)$ by a data-driven criterion $\mathcal{C}_n(m)$ and choose

$$\widehat{m} = \arg\min_{m \in \mathcal{M}_n} \mathcal{C}_n(m) \ .$$

The basic estimator $P_n\gamma(\widehat{f}_m)$ for $P\gamma(\widehat{f}_m)$ has poor performances in general since the same data are used to build the estimator $\widehat{f}_m$ and to estimate its loss. This is why the selection step usually requires some refined statistical approach such as penalization [Aka70, Mal73], resampling [Efr83] or cross-validation [Rud82]. In the remaining part of this chapter, we will present these different approaches.

Let us illustrate the solution provided by penalization in an elementary example. In least-squares density estimation, for regular histograms, the estimator $P_n\gamma(\widehat{f}_m)$ of the ideal criterion $\mathcal{C}_{\mathrm{id}}(m) = P\gamma(\widehat{f}_m)$ satisfies

$$\|f\|^2 + \mathbb{E}P_n\gamma(\widehat{f}_d) = \|f\|^2 - \|f_d\|^2 - \frac{d - \|f_d\|^2}{n} = \|f - f_d\|^2 - \frac{d - \|f_d\|^2}{n} \ , \tag{1.4}$$

where f_d is the expectation of $\widehat{f}_d$:

$$f_d = d \sum_{A \in \mathcal{A}_d^r} (P\mathbf{1}_A)\mathbf{1}_A \ .$$

Neglecting the term $\|f_d\|^2/n \leq \|f\|^2/n = O(1/n)$ in (1.4), both $\|f - f_d\|^2$ and $-d/n$ are non-increasing functions of d so the minimizer of the expectation of the empirical loss $P_n\gamma(\widehat{f}_d)$ with respect to d is the largest value of $d = n$. Since

$$\forall d \in \mathcal{M}_n^r, \qquad \mathbb{E}\|\widehat{f}_d - f\|^2 = \|f_d - f\|^2 + \frac{d}{n} - \frac{\|f_d\|^2}{n} \ ,$$

for $d = n$, one has $\mathbb{E}\|\widehat{f}_d - f\|^2 \geq 1 + O(1/n)$. If some estimator in the collection $(\widehat{f}_d)_{d\in\mathcal{M}_n^r}$ is consistent, this proves that minimizing the empirical loss does not yield to oracle inequalities.
On the other hand, the same computations show that

$$\mathbb{E}\left[P_n\gamma(\widehat{f}_d) + 2\frac{d}{n}\right] = \mathbb{E}\|\widehat{f}_d - f\|^2 - \|f\|^2 + \frac{\|f_d\|^2}{n} \ .$$

Hence, the *penalized* empirical loss

$$\mathcal{C}_{\text{pen}}(d) = P_n\gamma(\widehat{f}_d) + \text{pen}(d), \qquad \text{with} \qquad \text{pen}(d) = 2\frac{d}{n} \ ,$$

is essentially an unbiased estimator of the loss. To prove that a minimizer of $\mathcal{C}_{\text{pen}}$ satisfies an oracle inequality, we prove that, for any $d \in \mathcal{M}_n^r$, $P_n\gamma(\widehat{f}_d) + \text{pen}(d)$ is close to $P\gamma(\widehat{f}_d)$. This is where we use *concentration inequalities.* These provide, for each d and n, functions $\Delta_{d,n}$ such that,

$$\forall x > 0, \qquad \mathbb{P}\left(|P_n\gamma(\widehat{f}_d) + \text{pen}(d) - P\gamma(\widehat{f}_d)| \leq \Delta_{d,n}(x)\right) \geq 1 - e^{-x} \ .$$

The uniform control over all $d \in \mathcal{M}_n^r$ is obtained using a union bound. At the end, one can prove that there exists a constant C which is independent of n and d such that

$$\mathbb{E}\left[\sup_{d\in\mathcal{M}_n^r}\left(\left|(P_n - P)\gamma(\widehat{f}_d) + \text{pen}(d)\right| - \Delta_{d,n}(\log n)\right)_+\right] \leq \frac{C}{n} \ , \tag{1.5}$$

where the $\log n$ term comes from $\log|\mathcal{M}_n^r|$. More precisely, (1.5) is obtained using two concentration inequalities, the first is a weak version of Bernstein's inequality which says that, for any bounded function g,

$$\forall x > 0, \qquad \mathbb{P}\left(|(P_n - P)g| > \sqrt{\frac{2Pg^2x}{n}} + \frac{\|g\|_\infty x}{3n}\right) \leq 2e^{-x} \ . \tag{1.6}$$

The second one is a concentration inequality for totally degenerate U-statistics derived from [HRB03]. The reason is that all our quantities of interest can be decomposed as a sum of a centered empirical mean and a totally degenerate U-statistics of order 2. For example, one can show that, for regular histograms

$$\begin{aligned} &P_n\gamma(\widehat{f}_d) + \text{pen}(d) - P\gamma(\widehat{f}_d) \\ &\quad = \left(1 - \frac{1}{n}\right)(P_n - P)\gamma(f_d) + \frac{d}{n^2}\sum_{1\leq i\neq j\leq n}\sum_{A\in\mathcal{A}_d^r}(\mathbf{1}_{X_i\in A} - P\mathbf{1}_A)(\mathbf{1}_{X_j\in A} - P\mathbf{1}_A) \ . \end{aligned} \tag{1.7}$$

Using these concentration inequalities, we show that $(P_n - P)\gamma(\widehat{f}_d) + \text{pen}(d)$ is bounded from above by $\Delta_{d,n}(x) = \frac{C}{n}(\sqrt{dx} + \frac{dx}{n})$. Now, using (1.5) and elementary algebraic computations, we get that the minimizer $\hat{d}$ of $\mathcal{C}_{\text{pen}}(d)$ satisfies, for any $\epsilon \in (0, 1)$,

$$(1 - \epsilon)\mathbb{E}\|f - \widehat{f}_{\hat{d}}\|^2 \leq (1 + \epsilon)\inf_{d\in\mathcal{M}_n^r}\left\{\mathbb{E}\|f - \widehat{f}_d\|^2 + \frac{1}{\epsilon^3}\Delta_{d,n}(\log n)\right\} + \frac{C}{n} \ . \tag{1.8}$$

This is an oracle inequality, see (1.2), if $\inf_{d\in\mathcal{M}_n^r}\mathbb{E}\|f-\widehat{f}_d\|^2\geq\log(n)/n$ and an asymptotically optimal inequality, see (1.3), if $\inf_{d\in\mathcal{M}_n^r}\mathbb{E}\|f-\widehat{f}_d\|^2\gg\log(n)/n$.

Let us now analyse the previous sketch of proof to emphasize its main features.

- First, the penalty term $\mathrm{pen}(m)$ should be a data-driven quantity satisfying
$$\mathrm{pen}(m)\geq\mathbb{E}\left[(P-P_n)\gamma(\widehat{f}_m)\right]\ ,$$
to get an oracle inequality. Moreover, this upper bound should be sharp for the inequality to be asymptotically optimal.
- Second, the collection $\mathcal{M}_n$ should not be too rich for the remainder term $(\Delta_{d,n}(\log(|\mathcal{M}_n^r|))\approx\sqrt{d(\log n)/n}$ in the example) to be bounded from above, or even negligible compared to the risk $(\|f\|^2+\mathbb{E}\left[P\gamma(\widehat{f}_d)\right]\geq d/n$ in the example) for a reasonable range of $m\in\mathcal{M}_n$. Typically, this holds when $\log(|\mathcal{M}_n|)$ is smaller than a power of $\log n$, which is the case for the collections $\mathcal{M}_n^r$ and $\mathcal{M}_n^{1d}$ but not for the collection $\mathcal{M}_n^H$ which should be handled using a different strategy that I will sketch in the following subsection.

1.1.3 Selection of a subset of the Haar basis

As the collection $\mathcal{M}_n^H$ is very large, $\log_2(|\mathcal{M}_n^H|)\approx\frac{n}{\log n}$, one cannot use the rough union bound among all the models of the previous section. Recall that we want to compute a penalty term $\mathrm{pen}(m)$ such that the estimator

$$\widehat{f}_{\widehat{m}},\qquad\text{where}\qquad\widehat{m}\in\arg\min_{m\in\mathcal{M}_n^H}\left\{P_n\gamma(\widehat{f}_m)+\mathrm{pen}(m)\right\}\tag{1.9}$$

satisfies (1.2). Barron, Birgé and Massart [BBM99, BM97, BM01] show that this is not possible and that there is always a logarithmic loss. Therefore, the goal is to prove that there exists a constant $C>0$ such that

$$R_f(\widehat{f}_{\widehat{m}})\leq C(\log n)\inf_{m\in\mathcal{M}_n}R_f(\widehat{f}_m)\ .\tag{1.10}$$

To achieve (1.10), we look for penalties of the form $\mathrm{pen}(m)=\lambda^2|m|$ for some $\lambda>0$. The first reason is that the minimization problem defining $\widehat{m}$, a priori intractable, is actually easily solved here. Actually,

$$\forall m\in\mathcal{M}_n^H,\qquad P_n\gamma(\widehat{f}_m)+\lambda^2|m|=\sum_{(j,k)\in m}(\lambda^2-(P_n\psi_{j,k})^2)\ ,$$

thus, the minimizer is the set $\widehat{m}=\{(j,k),\text{ s.t. }|P_n\psi_{j,k}|>\lambda\}$ and the selected estimator is the *hard thresholded estimator* of [DJKP96]

$$\widehat{f}_{\widehat{m}}=1+\sum_{(j,k)\in M_n}(P_n\psi_{j,k}\mathbf{1}_{|P_n\psi_{j,k}|>\lambda})\psi_{j,k}\ .\tag{1.11}$$

Moreover, by definition of $\widehat{m}$,

$$\forall m\in\mathcal{M}_n^H,\qquad P_n\gamma(\widehat{f}_{\widehat{m}})+\lambda^2|\widehat{m}|\leq P_n\gamma(\widehat{f}_m)+\lambda^2|m|\ ,$$

therefore,

$$P\gamma(\widehat{f}_{\widehat{m}}) \le P\gamma(\widehat{f}_m) + \Big((P_n - P)\gamma(\widehat{f}_m) + \lambda^2|m|\Big) + \Big((P - P_n)\gamma(\widehat{f}_{\widehat{m}}) - \lambda^2|\widehat{m}|\Big) \ , \tag{1.12}$$

Now, we decompose,

$$\begin{aligned}(P-P_n)\gamma(\widehat{f}_{\widehat{m}}) &= (P-P_n)(\gamma(\widehat{f}_{\widehat{m}}) - \gamma(f_{\widehat{m}})) + (P-P_n)\gamma(f_{\widehat{m}})\\ &= 2\sum_{(j,k)\in\widehat{m}}((P_n-P)\psi_{j,k})^2 + (P-P_n)\gamma(f_{\widehat{m}}) \ . \end{aligned} \tag{1.13}$$

The trick to handle these large collections is to use a union bound to control all the differences $|(P_n - P)\psi_{j,k}|$ for all (j,k) in M_n instead of a union bound for all models $\mathcal{M}_n^H$ as in the previous subsection because the cardinality of M_n is much smaller. As $\|\psi_{j,k}\|_\infty \le 2^{j/2} \le \sqrt{\frac{n}{\log n}}$ and $\mathbb{E}\psi_{j,k}^2 = 2^j\mathbb{E}\mathbf{1}_{[k/2^j,(k+1)/2^j]} \le \|f_{M_n}\|_\infty$, inequality (1.6) gives, for any $x > 0$, $\mathbb{P}(\Omega(x)) \ge 1 - e^{-x}$, where, for some absolute constant $C > 0$,

$$\Omega(x) = \left\{\forall (j,k) \in M_n, \qquad |(P_n - P)\psi_{j,k}| \le C\frac{\|f_{M_n}\|_\infty + x + \sqrt{\log 2n}}{\sqrt{n}}\right\} \ .$$

Suppose that some (deterministic) upper bound L_n on $\|f_{M_n}\|_\infty$ is available to the statistician and define $\ell(x) = \sqrt{2}C(L_n + \sqrt{\log 2n} + x)$. It follows from (1.12) and (1.13) that, if the threshold $\lambda = \ell(x)/\sqrt{n}$ and the penalty $\mathrm{pen}(m) = \lambda^2|m|$, the estimator (1.9) satisfies, on $\Omega(x)$,

$$\forall m \in \mathcal{M}_n^H, \qquad \|\widehat{f}_{\widehat{m}} - f\|^2 \le \|\widehat{f}_m - f\|^2 + \ell(x)^2\frac{|m|}{n} \le \|f_m - f\|^2 + 2\ell(x)^2\frac{|m|}{n} \ .$$

Integrating this inequality gives finally that, when, $\|f_{M_n}\|_\infty \le L_n$ and the threshold $\lambda = C(L_n + \sqrt{\log n})/\sqrt{n}$, the hard thresholded estimator (1.11) satisfies

$$R_f(\widehat{f}_{\widehat{m}}) \le C(L_n^2 + \log n)\inf_{m\in\mathcal{M}_n} R_f(\widehat{f}_m) \ . \tag{1.14}$$

Interestingly, the key ingredients to perform this analysis are quite simple.

- We only had to estimate the means $P\psi_{j,k}$ of real-valued random variables by the empirical means $P_n\psi_{j,k}$.
- Then, to obtain the pseudo oracle bound (1.14), we only have to bound above the deviations of these estimators which derives via Berstein's inequality from a control of $P\psi_{j,k}^2$ by a constant and of its infinite norm $\|\psi_{j,k}\|_\infty$ by $\sqrt{n/\log n}$. However, as many procedures for least-squares density estimation, see [Bir14], this one depends on an upper bound L_n of $\|f\|_\infty$ that is in general *not available to the statistician.*

1.2 Cross-validation estimators for model selection

1.2.1 Random procedures for model selection

The selection among regular histograms is easily performed using the penalty $\mathrm{pen}(d) = 2d/n$. However, when dealing with the 1-breakpoint histograms $m = (c, d_1, d_2)$, one can check that

$$\frac{1}{2}\mathbb{E}\left[(P_n-P)\gamma(\widehat{f}_m)\right]=\frac{F(c)}{c}\frac{d_1}{n}+\frac{1-F(c)}{1-c}\frac{d_2}{n}-\frac{\|f_m\|^2}{n}\ ,$$

where F denotes the c.d.f. of P. This expectation cannot be sharply bounded above by deterministic penalties in general, in fact, it is only possible when the 1-breakpoint histogram is close to a regular histogram, that is, when $d_1/c\approx d_2/(1-c)$. However, these histograms are preferred to regular histograms when the regularity of f is assumed to be very different on an interval $[0,c]$ than on $[c,1]$ and an oracle has different bin sizes on $[0,c]$ than on $[c,1]$, that is d_1/c and $d_2/(1-c)$ are not comparable. In that case, the obvious upper bound

$$\frac{d_1}{cn}+\frac{d_2}{(1-c)n}$$

leads to oracle inequalities, using similar arguments as those of the previous section, but they are not asymptotically optimal. On the other hand, random strategies can be used to obtain sharper bounds. In this example, the c.d.f. F can be estimated by the empirical c.d.f. $\widehat{F}$, the difference can be uniformly bounded above by the DKW inequality [Mas90] and we would obtain an asymptotically optimal oracle inequality. But the problem at hand is also sufficiently simple to study more general random procedures that are widely used in a variety of practical problems without strong non-asymptotic guarantees. For example, this is the case of resampling estimators of the penalty suggested by [Efr83, Arl09]. Given a vector $(W_1,\dots,W_n)$ of non-negative random variables independent of the observation such that $\sum_{i=1}^n W_i=n$, one defines the resampling empirical process for any function g by $P_n^W g=\frac{1}{n}\sum_{i=1}^n W_i g(X_i)$. Efron's resampling heuristic [Efr79] states that the distribution of any functional $F(P_n,P)$ should be close to the one of its resampling counterpart $F(P_n^W,P_n)$ conditionally on the data. In particular,

$$\mathbb{E}\left[(P-P_n)\gamma(\widehat{f}_m)\right]\approx\mathrm{pen}_W(m)=C_W\mathbb{E}_W\left[(P_n-P_n^W)\gamma(\widehat{f}_m^W)\right]\ , \tag{1.15}$$

where $\mathbb{E}_W$ denotes the expectation conditionally on the data and $\widehat{f}_m^W$ is the resampling estimator of f_m, $\widehat{f}_m^W=\sum_{i\in\mathcal{I}_m}(P_n^W\varphi_i)\varphi_i$ and C_W is a constant. These penalties have been analyzed in my PhD article [10] under the assumption that the weights W are exchangeables, meaning that their distribution is invariant under any permutation of the coordinates. In [3, Lemma 1], we show that resampling penalties based on an exchangeable resampling vector are in fact particular cases of V-fold cross-validation penalties. Therefore, I will not discuss the performances of resampling penalties here and rather present cross-validation procedures before giving the results.

1.2.2 Cross-validation procedures

Cross-validation is one of the most classical methods of model selection. It was introduced in the 70's [All74, Gey76, Sto74] and extended to density estimation independently by Rudemo [Rud82] and Bowman [Bow84]. These methods require few assumptions on the unknown density f in general [Loa99]. The asymptotic properties of cross-validation estimators are well known, see for example [Hal83,

HM87, ST87, Sto84] for proofs of its consistency, its asymptotic normality and its efficiency. The original cross-validation schemes have then been extended to improve practical performances, among others, let us mention for example “biased" cross-validation [ST87], “corrected" cross-validation of Burman [Bur89], “trimmed" cross-validation [FK92], “modified" cross-validation [Stu92] or indirect cross-validation [SHS10]. See also the survey [AC10] for a recent overview on cross-validation for model selection.
The validation principle in statistics is an alternative approach to resampling for estimating quantities like $F(P_n, P)$ for some known functionals F, typically the risk or the loss $P\gamma(\widehat{f}_m)$. The idea is to split the data in two pieces, a training set $(X_i)_{i\in T}$ and a validation set $(X_i)_{i\in T^c}$, where both $T \subset \{1,\ldots,n\}$ and $T^c = \{1,\ldots,n\} \setminus T$ are non-empty. These sets are used to build the training empirical process $P_T g = \frac{1}{|T|}\sum_{i\in T} g(X_i)$ and the validation empirical process $P_{T^c} g = \frac{1}{|T^c|}\sum_{i\in T^c} g(X_i)$ which are used to estimate $F(P_n, P)$ by the *hold-out* estimator

$$\widehat{F}^{ho}_T = F(P_T, P_{T^c}) \ .$$

In practice, hold-out estimators are dependent on the choice of T and to reduce this variability, the cross-validation idea is to use a collection of training sets $T \in \mathcal{E}$ and compute the *cross-validation estimator*

$$\widehat{F}^{cv}_{\mathcal{E}} = \frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}} \widehat{F}^{ho}_T \ .$$

Hereafter, I will call *cross-validation scheme* a collection $\mathcal{E}$ of training sets. In our analysis, we discuss the performances of the following cross-validation schemes.

1. The trivial collection $\mathcal{E}^{ho} = \{T\}$, the corresponding cross-validation estimator $\widehat{F}^{cv}_{\mathcal{E}^{ho}} = \widehat{F}^{ho}_T$ is the hold-out estimator.

2. The complete collection $\mathcal{E}_p$ of all subsets $T \subset \{1,\ldots,n\}$ with cardinality $n-p \in \{1,\ldots,n-1\}$. The corresponding cross-validation estimator is known as the *leave-p-out* estimator.

3. Let $B \in \mathbb{N}\backslash\{0\}$ and let $T_1,\ldots,T_B$ denote B subsets of $\mathcal{E}_p$ chosen independently and uniformly, independently of the data. The collection $\mathcal{E}^{mc}_{p,B} = \{T_1,\ldots,T_B\}$ is our third collection of interest, the corresponding cross-validation procedures are called Monte-Carlo cross-validation [PC84].

4. Let V be a divisor of n and let $T^c_1,\ldots,T^c_V$ denote a deterministic partition of $\{1,\ldots,n\}$ of sets of cardinality n/V. The collection $\mathcal{E}^{vf}_V = \{T_1,\ldots,T_V\}$ provides the V-fold cross-validation estimators [BFOS84].

