__N^o D'ORDRE__: **9302**

UNIVERSITÉ PARIS–SUD — FACULTÉ DES SCIENCES D'ORSAY

THÈSE

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'UNIVERSITÉ PARIS XI

Spécialité : **Mathématiques**

par

Alain CELISSE

Model selection *via* cross-validation in density estimation, regression, and change-points detection

---

Soutenue publiquement le **9 décembre 2008** devant la commission d'examen:

M. Yannick	**BARAUD**	Université Nice Sophia-Antipolis	Examinateur
M. Philippe	**BESSE**	INSA de Toulouse	Examinateur
M. Gérard	**BIAU**	Université Paris–VI	Rapporteur
M. Stéphane	**BOUCHERON**	Université Paris–VII	Examinateur
M. Pascal	**MASSART**	Université Paris–XI, Orsay	Président
M. Stéphane	**ROBIN**	INRA et Agroparitech, Paris	Directeur

Rapporteurs :	M. Gérard	**BIAU**	Université Paris–VI
	M. Yuhong	**YANG**	University of Minnesota

Remerciements

First of all, I sincerely would like to thank Gérard Biau and Yuhong Yang for accepting to report on this Ph.D. manuscript, as well as Yannick Baraud, Philippe Besse, Stéphane Boucheron, and Pascal Massart for their kind participation as examiners in the Ph.D. defense.

Mes remerciements vont ensuite à mon directeur de thèse, Stéphane Robin, qui m'a initié au monde de la recherche. Tu m'as guidé et soutenu lorsque j'en avais besoin. Nos discussions scientifiques ont toujours été franches, constructives et m'ont ainsi beaucoup apporté. De plus, tu m'as offert l'opportunité de visiter l'ETHZ pendant un mois, ce qui m'a permis de faire des rencontres importantes et de prendre conscience de nombreuses choses. Enfin, tes qualités humaines et scientifiques sont pour moi un exemple que j'espère pouvoir suivre tout au long des années à venir. Pour tout cela (et d'autres choses encore), je te suis profondément reconnaissant.

Je ne puis écrire cette page sans mentionner également l'incomparable qualité de l'environnement dans lequel ce travail a vu le jour. Travailler au sein de l'équipe Statistique et génome à AgroParistech a été un réel bonheur. L'entourage scientifique dont j'ai bénéficié ainsi que les repas et (trop courtes) pauses que nous avons partagés ont eu (et ont encore) une grande importance pour moi. Je souhaite à tout thésard de pouvoir bénéficier d'un tel environnement privilégié. À tous Merci.

À l'occasion de divers congrès, groupes de travail, cours,... auxquels j'ai pris part, j'ai fait la connaissance de plusieurs personnes avec lesquelles j'ai eu le plaisir de pouvoir partager à la fois sur un plan scientifique et humain. Au premier chef, je remercie Sylvain Arlot qui m'a soutenu en des temps tourmentés et avec qui j'ai le plus grand plaisir à travailler. Nos longues discussions tardives ainsi que la confrontation de nos points de vue sont une expérience ô combien enrichissante. Je remercie également Étienne Roquain et Fanny Villers pour leur amitié et nos discussions scientifiques endiablées. Il me tarde que nous entamions notre collaboration qui, je n'en doute pas, sera fructueuse. Mes pensées vont également à Tristan Mary-Huard et Émilie Lebarbier auprès desquels j'ai pu longuement apprendre (et je ne parle pas seulement des type-I ou type-III) sur les fondements de la démarche scientifique. J'ai une pensée émue à l'idée de cette grande fraternité des débuts de soirée, à l'occasion de la fermeture des portes du labo...

Enfin, un grand merci à toutes les âmes charitables ayant contribué par leur relecture (intégrale ou partielle) de cette thèse : Stéphane, Sylvain, Liliane, Caroline et Tristan.

Je manquerais vraiment à tous mes devoirs si je ne mentionnais pas ici une grande partie des personnes qui ont pleinement pris part à ce travail par leur amitié et leur soutien. Un grand merci donc à : Michel avec qui je partage l'amour de la bonne musique (et accessoirement le bureau), Émilie pour m'avoir fait découvrir Glacière (surtout le bas des escaliers), Marie-Laure et Fred pour leur attention, leur écoute toutes particulières et le maniement des jetons, Caroline B. à qui je veux bien pardonner les origines déplorables pour sa joie de vivre et quelques unes de ces pâtisseries, Liliane pour son caractère, son infinie gentillesse et son aide salvatrice lors d'un moment de détresse profonde, et enfin Juliette, Max et Fred B., plusieurs "Ch'ti" (au moins d'adoption) avec qui j'ai passé des soirées inoubliables, tiraillé entre champions league et repassage...

Un grand merci également à tous ceux que je ne peux citer ici faute de place, mais auxquels je dois beaucoup. Ils me pardonneront sans doute, je l'espère, ce raccourci injuste.

Pour conclure, je remercie du fond du cœur mes parents qui se sont dévoués littéralement corps et âme pour leurs enfants et ont su leur transmettre l'amour du travail, ma petite sœur parce qu'elle est petite et que c'est ma sœur ! et enfin Caroline pour sa patience et de façon générale, pour ce qu'il y a de meilleur dans l'avenir...

Merci !

Contents

Foreword

At the core of this manuscript is the study of cross-validation algorithms. However from a chapter to another, this study is carried out in various frameworks and different aspects are dealt with. Furthermore, these chapters differ on their nature: State-of-the art, published or submitted papers, or simply the present version of some research works still in progress.

Except Chapter 7 which results from a joint work initiated since January 2008 with Sylvain Arlot, all the other chapters are the result of a personal work.

In this thesis, most of the chapters can be read separately, except Chapters 4 and 5 which are closely connected.

Since several chapters of this manuscript should be submitted in a few months, the same results may appear several times along the manuscript. Nevertheless to avoid too numerous repetitions, the proofs of these results have been given only once, which induces some dependence between chapters.

The manuscript is organized as follows:

- Chapter 1 is a brief introduction to the main ideas involved in this thesis and gives a short overview of the main aspects of the present work,
- Chapter 2 originates with the need for assessing my understanding of model selection. I hope that after some enhancements, it could serve as an introduction to this topic,
- The first part of Chapter 3 is the state-of-the art about cross-validation, while its second part is dedicated to new results such as closed-form expressions for the leave-p-out risk estimator in both density estimation and regression. These are recent results and this work is still in progress,
- Chapter 4 constitutes a paper published in Comput. Statist. and Data Analysis in 2008. It addresses the problem of optimal risk estimation in the density estimation setup. Since it has been published, the adopted notations has changed. We apologize for this,
- Chapter 5 corresponds to a submitted paper. Its main concern is the estimation of the proportion of true null hypotheses in the multiple testing problem. It relies on the application of the strategy described in Chapter 4, for which new theoretical results are derived,
- Chapter 6 assesses the quality of the model selection procedure *via* leave-p-out in the density estimation framework with respect to the complexity of the collection of models in hand. This work should be extended in the forthcoming months,
- Chapter 7 tackles the problem of change-points detection *via* several resampling algorithms with heteroscedastic data. An extensive simulation study is carried out, based on some theoretical considerations. This work has been initiated since January 2008.

Chapter 1

Introduction

Abstractness, sometimes hurled as a reproach at mathematics, is its chief glory and its surest title to practical usefulness. – Bell, Eric T.