A common heuristic in cross-validation is that, the larger $\mathcal{E}$, the better the cross-validation estimator should be, but the harder it is to compute in practice. Therefore, the following questions on the comparison of cross-validation schemes naturally arise.

- Can we compare performances of estimators built with different cross-validation schemes? For example, can one prove that leave-p-out estimators are better than hold-out estimators built with one T with cardinality $n-p$? Can we compare these with intermediate estimators like Monte-Carlo cross-validation estimators, which give hold-out when $B = 1$ and leave-p-out when $B \to \infty$? Is it better to perform deterministic splits as in V-fold cross-validation or random ones like Monte-Carlo with $p = n/V$ and $B = V$?

- Is there a cross-validation scheme with a reasonable cardinality ensuring performances close to the optimal? For example, does V-fold cross-validation performs as well as leave-p-out for a reasonable value of V, like $V = 5$ or $V = 10$?

For least-squares density estimation, there are two natural functionals one may estimate to build a model selection criterion. The first and most classical one is the loss itself $P\gamma(\widehat{f}_m)$. The corresponding cross-validation estimators are called cross-validation criteria and are defined by

$$\mathcal{C}^{cv}_{\mathcal{E}}(m) = \frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}} P_{T^c}\gamma(\widehat{f}^T_m), \qquad \text{where} \qquad \widehat{f}^T_m = \sum_{i\in\mathcal{I}_m}(P_T\varphi_i)\varphi_i \ . \tag{1.16}$$

The second one is the *ideal penalty* as defined by Arlot [Arl09] which is equal to

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{f}_m) \ .$$

It is called ideal because the empirical loss penalized by the ideal penalty is the ideal criterion $\mathcal{C}_{\mathrm{id}}(m) = P\gamma(\widehat{f}_m)$. The associated cross-validation estimators are called cross-validation penalties

$$\mathrm{pen}_{\mathcal{E}}(m) = \frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}}(P_T - P_{T^c})\gamma(\widehat{f}^T_m) \ .$$

These are used to define the cross-validation penalized criteria, defined for a constant $C > 0$ by

$$\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E},C}(m) = P_n\gamma(\widehat{f}_m) + C\mathrm{pen}_{\mathcal{E}}(m) \ .$$

In any case, we study the performance of the estimator

$$\widehat{f}_{\widehat{m}} \qquad \text{where} \qquad \widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \mathcal{C}(m) \ . \tag{1.17}$$

Notice that cross-validation is only used to *select* the estimators, those are built using all the data-set.

What makes a general study of cross-validation schemes difficult is usually the dependence between hold-out estimators $\widehat{F}^{ho}_T$ for different T. Least-squares density estimation is a particularly nice framework because, as the ideal penalty for regular-histograms selection (1.7), all cross-validation criteria and penalization can be decomposed into sums of centered empirical means and totally degenerate U-statistics of order 2. This is central in our analysis. In particular, oracle inequalities derive from the same concentration inequalities as those of the preceding section.

1.2.3 Oracle properties of cross-validation selectors

Our first series of results on cross-validation algorithms concern V-fold penalization. The main reason is that several cross-validation schemes lead to V-fold penalized criteria $\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}^{vf}_V,C}$, for different values of V and C. In [3, Lemma 1], we prove that all leave p-out penalties $\mathrm{pen}_{\mathcal{E}_p}$ and all resampling penalties pen_W based on exchangeable resampling weights W (see (1.15) and [Arl09] for a general definition) are n-fold penalties $\mathrm{pen}_{\mathcal{E}_n^{vf}}$ multiplied by some constant C and that both leave-p-out criteria $\mathcal{C}_{\mathcal{E}_p}$ and V-fold criteria $\mathcal{C}_{\mathcal{E}_V^{vf}}$ are particular instances of $\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}_V^{vf},C}$. The only differences between these criteria are the parameters V and C to use. Therefore, by studying general V-fold penalized criteria $\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}_V^{vf},C}$ for all values of C and V, we study a much broader class of cross-validation and resampling procedures.

We prove in [3, Theorem 5] an oracle inequality for the estimator (1.17) with $\widehat{m}$ minimizing $\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}_V^{vf},C}$, that is valid for any value of V and any constant $C > (V-1)/2$ in front of the penalty. More precisely, we show that, under some classical assumptions on the collection of models (as for example those of [BBM99]), for any $\epsilon \in (0,1)$,

$$\frac{1-\delta_- - \epsilon}{1+\delta_+ + \epsilon} R_f(\widehat{f}_{\widehat{m}}) \leq \inf_{m \in \mathcal{M}_n} R_f(\widehat{f}_m) + \Delta\left(\epsilon, f, \log(|\mathcal{M}_n|)\right) \ , \tag{1.18}$$

where $\delta = 2(\frac{C}{V-1} - 1)$, $\delta_- = (-\delta) \vee 0$ and $\delta_+ = \delta \vee 0$. The term δ in this inequality measures the bias of the criterion. More precisely, recall that $\mathcal{C}_{\mathrm{id}}(m) = P\gamma(\widehat{f}_m)$. The bias of the criterion is the difference $\mathbb{E}\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}_V^{vf},C}(m) - \mathbb{E}\mathcal{C}_{\mathrm{id}}(m)$ and one can show that

$$\mathbb{E}\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}_V^{vf},C}(m) - \mathbb{E}\mathcal{C}_{\mathrm{id}}(m) = \delta \mathbb{E}\left\|f_m - \widehat{f}_m\right\|^2 \ . \tag{1.19}$$

The inequality (1.18) is only interesting when $C > (V-1)/2$ and it is then an oracle inequality if ϵ is small enough. Moreover, it yields to an asymptotically optimal oracle inequality if $C = V - 1$ and $\Delta\left(\epsilon, f, \log(|\mathcal{M}_n|)\right) = o\left(\inf_{m \in \mathcal{M}_n} R_f(\widehat{f}_m)\right)$. These results are coherent with minimal penalty results (see Section 1.3.2 or [10, Theorem 2.2]).

The oracle inequality (1.18) is new for the following reasons. First, it is, to our knowledge, the first non-asymptotic oracle inequality proved for V-fold cross-validation penalization in least-squares density estimation. Second, it is, in any framework, the first oracle inequality valid for V-fold methods that holds for *any value of* V and in particular, that provides an asymptotically optimal oracle inequality without loss when V (i.e. the number of splits) increases. Previous bounds [Arl08] were obtained by concentration of the hold-out estimators and a union bound. They were thus valid only for small V and were deteriorating with V.

The oracle inequality is asymptotically optimal only when the bias is asymptotically null, which is coherent with the following comments made on cross-validation in other frameworks. First, corrected V-fold cross-validation criteria, which are unbiased, yield asymptotically optimal oracle inequalities, whatever the value of V. Second, V-fold criteria $\mathcal{C}_{\mathcal{E}_V^{vf}}(m)$, which are equal to $\mathbb{E}\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}_V^{vf},C}(m)$ with $C = V - 1/2$ satisfy $\delta = 2(\frac{C}{V-1} - 1) = \frac{1}{V-1}$ and are therefore only asymptotically optimal when $V \to \infty$. The drawback of this nice performance is that oracle bounds cannot be used to distinguish criteria with the same bias based on different values of V, and more generally between different cross-validation schemes. This was expected, since an oracle bound is always only an upper bound on the performance of a scheme; besides, the bias of a method is the only parameter that matters at first order in an oracle inequality and the bias (1.19) of $\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}^{vf}_V,C}$ can always be chosen equal to 0 for any V using $C = V - 1$.

1.2.4 Comparing model selection criteria : a heuristic with application to cross-validation schemes

If all V-fold cross-validation methods have the same performances, one should choose $V = 2$ to minimize the computation time of the estimators. However, one finds in the literature, see for example [BS92, HTF09] the advice that choosing $V = 5$ or $V = 10$ is much better than $V = 2$ from a practical point of view. To understand why, we present in [3, Section 4] a heuristic to differentiate two criteria of model selection. Let $m^* \in \arg\min_{m\in\mathcal{M}_n} R_f(m)$ (the oracle model) and let $\mathcal{C}_1$ and $\mathcal{C}_2$ denote two criteria, that is data-driven functions $\mathcal{M}_n \to \mathbb{R}$. Assume that both criteria have the same bias, i.e. that there exists a “constant" $A \in \mathbb{R}$ (independent of m), satisfying $\mathbb{E}\mathcal{C}_1(m) = \mathbb{E}\mathcal{C}_2(m) + A$ for any $m \in \mathcal{M}_n$. For any $\epsilon > 0$, denote by $\mathcal{M}^*_\epsilon = \{m \in \mathcal{M}_n,\ \text{s.t.}\ R_f(m) \le (1+\epsilon)R_f(m^*)\}$. The heuristic goes as follows : $\mathcal{C}_1$ is better than $\mathcal{C}_2$ if, for any “small" $\epsilon > 0$,

$$\mathbb{P}\left(\widehat{m}_1 \in \mathcal{M}^*_\epsilon\right) > \mathbb{P}\left(\widehat{m}_2 \in \mathcal{M}^*_\epsilon\right) \ ,$$

where $\widehat{m}_i = \arg\min_{m\in\mathcal{M}_n} \mathcal{C}_i(m)$. Now

$$\begin{aligned}\mathbb{P}\left(\widehat{m}_i \in \mathcal{M}^*_\epsilon\right) &= \mathbb{P}\left(\forall m \notin \mathcal{M}^*_\epsilon,\ \inf_{m'\in\mathcal{M}^*_\epsilon} \mathcal{C}_i(m') < \mathcal{C}_i(m)\right)\\ &\approx \mathbb{P}\left(\forall m \notin \mathcal{M}^*_\epsilon,\ \mathcal{C}_i(m^*) < \mathcal{C}_i(m)\right)\end{aligned}$$

Using a Gaussian approximation

$$\mathcal{C}_i(m) - \mathcal{C}_i(m^*) \approx \mathbb{E}\left[\mathcal{C}_i(m) - \mathcal{C}_i(m^*)\right] + N_m\sqrt{\mathrm{Var}\left[\mathcal{C}_i(m) - \mathcal{C}_i(m^*)\right]}$$

for some $N_m \sim \mathcal{N}(0,1)$, we get

$$\mathbb{P}\left(\widehat{m}_i \in \mathcal{M}^*_\epsilon\right) \approx \mathbb{P}\left(\forall m \notin \mathcal{M}^*_\epsilon,\ N_m > -\frac{\mathbb{E}\left[\mathcal{C}_i(m) - \mathcal{C}_i(m^*)\right]}{\sqrt{\mathrm{Var}\left[\mathcal{C}_i(m) - \mathcal{C}_i(m^*)\right]}}\right) \ . \tag{1.20}$$

Under our assumption on the bias, $\forall m$, the numerators $\mathbb{E}\left[\mathcal{C}_i(m) - \mathcal{C}_i(m^*)\right]$ do not depend on $i \in \{1,2\}$. Moreover, if this bias is “small", $\mathbb{E}\left[\mathcal{C}_i(m) - \mathcal{C}_i(m^*)\right]$ is non negative for any $m \notin \mathcal{M}^*_\epsilon$. Therefore, the proxy (1.20) for $\mathbb{P}\left(\widehat{m}_i \in \mathcal{M}^*_\epsilon\right)$ is larger for $i = 1$ than for $i = 2$ if

$$\forall m \notin \mathcal{M}^*_\epsilon, \qquad \mathrm{Var}\left[\mathcal{C}_1(m) - \mathcal{C}_1(m^*)\right] < \mathrm{Var}\left[\mathcal{C}_2(m) - \mathcal{C}_2(m^*)\right] \ . \tag{\textbf{HCC}}$$

V-fold penalized criteria satisfy

$$\mathbb{E}\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}^{vf}_V,C}(m) = \|f_m - f\|^2 + \left(\frac{2C}{V-1} - 1\right)\mathbb{E}\|f_m - \widehat{f}_m\|^2 + A \ .$$

We use our heuristic to compare V-fold criteria for different values of V. For any V, choose $C = V - 1$, so all criteria are unbiased. We prove in [3, Theorem 6] that, when $\mathcal{M}_n$ is the collection of regular histograms with bin sizes $d = 2^\ell$, for $\ell = 1, \ldots, \lfloor \log_2 n \rfloor$, the variances of the increments is proportional to $1 + 4/(V-1)$. This result supports therefore the following remarks made by practical users of V-fold methods :

1. when the bias (1.19) is fixed (and small), enlarging V improves the model selection performances of V-fold procedures,
2. this improvement is larger when V grows from 2 to 5 or 10 than from 10 to n.

Moreover, V appears as a first order term in the asymptotic development of the variance of the increment $\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}^{vf}_V,C}(m)-\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}^{vf}_V,C}(m^*)$ while it is only involved in smaller order terms in the variance of the criteria themselves. Previous articles only focused on the variances of the criteria [Bur89, BG05, Cel08, Cel14, CR08] and could not therefore explain as well the practical differences. Our results confirm the common advice that choosing $V=5$ or 10 will, for a reasonable computation time, yield to performances close to the optimal. Let us nevertheless recall to conclude this paragraph that choosing a larger V only improves the model selection performance *when the bias* (1.19) *is fixed and small.* For V-fold cross-validation criteria $\mathcal{C}^{cv}_{\mathcal{E}^{vf}_V}$, choosing a larger V also reduces the bias so, even if the asymptotic performances of such criteria are always better, they may be much worse for finite n, see for example [Cel14, LY11].

1.2.5 Results for general cross-validation criteria

Our second series of results covers more general cross-validation criteria, in particular Monte-Carlo cross-validation. Actually, any $\mathcal{E}$ among $\mathcal{E}^{ho}_p$, $\mathcal{E}_p$, $\mathcal{E}^{mc}_{p,B}$ and $\mathcal{E}^{vf}_V$ satisfies

$$\exists p\in\{1,\dots,n-1\}\qquad \text{s.t.}\qquad \mathcal{E}\subset\mathcal{E}_p\ . \tag{\textbf{SC}}$$

$$\mathcal{E}\text{ is independent of }X_1,\dots,X_n\ . \tag{\textbf{Ind}}$$

For any criterion satisfying (**SC**) and (**Ind**), we prove in [3, Theorem 9] an oracle inequality for the estimator (1.17). We show that, for all $p\in\{1,\dots,n-1\}$, for $A_p=\frac{n}{n-p}>1$, for any $\mathcal{E}\subset\mathcal{E}_p$,

$$\forall m\in\mathcal{M},\qquad \mathbb{E}\mathcal{C}_{\mathcal{E}}(m)=\|f-f_m\|^2+A_p\mathbb{E}\|f_m-\widehat{f}_m\|^2+A\ .$$

This means that all criteria $\mathcal{C}^{cv}_{\mathcal{E}}$ with $\mathcal{E}\subset\mathcal{E}_p$ have the same bias and therefore, we show in [3, Theorem 9] that the selected $\widehat{m}_{\mathcal{E}}$ satisfies, for $\delta_p=A_p-1>0$

$$\forall\epsilon>0,\qquad R_f(\widehat{m}_{\mathcal{E}})\le(1+\delta_p+\epsilon)\inf_{m\in\mathcal{M}}R_f(m)+\Delta(\mathcal{E},\epsilon)\ . \tag{1.21}$$

The leading function $1+\delta_p+\epsilon$ is common to all schemes, the difference is that the remainder term $\Delta(\mathcal{E},\epsilon)$ can be written

$$\Delta(\mathcal{E},\epsilon)=\left(1+\pi^{\mathcal{E}}\frac{n}{p}\right)\Delta(\epsilon),\quad\text{with}\quad \pi^{\mathcal{E}}=\max_{i\in\{1,\dots,n\}}\frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}}\mathbf{1}_{i\in T}\ .$$

It is interesting to notice that, for V-fold $\mathcal{E}^{vf}_V$ with $V=n/p$ and leave-p-out $\mathcal{E}_p$ schemes, $\pi^{\mathcal{E}}=\frac{p}{n}$, so the upper bound (1.21) on $\Delta(\mathcal{E},\epsilon)$ has the smallest order possible for these schemes. The worst case is achieved by hold-out criteria where $\pi^{\mathcal{E}}=1$, leading to a loss of order n/p in the remainder term of (1.21). Even if it is only an upper bound, it is coherent with the idea that multiplying the number of splits improves the stability of the resulting criterion. Finally, for Monte-Carlo cross-validation, $\pi^{\mathcal{E}}$ concentrates around p/n with a non-increasing remainder term (as a function of B), bounded by 1 and behaving as $\log n/B$ for large B. This is coherent with the fact that Monte-Carlo cross-validation corresponds to hold-out when $B = 1$ and to leave-p-out when $B \to \infty$. More precisely, this shows that the remaining term in (1.21) has the smallest order possible once $B \geq n \log n/p$. However, when $B = n/p$, the remainder term is larger than the one obtained for $V = (n/p)$-fold procedure (which also uses B splits) by a logarithmic factor, suggesting that it might be preferable to perform B well-chosen deterministic splits than random ones. To investigate further the comparisons between these criteria, we also compute in [3, Theorem 10] the variances of the increments in our general cross-validation scheme. We prove that these variances are always convex combinations of the variance of hold-out criteria (HO), which is the largest one, and the leave-p-out (LPO), which is the smallest one. For Monte-Carlo cross-validation (MC), the result can informally be stated as

$$\mathrm{Var}(\mathrm{MC}) = \frac{1}{B}\mathrm{Var}(\mathrm{HO}) + \left(1 - \frac{1}{B}\right)\mathrm{Var}(\mathrm{LPO}) \enspace .$$

This result, which is valid more generally for Monte-Carlo cross-validation estimators, confirms the remarks done after the oracle inequality. Moreover, when $B = V$, it allows to compare more precisely Monte-Carlo and V-fold cross-validation, showing that the variance of V-fold is the one of Monte-Carlo divided by a factor that is as $n \to \infty$ (depending on V) between 2 and 3. Again, this confirms the impression that performing deterministic splitting is preferable, but the improvement is probably much smaller than a log factor. The superiority of V-fold over Monte-Carlo cross-validation was already suggested by the variance computations of Burman [Bur89].

From a technical point of view, the study of cross-validation is possible in density estimation thanks to the closed formula we get for all penalties and criteria. The oracle inequalities are proved using the approach introduced in Section 1.1.2 and rely on concentration inequalities for the empirical mean and U-statistics of order 2. Closed formulas are also central to compute variances. General lemmas are provided to compute covariances of sums of U-statistics and empirical means, they are, up to our knowledge, all new. Finally, the heuristic, based on Gaussian approximations, is new but the increments of criteria already appeared in the relative bounds of [Cat07] which can be used for model selection, see [Aud04].

1.3 Linear estimator selection

This section presents some extensions of model selection theory to linear estimator selection in least-squares density estimation. Given a function $m : \mathbb{X}^2 \to \mathbb{R}$, the linear estimator of f associated to m is defined by

$$\forall x \in \mathbb{X}, \qquad \widehat{f}_m(x) = \frac{1}{n}\sum_{k=1}^{n} m(X_i, x) \enspace .$$

Linear estimators have been introduced in [WB79] under the name delta-sequences and are referred to as additive estimators in [DL01]. We get interested in these since several classical density estimators are linear, as shown in the following examples.

Example 1 (Projection estimators)**.** *Let S denote a finite-dimensional linear subspace of $L^2(\mu)$ and let $(\varphi_i)_{i\in\mathcal{I}}$ denote an orthonormal basis of S, the projection* *estimators presented in Section 1.1 are defined by*

$$\widehat{f_S} = \sum_{i\in\mathcal{I}} (P_n\varphi_i)\varphi_i \ ,$$

they are linear estimators associated to the function m_S, where

$$\forall (x,y)\in\mathbb{X}^2, \qquad m_S(x,y) = \sum_{i\in\mathcal{I}} \varphi_i(x)\varphi_i(y) \ .$$

Example 2 (Parzen-Rosenblatt estimators). *Let $\mathbb{X}=\mathbb{R}$ and $k:\mathbb{X}\to\mathbb{R}$, $h>0$ and $m_{k,h}(x,y)=\frac{1}{h}k(\frac{x-y}{h})$. The linear estimator, defined by*

$$\forall x\in\mathbb{X}, \qquad \widehat{f_{k,h}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right)$$

is called the Parzen estimator in reference to [Par62], it has been introduced by Rosenblatt [Ros56]. It is still one of the most famous density estimators. The function k is called the kernel and h the bandwidth. Among famous kernels, one can mention here the Gaussian kernel, the Epanechnikov kernel, Wasserman's kernel, splines kernels or Pinsker's kernels and refer to Tsybakov's book [Tsy09] for definitions.