Actually, theoretical and applied statistics are strongly related to each other. The last decade provides several illustrations of the tight connection between theory and practice such as the recent rise of high-throughput data where $p \ll n$ in many applied areas, which is at the origin of a large amount of theoretical literature on the matter with lots of new refinements.

cDNA microarray data with several thousands of genes (p large) on a single array and only a few replications (n small) are the perfect illustration of this fact. Indeed in such an experiment, the biologist is interested in relating the phenomenon he observes to the over- or under-expression of a set of genes. This problem may be addressed through several aspects, depending on the precise biological question.
If the problem is to detect a set of "representative" genes which are likely involved in the studied mechanism, the variable selection setup [37] may provide a satisfactory answer. Note that we do not look for the exhaustive set of genes involved in the phenomenon, but rather try to pick out a smaller set of variables, which can be more easily studied and interpreted.
On the contrary provided the question is to recover the whole set of induced genes between two experimental conditions, then some multiple testing tools can be used [8, 27, 54].
Finally, if biologists are more interested in the inference of relationships between genes, graphical models turn out to be an appropriate tool towards a possible answer [56].

Conversely, theory turns out to be essential to the practice since it provides some understanding of the applied methods and even guidelines towards a better practice. For instance, theoretical calculations have demonstrated that the resubstitution error, widespread in the practitioner community, used to provide a very poor estimate of the true error. Another illustration is given by the deficiencies suffered by the maximum likelihood estimator (MLE), which have been proved to be inconsistent in some situations and may even provide extremely bad estimations. At least, such theoretical results provide guidelines towards a cautious use of some already widespread tools.

Besides, theoretical statistics help to get more insight into the understanding. If we come back to the problem of variable selection in a gene expression analysis, the resulting set of selected explanatory variables (genes) will be more easily interpretable than an exhaustive one. Actually, with variable selection comes a "model", which approximates, and this way, simplifies the truth in order to make it more understandable.
Another example is given by the CGH (Comparative Genomic Hybridization) microarray data, for which the problem is to detect regions with homogeneous copy numbers along the chromosomes. Thus, the segmentation (or change-points detection) methods [33, 41, 5] attempt to yield such an answer, so that they enable to detect the association of a disease (for instance a cancer) with some physical changes in the chromosome structure (loss or gain of a piece of chromosome).

After a modeling step, which aims at describing the observed phenomenon, model selection is one of the possible frameworks, which provide the tools and their justifications to understand the problem in hand.

1.1 Model selection *via* penalization

1.1.1 Model selection strategy

In the microarray data analysis, the interpretation of the gene expressions as some explanatory variables of an observed signal to which they are linked (through a "given relationship") already constitutes a "model" of the real situation. Several such models may be formalized, which are more or less complex and the main concern of the practitioner is to choose the appropriate model with respect to its purposes. For instance, we may imagine that the relationship between the observed signal and the explanatory variables is known (a linear regression model), and that the main interest lies in the subset of variables which explain the best what is observed. In this case, model selection considers all the subsets of variables from the original ones and chooses the best one among all of them.

Let us assume we would like to estimate a parameter of interest s. From our preliminary knowledge of the problem in hand or simply on the basis on some arbitrary assumptions, we choose a family $(S_m)_{m\in\mathcal{M}_n}$, which is expected to be close to the target s. Each element S_m of this family is named a model, and with each model is associated an estimator $\widehat{s}_m$ of s. Since these estimators all depend on an index m, the purpose of model selection is to choose the best possible index $\widehat{m}$ according to a given criterion depending on our goal. To reach this goal in model selection *via* penalization, a penalized criterion is designed which is defined by

$$\mathrm{crit}(m) = P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m),$$

where $P_n\gamma(\widehat{s}_m)$ denotes the empirical risk of the estimator $\widehat{s}_m$ and pen : $\mathcal{M}_n \to \mathbb{R}_+$ denotes a penalty function. The function pen aims at penalizing too large models S_ms. A detailed description of the model selection strategy can be found in [6, 38].

1.1.2 Efficiency and consistency

Depending on our goal, the choice of the best model may result from very different model selection strategies. In the CGH microarray data, the biological interpretation of the signal as the ratio of the number of copies in a region of the chromosome entails that the candidate functions to explain the signal are chosen piecewise constants. Provided we have good reasons to think that a true function exists and belongs to the set of candidates we consider, we may try to recover it. This is the *identification purpose*. Another possibility is to consider that with the few observations we have in hand, recovering the true function is not realistic. Then, we have to content ourselves with an estimation of the truth, it is the *estimation* point of view.

Model selection *via* penalization was born with the ideas of Akaike and Mallows and their respective criteria AIC [1] and C_p [36]. The latter attempts to unbiasedly estimate the risk of an estimator, while the former pursue the same purpose from an asymptotic point of view. A few years later, Schwarz [45] introduces the BIC criterion, derived from the Laplace approximation. All these criteria differ from one another by their purposes. Whereas AIC and C_p intend to provide an estimator of the target parameter s with the same asymptotic risk as the smallest one among all the considered estimators [48, 13], BIC is rather dedicated to recover s, or at least the best approximation of it [39, 34]. The first approach is named the asymptotic *efficiency* viewpoint, while the last one is called *consistency* [13].

1.1.3 Complexity of the collection of models

In the typical example of variable selection, two situations can be distinguished. The first one is the situation when the variables are ordered. A variable may only belong to the final model provided all the preceding ones do. It is called *ordered variable selection*. The second one applies when variables are

unordered and we choose the subset of candidates among all the possible subsets of the original ones. This is the named *all subset selection*.
The first example, we have at most the same number of candidate subsets as the number of original variables, whereas in the second setting, the number of candidates is exponentially large.

That is why we distinguish two broad classes of examples, which are said to have respectively a polynomial complexity (ordered variable selection), or an exponential complexity (unordered variable selection).
Corresponding to these two classes of examples, different types of penalties have been designed. For instance whereas AIC and C_p-like penalties apply to the polynomial framework [4, 12, 19], they do not apply anymore to the exponential setting in which they may be definitely misleading [15, 13]. Some other penalties have been specially designed to the exponential framework [5, 13, 44]. Like the unordered variable selection, the change-points detection problem is an illustration of such a framework since the number of models of dimension D is equals to $\binom{n-1}{D-1}$ (see [33, 5]), which is very huge even with small values of n.

1.2 Resampling and cross-validation

1.2.1 Risk estimation

Some of the above penalized criteria either derive from an asymptotic heuristics or exhibits some asymptotic optimality properties. The others [4, 12, 5, 13] result from some called concentration inequalities and may therefore be understood as upper bounds of the risk of an estimator $\widehat{s}_m$. Moreover, these penalties depend on some unknown universal constants that must be determined by an intensive simulation step for instance [33].

Another strategy consists in estimating rather than upper bounding the risk of $\widehat{s}_m$. This goal may be reached thanks to *resampling strategies*, the idea of which may be summarized as follows. Though we do not know P the distribution of the observations, we observe the empirical distribution P_n yet. Then, we use the original observations to generate some new samples (the *resamples*) according to a known distribution conditionally to P_n. The rationale is thus to mimic the relationship between P and P_n thanks to that of P_n and P_n^W, the empirical distribution of the resample.
Several types of resampling schemes have emerged, one of the most famous being the bootstrap [22, 43, 49, 23]. Provided we draw resamples of cardinality m lower than n (that of the original data), we call the corresponding strategy *subsampling*. The jackknife introduced by Quenouille [42] is one of them as well as cross-validation algorithms such as the *leave-one-out* (Loo) [51, 52, 50], the *V-fold cross-validation* (VFCV) [2, 25, 26, 18] or the *leave-p-out* (Lpo) [17, 46, 58, 20].
There are some differences between bootstrap and cross-validation strategies. Cross-validation (CV) has been designed actually to compute the risk (or any analogous quantity) of an estimator. It is mainly based on the independence property between the original data used to compute the estimator and a new independent one devoted to the assessment of its accuracy. Unlike cross-validation, bootstrap is not dedicated to the risk estimation. It rather attempts to approximate the unknown distribution of a statistic of interest. In particular, it is not based on the above independence property. The risk estimation is then a by-product of the bootstrap scheme.