Example 3 (Weighted sequences). *Let $(\varphi_i)_{i\in\mathbb{N}}$ denote an orthonormal family in $L^2(\mu)$ and let $\omega=(\omega_i)_{i\in\mathbb{N}}$ denote a sequence (of weights) in ℓ^2. The weighted estimator*

$$\widehat{f_\omega} = \sum_{i=0}^{+\infty} \omega_i (P_n\varphi_i)\varphi_i \ ,$$

is a linear estimator associated to the function

$$\forall (x,y)\in\mathbb{X}^2, \qquad m_\omega(x,y) = \sum_{i=0}^{+\infty} \omega_i\varphi_i(x)\varphi_i(y) \ .$$

Examples of weighted estimators are projection estimators and Pinsker's estimators [Pin80]. Pinsker's estimators have been successfully used in density estimation in [Efr85, Efr00, Efr05, Gol92, Rig06, RT07], on the Fourier basis with Pinsker's weights where they were proved to be sharp minimax on Sobolev Spaces, see Section 1.3.4 for details.

Given a collection of linear estimators $(\widehat{f_m})_{m\in\mathcal{M}_n}$ we want to select one with a least-squares risk as close as possible to the minimal one. The linear estimator selection framework is sufficiently rich to cover the problems of model selection studied by Barron, Birgé and Massart [BBM99, BM97, BM01, Mas07] by example 1, or the problem of the selection of the bandwith and/or of the kernel for Parzen estimators by example 2, which has been widely studied both theoretically [DL01, Tsy09, GL11] and practically [Sil86, WJ95, JMS96]. It also authorizes the competitions between these methods. However, it does not cover more general estimator selection problems like the selection of estimators issued from a first phase of selection : thresholded estimators or Lasso estimators, as does the construction of Baraud [Bar11], or the choice between large collections of fixed functions as does Birgé [Bir06b].

1.3.1 Optimal selection

This section presents the results of [4] on asymptotically optimal selection of linear estimators. Let pen denote some real-valued function defined on $\mathcal{M}_n$, our estimator is defined by $\widehat{f}_{\widehat{m}}$, where

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma(\widehat{f}_m) + \mathrm{pen}(m) \right\} \ , \tag{1.22}$$

Recall that $R_f(\widehat{f}_m) = \mathbb{E}\|\widehat{f}_m - f\|^2$ denotes the least-squares risk of $\widehat{f}_m$. We want to prove that our estimator satisfies an oracle inequality, which means as before that, for some sequence $\epsilon_n \geq 0$,

$$R_f(\widehat{f}_{\widehat{m}}) \leq (1+\epsilon_n) \inf_{m\in\mathcal{M}_n} R_f(\widehat{f}_m) + \Delta(\epsilon_n, \mathcal{M}_n) \ .$$

We prove in [4, Theorem 3.1] that the penalty $\mathrm{pen}_{\mathrm{opt}}(m) = \frac{2}{n}\mathbb{E}\left[m(X,X)\right]$ yields an asymptotically optimal oracle inequality, under technical assumptions on the collection $\mathcal{M}_n$. These are satisfied in all our examples when the cardinality of $\mathcal{M}_n$ is asymptotically not larger than the exponential of some power of $\log n$. This penalty is not always directly computable in practice but can be estimated by its empirical counterpart $\frac{2}{n^2}\sum_{i=1}^n m(X_i, X_i)$ without affecting its theoretical performances. Moreover, for projection estimators onto regular histogram spaces, $m_S(x,x) = d$ for any x so this optimal penalty coincide with Mallow's penalty studied in Section 1.1.1 for the optimal selection of the bin size of a projection estimator on a regular histogram. For Parzen's estimators $m_{k,h}(x,x) = \frac{1}{h}k(0)$ for any x, so $\mathrm{pen}_{\mathrm{opt}}(m_{k,h}) = 2k(0)/(nh)$ can also be directly be computed and used for defining $\widehat{m}$.

The paper introduces the new proof of oracle inequalities in least-squares density estimation that only relies on Bernstein's concentration inequality for the empirical mean (see (1.6) and [Mas07] for a proof) and a concentration inequality for totally degenerate U-statistics of order 2. This inequality was first proved by Giné, Latala and Zinn in [GLZ99]. Exact constants were obtained in [HRB03], but only for real-valued random variables. The most general result is now the one of Adamczak [Ada06], that is valid for U-statistics of any order. We use a version given in Giné and Nickl's book [GN15, Theorem 3.4.8]. The asymptotic behavior of the main quantities involved in our proofs have also recently been studied, for example in [DO13, MS11].

1.3.2 Minimal penalties

We are also interested in minimal penalties for linear estimator selection. Following Birgé and Massart [BM07], a minimal penalty is defined as a function $\mathrm{pen}_{\mathrm{min}} : \mathcal{M}_n \to \mathbb{R}$ such that the estimator selected by a penalty equal to $u\mathrm{pen}_{\mathrm{min}}$ does not satisfy an oracle inequality when $u < 1$ but does satisfy one when $u > 1$. Since we minimize $P_n\gamma(\widehat{f}_m) + \mathrm{pen}(m)$ instead of $P\gamma(\widehat{f}_m)$, the ideal penalty is $\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{f}_m)$. Arlot and Massart [AM08] decompose the ideal penalty in three terms $p_1(m) + p_2(m) + \delta(m)$, with

$$p_1(m) = P(\gamma(\widehat{f}_m) - \gamma(f_m)), \quad p_2(m) = P_n(\gamma(f_m) - \gamma(\widehat{f}_m)), \quad \delta(m) = (P - P_n)\gamma(f_m) \ ,$$

and show that the centered term $\delta(m)$ does not matter so an optimal penalty is

$$\mathrm{pen}_{\mathrm{opt}}(m) \approx p_1(m) + p_2(m) \ .$$

On the other hand, choosing the “penalty" $\mathrm{pen}_{\mathrm{min}}(m) = p_2(m)$ leads to a criterion that is close to the bias $\|f - f_m\|^2$. The selected estimator is the minimizer of the bias and therefore a very complex m which has disastrous performances, see the regular histogram example (1.4). This heuristic applies for linear estimator selection as well and we prove in [4, Theorem 4.3] that the minimal penalty is given by

$$\mathrm{pen}_{\mathrm{min}}(m) = \frac{1}{n}\left(2\mathbb{E}\left[m(X,X)\right] - \mathbb{E}\left[A_m(X,X)\right]\right) \ ,$$

where

$$\forall (x,y) \in \mathbb{X}^2, \qquad A_m(x,y) = \int_{\mathbb{X}} m(x,z)m(z,y)d\mu(z) \ . \tag{1.23}$$

The analysis of this last result is richer than usual. First, for projection estimators, we have by construction $A_{m_S} = m_S$, which implies that

$$\mathrm{pen}_{\mathrm{min}}(m_S) = \frac{1}{2}\mathrm{pen}_{\mathrm{opt}}(m_S) \ .$$

This property is called “slope heuristic" in [BM07] and was already proven for the selection of projection estimators in my PhD article [10]. It is linked with Arlot and Massart’s approach via the relation

$$p_1(m_S) = p_2(m_S) = \left\|\widehat{f}_{m_S} - f_{m_S}\right\|^2 \ ,$$

which implies that

$$\mathrm{pen}_{\mathrm{opt}}(m) \approx p_1(m) + p_2(m) \approx 2p_2(m) \approx 2\mathrm{pen}_{\mathrm{min}}(m) \ .$$

It is here a corollary of our general result. The situation is different for Parzen estimators since $A_{m_{k,h}}(x,x) = \frac{\|k\|^2}{h}$, and therefore

$$\mathrm{pen}_{\mathrm{min}}(m_{k,h}) = \frac{2k(0) - \|k\|^2}{nh}, \qquad \mathrm{pen}_{\mathrm{opt}}(m_{k,h}) = \frac{2k(0)}{nh} \ .$$

There exists a relationship between the minimal and the optimal penalty, but it is not universal anymore since it depends on the kernel k. A similar situation was described by Arlot and Bach [AB09] in a regression setting. Interestingly, one can notice that, if $\|k\|^2 > 2k(0)$ the minimal penalty is negative, which implies that minimizing an unpenalized empirical loss yields an oracle inequality. To the best of our knowledge, such phenomenon was only observed in a very particular classification setting in [FT06].

The slope heuristic is used in practice to *calibrate* a penalty when the *shape* of $\mathrm{pen}_{\mathrm{opt}}$ is known, that is when a function $\mathsf{p} : \mathcal{M}_n \to \mathbb{R}$ is known such that $\mathrm{pen}_{\mathrm{opt}} = \digamma\mathsf{p}$ for an unknown constant $\digamma$. In this situation, Arlot and Massart [AM08] proposed to evaluate $\digamma$ using the following *slope algorithm.* For any $u \in \mathbb{R}$, compute the function $\widehat{m}_u$ selected by the penalty $\mathrm{pen} = u\mathsf{p}$. Call $\widehat{u}$ a constant such that the complexity (for example d in regular histograms or $1/h$ for Parzen estimators) of $\widehat{m}_u$ is very large when $u < \widehat{u}$ and much smaller when $u > \widehat{u}$. Choose finally the constant $u = 2\widehat{u}$ to calibrate the penalty. The idea is that $\widehat{u}\mathsf{p}$ should be close to $\mathrm{pen}_{\mathrm{min}}$ to observe a transition in the complexity of $\widehat{m}_u$ so $2\widehat{u}\mathsf{p} \approx 2\mathrm{pen}_{\mathrm{min}}(m) \approx \mathrm{pen}_{\mathrm{opt}}(m)$ by the slope heuristic. In any case, the complexity of $\widehat{m}_u$, that is a non-increasing function of u, is guaranteed to remain reasonable when one uses the slope algorithm with $u > 0$. This first algorithm will not work anymore for the selection of Parzen kernels.

1.3.3 Cross-validation selection of linear estimators

This section presents the results of the article [17]. Our purpose was to extend some results of [3], which are recalled in Section 1.2, to linear estimator selection.
The first result that we presented in Section 1.2.3, [3, Lemma 1], is a lemma proving that many cross-validation selectors based on leave-p-out or V-fold schemes could be studied simultaneously by studying V-fold penalization criteria

$$\mathcal{C}^{cv}_{\mathrm{pen},\mathcal{E}^{vf}_V,C}(m) = P_n\gamma(\widehat{f}_m) + C\mathrm{pen}_{\mathcal{E}^{vf}_V}(m) \enspace ,$$

for different values of V and C. This lemma is no longer true for general linear estimators. In fact, it is still true that leave-p-out penalties (and resampling ones based on an exchangeable resampling vector) are proportional to n-fold penalties. These penalties are still a linear combination of a centered empirical mean and the totally degenerate U-statistic based on m, that is

$$U_m = \sum_{1\le i\neq j\le n} m(X_i,X_j) - \mathbb{E}\left[m(X_i,X_j)|X_i\right] - \mathbb{E}\left[m(X_i,X_j)|X_j\right] + \mathbb{E}\left[m(X_i,X_j)\right] \enspace .$$

The difference is that the decomposition of cross-validation criteria

$$\mathcal{C}_{\mathcal{E}}(m) = \frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}} P_{T^c}\gamma(\widehat{f}^T_m)$$

also involves totally degenerate U-statistics associated to functions A_m defined in (1.23). The simplification in Section 1.2 was due to the relation $m = A_m$ which only holds for projection estimators. However, our general method of proof can still be applied to study both selectors.
This is why we have been able to extend all oracle results of [3] to linear estimators, showing in particular oracle inequalities for any V-fold cross-validation (penalized) criteria and any leave-p-out (penalized) criteria. Moreover, using the concentration for U-statistics of [GN15] instead of [HRB03], we extend in [17, Theorem 5] the results [3, Theorems 5 and 9] to any Polish-space valued observations, in particular $\mathbb{R}^d$-valued observations.
The heuristic (**HCC**) for comparison of criteria of Section 1.2 remains valid in linear estimator selection. It is an interesting problem, though technically challenging (see [Mag15, Chapter 4]), to see if the conclusions in the end of Section 1.2 remain valid.

1.3.4 From linear estimator selection to sharp adaptive estimators

The material of this section is also borrowed from the paper [17] that I am preparing in collaboration with S. Arlot and N. Magalhães. We apply asymptotically optimal oracle inequalities for linear estimator selection to derive adaptive estimators. Given a class $\mathcal{F}$ of densities and an estimator $\widehat{f}$, the maximal risk of $\widehat{f}$ over $\mathcal{F}$ is defined by

$$\overline{R}_{\mathcal{F}}(\widehat{f}) = \sup_{f\in\mathcal{F}} \mathbb{E}_f\|f-\widehat{f}\|^2 \enspace .$$

The *minimax* risk over $\mathcal{F}$ is defined as the minimum, over all estimators, of the maximal risks,

$$\overline{R}(\mathcal{F}) = \inf_{\widehat{f}} \overline{R}_{\mathcal{F}}(\widehat{f}) \enspace .$$

An estimator $\widehat{f}$ is called asymptotically minimax over $\mathcal{F}$ when there exists a constant $\epsilon \geq 0$ such that

$$\overline{R}_{\mathcal{F}}(\widehat{f}) \leq (1+\epsilon)\overline{R}(\mathcal{F}) \ .$$

It is called *sharp* minimax over $\mathcal{F}$, when there exists $\epsilon_n \to 0$ such that

$$\forall n \geq 1, \qquad \overline{R}_{\mathcal{F}}(\widehat{f}) \leq (1+\epsilon_n)\overline{R}(\mathcal{F}) \ . \tag{1.24}$$

To apply our result, we have to define a class of functions $\mathcal{F}$ and a proper collection of estimators. On $\mathbb{X} = [0,1]$, let $(\varphi_i)_{i\in\mathbb{N}}$ denote the Fourier basis defined for any $x \in \mathbb{X}$ by $\varphi_0(x) = 1$ and

$$\forall k \in \mathbb{N}\setminus\{0\}\,, \qquad \varphi_{2k-1}(x) = \sqrt{2}\cos(2\pi kx), \qquad \varphi_{2k}(x) = \sqrt{2}\sin(2\pi kx) \ . \tag{1.25}$$

Given two positive real numbers β and L, define the ellipsoid

$$\mathcal{E}_{\beta,L} = \left\{ (a_k)_{k\in\mathbb{N}}, \ \text{s.t.} \ \sum_{k\in\mathbb{N}} k^{2\beta}a_k^2 \leq L^2 \right\} \ .$$

Denote by $\langle .,. \rangle$ the inner product in $L^2(\mu)$. Our class of interest is the Sobolev class $\mathcal{F}_{\beta,L}$ of functions $g : [0,1] \to \mathbb{R}$ such that the Fourier coefficients $(\langle g, \varphi_i\rangle)_{i\in\mathbb{N}}$ satisfy $(a_k)_{k\in\mathbb{N}} \in \mathcal{E}_{\beta,L}$, where $a_0 = P\varphi_0$ and

$$\forall k \geq 1, \qquad a_k = \sqrt{\langle g, \varphi_{2k-1}\rangle^2 + \langle g, \varphi_{2k}\rangle^2} \ .$$

The minimax risk over $\mathcal{F}_{\beta,L}$ was computed by Pinsker [Pin80] who evaluated C_β such that, as $n \to \infty$,

$$\overline{R}(\mathcal{F}_{\beta,L}) = (1+o(1))C_\beta L^{\frac{1}{2\beta+1}} n^{-\frac{\beta}{1+2\beta}} \ .$$

Pinsker also built estimators $\widehat{f}_{\beta,L}$ sharp minimax over $\mathcal{F}_{\beta,L}$. In this presentation we only need to precise that there exist an integer $N_{\beta,L} \leq n$ and a sequence $(\omega_{\beta,L,i})_{i\in\mathbb{N}}$ in $[0,1]^{\mathbb{N}}$ such that

$$\widehat{f}_{\beta,L} = \sum_{i=0}^{N_{\beta,L}} \omega_{\beta,L,i}(P_n\varphi_i)\varphi_i \ .$$

Hence, Pinsker's estimators are linear and can be selected with the different optimal selection criteria that we presented. For example, the optimal penalty of Section 1.3.1 can be used since the weights satisfy, for any $k \geq 1$, $\omega_{\beta,L,2k-1} = \omega_{\beta,L,2k}$, therefore

$$\text{pen}_{\text{opt}}(m_\omega) = \sum_{i=1}^{N_{\beta,L}} \omega_{\beta,L,i} \ .$$

One should also mention that projection estimators can be minimax but *they are not* sharp minimax in general.

Pinsker's minimax estimators depend on parameters β and L that are typically unknown to the statistician. Therefore, one looks for estimators $\widehat{f}$ which are minimax on $S_{\beta,L}$ for any value of these parameters in intervals $\beta \in \mathcal{B}_n$ and $L \in \mathcal{L}_n$, without knowing in advance to which $S_{\beta,L}$ f belongs. Such estimators are called adaptive to β and L to emphasize that they behave as the best "estimator" one could built if we knew them in advance. They are called sharp adaptive when they are simultaneously sharp minimax for all $\beta \in \mathcal{B}_n$ and $L \in \mathcal{L}_n$. Estimators sharp adaptive to β and L have been built by Golubev [Gol92], Efromovitch [Efr05], Rigollet [Rig06], Rigollet and Tsybakov [RT07]. These used blockwise Stein methods or aggregation procedures, our estimators are, to the best of our knowledge, the first obtained by minimization of a penalized empirical loss.

To apply our selection procedures, we cannot use all Pinsker's estimators, we first discretize properly the intervals $\mathcal{B}_n$ and $\mathcal{L}_n$ to produce a finite collection of estimators. The discretized sets should be sufficiently refined to approach any point in the original interval at rate $1/n$ when $n \to \infty$, and it shall not be too large for the size of the collection to remain polynomial in n. Taking care of both constraints, we build sharp adaptive estimators over Sobolev classes with parameters $\beta \in \mathcal{B}_n$ and $L \in \mathcal{L}_n$ with Lebesgue measures $\mu(\mathcal{B}_n) = O(n)$ and $\mu(\mathcal{L}_n) = O(\log n)$. The rate of convergence of the sequence ϵ_n in (1.24) is $n^{-\gamma(\beta)}$ for some function γ. It improves upon previous results where this rate was always larger than $1/\log n$. However, it is not optimal since one can check that Pinsker's estimators achieve $n^{-4\gamma(\beta)}$. This 4 is for now our price to pay for adaptivity.

Pinsker's weights $(\omega_{\beta,L,i})_{i\in\mathbb{N}\setminus\{0\}}$ belong to the collection $\mathcal{W}_{ni}$ of non increasing sequences of real numbers in $[0,1]$. Sharp minimax adaptivity over Sobolev balls and other classes of interest (see [CT02] for examples) can therefore also be deduced from an oracle inequality over the class of estimators $(\widehat{f_\omega})_{\omega\in\mathcal{W}_{ni}}$, where for any ω,

$$\widehat{f_\omega} = \sum_{i=0}^{+\infty} \omega_i (P_n\varphi_i)\varphi_i \ .$$

This is, for example, the point of view adopted by Cavalier and Tsybakov [CT02] for inverse problems or Rigollet [Rig06] in least-squares density estimation. Of course, the collection $\mathcal{W}_{ni}$ is much too large to apply directly our selection results, but it can be discretized by a finite family $\mathcal{W}_{ni}^n$ of cardinality growing as the exponential of a power of $\log n$. This discretization uses the same weakly geometrically increasing blocks as blockwise Stein's construction. The maximal risk of the infimum over all estimators $(\widehat{f_\omega})_{\omega\in\mathcal{W}_{ni}^n}$ is not larger than $1 + 1/\sqrt{\log n}$ times the maximal risk of the infimum over all estimators $(\widehat{f_\omega})_{\omega\in\mathcal{W}_{ni}}$. To build the collection $\mathcal{W}_{ni}^n$, we adapt the construction of Rigollet [Rig06] and prove that our methods of selection can thus be used to obtain estimators with maximal risks not larger than $1 + O(1/\sqrt{\log n})$ times the maximal risk over all $(\widehat{f_\omega})_{\omega\in\mathcal{W}_{ni}}$, providing in particular other sharp adaptive estimators over Sobolev balls.

1.4 Interaction neighborhood selection in discrete random fields

This section presents some applications of model selection in discrete random fields that I developed with D. Y. Takahashi in the articles [6, 9]. We were interested in discrete random fields as natural models for brain activity in neuroscience.

1.4.1 Position of the problem

A discrete random field is a triplet (S, A, P) where S a finite set (large), A a finite set (small, bounded), and P is a probability measure on the set of *configurations* $\mathbb{X} = A^S$. We are interested in the estimation of the conditional probabilities

$$\forall x \in \mathbb{X}, \forall i \in S, \qquad P\left(x(i)|x(j), j \neq i\right) \ ,$$

based on the observation of a sequence $X_1, \ldots, X_n$ of independent configurations distributed according to P.
Discrete random fields are used in neuroscience to model brain activity see [SBSB06]. The set S is the set of neurons, these neurons communicate via electric signals (spikes) extremely localized in time that have always the same frequency and intensity, so the activity of neuron i can be either on (a spike is emitted) or off (no spike is emitted), therefore we use the finite alphabet $A = \{-1, 1\}$ to describe at each time the activity of i. The conditional probabilities encode some *functional dependencies* between neurons and this is why they are our object of interest. More precisely, [SBSB06] use Ising measures to model P. Let $J : S^2 \to \mathbb{R}$ denote a function called *potential*, the probability of the configuration $x \in \mathbb{X}$ is

$$P_J(x) = \frac{1}{Z_J} e^{\sum_{(i,j)\in\mathbb{X}^2} x(i)x(j)J_{i,j}} \ ,$$

in particular, for any $i \in S$,

$$P_J(x(i) = 1|x(j), j \neq i) = \frac{1}{1 + e^{-2\sum_{j\neq i} J_{i,j}x(j)}} \ .$$

Therefore, $J_{i,j} > 0$ means that a spike of j favors a spike in i, we say that j has an excitatory influence on i, $J_{i,j} < 0$ means that a spike of j prevents a spike in i, we say that j has an inhibitory influence on i and $J_{i,j} = 0$ means that j has no influence on i. The absolute value of $J_{i,j}$ measures the strength of these influences.
Inference in Ising models has already been studied, see among others [BMS08, CT06, RWL10] who proposed nice efficient algorithms to recover the set of edges $\mathcal{E} = \{(i, j) \in S^2, \text{ s.t. } J_{i,j} \neq 0\}$ of the interaction graph $(S, \mathcal{E})$. This set is a natural object of interest in neuroscience, which makes the estimators of [BMS08, RWL10] of particular interests, since they can be efficiently computed on very large graphs under sparsity assumptions. However, by taking an exact recovery approach, theoretical properties of these estimators such as the consistency results, are only verified under severe restrictions on the underlying graph. In particular, the incoherence assumption for ℓ_1-penalization methods of [RWL10] cannot be checked in practice. No risk bounds for the estimators are available and the properties of the estimators are unknown when these assumptions fail, or when the Ising model is not true.
In [6, 9], we build estimators for the conditional probabilities and provide risk bounds for these estimators. We work in the Ising model in [9] and in general discrete random fields in [6]. In both papers, our approach is based on model selection. We fix some site $i \in S$ and focus on the estimation of the family of conditional probabilities $(P_i(x))_{x\in\mathbb{X}}$, where $P_i(x) = P\left(x(i)|x(j), j \neq i\right)$ that, by some abuse of notation, we denote P_i. To assess the performances of an estimator $\widehat{P}_i = (\widehat{P}_i(x))_{x\in\mathbb{X}}$, we study two losses. In [9], we studied the maximal loss

$$\ell_{i,\infty,P}(\widehat{P}_i) = \left\|\widehat{P}_i - P_i\right\|_\infty, \qquad \text{where} \qquad \|f\|_\infty = \max_{x\in\mathbb{X}} |f(x)| \ .$$

In [6], we studied the least-squares loss

$$\ell_{i,2,P}(\widehat{P}_i) = \left\|\widehat{P}_i - P_i\right\|_P^2, \qquad \text{where} \qquad \|f\|_Q^2 = \sum_{x\in\mathbb{X}} f(x)^2 Q(x) \ .$$

Remark that these losses are always upper bounded by 1. In both cases, the risk of $\widehat{P}$ is measured by the integrated loss $R_{i,P}(\widehat{P}_i) = \mathbb{E}_P(\ell_{i,P}(\widehat{P}))$. To build our estimators, we assume that we observe a sample $X_1, \ldots, X_n$ of i.i.d. random configurations distributed according to P, which is used to build the empirical measure $\widehat{P}_n$ defined for any function $f : \mathbb{X} \to \mathbb{R}$ by $\widehat{P}_n f = \frac{1}{n}\sum_{t=1}^n f(X_t)$. Given a subset $m \subset S$, we define the conditional probabilities $\widehat{P}_{i,n,m} = (\widehat{P}_{i,n,m}(x))_{x\in\mathbb{X}}$ and $P_{i,m} = (P_{i,m}(x))_{x\in\mathbb{X}}$, where $\widehat{P}_{i,n,m}(x) = \widehat{P}_n(x(i)|x(j), j \in m\backslash\{i\})$ and $P_{i,m}(x) = P(x(i)|x(j), j \in m\backslash\{i\})$. Given a collection $\mathcal{M}_n$ of subsets $m \subset S$, our purpose is to select an estimator $\widehat{m} \in \mathcal{M}_n$ such that the estimator $\widehat{P}_{i,n,\widehat{m}}$ has a risk as close as possible to the infimum $\inf_{m\in\mathcal{M}_n} R_{i,P}(\widehat{P}_{i,n,m})$. We look therefore for a subset $m \subset S$ that predicts as well as possible the conditional probabilities P_i.

1.4.2 Oracle inequalities for neighborhood selection

As the results for least-squares risks are easier to expose, we now focus on this risk. Since i is fixed, we will moreover not mention it in the notation of the loss and denote by $\ell_P(\widehat{P}_{i,n,m}) = \ell_{i,2,P}(\widehat{P}_{i,n,m})$, $R_P(\widehat{P}_{i,n,m}) = \mathbb{E}_P\left[\ell_P(\widehat{P}_{i,n,m})\right]$, results for the maximal loss can be found in [9].
The starting point of our analysis is the following computation of the risk of $\widehat{P}_{i,n,m}$. We show in [6, Theorem 3.1] the following bound.

$$\forall m \subset S, \qquad \mathbb{E}_P\left[\ell_P(\widehat{P}_{i,n,m})\right] \le \left[\|P_i - P_{i,m}\|_P^2 + 6\frac{|A|^{|m|}}{n}\right] \wedge 1 \ . \tag{1.26}$$

The risk of $\widehat{P}_{i,n,m}$ is decomposed as a sum of an approximation term $\|P_i - P_{i,m}\|_P^2$ measuring the *bias* of the estimator and a variance term $C\frac{|A|^{|m|}}{n}$ measuring the complexity to estimate $P_{i,m}$. The variance term has the expected form since $|A|^{|m|}$ is the number of conditional probabilities to estimate when we use the subset m and therefore, the number of parameters in the corresponding statistical model.
Now, following the general approach for model selection of Section 1.1.1, we estimate the loss of $\widehat{P}_{i,n,m}$ to find a data-driven criterion $\mathcal{C}(m)$ and choose as a final estimator

$$\widehat{P}_{i,n,\widehat{m}}, \qquad \text{where} \qquad \widehat{m} = \arg\min_{m\in\mathcal{M}_n} \mathcal{C}(m) \ .$$

From (1.26), we can restrict $\mathcal{M}_n$ to subsets $m \subset \mathbb{X}$ with cardinality smaller than $\ell_n = \log_{|A|} n$ so we can guarantee a non trivial risk bound for any $\widehat{P}_{i,n,m}$, with $m \in \mathcal{M}_n$. This helps to reduce drastically the practical implementation of the estimator, making it actually computable for the analysis of a real data-set from neuroscience (see Section 7 in [6]).
As the variance part of the risk is computable, we only have to estimate the bias term. We prove [6, Proposition A.11 in the supplementary material] a Pythagoras relation for the least-squares loss

$$\|P_i - P_{i,m}\|_P^2 = \|P_i\|_P^2 - \|P_{i,m}\|_P^2 \ .$$

Finally, it is a corollary of [6, Theorem 3.1] that

$$\forall m \in \mathcal{M}_n, \qquad \mathbb{E}_P\left[-\left\|\widehat{P}_{i,n,m}\right\|_{\widehat{P}}^2\right] \geq -\left\|P_{i,m}\right\|_P^2 - 6\frac{|A|^{|m|}}{n} \ .$$

We propose the following criterion

$$\mathcal{C}(m) = -\left\|\widehat{P}_{i,n,m}\right\|_{\widehat{P}}^2 + C\frac{|A|^{|m|}}{n} \ . \tag{1.27}$$

Using concentration inequalities as in Section 1.1.1, we prove in [6, Theorem 3.2] that, if $C \geq 12$, there exists a function $c = c_{|A|} > 0$ such that, if $\mathcal{M}_n = \{m \subset S, \text{ s.t. } |m| \leq \ell_n\}$, then

$$cR_P(\widehat{P}_{i,n,\widehat{m}}) \leq \inf_{m\in\mathcal{M}_n}\left\{\left\|P_i - P_{i,m}\right\|_P^2 + \frac{|A|^{|m|}}{n}\right\} + \frac{(\ell_n \log|S|)^2}{n} \ .$$

In other words, the selected $\widehat{m}$ optimizes the risk bound (1.26) up to the remainder term $\frac{(\ell_n \log|S|)^2}{n}$ that is a price to pay for the selection step. A very interesting feature of this result for analyzing neuroscience data-sets is that it holds without restrictions on the shape of P which can be Ising without restriction on the interaction graph, and can even not be Ising.

To go further in our analysis and explain how we derive rates of convergence, we shall however now restrict the discussion of the end of the paragraph to Ising measures where the control of the bias term is particularly elementary. Actually, we always have $\left\|P - P_{i,m}\right\|_P^2 \leq 2\left\|P - P_{i,m}\right\|_\infty$ and we show in [9, Proposition 4.4] that there exists a function $C = C(\beta)$, where $\beta = \max_{i\in S}\sum_{j\in S}|J_{i,j}|$ is the inverse of the temperature in the Ising Model (see [Geo88]), such that

$$\left\|P - P_{i,m}\right\|_P^2 \leq C(\beta)\sum_{j\notin m}|J_{i,j}| \ .$$

We deduce that, for example, when the $|J_{i,j}|$ (arranged by non increasing absolute values) decrease exponentially, and provided that $|S|$ is at most polynomial in n, $\widehat{P}_{i,n,\widehat{m}}$ converges to P_i at a rate $n^{-\gamma}$. Other rates, and other examples of more general Gibbs measures are presented in [6, Section 4.2].

1.4.3 Minimal penalty phenomenon and slope heuristic for neighborhood estimation

As for the selection of linear estimators, we prove in [6, Theorem 5.1] that there is a minimal penalty for neighborhood selection, which is given by the following equivalent of $p_2(m)$ (see Section 1.3.2).

$$p_2(m) = \left\|\widehat{P}_{i,n,m} - P_{i,m}\right\|_{\widehat{P}}^2 \ .$$

Moreover, we also prove in [6, Theorem 5.2] that $p_1(m) + p_2(m)$ is an ideal penalty (if the cardinality $|\mathcal{M}_n|$ is not larger than the exponential of some power of $\log n$), where

$$p_1(m) = \left\|\widehat{P}_{i,n,m} - P_{i,m}\right\|_P^2 \ .$$

Of course, none of these quantities can be computed in practice, so no actual "slope algorithm" can be deduced from them. Nevertheless, we show [6, Lemma A.5 in the supplementary material] that

$$p_1(m) \approx p_2(m) \ ,$$

and thus the relationship

$$\mathrm{pen}_{\mathrm{min}}(m) = \frac{1}{2}\mathrm{pen}_{\mathrm{opt}}(m) \ .$$

We use this result to suggest to *calibrate* in practice the constant C of the penalty defining our selection criterion (1.27) by the slope algorithm of Section 1.3.2. We compare in simulated data-sets this choice of C with the constant 12 deduced from the risk bound (1.26). It turns out that the theoretical bound is extremely pessimistic and leads to very small neighborhoods while the slope algorithm provides a much more reasonable choice. This was even more interesting in the analysis of a real data-set. All these experiments can be found in [6, Sections 6 and 7] and the set of MATLAB routines that we used can be downloaded from www.princeton.edu/∼dtakahas/publications/LT11routines.zip.

Chapter 2

Subgaussian Estimators

2.1 Position of the problem

I present in this section my article with L. Devroye, G. Lugosi and R.I. Oliveira [1] on the estimation of the mean $\nu(P) \in \mathbb{R}$ of a probability distribution P assuming it has finite variance $\sigma^2(P)$, based on the observation of an i.i.d. sample $X_1, \dots, X_n$ with common distribution P. By the central limit theorem, the empirical mean $\nu_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies,

$$\forall x > 0, \qquad \sup_{P \in \mathcal{P}_2} \lim_{n \to \infty} P\left(\nu_n - \nu(P) > \sigma(P)\sqrt{\frac{2x}{n}}\right) = \Phi(1 - \sqrt{2x}) \ ,$$

where Φ denotes the c.d.f. of the standard Gaussian random variable and $\mathcal{P}_2$ denote the set of all distributions such that $\mathbb{E}_P\left[X^2\right] < \infty$. This result is essentially optimal, Catoni [Cat12a] proved that for any set $\mathcal{P}$ of distribution containing $(\mathcal{N}(m, 1))_{m \in \mathbb{R}}$, for any $x > 0$ and any estimator $\widehat{\nu}_x$ possibly depending on x,

$$\sup_{P \in \mathcal{P}} P\left(\nu_x - \nu(P) > \sigma(P)\sqrt{\frac{r(x)}{n}}\right) \le \Phi(1 - \sqrt{2x}) \ ,$$

only if $r(x) \ge 2x$. As $\Phi(1 - \sqrt{2x}) \le e^{-x}$, the central limit theorem implies that the empirical mean is asymptotically subgaussian, that is

$$\forall x > 0, \qquad \sup_{P \in \mathcal{P}_2} \limsup_{n \to \infty} P\left(\nu_n - \nu(P) > \sigma(P)\sqrt{\frac{2x}{n}}\right) \le e^{-x} \ . \tag{2.1}$$

We are interested in the construction of estimators satisfying this property non asymptotically. More precisely, let us introduce the following definitions.

Definition 1. *Let n be a positive integer, $L > 0$, $x_n > 0$. Let $\mathcal{P} \subset \mathcal{P}_2$ be a family of probability distributions over $\mathbb{R}$.*

1. **single-x subgaussian estimation:** *a single-x L-subgaussian estimator for $(\mathcal{P}, x_n)$ is a measurable mapping $\widehat{\nu} : \mathbb{R}^n \times [0, x_n] \to \mathbb{R}$ such that if $P \in \mathcal{P}$, $x \le x_n$, and $X_1^n = (X_1, \dots, X_n)$ is a sample of i.i.d. random variables distributed as P, then*

$$P\left(|\widehat{\nu}(X_1^n, x) - \nu(P)| > L\,\sigma(P)\sqrt{\frac{1+x}{n}}\right) \le e^{-x} \ . \tag{2.2}$$

We also write $\widehat{\nu}_x(\cdot)$ for $\widehat{\nu}(\cdot, x)$.

2. **multiple-x subgaussian estimation:** *a multiple-x L-subgaussian estimator for $(\mathcal{P}, x_n)$ is a measurable mapping $\widehat{\nu} : \mathbb{R}^n \to \mathbb{R}$ such that, for each $x \leq x_n$, $P \in \mathcal{P}$ and i.i.d. sample $X_1^n = (X_1, \dots, X_n)$ distributed as P,*

$$P\left(|\widehat{\nu}(X_1^n) - \nu(P)| > L\,\sigma(P)\sqrt{\frac{1+x}{n}} \right) \leq e^{-x} \ . \tag{2.3}$$

The difference between (2.2) or (2.3) and the original subgaussian bound (2.1) is that we replaced the optimal $\sqrt{2x}$ by $L\sqrt{1+x}$, with some possibly suboptimal constant $L > \sqrt{2}$. We shall discuss the possibility to reach the optimal $L = \sqrt{2} + o(1)$ in (2.2) or (2.3), replacing x by $1 + x$ is a technical convenience to avoid unnecessary discussions when $x \leq 1$.
It transpires from these definitions that multiple-x estimators are preferable whenever they are available, because they combine good typical behavior with nearly optimal bounds under extremely rare events. By contrast, the need to commit to a x in advance means that single-x estimators may be too pessimistic when a large x is desired. The main problem addressed in this section is the following:

Given a family $\mathcal{P}$ (or more generally a sequence of families $\mathcal{P}_n$), find the largest possible sequence x_n such that multiple-x L-subgaussian estimators for $(\mathcal{P}, x_n)$ exist for all large n, and with a constant L that does not depend on n.

2.2 Main results

We discuss in these notes the examples where $\mathcal{P}$ is the class $\mathcal{P}_2$ of all distributions with finite second moment, the class $\mathcal{P}_2^\sigma, \mathcal{P}_2^{\leq\sigma} \subset \mathcal{P}_2$ of distributions respectively with variance equal to σ^2 and smaller than σ^2 and the class $\mathcal{P}_{\mathrm{krt}\leq\kappa}$ of distributions having a fourth moment and a kurtosis $\frac{\sqrt{\mathbb{E}[(X-\nu(P))^4]}}{\sigma(P)^2} \leq \kappa$, other examples can be found in [1].

The most classical estimator is the empirical mean ν_n and we first recall its performances to explain why our estimators outperform it. Catoni [Cat12a] proved that (2.3) is satisfied by ν_n over $\mathcal{P}_{\mathrm{krt}\leq\kappa}$ for any sequence x_n such that $e^{x_n} = o(n)$. Bernstein inequality proves that, under exponential moment assumptions on P, (2.3) is still true for ν_n for any $x_n = o(n)$, with optimal $L = \sqrt{2} + o(1)$, a result that cannot be improved in general by the Gärtner-Ellis theorem (see [DZ02]), unless P has subgaussian tails.
We prove that there exists a constant c such that multiple-x subgaussian estimators exist for $(\mathcal{P}, cn)$ and $\mathcal{P} = \mathcal{P}_2^\sigma$ [1, Theorem 3.2] or $\mathcal{P} = \mathcal{P}_{\mathrm{krt}\leq\kappa}$ [1, Theorem 3.6]. Moreover, on $\mathcal{P}_{\mathrm{krt}\leq\kappa}$, we can even reach the optimal $L = \sqrt{2} + o(1)$ for any $x_n = o((n/\kappa)^{2/3})$. We also prove (Theorems 3.2 and 3.6) that, for these classes, there exists a constant C such that no single-x subgaussian estimators exist for $(\mathcal{P}, Cn)$. Notice that these estimators clearly outperform the empirical mean even on classes of distributions having exponential moments.
Catoni [Cat12a] already built single-x subgaussian estimators achieving (2.2) for $(\mathcal{P}_2, \frac{n}{2} - 1)$. In contrast to Catoni's result, we prove (Theorem 3.2) that there

doesn't exist multiple-x estimators over $\mathcal{P}_2$ for any sequence $x_n \to \infty$ and therefore that the notions of single-x and multiple-x estimators are distincts. In fact we show (Theorem 3.2) that there exist multiple-x estimators over all distributions with fixed variance σ^2, and actually over all distributions with variance σ^2 between two bounds $0 < \sigma_1^2 \leq \sigma^2 \leq \sigma_2^2 < \infty$, but there doesn't exist any when the ratio σ_2/σ_1 is unbounded.
In particular, one cannot build subgaussian estimators on the class $\mathcal{P}_2^{\leq\sigma}$ of distribution with variance bounded by σ. Actually, it is not even possible on the class $(B(p))_{p\in(0,1)}$ of Bernoulli distributions. On the other hand, one can build estimators $\widehat{\nu}_\sigma$ such that

$$\forall x \in (0, cn), \forall P \in \mathcal{P}_2^{\leq\sigma}, \qquad P\left(\widehat{\nu}_\sigma - \nu(P) > L\sigma\sqrt{\frac{1+x}{n}}\right) \leq e^{-x} \ , \tag{2.4}$$

The difference is that σ in (2.4) is only an upper bound on the variance $\sigma(P)$ in (2.3). Such *weakly* subgaussian estimators are however sufficient to perform the estimator selection procedure that are presented in Section 2.5.
Consider, for some $\alpha \in (0, 1)$ and $M > 0$, the class $\mathcal{P}_{1+\alpha}^M$ of all distributions satisfying

$$\mathbb{E}\left[|X - \nu(P)|^{1+\alpha}\right] \leq M \ .$$

We prove that, for any $n \geq 1$, $x > \log 2$ and any estimator $\widehat{\nu}_x$ there exists $P \in \mathcal{P}_{1+\alpha}^M$ such that, with probability larger than e^{-x} the difference $\widehat{\nu}_x - \nu(P)$ is larger than $M^{1/(1+\alpha)}\left(\frac{x}{n}\right)^{\alpha/(1+\alpha)}$. This result proves that neither (2.2) (and thus (2.3)) nor (2.4) can be achieved over $\mathcal{P}_{1+\alpha}^M$. In particular, we cannot build confidence intervals for $\nu(P)$ of size $O(n^{-1/2})$ and the rate $n^{-\alpha/(1+\alpha)}$ achieved by the empirical mean when P lies in the domain of attraction of $(1 + \alpha)$-stable distributions (see Feller [Fel71] for definitions and theorems) is optimal.