Throughout this manuscript, we focus on the study of the Lpo algorithm. Our idea is that a better understanding of the behaviour of this resampling scheme would provide more insight into all the CV algorithms. Indeed, these algorithms may be understood as some "approximations" to the ideal but unachievable Lpo estimator, since the latter usually turns out to be too computationally demanding. For instance when $p = 1$, Lpo amounts to the famous Loo algorithm. Thus when n is large, the computational burden of the Loo may be too expensive, which is the reason of the introduction of the VFCV algorithm [17].

1.2.2 Choice of p

A crucial issue of the CV-based strategies is the determination of the size of the test set, denoted by p in the Lpo strategy. Indeed, this parameter does not only drive the amount of computational complexity, but it also determines how biased and variable the resulting Lpo estimator of the risk will be [17, 24, 3, 20].
Several asymptotic results exist about the optimal choice of p with respect to the efficiency or consistency points of view, most of them concerning the regression framework. Indeed, the asymptotic equivalence between the Loo estimator and AIC [52] or Mallows' C_p [35] and other penalized criteria [58] indicate the asymptotic optimality of the Loo in terms of efficiency in some frameworks. As for the consistency viewpoint, Shao [46, 47] has established that consistency may be reached provided $p/n \to 1$. A more refined result has been proven by Yang [57], who relates the required magnitude of p/n to the convergence properties of the involved estimators.
From a non-asymptotic viewpoint, very few results exist to the best of our knowledge. Recently, Arlot [3] proposed a strategy relying on resampling penalties, which enables to control the amount of bias (overpenalization) independently from the computational cost. He studies the VFCV algorithm, but do not yet provide any real guideline to choose V.

1.2.3 Closed-form expressions

The first results of this manuscript (Chapter 3) are closed-form formulas for the Lpo estimator of the quadratic risk (Section 3.3) for a wide range of estimators. These results constitute an enhancement in several respects.
Indeed, they first allow the effective use of the Lpo risk estimator, which used to be considered as intractable. Furthermore, the computational cost of the Lpo estimator therefore becomes less expensive than that of the usual V-fold estimator, which moreover used to introduce some additional variability due to the preliminary random partitioning of the data.
A straightforward consequence of the computational cost drop is the opportunity to choose the parameter p independently from any computational consideration. Closed-form expressions are derived for several broad classes of estimators, which are extensively used both in the density estimation as well as in the regression frameworks. These estimators are projection estimators in the density estimation setup as well as in the regression context, some kernel estimators and the specific regressogram in the regression framework.
Remark: When deriving closed-form expressions necessitates to weaken the level of generality of the calculations, we describe the precise reasons why more general results cannot be achieved.

In this entire manuscript, our goal is the study of the Lpo algorithm (as a generic and representative CV procedure) in various settings such as density estimation, regression and change-points detection. The behaviour of this algorithm with respect to p is analyzed in the model selection framework, mainly from a non-asymptotic viewpoint.

1.3 Optimal risk estimation

In Chapter 4, a first approach of the problem of model selection in the density estimation context is made through the "optimal" risk estimation. Along this thesis, we use the quadratic loss function denoted by

$$\ell\left(s, \widehat{s}\,\right) = \left\| s - \widehat{s}\,\right\|^2,$$

where s the target function (density or regression function) and $\widehat{s}$, any estimator.

In the density estimation framework, let us assume that s is an unknown density on $[0,1]$ we would like to estimate, and let $(\widehat{s}_m)_{m\in\mathcal{M}_n}$ denote the family of candidate estimators.
For a given $1 \leq p \leq n-1$, our strategy consists in defining (provided it exists and is unique)

$$\widehat{m} := \mathrm{Argmin}_{m\in\mathcal{M}_n} \widehat{R}_p(\widehat{s}_m), \tag{1.1}$$

where $\widehat{R}_p(\widehat{s}_m)$ denotes the Lpo based estimator of the quadratic risk of $\widehat{s}_m$. Then the final density estimator is $\widetilde{s} := \widehat{s}_{\widehat{m}}$.