2.3 Main proof ideas

This section gathers the main ideas underlying the proofs of our results. We present the median of means principle and the method of confidence intervals for the construction of subgaussian estimators in Section 2.3.1. Section 2.3.2 presents the 1-parameter family method to build minimax lower bounds in our problem.

2.3.1 Upper bounds

Our *constructions of estimators* use two main ideas. The first one is the *median of mean principle*. Suppose to simplify that x is a divisor of n. As for the construction of V-fold estimators, following Hsu [Hsu10], we divide $\{1, \dots, n\}$ into x disjoint blocks B_j with the same cardinality $|B_j| = n/x$ and build the vectors of (empirical) *means* $Y = (Y_i)_{i=1}^x$, where $Y_i = \frac{x}{n}\sum_{j\in B_i} X_j$, our preliminary estimator is the median $\overline{Y}_x$ of the vector Y (the median of means). Using Tchebycheff's inequality to bound the probability that $Y_i \notin [\nu(P) \pm 2e\sigma(P)\sqrt{\frac{x}{n}}]$ by e^{-1} and a basic concentration bound for binomial random variables, we get that

$$P\left(|\overline{Y}_x - \nu(P)| > 2e\sigma(P)\sqrt{\frac{x}{n}}\right) = P\left(\sum_{i=1}^{x} \mathbf{1}_{Y_i \notin [\nu(P) \pm 2e\sigma(P)\sqrt{\frac{x}{n}}]} \geq \frac{n}{2}\right) \leq e^{-x} \ .$$

In other words, $\overline{Y}_x$ is a single-x subgaussian estimator over $\mathcal{P}_2$ with $L_{\mathcal{P}_2} = 2e$ and $x_n = n$. Now, one cannot turn single-x into multiple-x estimators, but one *can* build multiple-x estimators from the slightly stronger concept of *subgaussian confidence intervals*, that is, roughly speaking, an empirical confidence interval for $\nu(P)$ with “subgaussian length". On the class $\mathcal{P}_2^{\le\sigma}$, one deduces from our collection of medians of mean estimators $(\overline{Y}_x)_{x\in(0,n)}$ the collection of confidence interval $(\widehat{I}_x)_{x\in(0,n)}$ where $\widehat{I}_x = [\overline{Y}_x \pm 2e\sigma\sqrt{\frac{x}{n}}]$.

Now we combine these intervals to produce a multiple-x estimator. For any $k \le n$, the interval $\widehat{I}_k$ has length $4e\sigma\sqrt{\frac{k}{n}}$ and satisfy

$$P(\nu(P) \in \widehat{I}_k) \ge 1 - e^{-k} \quad .$$

Now define $\widehat{k}_n$ as the minimal $k \le n-2$ such that $\cap_{j=k}^{n-2}\widehat{I}_j \neq \emptyset$, the set $\cap_{j=\widehat{k}_n}^{n-2}\widehat{I}_j$ is a non-empty closed interval and our final estimator $\widehat{\nu}$ is its midpoint. To conclude, let $x \le n-4$ and $k = \lfloor x \rfloor + 2$, then

1. By a union bound $P(\nu(P) \in \cap_{j=k}^{n-2}\widehat{I}_j) \ge 1 - e^{1-k} \ge 1 - e^{-x}$.

2. When $\left\{\nu(P) \in \cap_{j=k}^{n-2}\widehat{I}_j\right\}$ holds, $\widehat{k}_n \le k$.

3. When $\widehat{k}_n \le k$, $\widehat{\nu} \in \widehat{I}_k$.

4. When both $\widehat{\nu}$ and $\nu(P)$ belong to $\widehat{I}_k$,

$$|\widehat{\nu} - \nu(P)| \le = 4e\sqrt{\frac{k}{n}} = 4e\sqrt{\frac{x+2}{n}} \le 4\sqrt{2}e\sqrt{\frac{x+1}{n}} \quad .$$

Hence, $\widehat{\nu}$ is a x-multiple weakly $4e\sqrt{2}$-subgaussian estimator over $\mathcal{P}_2^{\le\sigma}$ with $x_n = c_{\mathcal{P}_2^{\le\sigma}} n$ and a subgaussian one over $\mathcal{P}_2^{\sigma}$.

Although general, the method of confidence intervals looses constant factors. Our second idea for building estimators, which is specific to the bounded kurtosis case, is to use a data-driven truncation of the data to improve the performances of the empirical mean. By using preliminary estimators of the mean and variance (also based on the median of means principle), we truncate the random variables in the sample and obtain a Bennett-type concentration inequality with sharp constant $L = \sqrt{2} + o(1)$. A crucial point in this analysis is to show that our truncation mechanism is fairly insensitive to the preliminary estimators being used.

2.3.2 Lower bounds

The *negative results* are minimax lower bounds over simple families of distributions such as Bernoulli distributions, Laplace distributions with fixed scaling parameter for single-x, and the Poisson family for multiple-x estimators. The main point is that it is easy to compare the probabilities of an event for different values of the parameter. Interestingly, Catoni [Cat12a] also derives his lower bounds using a one dimensional family (in his case, Gaussians with fixed variance $\sigma^2 > 0$).

As an example, consider the class of *all Laplace distibutions with scale parameter equal to* 1. To define such a distribution, let $\lambda \in \mathbb{R}$ and let La_λ be the probability measure on $\mathbb{R}$ with density

$$\frac{d\mathsf{La}_\lambda}{dx}(x) = \frac{e^{-|x-\lambda|}}{2} \;.$$

Denote by $\mathcal{P}_{\mathsf{La}} = \{\mathsf{La}_\lambda \,:\, \lambda \in \mathbb{R}\}$ the class of all such distributions.
A simple calculation reveals that for all $\lambda \in \mathbb{R}$, the mean, variance, and central third moment are $\nu_{\mathsf{La}_\lambda} = \lambda$, $\sigma^2_{\mathsf{La}_\lambda} = 2$ and $\mathsf{La}_\lambda|X-\lambda|^3 = 6 \le (\eta\,\sigma_{\mathsf{La}_\lambda})^3$ with $\eta = 3^{1/3}\,2^{1/6}$. The next result proves that single-x L-subgaussian estimators are limited to linear x_n even over the one-dimensional family $\mathcal{P}_{\mathsf{La}}$.

Theorem 1. *If $n \ge 2$ then, for any constant $L \ge \sqrt{2}$, there are no single-x L-subgaussian estimators for $(\mathcal{P}_{\mathsf{La}}, 9L^2n-1)$.*

Proof. We proceed by contradiction, assuming that there exist L-subgaussian single-x estimators $\widehat{E}_x$ for $(\mathcal{P}_{\mathsf{La}}, x)$ where $x = 9L^2n-1$. We set

$$\lambda = 2L\sqrt{2\,(1+x)/n}$$

and consider $X_1^n =_d \mathsf{La}_0^{\otimes n}$ and $Y_1^n =_d \mathsf{La}_\lambda^{\otimes n}$. The triangle inequality applied to the exponents of $d\mathsf{La}_\lambda/dy$ and $d\mathsf{La}_0/dy$ shows that the densities of the two product measures satisfy, for all $y_1^n = (y_1,\dots,y_n) \in \mathbb{R}^n$

$$\frac{d\mathsf{La}_0}{dy_1^n}(y_1^n) \ge e^{-\lambda n}\,\frac{d\mathsf{La}_\lambda}{dy_1^n}(y_1^n)\;,$$

and therefore,

$$\mathbb{P}\left(\widehat{E}_x(X_1^n) \ge \frac{\lambda}{2}\right) \ge e^{-\lambda n}\mathbb{P}\left(\widehat{E}_x(Y_1^n) \ge \frac{\lambda}{2}\right). \tag{2.5}$$

Using the definition of λ and the fact that $\nu_{\mathsf{La}_\lambda} = \lambda$ and $\sigma^2_{\mathsf{La}_\lambda} = 2$, we see that the right-hand side above is simply

$$e^{-\lambda n}\,\mathbb{P}\left(\widehat{E}_x(Y_1^n) \ge \nu_{\mathsf{La}_\lambda} - L\,\sigma_{\mathsf{La}_\lambda}\sqrt{\frac{1+x}{n}}\right) \ge e^{-\lambda n}\,(1-e^{-x}).$$

On the other hand, the left-hand side in (2.5) is

$$\mathbb{P}\left(\widehat{E}_x(X_1^n) \ge \nu_{\mathsf{La}_0} + L\sigma_{\mathsf{La}_0}\sqrt{\frac{1+x}{n}}\right) \le e^{-x}.$$

We deduce

$$e^{-\lambda n} \le \frac{e^{-x}}{1-e^{-x}} \le 2e^{-x}.$$

If we use again the definition of λ, we see that

$$e^{-2L\sqrt{n\,2(1+x)}} \le 2\,e^{-x},$$

or

$$e^{-6\sqrt{2}\,L^2\,n} \le 2e^{1-9L^2n} \Rightarrow n \le \frac{1+\ln 2}{L^2\,(9-6\sqrt{2})}\;.$$

For $L \ge \sqrt{2}$, some simple estimates show that this leads to a contradiction when $n \ge 2$. □

2.4 Bibliographic remarks

The explicit distinction between single-x and multiple-x subgaussian estimators, and our constructions of multiple-x subgaussian estimators for large x_n, are all new, although the method of confidence intervals might be related at a high level to Lepskii's adaptation method [Lep90, Lep91]. On the other hand, constructions of single-x estimators are implicit in older works on stochastic optimization of Nemirovsky and Yudin [NY83] (see also Levin [Lev05] and Hsu [Hsu10]), sampling from large discrete structures by Jerrum, Valiant, and Vazirani [JVV86], and sketching algorithms, Alon, Matias, and Szegedy [AMS96]. Besides the estimator selection problem that we mentioned earlier, subgaussian estimators have received many attention recently, as well as their generalizations to multivariate settings, and their applications in a variety of statistical learning problems where heavy-tailed distributions may be present, see Catoni [Cat12a], Hsu and Sabato [HS14], Brownlees, Joly, and Lugosi [BJL15], Minsker [Min13], Audibert and Catoni [AC11], Bubeck, Cesa-Bianchi, and Lugosi [BCBL13], or my paper with R. I. Oliveira [19]. Most of these papers use single x subgaussian estimators. Catoni's paper [Cat12a] is the closest in spirit to ours, as it focuses on subgaussian mean estimation as a fundamental problem. That paper presents single x subgaussian estimators with nearly optimal $L = \sqrt{2} + o\,(1)$ for a wide range of x and the classes $\mathcal{P}_2^{\sigma}$ and $\mathcal{P}_{\mathrm{krt}\leq\kappa}$. The single x subgaussian estimator introduced by [Cat12a] may be converted into a multiple-x estimators with subexponential (instead of subgaussian) tails for $\mathcal{P}_2^{\sigma}$ by choosing the single parameter of the estimator appropriately. Loosely speaking, this corresponds to squaring the term $\sqrt{x}$ in (2.3). Catoni also obtains multiple-x estimators for $\mathcal{P}_2$ with subexponential tails. These ideas are strongly related to Audibert and Catoni's paper on robust least-squares linear regression [AC11].

2.5 An application to estimator selection

To conclude this section, I would like to present an application of subgaussian estimators to estimator selection. The main ideas of this paragraph come from the paper of Baraud [Bar11]. It is an adaptation of his general construction to the least-squares density estimation framework that was originally developed in my paper with R. I. Oliveira [19].

As in Section 1.1.1, we assume that we want to estimate the density f with respect to a known measure μ of a probability measure P based on the observation of an i.i.d. sample $X_1, \ldots, X_n$. To this purpose, we are given a collection $(\widehat{f}_m)_{m\in\mathcal{M}_n}$ of estimators, for example, thresholded estimators with various thresholds λ, projection estimators, linear estimators or even “black box" estimators provided by some “experts". Let S_p be a linear space with finite dimension p that may grow with n. The space S_p is typically a large dimensional approximation space for all the estimators $(\widehat{f}_m)_{m\in\mathcal{M}_n}$. To avoid technicalities in this presentation, assume now that all $(\widehat{f}_m)_{m\in\mathcal{M}_n}$ actually belong to S_p. Let $(\varphi_i)_{i\in\{1,\ldots,p\}}$ denote an orthonormal basis of S_p so each estimator $\widehat{f}_m$ is equal to

$$\widehat{f}_m = \sum_{j=1}^{p} \widehat{\beta}_{m,j}\varphi_j \ ,$$

its loss $P\gamma(\widehat{f}_m)$ is equal to

$$P\gamma(\widehat{f}_m) \le \sum_{j=1}^{p} \widehat{\beta}_{m,j}^2 - 2\widehat{\beta}_{m,j}P\varphi_j \enspace .$$

To estimate this loss and build a selection criterion, it is therefore sufficient to estimate, as in Section 1.1.3, the expectations $P\varphi_j$, for all $j \in \{1,\dots,p\}$. Assume that some deterministic upper bound $\sigma_n^2 \ge \|f\|_\infty$ is available to the statistician, then, for any function φ_j, $P\varphi_j^2 \le \|f\|_\infty \|\varphi_j\|^2 \le \sigma_n^2$. Therefore, one can build a weakly subgaussian estimator $\widehat{\varphi}_j$ of $P\varphi_j$ satisfying, for some absolute constants C and L,

$$\forall x \in [0, Cn], \qquad \mathbb{P}\left(|\widehat{\varphi}_j - P\varphi_j| > L\sigma_n\sqrt{\frac{1+x}{n}} \right) \le e^{-x} \enspace . \tag{2.6}$$

Moreover, (2.6) holds even when $\varphi_j(X)$ may have heavy-tailed distributions. Define the functions $\ell(x) = L\sigma_n\sqrt{1+x+\log p}$ and the event

$$\Omega(x) = \left\{ \forall j \in \{1,\dots,p\}, \qquad |\widehat{\varphi}_j - P\varphi_j| \le \frac{\ell(x)}{\sqrt{n}} \right\} \enspace .$$

For any x such that $x + \log p \le Cn$, $\mathbb{P}(\Omega(x)) \ge 1 - e^{-x}$. Let $\mathcal{P}_p$ denote the collection of all subsets of $\{1,\dots,p\}$ and for any $E \in \mathcal{P}_p$, let S_E denote the linear span of $(\varphi_i)_{i\in E}$. To any $m \in \mathcal{M}_n$, we associate the collection $(\widetilde{f}_{m,E})_{E\in\mathcal{P}_p}$ of projections $\widetilde{f}_{m,E} = \sum_{j\in E}\widehat{\beta}_{m,j}\varphi_j$ of $\widehat{f}_m$ onto S_E. We finally define

$$\mathcal{C}(m) = \min_{E\in\mathcal{P}_p}\left\{ \left\|\widetilde{f}_{m,E}\right\|^2 - 2\sum_{j\in E}\widehat{\beta}_{m,j}\widehat{\varphi}_j + \frac{1}{2}\left\|\widetilde{f}_{m,E} - \widehat{f}_m\right\|^2 + \mathrm{pen}(E) \right\} \enspace ,$$

and the estimator

$$\widehat{f}_{\widehat{m}}, \qquad \text{where} \qquad \widehat{m} = \arg\min_{m\in\mathcal{M}_n} \mathcal{C}(m) \enspace .$$

As in Section 1.1, the penalty pen compensates the fluctuations of the estimators. Without strong assumptions on $\widehat{f}_m$, we would have to choose penalties proportional to p/n which would yield to poor risk bounds for the resulting estimators. This is why we project the estimators on linear subspaces S_E and build penalties $\mathrm{pen}(E)$ to get better risk bounds when $\widehat{f}_m$ is often close to S_E without having to assume it beforehand. Notice that

$$\begin{aligned}\left\|\widetilde{f}_{m,E}\right\|^2 - 2\sum_{j\in E}\widehat{\beta}_{m,j}\widehat{\varphi}_j = P\gamma(\widetilde{f}_{m,E}) &- 2\sum_{j\in E}(\widehat{\beta}_{m,j} - P\varphi_j)(\widehat{\varphi}_j - P\varphi_j) \\ &- 2\sum_{j\in E}(P\varphi_j)(\widehat{\varphi}_j - P\varphi_j) \enspace .\end{aligned}$$

The last term in the right hand side of this equation can be forgotten in a first analysis. The second term is controlled on $\Omega(x)$ by

$$\forall E \subset \{1,\dots,p\}, \qquad \left| 2\sum_{j\in E}(\widehat{\beta}_{m,j} - P\varphi_j)(\widehat{\varphi}_j - P\varphi_j) \right| \le \frac{1}{2}\left\|\widetilde{f}_{m,E} - f_E\right\|^2 + 2\ell^2(x)\frac{|E|}{n} \enspace ,$$

where f_E is the orthogonal projection of f onto S_E. Therefore, if $\mathrm{pen}(E) = 2\ell^2(x)\frac{|E|}{n}$, one deduces the following bounds

$$\forall E \subset \{1,\dots,p\}\,, \qquad \frac{1}{4}\left\|f-\widehat{f}_m\right\|^2 \leq \|f-f_E\|^2+\frac{1}{2}\|f_E-\widetilde{f}_{m,E}\|^2+\frac{1}{2}\left\|\widetilde{f}_{m,E}-\widehat{f}_m\right\|^2 \lesssim \mathcal{C}(m)$$

and finally, for all $E \subset \{1,\dots,p\}$,

$$\begin{aligned}\mathcal{C}(m) &\lesssim \|f-f_E\|^2+\frac{3}{2}\|f_E-\widetilde{f}_{m,E}\|^2+\frac{1}{2}\left\|\widetilde{f}_{m,E}-\widehat{f}_m\right\|^2+2\mathrm{pen}(E)\\ &\leq 3\left\|f-\widehat{f}_m\right\|^2+\frac{7}{2}\left\|\widehat{f}_m-\widetilde{f}_{m,E}\right\|^2+2\mathrm{pen}(E)\ .\end{aligned}$$

Thus, $\mathcal{C}(m)$ is a proxy for the loss of the estimators. In fact, using basic algebra, we get that there exists an absolute constant $0 < c < 1$ such that, on $\Omega(x)$,

$$c\left\|\widehat{f}_{\widehat{m}}-f\right\|^2 \leq \inf_{m\in\mathcal{M}_n}\left\{\left\|\widehat{f}_m-f\right\|^2+\min_{E\in\mathcal{P}_p}\left\{\left\|\widetilde{f}_{m,E}-\widehat{f}_m\right\|^2+\mathrm{pen}(E)\right\}\right\}\ .$$

This inequality has the flavor of an oracle inequality as it proves some optimality of the selected estimator via the comparison of its loss $\|\widehat{f}_{\widehat{m}}-f\|^2$ with the infimum of the losses $\|\widehat{f}_m-f\|^2$. As already mentioned, the bound is good when at least one estimator is simultaneously close to f and to a small dimensional subspace S_E of S. What is interesting is that Baraud's construction allows to derive such powerful results from a collection of subgaussian estimators $(\widehat{\varphi}_j)_{j\in\{1,\dots,p\}}$. Besides, Baraud's estimators can be computed efficiently, at least when we start with a family of easily computable estimators in a particular framework of Gaussian regression [BGH14].