1.3.1 Bias-variance tradeoff

As we can see in (1.1) for a given m, we have as many risk estimators as possible values of p, that is $n-1$. The question is to choose $\widehat{p}$ which provides the "best" risk estimator.
The intuition about CV (which is confirmed for VFCV in the classification context for instance [24]) is that when p is small, we use most of the data $(n-1)$ to compute the estimator and only one (the remaining observation) to assess the performance of the latter. Thus whereas this estimator is expected to be close to the original one (computed from n observations), its performance assessment should be rough since only one observation has been used. The resulting risk estimator should be nearly unbiased, but quite variable. Conversely, if we use only a few data to compute the estimator, it could be far from the original one (and the resulting risk estimator could be biased), while its performance assessment is made from a large amount of data, resulting in a slightly variable estimator of the risk.
Following these considerations in Section 4.4.2, we choose the value of p that reaches the best bias-variance tradeoff, that is the minimizer of the mean square error (MSE) of the Lpo risk estimator. We recall that the MSE is the sum of the square of the bias of $\widehat{R}_p$ and the variance of $\widehat{R}_p$. Thus for each m:

$$\widehat{p}(m) = \mathrm{Argmin}_{1\leq p\leq n-1}\widehat{MSE}(\widehat{R}_p(\widehat{s}_m)),$$

where $\widehat{MSE}(\widehat{R}_p(\widehat{s}_m))$ denotes an estimator of the MSE.

Remarks:

- Note that the computational burden of all the procedure remains reasonable thanks to closed-form expressions for the Lpo risk estimator as well as for its bias and variance.
- In Chapter 4 and Chapter 5, we apply this strategy to histograms. We point out that the same strategy may be straightforwardly followed with kernel density estimators as well as any projection estimators.

1.3.2 Multiple testing

In biology for instance, the development of new technologies with high throughput data such as cDNA microarrays has generated a great interest in the statistical community, with the challenging problem of *multiple testing*. From a microarray on which thousands of genes are hybridized, the question is the following: How many genes have been induced between the condition 1 and condition 2? Thanks to a few technical replicates of this microarray, we have to perform a test for each gene, which results in thousands of test statistics $(T_i)_{1\leq i\leq n}$. A question arises: Is there a relevant way to distinguish between induced and non-induced genes in presence of so many genes?

We first explain the specificity of the multiple testing problem. Let us assume we are given a set of null hypotheses against their alternatives from which we aim at designing a decision rule resulting in the rejection of a subset of them for a controlled error. When this set is reduced to a single null hypothesis, the classical statistical test theory applies and for a given level $0 < \alpha < 1$ of the type-I error, that is the error of making a *false positive* (= a false rejection), we build the decision rule which maximizes the power of the test.
However, the problem turns out to be more intricate when the number n of tested hypotheses is larger. Indeed, testing each null hypothesis at level α and rejecting $0 \leq k \leq n$ of them potentially leads to a type-I error as large as $k\alpha$. Moreover since we do not know how many of them should be rejected and provided we decide to keep the error lower than $0 < \beta < 1$, a simple idea is to set $\alpha = \beta/n$, which provides the desired control. However, the resulting individual level at which we actually test each hypothesis is decreasing with n. The procedure becomes more and more *conservative* independently from the truth about each hypothesis, which means that less and less null hypotheses will be rejected as n grows, no matter whether they are false. Thus, a sufficient condition to reject no hypothesis is to add enough true null hypotheses, which is unsatisfactory.

This reasoning highlights that we need for some multiple testing procedures (MTP), which would take into account all the test statistic values to design the final rejection rule, instead of only that of each hypothesis, independently from the others.