Chapter 3

Separation rates for multiple testing

This chapter is devoted to the presentation of my article with M. Fromont and P. Reynaud-Bouret [2]. Based on relationships between the theories of aggregated tests and multiple testing, we adapt the definition of separation rates to multiple testing and develop a minimax theory to measure the performances of multiple testing procedures.

3.1 Gaussian regression framework

Our article is quite general, but we shall focus in these notes on the elementary following Gaussian regression model where one observes a vector $Y \in \mathbb{R}^n$ such that, for some unknown signal $f \in \mathbb{R}^n$ of interest and some unknown standard Gaussian vector $\varepsilon \sim \mathcal{N}(0, I_n)$

$$Y = f + \varepsilon \ .$$

The vector Y is an n-dimensional Gaussian vector with distribution $P_f \equiv \mathcal{N}(f, I_n)$.

3.2 Simple tests

Given a linear subspace $S_0 \subset \mathbb{R}^n$ with dimension d_0, we first test the assumption $H_0 : f \in S_0$ with the χ^2-statistic $T_0 = \|\Pi_0 Y\|^2$ where Π_0 is the orthogonal projection onto the orthogonal of S_0 and $\|.\|$ denotes the Euclidean norm in $\mathbb{R}^n$. For any $\alpha \in [0,1]$, let $q_{n,d_0,\alpha}$ the $1-\alpha$ quantile of the χ^2 distribution with $n - d_0$ degrees of freedom. For any $\alpha \in (0,1)$, the test $\phi_{n,0,\alpha} = \mathbf{1}_{T_0 > q_{n,d_0,\alpha}}$ of H_0 against $H_n : f \in \mathbb{R}^n \backslash S_0$ has level $\mathrm{ER}(\phi_{n,0,\alpha}, S_0) = \alpha$.
More generally, for any integer d_m such that $d_0 \leq d_m \leq n$ and any $\alpha \in [0,1]$, let $q_{d_m,d_0,\alpha}$ the $1-\alpha$ quantile of the χ^2 distribution with $d_m - d_0$ degrees of freedom. Then, given a linear space A_m with dimension d_m such that $S_0 \subset A_m \subset \mathbb{R}^n$, one can consider the projection $\Pi_{m,0}Y$ of Y onto the orthogonal of S_0 into A_m and the statistic $T_{m,0} = \|\Pi_{m,0}Y\|^2$. The test $\phi_{m,0,\alpha} = \mathbf{1}_{T_{m,0} > q_{d_m,d_0,\alpha}}$ of H_0 against $H_{1,m}$: $f \in A_m \setminus S_0$ has level $\mathrm{ER}(\phi_{m,0,\alpha}, S_0) = \alpha$. Now a basic remark for our extension to multiple testing is that $\Pi_{m,0}Y$ is also the projection of Y onto the orthogonal of $S_m = S_0 + A_m^{\perp}$ so $T_{m,0}$ is also a test of $H_m : f \in S_m$ against $H_{n,m} : f \in \mathbb{R}^n \setminus S_m$.
The performances of tests of a single hypothesis are evaluated through the notion of separation rates. The following definition, due to Baraud [Bar02], can be viewed as

a non-asymptotic version of Ingster's original work [Ing93]. Let d denote a distance on $\mathbb{R}^n$. For any $g \in \mathbb{R}^n$ and any subset $S \subset \mathbb{R}^n$, let

$$d(g,S) := \inf_{h\in S} d(g,h) \ .$$

Given β in $(0,1)$, two linear subspaces $S_0 \subset A_m \subset \mathbb{R}^n$, and a test $\overline{\Phi}$ of $H_0 : f \in S_0$ against the alternative $H_{1,m} : f \in A_m \setminus S_0$, the uniform separation rate of $\overline{\Phi}$ with prescribed second kind error rate β is defined by

$$\mathrm{SR}_d^{\beta}\left(\overline{\Phi}, A_m, S_0\right) \ = \ \inf\left\{ r>0, \sup_{f\in A_m,\ d(f,S_0)\geq r} P_f(\overline{\Phi}=0) \leq \beta \right\} \ .$$

In words, $\mathrm{SR}_d^{\beta}\left(\overline{\Phi}, A_m, S_0\right)$ is the minimal distance r such that the test $\overline{\Phi}$ of H_0 against $H_{1,m,r} \ : \ f \ \in \ \{g \in A_m, \ \text{s.t.} \ d(g,S_0) \geq r\}$ has second kind error upper bounded by β.

The minimax separation rate over A_m with prescribed level α and second kind error β is defined as

$$\mathrm{mSR}_d^{\alpha,\beta}\left(A_m, S_0\right) = \inf_{\overline{\Phi}} \mathrm{SR}_d^{\beta}\left(\overline{\Phi}, A_m, S_0\right) \ ,$$

where the infimum is taken over all possible level-α tests. Then, a level-α test $\overline{\Phi}$ is called minimax over A_m if there exists $C_{\alpha,\beta}$ such that $\mathrm{SR}_d^{\beta}\left(\overline{\Phi}, A_m, S_0\right) \leq C_{\alpha,\beta}\mathrm{mSR}_d^{\alpha,\beta}\left(A_m, S_0\right)$. Finally, it is called adaptive in the minimax sense over a collection $\mathcal{A} = (A_m)_{m\in\mathcal{M}_n}$ of classes A_m of alternatives if it is simultaneously minimax over all $A_m \in \mathcal{A}$, without knowing in advance the subset A_m to which f belongs.

It is not very hard to prove that the tests $\phi_{m,0,\alpha}$ are minimax. To build adaptive tests, the idea of Baraud [Bar02] is to aggregate these, that is, to reject H_0 if one of the tests $H_0 : f \in S_0$ against $H_{1,m} : f \in A_m \setminus S_0$ is rejected. Recall that this also means that one of the tests of H_m against $H_{n,m}$ is rejected. Of course, to obtain a final test with level α, the level of each individual tests has to be corrected. More precisely, we shall choose a collection $(\alpha_m)_{m\in\mathcal{M}_n}$ of levels and reject H_0 if $\overline{\Phi}_{\mathcal{M}_n} = 1$, where

$$\overline{\Phi}_{\mathcal{M}_n} = \max_{m\in\mathcal{M}_n} \phi_{m,\alpha_m} \ .$$

To ensure a level-α test, we have to choose $(\alpha_m)_{m\in\mathcal{M}_n}$ such that

$$\sup_{f\in S_0} \mathbb{E}_f\left[\sup_{m\in\mathcal{M}_n} \phi_{m,\alpha_m}\right] \leq \alpha \ . \tag{3.1}$$

Classical choices of $(\alpha_m)_{m\in\mathcal{M}_n}$ are based on union bounds such as Bonferroni weights $\alpha_m = \alpha/|\mathcal{M}_n|$ for any $m \in \mathcal{M}_n$. A refined strategy has been proposed by Baraud, Huet and Laurent [BHL03]: given $(w_m)_{m\in\mathcal{M}_n}$ such that $\sum_{m\in\mathcal{M}_n} w_m \leq 1$, choose $\alpha_m = w_m u_\alpha$, where

$$u_\alpha = \sup\left\{ u>0, \ \text{s.t.} \ \sup_{f\in S_0} \mathbb{E}_f\left[\sup_{m\in\mathcal{M}_n} \phi_{m,w_m u}\right] \leq \alpha \right\} \ .$$

Adaptive properties of such aggregated tests have been studied in many frameworks, among them of course Gaussian regression frameworks with various classes of alternatives [Spo96, Bar02, BHL03, LLM12, DS01], density or Poisson processes frameworks [Ing00, FL06, FLRB11], or more complex ones corresponding to two-sample type problems [8, FLRB13, CD15].

3.3 Multiple testing

Let us now present the multiple testing framework. Given a collection of linear subspaces $(S_m)_{m\in\mathcal{M}_n}$ of $\mathbb{R}^n$, we denote, for any $m \in \mathcal{M}_n$, by H_m the hypothesis $H_m : f \in S_m$. Our goal is to test simultaneously all the assumptions $(H_m)_{m\in\mathcal{M}_n}$. For this purpose, we also start with the collection of elementary tests $\Phi_{\mathcal{M}_n} = (\phi_{m,\alpha})_{m\in\mathcal{M}_n,\alpha\in(0,1)}$ of the previous section and recall that each test $\phi_{m,\alpha}$ of H_m against $H_{n,m}$ has level α.

We want to build a multiple test which should infer which assumptions are true and which assumptions are false. Following Goeman and Solari [GS10], we define the set of *false hypotheses* by

$$\forall f \in \mathbb{R}^n, \qquad F(f) = \{m \in \mathcal{M}_n, \text{ s.t. } f \notin S_m\} \ ,$$

and the set of true hypotheses by $T(f) = \mathcal{M}_n \setminus F(f)$. A multiple test R_α is a data-driven subset of $\mathcal{M}_n$ of *rejected hypotheses* whose aim is to infer the set of *false hypotheses*.

In the following, we use the family-wise error rate FWER as a measure the first kind error of a multiple test. It is defined by

$$\text{FWER}(R_\alpha) = \sup_{f\in\mathbb{R}^n} P_f\left(R_\alpha \cap T(f) \neq \emptyset\right) \ ,$$

We seek for multiple tests such that $\text{FWER}(R_\alpha) \le \alpha$, that is we want to be sure that, except on a set with prescribed probability, all rejected assumptions are false. An elementary way to achieve this goal is to set $R_\alpha = \{m \in \mathcal{M}_n, \ \text{s.t.} \ \phi_{m,\alpha_m} = 1\}$. Actually, these tests satisfy

$$\text{FWER}(R_\alpha) = \sup_{f\in\mathbb{R}^n} \mathbb{E}_f\left[\sup_{m\in T(f)} \phi_{m,\alpha_m}\right] \ ,$$

and the condition $\text{FWER}(R_\alpha) \le \alpha$ reminds the condition (3.1) required to ensure a level α for the aggregated test. Actually, one can check for check for example that the multiple tests derived from Bonferroni procedures or Baraud, Huet and Laurent correction of the level satisfy $\text{FWER}(R_\alpha) \le \alpha$, the proof is quite elementary.

A refined strategy, still controlling the FWER is given by the sequential Holm's strategy [Hol79]. Denote by $R_\alpha^0 = \emptyset$ and define recursively

$$\forall k \ge 0, \qquad R_\alpha^{k+1} = R_\alpha^k \cup \left\{m \in \mathcal{M}_n, \text{ s.t. } \phi_{m,\frac{\alpha}{|\mathcal{M}_n|-|R_\alpha^k|}} = 1\right\} \ ,$$

the set of hypotheses that are rejected using a Bonferroni procedure on the remaining hypotheses after k-steps. The sequence converges in at most $|\mathcal{M}_n|$ steps and the final multiple test of Holm is defined by $R_\alpha = R_\alpha^{|\mathcal{M}_n|}$. Baraud, Laurent and Huet initial collection of rejected assumptions can also be incremented using a recursive procedure. The resulting test is the same as the one built with a min $-p$ procedure (see [DvdL07] for example). Both recursive algorithms are step-down procedures and their FWER is controlled by α thanks to [GS10, Theorem 1]. The assumption called "one step" in this last theorem is verified using the analogy with the control (3.1) of the level of an aggregated test.

The take-home message here is that both aggregated tests and multiple tests are based on elementary tests $(\phi_{m,\alpha})$ of a collection of hypotheses H_m, $m \in \mathcal{M}_n$. They differ in their objectives, while aggregated tests are only concerned with one assumption H_0 included in the intersection $\cap_{m\in\mathcal{M}_n} H_m$, multiple tests are concerned with the collection of all false hypotheses $\{H_m, m \in F(f)\}$. The consequence is that an aggregated test only requires that the first step $R^1_\alpha = \{m \in \mathcal{M}_n,\ \text{s.t. } \phi_{m,\alpha_m} = 1\}$ is non-empty to reject to reject H_0. However, using the step-down strategy of [GS10], one can often use the calibration of the level for an aggregation strategy to build multiple tests $R^{|\mathcal{M}_n|}_\alpha$, still controlling the FWER, and rejecting more assumptions than R^1_α.

3.4 Weak family-wise separation rates for multiple tests

Let us now consider a multiple test R_α and the aggregated test $\overline{\Phi}(R_\alpha) = \mathbf{1}_{R_\alpha \neq \emptyset}$ derived from it. Given β in $(0,1)$ and $S \subset \mathbb{R}^n$, the uniform separation rate of $\overline{\Phi}(R_\alpha)$ over S with prescribed second kind error rate β and distance d is $\mathrm{SR}^\beta_d\left(\overline{\Phi}(R_\alpha), S, \cap_{m\in\mathcal{M}_n} S_m\right)$. This quantity is closely related to the maximin optimality criterion of Romano, Shaikh and Wolf [RSW11, Theorem 4.1] which consists in maximizing the power

$$\inf_{f\in S\subset\mathbb{R}^n\setminus\bigcap_{m\in\mathcal{M}_n} S_m} P_f\left(R_\alpha \neq \emptyset\right) \ .$$

The difference is that we look for a minimal distance r between f (in S) and $\cap_{m\in\mathcal{M}_n} S_m$ which guarantees a fixed minimal level of power $(1-\beta)$ for a given procedure. This notion of minimal distance r is considered as a rate of testing (in the spirit of the rates of estimation) used to compare the performance of two testing procedures.

We could naturally define the weak family-wise separation rate as $\mathrm{SR}^\beta_d\left(\overline{\Phi}(R_\alpha), S, \cap_{m\in\mathcal{M}_n} S_m\right)$. However, in this second kind error criterion, only alternatives which deviate from the intersection $\cap_{m\in\mathcal{M}_n} S_m$ are taken into account. Considering such a definition would thus amount to confuse multiple tests with their corresponding aggregated tests, seeing all the tested hypotheses as only intermediate hypotheses to an ultimate one: $f \in \cap_{m\in\mathcal{M}_n} S_m$. This would depart from the multiple testing philosophy, where each tested hypothesis has its own significance and has to be taken into account by itself. To address this requirement, instead of alternatives f in S such that "$d\left(f, \cap_{m\in\mathcal{M}_n} S_m\right) \geq r$" (for $r > 0$), we consider alternatives f in S such that "$\exists m \in \mathcal{M}_n,\ d\left(f, S_m\right) \geq r$". This leads us to consider the set of false hypotheses under P_f at least at distance r from f, that is

$$F_r(f) = \{m \in \mathcal{M}_n,\ d(f, S_m) \geq r\} \ .$$

Note that $F_r(f) \neq \emptyset$ implies that $d(f, \cap_{m\in\mathcal{M}_n} S_m) \geq r$ but the converse is false, see Figure 3.1. We can now introduce the following definition.

Given β in $(0,1)$ and a class $S \subset \mathbb{R}^n$, the weak family-wise separation rate of a multiple test R_α over S with prescribed second kind error rate β is defined by

$$\mathrm{wFWSR}^\beta_d\left(R_\alpha, S\right) \ = \ \inf\left\{ r > 0, \sup_{f\in S,\ F_r(f)\neq\emptyset} P_f\left(R_\alpha = \emptyset\right) \leq \beta \right\} \ .$$

This is the minimal radius r such that, except on a set with prescribed probability, if $f \in S$ departs from at least one assumption by a distance r, one assumption is rejected. This notion of weak family-wise separation rate is related to uniform separation rate thanks to the following result see [2, Proposition 3].

Proposition 2. *For any subset S of $\mathbb{R}^n$ and β in $(0,1)$,*

$$\mathrm{wFWSR}_d^\beta\left(R_\alpha, S\right) \leq \mathrm{SR}_d^\beta\left(\overline{\Phi}(R_\alpha), S, \bigcap_{m\in\mathcal{M}_n} S_m\right) ,$$

with an equality if the collection of hypotheses $(S_m)_{m\in\mathcal{M}_n}$ and the distance d satisfy

$$\forall r > 0, \forall f \in \mathbb{R}^n, \quad [F_r(f) \neq \emptyset] \quad \textit{if and only if} \quad \left[d\left(f, \bigcap_{m\in\mathcal{M}_n} S_m\right) \geq r\right] . \tag{3.2}$$

The necessity of condition (3.2) can be understood in the example drawn in Figure 3.1. Point c is a possible value for f in $\mathrm{SR}_d^\beta\left(\overline{\Phi}(R_\alpha), S, \cap_{m\in\mathcal{M}_n} S_m\right)$ but not in $\mathrm{wFWSR}_d^\beta\left(R_\alpha, S\right)$. The alternative set considered for $\mathrm{wFWSR}_d^\beta\left(R_\alpha, S\right)$ is thus strictly included in the alternatives considered for $\mathrm{SR}_d^\beta\left(\overline{\Phi}(R_\alpha), S, \cap_{m\in\mathcal{M}_n} S_m\right)$, it may therefore be strictly more difficult to control in this example the separation rate in the latter case.

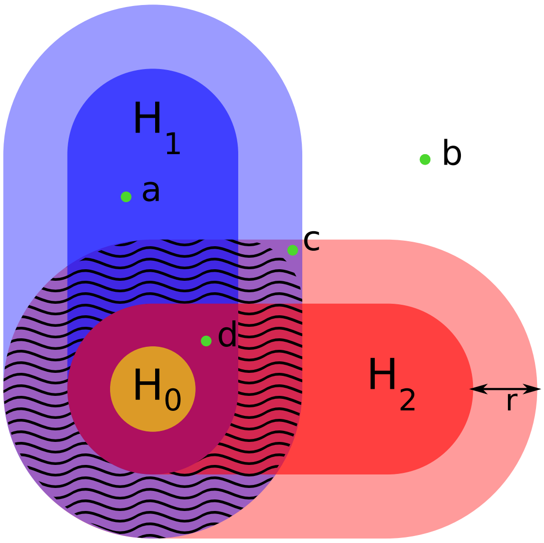

Figure 3.1: Visualization of a multiple testing problem with two hypotheses H_1 and H_2 represented with darker colors. Their r-neighborhoods are of lighter shade. The r-neighborhood of $S_1 \cap S_2$ is hatched. The hypothesis H_0 is strictly included in $H_1 \cap H_2$. Point a corresponds to a f such that $T(f) = \{1\}$ and $F(f) = F_r(f) = \{2\}$. Point b corresponds to a f such that $T(f) = \emptyset$ and $F(f) = F_r(f) = \{1,2\}$. Point c corresponds to a f such that $T(f) = \emptyset$, $F(f) = \{1,2\}$, $F_r(f) = \emptyset$ but $d(f, S_1 \cap S_2) \geq r$. Point d corresponds to a f such that $T(f) = \{1,2\}$ and $F(f) = F_r(f) = \emptyset$ but $f \notin S_0$.

Condition (3.2) is satisfied under different classical assumptions on the collection of hypotheses. For example, if the collection $(S_m)_{m\in\mathcal{M}_n}$ is closed (under intersection), that is

$$\forall (m,m') \in \mathcal{M}_n^2, \qquad S_m \cap S_{m'} \in (S_m)_{m\in\mathcal{M}_n} ,$$

then condition (3.2) is always satisfied. For instance, the collection $(S_m)_{m\in\{1,\dots,n\}}$, where $S_n = \mathbb{R}^n$ and, for all $m \in \{1, \dots, n-1\}$,

$$S_m = \{ f \in \mathbb{R}^n, \text{ s.t. } f_{m+1} = \dots = f_n = 0 \}$$

is closed and therefore condition (3.2) is satisfied.
Furthermore, consider the collection $(S'_m)_{m\in\{1,\dots,n\}}$, where, for all $m \in \{1, \dots, n\}$,

$$S'_m = \{ f \in \mathbb{R}^n, \text{ s.t. } f_m = 0 \}, \qquad \text{(which is not closed)} \ .$$

It satisfies condition (3.2) with the supremum distance $d = d_\infty$, that is

$$d_\infty(f,g) = \max_{i=1,\dots,n} |f_i - g_i| \ , \tag{3.3}$$

but not with any other distance d_s for $s \geq 1$ defined by

$$d_s(f,g) = \left(\sum_{i=1}^n |f_i - g_i|^s \right)^{1/s} \ . \tag{3.4}$$

In this kind of examples, the following more general result can be used.

Proposition 3. *[2, Proposition 4] Let d be a distance on $\mathbb{R}^n$, and S be a subset of $\mathbb{R}^n$. If there exists some distance d' on $\mathbb{R}^n$ such that:*

$$\forall f \in \mathbb{R}^n, \ \forall r > 0, \quad [F_r(f) \neq \emptyset] \quad \textit{if and only if} \quad \left[d'\left(f, \bigcap_{m\in\mathcal{M}_n} S_m \right) \geq r \right] \ , \tag{3.5}$$

then for every $\beta \in (0,1)$,

$$\mathrm{wFWSR}_d^\beta(R_\alpha, S) = \mathrm{SR}_{d'}^\beta \left(\overline{\Phi}(R_\alpha), S, \bigcap_{m\in\mathcal{M}_n} S_m \right) \ .$$

This result applies to the collection $(S'_m, \ m = 1, ..., n)$ and any distance d, for example any distance d_s ($s \geq 1$) defined by (3.4), such that condition (3.5) is satisfied with $d' = d_\infty$. Thus, for every multiple test R_α of $(S'_m, \ m = 1, ..., n)$, every subset S of $\mathbb{R}^n$, and every distance d,

$$\mathrm{wFWSR}_d^\beta(R_\alpha, S) = \mathrm{SR}_{d_\infty}^\beta \left(\overline{\Phi}(R_\alpha), S, \{0\} \right) \ .$$

3.5 (Strong) Separation rates for multiple testing

We now introduce the following stronger notion of family-wise separation rate.
Given β in $(0,1)$ and $S \subset \mathbb{R}^n$, the family-wise separation rate of a multiple test R_α over S with prescribed second kind error rate β is defined by

$$\mathrm{FWSR}_d^\beta(R_\alpha, S) \ = \ \inf \left\{ r > 0, \ \sup_{f\in S} P_f \left(F_r(f) \cap (\mathcal{M}_n \setminus R_\alpha) \neq \emptyset \right) \leq \beta \right\} \ .$$

This is the minimal radius r ensuring that, except on an event with prescribed probability, all false assumptions are rejected. The family-wise separation rate is a stronger quality criterion than the weak family-wise separation rate, which is formalized in the following result.

Proposition 4. *[2, Proposition 5] For any distance d, any subset S of $\mathbb{R}^n$, and any β in* $(0,1)$,

$$\mathrm{wFWSR}_d^{\beta}\left(R_\alpha, S\right) \leq \mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right) \ .$$

By definition, for fixed S, $\mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right)$ is monotonous in R_α, i.e. if $R_\alpha \subset R'_\alpha$ a.s.,

$$\mathrm{FWSR}_d^{\beta}\left(R'_\alpha, S\right) \leq \mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right) \ . \tag{3.6}$$

In particular, using the step-down procedure instead of its first step improves the performances of a multiple test. Likewise, for fixed R_α, $\mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right)$ is monotonous in S: if $S \subset S'$ then $\mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right) \leq \mathrm{FWSR}_d^{\beta}\left(R_\alpha, S'\right)$. In words, reducing the set of alternatives improves the performances of any multiple test. Finally, remark that wFWSR has the same properties of monotonicity.

This stronger notion of separation rates is used to define a minimax approach for multiple tests.
Given α and β in $(0,1)$, $S \subset \mathbb{R}^n$, the minimax family-wise separation rate over S with prescribed FWER α and prescribed second kind error rate β is defined by

$$\mathrm{mFWSR}_d^{\alpha,\beta}\left(S\right) = \inf_{R_\alpha} \mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right) \ ,$$

where the infimum is taken over all possible multiple tests with a FWER controlled by α. A multiple test R_α, whose FWER is controlled by α, is called minimax over S if there exists $C_{\alpha,\beta}$ such that $\mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right) \leq C_{\alpha,\beta}\mathrm{mFWSR}_d^{\alpha,\beta}\left(S\right)$. It is called adaptive in the minimax sense over a collection $\mathcal{S}$ of subsets S if it is simultaneously minimax over all $S \in \mathcal{S}$, without knowing in advance the S to which f belongs.
From the monotonicity properties of FWSR_d^{β}, we deduce that $\mathrm{mFWSR}_d^{\alpha,\beta}$ is non decreasing: if $S \subset S'$ then $\mathrm{mFWSR}_d^{\alpha,\beta}\left(S\right) \leq \mathrm{mFWSR}_d^{\alpha,\beta}\left(S'\right)$. Furthermore, when $(H_m)_{m\in\mathcal{M}_n}$ is reduced to a single hypothesis $H_0 : f \in S_0$, that is if all $(S_m)_{m\in\mathcal{M}_n} = \{S_0\}$, one has that for any multiple test R_α and any subset S of $\mathbb{R}^n$,

$$\begin{cases} \mathrm{wFWER}\left(R_\alpha\right) = \mathrm{FWER}\left(R_\alpha\right) = \mathrm{ER}\left(\overline{\Phi}(R_\alpha), S_0\right) \ , \\ \mathrm{wFWSR}_d^{\beta}\left(R_\alpha, S\right) = \mathrm{FWSR}_d^{\beta}\left(R_\alpha, S\right) = \mathrm{SR}_d^{\beta}\left(\overline{\Phi}(R_\alpha), S, S_0\right) \ . \end{cases}$$

Conversely, for any single test $\overline{\Phi}$ of H_0 against the alternative $H_1 : f \in S$, one can build $R\left(\{\overline{\Phi}\}\right) = H_0$ if $\overline{\Phi} = 1$ and $R\left(\{\overline{\Phi}\}\right) = \emptyset$ otherwise. One can then check that

$$\begin{cases} \mathrm{ER}\left(\overline{\Phi}, S_0\right) = \mathrm{wFWER}\left(R\left(\{\overline{\Phi}\}\right)\right) = \mathrm{FWER}\left(R\left(\{\overline{\Phi}\}\right)\right) \ , \\ \mathrm{SR}_d^{\beta}\left(\overline{\Phi}, S, S_0\right) = \mathrm{wFWSR}_d^{\beta}\left(R\left(\{\overline{\Phi}\}\right), S\right) = \mathrm{FWSR}_d^{\beta}\left(R\left(\{\overline{\Phi}\}\right), S\right) \ . \end{cases}$$

It is also easy to prove that when $(H_m)_{m\in\mathcal{M}_n}$ is reduced to a single hypothesis $H_0 : f \in S_0$, for any subset S of $\mathbb{R}^n$,

$$\mathrm{mFWSR}_d^{\alpha,\beta}\left(S\right) = \mathrm{mSR}_d^{\alpha,\beta}\left(S, S_0\right) \ .$$

Our minimax approach for multiple tests is thus a generalization of the classical minimax theory for single hypothesis tests. Even when $(H_m)_{m\in\mathcal{M}_n}$ is not reduced to a single hypothesis H_0, both theories have close links established in the following result.

Theorem 5. *[2, Theorem 6] Let d be a distance on $\mathbb{R}^n$ and let $S \subset \mathbb{R}^n$. If there exists some distance d' on $\mathbb{R}^n$ satisfying* (3.5)*, then for every β in $(0, 1)$,*

$$\mathrm{mFWSR}_d^{\alpha,\beta}(S) \geq \mathrm{mSR}_{d'}^{\alpha,\beta}\left(S, \bigcap_{m \in \mathcal{M}_n} H_m\right) . \tag{3.7}$$

The main role of this result is to provide lower bounds for the minimax family-wise separation rate using the abundant literature on classical minimax testing.

3.6 Some examples

These lower bounds may be tight as shown by the following example. For any $f \in \mathbb{R}^n$, define $|f|_0 = |\{i \in \{1, \dots, n\},\ \text{s.t. } f_i \neq 0\}$ and, for any $k \in \{0, \dots, n\}$,

$$S_k = \{f \in \mathbb{R}^n,\ |f|_0 \leq k\} . \tag{3.8}$$

Baraud [Bar02] proved that

$$\mathrm{mSR}_{d_2}^{\alpha,\beta}(S_k, \{0\}) \geq \sigma \left(k \ln\left(1 + \frac{n}{k^2} \vee \sqrt{\frac{n}{k^2}}\right)\right)^{1/2} . \tag{3.9}$$

Moreover, he proved that this lower bound is tight by considering an aggregation of tests of the hypotheses

$$H_m : f \in S_m = \{f \in \mathbb{R}^n,\ \text{s.t. } f_{m+1} = \ldots = f_n = 0\} , \tag{3.10}$$

for all $m \in \{1, \dots, n\}$.

Theorem 6. *[2, Theorem 9] Given α in $(0, 1)$, there exists a multiple tests R_α of $(H_m)_{m \in \{1,\dots,n\}}$, where H_m is defined in* (3.10) *such that* $\mathrm{FWER}(R_\alpha) \leq \alpha$ *and for any k in $\{1, \dots, n\}$, β in $(0, 0.5)$,*

$$\mathrm{FWSR}_{d_2}^{\beta}(R_\alpha, S_k) \leq \sigma\sqrt{k}\left(\sqrt{2\ln(n/\alpha)} + \sqrt{-2\ln(2\beta)}\right) .$$

Therefore, for $k \approx n^\gamma$ for some $\gamma \in (0, 1/2)$ the lower bound in Theorem 5 is tight. Moreover, if (3.2) holds, for any subset S of $\mathbb{R}^n$ and any β in $(0, 1)$,

$$\mathrm{mFWSR}_d^{\alpha,\beta}(S) \geq \mathrm{mSR}_d^{\alpha,\beta}\left(S, \bigcap_{m \in \mathcal{M}_n} H_m\right) . \tag{3.11}$$

In words, under the assumption (3.2) that makes possible the comparison of alternatives, testing multiple hypotheses is more difficult than testing a single hypothesis. It is however not surprising that, when condition (3.2) fails, the inequality (3.11) may not hold. Actually, consider the set of assumptions $H_m : f \in S'_m$, where

$$S'_m = \{f \in \mathbb{R}^n,\ \text{s.t. } f_m = 0\} , \tag{3.12}$$

for all $m \in \{1, \dots, n\}$. Then following result holds.

Theorem 7. *[2, Theorem 7] For any α in $(0,1)$, there exists a multiple test R_α of $(H_m)_{m\in\{1,\dots,n\}}$ where $H_m : f \in S'_m$, with S'_m defined in (3.12) such that* $\mathrm{FWER}(R_\alpha) \leq \alpha$ *and, for all $k \in \{1,\dots,n\}$, and $\beta \in (0,1)$,*

$$\mathrm{FWSR}^{\beta}_{d_2}\left(R_\alpha, S_k\right) \leq \sigma\left(\sqrt{2\ln\left(\frac{k}{2\beta}\right)} + \sqrt{2\ln\left(\frac{n}{\alpha}\right)}\right) \quad .$$

Therefore, the minimax separation rate for the multiple tests problem is far smaller than the corresponding separation rate when $k \approx n^\gamma$ for $\gamma \in (0,1/2)$. Remarking that

$$\mathrm{mSR}^{\alpha,\beta}_{d_\infty}\left(S_k,\{0\}\right) = \mathrm{mSR}^{\alpha,\beta}_{d_\infty}\left(S_1,\{0\}\right) = \mathrm{mSR}^{\alpha,\beta}_{d_2}\left(S_1,\{0\}\right) \quad ,$$

Baraud's bound (3.9) proves that

$$\mathrm{mSR}^{\alpha,\beta}_{d_\infty}\left(S_k,\{0\}\right) \geq \sigma\sqrt{\ln(1+n)}, \tag{3.13}$$

Hence, by Theorem 5, the rate $\sigma\sqrt{\log(1+n)}$ in Theorem 7 is minimax.

One can use in Theorems 6 and 7 the multiple tests based on Bonferroni procedure, which is thus minimax in these very basic Gaussian regression frameworks. This result, that is a bit disappointing, is mainly due to the independence of the p-values of the individual tests of $H_m : f \in S'_m$. We also considered in [2] another Gaussian regression model, where p-values are roughly dependent, to show that Bonferroni procedures are clearly suboptimal from the minimax point of view in this context, while min-p procedures are proved to be adaptive in the minimax sense. The strong dependence structure enabled us to use again known results in the classical minimax theory for single hypothesis tests.

Chapter 4

Bradley-Terry model in random environment

I would like to conclude these notes by presenting an article recently submitted and written with my colleagues in Nice R. Diel and R. Chetrite [15]. We study a question of probabilistic nature on the Bradley-Terry model, but the probabilistic tools involved in many proofs are quite common in statistics in general and in model selection in particular. For example, we repeatedly use concentration inequalities and controls of the supremum of empirical processes.

4.1 Position of the problem

Paired comparisons is a general framework used in various applications such as sport competitions, chess tournaments, comparisons of medical treatments. It is at the core of the estimation by tests theory of Birgé [Bir06a] and used by Baraud [Bar11] to study estimator selection in general frameworks.

The Bradley-Terry model is a toy model of paired comparisons. A set of N players (teams, treatments, ...) called $1,\ldots,N$ face each other once by pairs with independent outcomes. When i faces j, the result is described by a Bernoulli random variable $X_{i,j}$ that is equal to 1 when i beats j, and of course, $X_{j,i} = 1 - X_{i,j}$. Each player has a value $V_i > 0$ modeling its "strength" or its "merit" that is used to define

$$\forall 1 \le i < j \le N, \qquad \mathbb{P}\left(X_{i,j} = 1 | V_1,\ldots,V_N \right) = \frac{V_i}{V_i + V_j} \ .$$

Finally, the score $S_i = \sum_{j=1, j\neq i}^{N} X_{i,j}$ of each player i is used to define its rank at the end of the championship, for example, a player i_N such that $S_{i_N} = \max_{i\in\{1,\ldots,N\}} S_i$ is called a winner. The vector $\mathbb{V}_1^N = (V_1,\ldots,V_N)$ is a random vector whose distribution is described as follows. Let $\mathbb{U}_1^N = (U_1,\ldots,U_N)$ be an i.i.d. sample and for any $i \in \{1,\ldots,N\}$, let $V_i = U_{(i)}$ be the i-th order statistic of $\mathbb{U}_1^N$.

This model has been introduced by Zermelo [Zer29] and rediscovered independently by Bradley and Terry [BT52]. It was later generalized to allow ties ([Dav70, RK67]) or to incorporate within-pair order effects ([DB77]). Despite its simplicity, it has been widely used in applications for example to model sport tournaments, reliability problems, ranking scientific journals,. . . (see [Cat12b] for a recent overview). The problem of estimating the strength in the Bradley-Terry models has also been widely

studied, see for example [Dav63, HT98, SY99, Hun04, GJ05, YYX12] and references therein.
Nevertheless, the Bradley-Terry model has rarely been associated to random environment (see however [SR09]) and, to the best of our knowledge, has never been studied mathematically in this context. The random environment seems however natural as it allows to manage the heterogeneity of strengths of players globally, without having to look at each one specifically. It is a method already used fruitfully in other areas such as continuous or discrete random walks (see [Zei12] or [DR14] for recent presentations). Moreover, the inference of the distribution of the strength might be much simpler than the inference of all strengths (one can imagine inference among a one or two parameters family of distributions), and prediction of statistics using only this distribution might therefore be simplified by the introduction of the random media. Our problem here is to understand how the choice of the distribution for the strengths of players influences the ranking of the players. In particular, does a player with the highest strength end up with the highest score? And if not, what is the number of potential winners?
Hereafter, let U denote a copy of U_1 independent of $\mathbb{U}_1^N$, let Q denote the tail distribution function of U, that is $Q(t) = \mathbb{P}\left(U > t\right)$ for all $t > 0$ and $\operatorname{supp} Q$ its support, let $\mathbb{P}$ denote the probability of an event with respect to the randomness of $\mathbb{V}_1^N$ and $(X_{i,j})_{1\leq i<j\leq N}$, it called the annealed probability. Let $\mathbb{P}_V$ denote the probability measure given $\mathbb{V}_1^N$, that is $\mathbb{P}\left(\,\cdot\,|\mathbb{V}_1^N\right)$, it is called the quenched probability. In particular,

$$\forall 1 \leq i < j \leq N, \qquad \mathbb{P}_V\left(X_{i,j} = 1\right) = \frac{V_i}{V_i + V_j} \ .$$

We are interested in the asymptotic probability that the “best" player wins, meaning that the player N with the largest strength V_N ends up with the best score S_N.

4.2 Main results

The first theorem is gives conditions under which this (annealed) probability is asymptotically 1 when the number of players $N \to \infty$.

Theorem 8. *Assume that there exist $\beta \in (0, 1/2)$ and $x_0 > 0$ in the interior of $\operatorname{supp} Q$ such that $Q^{1/2-\beta}$ is convex on $[x_0, \infty)$ and that $\mathbb{E}\left[U^2\right] < \infty$. Then,*

$$\mathbb{P}\left(\text{“the player } N \text{ wins"}\right) \geq \mathbb{P}\left(S_N > \max_{1\leq i\leq N-1} S_i\right) \xrightarrow[N\to\infty]{} 1 \ .$$

When $\operatorname{supp} Q = \mathbb{R}_+$, the convexity condition is not very restrictive as it is satisfied by standard continuous distributions with tails function $Q(x) \simeq e^{-x^a}$, $Q(x) \simeq x^{-b}$ or $Q(x) \simeq (\log x)^{-c}$. The moment condition $\mathbb{E}\left[U^2\right] < \infty$ is more restrictive but still allows natural distributions of the merits such as exponential or exponential of Gaussian. It is a technical convenience to control the explosion of maximal strengths allowing to avoid a lot of tedious computations.

When $\operatorname{supp} Q$ is compact, we can always assume that $\operatorname{supp} Q \subset [0, 1]$ and $1 \in \operatorname{supp}(Q)$, since the distribution of $(X_{i,j})_{1\leq i<j\leq N}$ given $\mathbb{V}_1^N$ is invariant by multiplication of the merits by a common $\lambda > 0$. From now on, we make this assumption. The moment condition is always satisfied and the only condition is the convexity

one, which is natural to forbid an accumulation of good players with strength close to 1.
Let us investigate the necessity of the convexity condition. Suppose that $Q(1-u) \sim u^{\alpha}$ when $u \to 0$, then the convexity condition holds only if $\alpha > 2$. To check the tightness of the bound 2, we introduce the following assumption.

Assumption : There exists $\alpha \in [0, 2)$ such that,

$$\log Q(1-u) = \alpha \log(u) + o(\log u) \quad \text{when } u \to 0. \tag{$\mathbf{A}$}$$

Notice that some standard distributions satisfy Assumption ($\mathbf{A}$), for example the uniform distribution satisfies ($\mathbf{A}$) with $\alpha = 1$, the Arcsine distribution satisfies it with $\alpha = 1/2$ and any Beta distribution $B(a, b)$ satisfies it as long as the parameter $b < 2$ with $\alpha = b$. The next quenched result studies, under Assumption ($\mathbf{A}$), the size of the set of possible winners.