The MTPs are specifically designed to provide the desired control of a given type-I error such as false discovery rate (FDR) for instance, defined as the expected ratio of the number of false positives over the total number of rejections:

$$\mathrm{FDR}\left[\,\mathcal{R}\left(T_1,\ldots,T_n\right)\right] := \mathbb{E}\left[\frac{FP\left(\mathcal{R}\left(T_1,\ldots,T_n\right)\right)}{1\vee R\left(\mathcal{R}\left(T_1,\ldots,T_n\right)\right)}\right],$$

where $R\left(\mathcal{R}\left(T_1,\ldots,T_n\right)\right)$ denotes the total number of rejected hypotheses from the rejection rule $\mathcal{R}\left(T_1,\ldots,T_n\right)$.
Most of them apply under an independence assumption [8, 10, 9], but some others still hold under some types of dependence [11, 14, 14]. The most famous MTP is the Benjamini-Hochberg (BH) procedure [8] which deals with p-values rather than test statistics. Let us define $(p_i)_{1\leq i\leq n}$ the set of p-values associated with each tested null hypothesis $(H_{0,i})_{1\leq i\leq n}$. Then for a given level $0<\alpha<1$, the BH procedure relies on

$$k := \max\left\{1\leq i\leq n \mid p_{(i)}\leq i\frac{\alpha}{n}\right\}$$

such that we reject all the null hypotheses $H_{0,(i)}$ with $i\leq k$, where $p_{(1)}\leq p_{(2)}\leq\ldots\leq p_{(n)}$ (see also Figure **??**). Under the independence assumption, Benjamini and Hochberg have proven that

$$\mathrm{FDR}\left[\,\mathcal{R}_{BH}\left(p_1,\ldots,p_n\right)\right] \leq \alpha\pi_0\leq\alpha,$$

where $\pi_0 = n_0/n$ and n_0 denotes the unknown number of true null hypotheses.
Actually, the control provided by the BH procedure is at level $\alpha\pi_0$. However since π_0 is unknown, we only control the FDR at level α. Moreover, we observe that the gap between these two quantities may be substantial provided π_0 is not that close to 1. A reasonable idea in this setting is to estimate π_0 so as we get a tighter upper bound. This is the rationale behind the named plug-in strategy described in [53, 55, 28] for instance.
It turns out that an estimator of π_0 may be derived through the density estimation methodology we described above in Section 1.3.1 with histograms. The resulting estimator exhibits a consistency property (Section 5.4) as well as it asymptotically yields the expected control of the FDR.

1.4 Leave-p-out risk estimator as a penalized criterion

1.4.1 Random penalty

Another aspect of model selection *via* Lpo is developed through the interpretation of the Lpo risk estimator as a penalized criterion. Indeed, the usual expression of penalized criteria is for each m

$$\mathrm{crit}(m) = P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m),$$

where $P_n\gamma(\widehat{s}_m)$ denotes the empirical contrast of the estimator $\widehat{s}_m$ and $\mathrm{pen}(\cdot)$ is the penalty term, which grows with the increasing complexity of the model S_m.

In Section 6.2.1, we point out that the Lpo risk estimator can be understood as a penalized criterion. For each $1\leq p\leq n-1$, we rewrite the Lpo estimator in the same way

$$\widehat{R}_p(\widehat{s}_m) = P_n\gamma(\widehat{s}_m) + \left(\widehat{R}_p(\widehat{s}_m) - P_n\gamma(\widehat{s}_m)\right),$$

where the second term in brackets is named the Lpo penalty. Still in Section 6.2.1, we establish that this random penalty overpenalizes for any projection estimator.

1.4.2 Oracle inequality and adaptivity

In Chapter 6, we pursue the estimation viewpoint and not the identification one, since we do not try to recover any true density, but rather look for a reliable estimation of it. Therefore, the quality measure of the procedure is the risk of the final estimator $\widetilde{s}$. Moreover, we focus on non-asymptotic results so that we let the model collection $(S_m)_{m\in\mathcal{M}_n}$ depend on the sample size n.