Theorem 9. *For any $r \in \mathbb{R}_+$, let $G_r = \{\lceil N-r \rceil + 1, \dots, N\}$ denote the set of the $\lfloor r \rfloor$ best players. If* ($\mathbf{A}$) *holds, for any $0 < \gamma < 1 - \alpha/2$ then, $\mathbb{P}$ almost-surely,*

$$\mathbb{P}_V\left(\text{“none of the } N^{\gamma} \text{ best players wins"}\right) = \mathbb{P}_V\left(\max_{i \in G_{N^{\gamma}}} S_i < \max_{i \in G_N} S_i\right) \to 1\ . \tag{4.1}$$

For any $\gamma > 1 - \alpha/2$ then, $\mathbb{P}$ almost-surely,

$$\mathbb{P}_V\left(\text{“one of the } N^{\gamma} \text{ best players wins"}\right) = \mathbb{P}_V\left(\max_{i \notin G_{N^{\gamma}}} S_i < \max_{i \in G_N} S_i\right) \to 1\ . \tag{4.2}$$

The first part of the theorem shows that, when $Q(1-u) \sim u^{\alpha}$, with $\alpha < 2$, then none of N^{γ} “best" players, for any $\gamma \in (0, 1-\alpha/2)$ wins the competition. In particular, the best one does not win either. In this sense, the bound 2 in the asymptotic development of Q around 1 is tight in Theorem 8.
The second result (4.2) in Theorem 9 shows the sharpness of the bound $1 - \alpha/2$ in (4.1). Informally, this theorem shows that, under Assumption ($\mathbf{A}$), $N^{1-\alpha/2}$ players can be champion.

Under Assumption ($\mathbf{A}$), the best player does not win the championship. Therefore, we may wonder what strength v_{N+1} an additional tagged player $N+1$ should have to win the competition against players distributed according to Q. The following quenched result discusses the asymptotic probability that player $N+1$ wins as a function of its strength v_{N+1}.

Theorem 10. *Assume* ($\mathbf{A}$) *and let*

$$\vartheta_U = \mathbb{E}\left[\frac{U}{(U+1)^2}\right] \qquad \text{and} \qquad \epsilon_N = \sqrt{\frac{2-\alpha}{\vartheta_U}\frac{\log N}{N}}\ .$$

If $\liminf_{N\to\infty} \frac{v_{N+1}-1}{\epsilon_N} > 1$, *then,* $\mathbb{P}$*-almost surely*

$$\mathbb{P}_V\left(\text{“player } N+1 \text{ wins"}\right) \geq \mathbb{P}_V\left(S_{N+1} > 1 + \max_{i=1,\dots,N} S_i\right) \to 1\ .$$

If $\limsup_{N\to\infty} \frac{v_{N+1}-1}{\epsilon_N} < 1$, *then,* $\mathbb{P}$*-almost surely*

$$\mathbb{P}_V\left(\text{“player } N+1 \text{ does not win"}\right) \geq \mathbb{P}_V\left(S_{N+1} < \max_{i=1,\dots,N} S_i\right) \to 1\ .$$

This result shows a cut-off phenomenon around $1+\epsilon_N$ for the asymptotic probability that player $N+1$ wins.
It is interesting to notice that, for a given α, ϵ_N is a non increasing function of ϑ_U. Therefore, when U is stochastically dominated by U', that is $\mathbb{P}(U \geq a) \leq \mathbb{P}(U' \geq a)$ for any $a \in [0,1]$, we have $\vartheta_U \leq \vartheta_{U'}$, hence $\epsilon_N^U \geq \epsilon_N^{U'}$. In other words, it is easier for the tagged player to win against opponents distributed as U' than as U even if the latter has a weaker mean than the former. This result may seem counter-intuitive at first sight. In the following example, it is easier for the additional player to win the competition in case 1 than in case 2, since both distributions satisfy (**A**) with $\alpha = 0$.

1. All players in $\{1,\dots,N\}$ have strength 1.

2. The players in $\{1,\dots,N\}$ have strength 1 with probability 1/2 and strength 1/2 with probability 1/2.

Actually the score of the tagged player is smaller when he faces stronger opponents as expected, but so is the best score of the other good players.
Remark that Theorem 8 is an annealed result while Theorems 9 and 10 are quenched. Indeed, the first theorem requires to control precisely the difference of strengths between the best player and the others when all the players are identically distributed, this seems complicated in the quenched case. This problem does not appear in the other results: for example, in Theorem 10, the strength of the tagged player is deterministic and the strengths of others are bounded by 1.

4.3 Main proofs ideas

To conclude, let us briefly present the main ideas underlying our proofs. To avoid redundancy, let us focus on the proof of Theorem 8. Let $Z_N = \max_{i\in\{1,\dots,N-1\}} S_i$. Our strategy is to build random bounds s^N and z^N depending only on $\mathbb{V}_1^N$ such that,

$$\mathbb{P}\left(S_N \geq s^N\right) \to 1, \quad \mathbb{P}\left(Z_N \leq z^N\right) \to 1 \quad \text{and} \quad \mathbb{P}\left(s^N > z^N\right) \to 1 \ . \tag{4.3}$$

It follows that,

$$\begin{aligned}\mathbb{P}\left(S_N > Z_N\right) &\geq \mathbb{P}\left(S_N \geq s^N,\ Z_N \leq z^N,\ s^N > z^N\right)\\ &\geq 1 - \mathbb{P}\left(S_N < s^N\right) - \mathbb{P}\left(Z_N > z^N\right) - \mathbb{P}\left(s^N < z^N\right) \to 1 \ .\end{aligned}$$

The construction of s^N and z^N is obtained thanks to concentration inequalities. The concentration of S_N is easy, the tricky part is to build z^N. First, we use the bounded difference inequality to concentrate Z_N around its expectation. To apply this inequality, we have to decompose the set of random variables $(X_{i,j})_{1\leq i<j\leq N}$ into N groups such that, when we change the value of one group, the score of the best player does not change by more than one victory. This decomposition is based on the round-robin algorithm. It may be funny to remark that this decomposition, which is mathematically a bit tricky, is on the other hand totally common in sport tournaments. For example, in European soccer championships, a group corresponds to the matchs played in a week. Next, we have to bound above the expectation

$\mathbb{E}_V\left[Z_N\right]$. First we use the argument used by Pisier [Pis83]

$$\forall\lambda > 0, \qquad \mathbb{E}_V\left[Z_N\right] \leq \frac{1}{\lambda}\log\mathbb{E}_V\left[e^{\lambda Z_N}\right] \leq \frac{1}{\lambda}\sum_{k=1}^{N-1}\log\mathbb{E}_V\left[e^{\lambda S_k}\right] \ .$$

Then we compute a sharp bound on the minimal value of a potential winner, giving finally a bound on the cardinality of the set of potential winners under Assumption (**A**). These bounds are used to cut the set $\{1,\dots,N-1\}$ in two parts, the set of "potential winners" P with a controlled cardinality and the complementary $P^c = \{1,\dots,N-1\}\setminus P$ where each player has a controlled maximal force. These are used to cut the sum

$$\sum_{k=1}^{N-1}\log\mathbb{E}_V\left[e^{\lambda S_k}\right] = \sum_{k\in P}\log\mathbb{E}_V\left[e^{\lambda S_k}\right] + \sum_{k\in P^c}\log\mathbb{E}_V\left[e^{\lambda S_k}\right] \ .$$

The first sum is bounded by $|P|\log\mathbb{E}_V\left[e^{\lambda S_{N-1}}\right]$ and the second one by $N\log\mathbb{E}_V\left[e^{\lambda S_{N_{P^c}}}\right]$, where $N_{P^c} = \max\{P^c\}$. This last sum is proved negligible. The first term is then handled by an analysis of the asymptotic of V_{N-1} and V_N, using tools for extreme values of an i.i.d. sample.

The bound on the minimal strength of a winner which is roughly equal to $(1 - N^{-1/2})V_N$ is actually a by-product of our results which is not stressed in our theorems. It might however be of interest in practical situations (a medical treatment champion is at worst $(1 - N^{-1/2})$ times as efficient as the actual best one). It is also very important in related pairwise comparison problems. For example, in estimation by tests, it is the central quantity to derive oracle inequalities.

Bibliography

[AB09] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing System*, 22:46–54, 2009.

[AC10] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79, 2010.

[AC11] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 10 2011.

[Ada06] R. Adamczak. Moment inequalities for u-statistics. *Ann. Probab.*, 34(6):2288–2314, 2006.

[Aka70] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[All74] D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[AM08] S. Arlot and P. Massart. Data-driven calibration of penalties for least squares regression. *J. Mach. Learn. Res.*, 10:245–279, 2008.

[AMS96] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 20–29, New York, NY, USA, 1996. ACM.

[Arl08] S. Arlot. V-fold cross-validation improved: V-fold penalization. *Preprint arXive:0802.0566*, 2008.

[Arl09] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[Aud04] J. Y. Audibert. A better variance control for pac-bayesian classification. Technical Report 905b, Laboratoire de Probabilités et modèles aléatoires, 2004.

[Bar02] Y. Baraud. Nonasymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.

[Bar11] Y. Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401, 2011.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BCBL13] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *Information Theory, IEEE Transactions on*, 59(11):7711–7717, Nov 2013.

[BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth, Belmont, California, 1984.

[BG05] Y. Bengio and Y. Grandvalet. *Bias in estimating the variance of K-fold cross-validation*, volume 1 of GERAD 25th anniv. Ser. of *Statistical modeling and analysis for complex problem*. Springer, New-York, 2005.

[BGH14] Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the gaussian setting. *Ann. Inst. H. Poincaré Probab. Statist.*, 50(3):1092–1119, 2014.

[BHL03] Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1):225–251, 2003.

[Bir06a] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[Bir06b] L. Birgé. Statistical estimation with model selection. *Indag. Math. (NS)*, 17(4):497–537, 2006.

[Bir14] Lucien Birgé. Model selection for density estimation with $\mathbb{L}_2$-loss. *Probab. Theory Related Fields*, 158(3-4):533–574, 2014.

[BJL15] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, to appear, 2015.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[BMS08] G. Bressler, E. Mossel, and A. Sly. *Approximation, randomization and combinatorial optimization*. Springer, 2008.

[Bow84] A. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

[BS92] L. Breiman and P. Spector. Submodel selection and evaluation in regression. the x-random case. *International Statistical Review*, 60(3):291–319, 1992.

[BT52] R. Bradley and M. Terry. Rank analysis of incomplete block designs. i. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.

[Bur89] P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing problem. *Biometrika*, 76(3):503–514, 1989.

[Cat07] O. Catoni. *Pac-Bayesian supervised classification, the thermodynamics of statistical learning*, volume 56 of *IMS Lecture Notes Monogr. Ser.* Institute of Mathematical Statistics, 2007.

[Cat12a] O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(4):1148–1185, 11 2012.

[Cat12b] M. Cattelan. Models for paired comparison data: a review with emphasis on dependent data. *Statist. Sci.*, 27(3):412–433, 2012.

[CD15] O. Collier and A. Dalalyan. Curve registration by nonparametric goodness-of-fit testing. *J. Statist. Plann. Inference*, 162:20–42, 2015.

[Cel08] A. Celisse. *Model selection via cross-validation in Density estimation, regression and change-points detection.* PhD thesis, Université Paris-Sud 11, 2008.

[Cel14] A. Celisse. Optimal cross-validation in density estimation with l^2-loss. *Ann. Statist.*, 42(5):1879–1910, 2014.

[CR08] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Comput. Statist. Data Anal.*, 52(5):2350–2368, 2008.

[CT02] L. Cavalier and A. B. Tsybakov. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354, 2002.

[CT06] I. Csiszár and Z. Talata. Consistent estimation of the basic neighborhood of markov random fields. *Ann. Statist.*, 34:123–145, 2006.

[Dav63] H. A. David. *The method of paired comparisons.* Hafner Publishing Co., New York, 1963.

[Dav70] R. Davidson. On extending the bradley-terry model to accomodate ties in paired comparison experiments. *J. Amer. Statist. Assoc.*, 65(329):317–328, 1970.

[DB77] R. Davidson and R. Beaver. On extending the bradley-terry model to incorporate within-pair order effects. *Biometrics*, 33:693–702, 1977.

[DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[DJKP96] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation.* Springer, 2001.

[DO13] P. Deheuvels and S. Ouadah. Uniform-in-bandwith functional limit laws. *J. Theoret. Probab.*, 26(3):697–721, 2013.

[DR14] A. Drewitz and A. F. Ramírez. Selected topics in random walks in random environment. In *Topics in percolative and disordered systems*, volume 69 of *Springer Proc. Math. Stat.*, pages 23–83. Springer, New York, 2014.

[DS01] L. Dümbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152, 2001.

[DvdL07] S. Dudoit and M. J. van der Laan. *Multiple testing procedures with applications to genomics.* Springer, 2007.

[DZ02] A. Dembo and O. Zeitouni. Large deviations and applications. In *Handbook of stochastic analysis and applications*, volume 163 of *Statist. Textbooks Monogr.*, pages 361–416. Dekker, New York, 2002.

[Efr79] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[Efr83] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[Efr85] S. Efromovich. Nonparametric estimation of a density function of unknown smoothness. *Theory Probab. Appl.*, 18(557-568), 1985.

[Efr00] S. Efromovich. On sharp adaptive estimation of multivariate curves. *Math. Methods Statist.*, 9:117–139, 2000.

[Efr05] S. Efromovich. Estimation of the density of regression errors. *Ann. Statist.*, 33:2194–2227, 2005.

[Fel71] W. Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney, 1971.

[FK92] W. Feluch and J. Koronacki. A note on modified cross-validation in density estimation. *Comput. Statist. Data Anal.*, 13(2):143–151, 1992.

[FL06] M. Fromont and B. Laurent. Adaptive goodness-of-fit tests in a density model. *Ann. Statist.*, 34(2):680–720, 2006.

[FLRB11] M. Fromont, B. Laurent, and P. Reynaud-Bouret. Adaptive tests of homogeneity for a poisson process. *Ann. Inst. Henri Poincaré P & S*, 47(1):176–213, 2011.

[FLRB13] M. Fromont, B. Laurent, and P. Reynaud-Bouret. The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *Ann. Statist.*, 41(3):1431–1461, 2013.

[FT06] M. Fromont and C. Tuleau. *Functional classification with margin conditions*, volume 4005 of *Lecture Notes in Comput. Sci.* Springer, Berlin, 2006.

[Geo88] H. O. Georgii. *Gibbs measure and phase transitions*, volume 9 of de Gruyter studies in mathematics. de Gruyter, Berlin, 1988.

[Gey76] S. Geysser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1976.

[GJ05] M. Glickman and S. Jensen. Adaptive paired comparison design. *J. Statist. Plann. Inference*, 127(1-2):279–293, 2005.

[GL11] A. Goldenshluger and O. Lepski. Bandwith selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

[GLZ99] E. Giné, R. Latala, and J. Zinn. Exponential and moment inequalities for u-statistics. *Prog. Probab.*, 47(High dimensional probability II):13–38, 1999.

[GN09] E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Ann. Probab.*, 37(4):1605–1646, 2009.

[GN15] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge University press, 2015.

[Gol92] G. K. Golubev. Nonparametric estimation of smooth probability densities in l_2. *Probl. Inf. Transm.*, 28(1):44–54, 1992.

[GS10] J. J. Goeman and A. Solari. The sequential rejection principle of familywise error control. *Ann. Statist.*, 38(6):3782–3810, 2010.

[Hal83] P. Hall. Large sample optimality of least-squares cross-validation in density estimation. *Ann. Statist.*, 11(4):1156–1174, 1983.

[HM87] P. Hall and J. S. Marron. Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields*, 74(4):567–581, 1987.

[Hol79] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70, 1979.

[HRB03] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

[HS14] D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 37–45. JMLR Workshop and Conference Proceedings, 2014.

[Hsu10] D. Hsu. Robust statistics. Available from `http://www.inherentuncertainty.org/2010/12/robust-statistics.html`, 2010.

[HT98] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *Ann. Statist.*, 26(2):451–471, 1998.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer, New-York, 2009.

[Hun04] D. R. Hunter. MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, 32(1):384–406, 2004.

[Ing93] Y. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III. *Math. Methods Statist.*, 2(2,3,4):85–114,171–189,249–268, 1993.

[Ing00] Y. I. Ingster. Adaptive chi-square tests. *J. Math. Science*, 99(2):1110–1119, 2000.

[JMS96] M. C. Jones, J. S. Marron, and S. J. Shealter. A brief survey of bandwith selection for density estimation. *J. Amer. Statist. Assoc.*, 91(433):401–407, 1996.

[JVV86] M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:186–188, 1986.

[Lep90] O. Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Appl.*, 36:454–466, 1990.

[Lep91] O. Lepski. Asymptotically minimax adaptive estimation i : Upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36(682-697), 1991.

[Lev05] L. A. Levin. Notes for miscellaneous lectures. *CoRR*, abs/cs/0503039, 2005.

[LLM12] B. Laurent, J.-M. Loubes, and C. Marteau. Nonasymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electron. J. Stat.*, 6:91–122, 2012.

[Loa99] R. Loader. Bandwith selection: classical or plug-in. *Ann. Statist.*, 1999.

[LY11] W. Liu and Y. Yang. Parametric or nonparametric? a parametricness index for model selection. *Ann. Statist.*, 39(4):2074–2102, 2011.

[Mag15] N. Magalhães. *Validation croisée et pénalisation pour l'estimation de densité.* PhD thesis, Université Paris Sud - Paris XI, <tel- 01164581>, 2015.

[Mal73] C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics.* Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Min13] S. Minsker. Geometric median and robust estimation in Banach spaces. *arXiv preprint*, 2013.

[MS11] D. M. Mason and J. W. H. Swanepoel. A general result on the uniform in bandwith consistency of kernel-type function estimators. *TEST*, 20(1):72–94, 2011.

[NY83] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley Interscience, 1983.

[Par62] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076, 1962.

[PC84] R. R. Picard and R. D. Cook. Cross-validation of regression models. *J. Amer. Statist. Assoc.*, 79(387):575–583, 1984.

[Pin80] M. S. Pinsker. Optimal filtration of square-integrable signals in gaussian noise. *Problems Inform. Transmission*, 16(2):52–68, 1980.

[Pis83] G. Pisier. Some applications of the metric entropy condition to harmonic analysis. In *Banach spaces, harmonic analysis, and probability theory (Storrs, Conn., 1980/1981)*, volume 995 of *Lecture Notes in Math.*, pages 123–154. Springer, Berlin, 1983.

[Rig06] P. Rigollet. Adaptive density estimation using the blockwise stein-method. *Bernoulli*, 12(2):351–370, 2006.

[RK67] P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: a generalization of the bradley-terry model. *J. Amer. Statist. Assoc.*, 62(317):194–204, 1967.

[Ros56] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.*, 27:832–837, 1956.

[RSW11] J. P. Romano, A. Shaikh, and M. Wolf. Consonance and the closure method in multiple testing. *Int. J. Biostatistics*, 7(1):1–25, 2011.

[RT07] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

[Rud82] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, 9(2):65–78, 1982.

[RWL10] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1 regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010.

[SBSB06] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in neural population. *Nature*, 440:1007–1012, 2006.

[SHS10] O. Y. Savchuk, J. D. Hart, and S. J. Shealter. Indirect cross-validation for density estimation. *J. Amer. Statist. Assoc.*, pages 415–423, 2010.

[Sil86] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC Press, 1986.

[Spo96] V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498, 1996.

[SR09] C. Sire and S. Redner. Understanding baseball team standings and streaks. *Eur. Phys. J. B*, 67:473–481, 2009.

[ST87] D. W. Scott and G. R. Terrel. Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.*, 82(400):1131–1146, 1987.

[Sto74] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[Sto84] C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12(4):1285–1297, 1984.

[Stu92] W. Stute. Modified cross-validation in density estimation. *J. Statist. Plann. Inference*, 30(3):293–305, 1992.

[SY99] G. Simons and Y.-C. Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *Ann. Statist.*, 27(3):1041–1060, 1999.

[Tsy09] A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer series in statistics. Springer, New-York, 2009.

[WB79] G. Walter and J. Blum. Probability density estimation using delta sequences. *Ann. Statist.*, 7(2):328–340, 1979.

[WJ95] M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on statistics and applied probability.* Chapman and Hall, London, 1995.

[YYX12] T. Yan, Y. Yang, and J. Xu. Sparse paired comparisons in the Bradley-Terry model. *Statist. Sinica*, 22(3):1305–1318, 2012.

[Zei12] O. Zeitouni. Random walks in random environment. In *Computational complexity. Vols. 1–6*, pages 2564–2577. Springer, New York, 2012.

[Zer29] E. Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.*, 29(1):436–460, 1929.