Following the efficiency viewpoint, we would like to design a procedure such that

$$\mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right] \approx \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] = \mathbb{E}\left[\|s-\widehat{s}_{m^*}\|^2\right],$$

where $\mathcal{S}_{m^*}$ denotes the *oracle* model, that is the model we would choose if we knew the distribution of the observations. Unfortunately, this goal can only approximately be reached and we have to content ourselves with an oracle inequality

$$\mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right] \leq C \inf_{m\in\mathcal{M}_n} \left\{\mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] + r(m,n)\right\},$$

where $C \geq 1$ denotes a constant independent from s and $r(m,n)$ is a remainder term with respect to $\mathbb{E}\left[\|s-\widehat{s}_m\|^2\right]$ [6, 13, 38]. This inequality simply says that the risk of the final estimator $\widetilde{s}$ is almost the same as the smallest one among all the estimators we consider, up to a constant C and a remainder term. Obviously, the closer C to 1 and the smaller the remainder, the better the model selection procedure.

Another desirable property of a model selection procedure is the *adaptivity in the minimax sense* with respect to some functional space [6]. This notion is strongly related to the approximation properties of functional spaces. We refer to [21] for an extensive study of a wide range of functional spaces.
Let assume that s belongs to such a space $\mathcal{H}(\theta^*)$ for $\theta^* \in \Theta$. An estimator is said to be *adaptive for* θ^* if, without knowing θ^*, it "works as well as" any estimator which would exploit this knowledge. For us, the effectiveness measure is the quadratic risk and an estimator enjoys such an adaptivity property if its risk is nearly the same (up to some constants) as the minimax risk with respect to $\mathcal{H}(\theta^*)$:

$$\inf_{\widehat{s}} \sup_{s\in\mathcal{H}(\theta^*)} \mathbb{E}\left[\|s-\widehat{s}\|^2\right] \leq \sup_{s\in\mathcal{H}(\theta^*)} \mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right] \leq C \inf_{\widehat{s}} \sup_{s\in\mathcal{H}(\theta^*)} \mathbb{E}\left[\|s-\widehat{s}\|^2\right],$$

where the infimum is taken over all possible estimators. Note that very often, $C \leq 1$ depends on the unknown parameters θ^*, but neither from s nor from n.
If this property holds for every parameters $\theta \in \Theta$, then $\widehat{s}_{\widehat{m}}$ is said to be *adaptive in the minimax sense with respect to the family* $\{\mathcal{H}(\theta)\}_{\theta\in\Theta}$.
For instance in the problem of density estimation on $[0,1]$ when s belongs to some Hölder space $\mathcal{H}(L,\alpha)$ $L > 0$ and $\alpha \in (0,1]$, it is known since the early 80s [29] that the minimax rate with respect to $\mathcal{H}(L,\alpha)$ for the quadratic risk is of order $L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}$, with any $L > 0$ and $\alpha > 0$.
Remark: Note that when the problem is the estimation over $\mathbb{R}$, things turn out to be very different. For instance, the minimax rate now depends on the value of regularity parameter α with respect to the parameter p of the L^p-norm used for the assessment [30].

In Chapter 6, we derive two oracle inequalities in the density estimation framework. The first one holds in the polynomial complexity setting for any projection estimator with an orthonormal basis. It is followed by an adaptivity result in the minimax sense with respect to Hölder balls for histogram estimators.
The second one is devoted to the exponential complexity setting. A preliminary illustration shows that the Lpo with small values of p suffers some overfitting due to the richness of the model collection. We propose to add a penalty term to the Lpo risk estimator, which takes the collection complexity into account.
Then, a simulation study highlights that the Loo estimator suffers some overfitting in the polynomial setting. However, it turns out that the Lpo algorithm with a choice of $p > 1$ balances this phenomenon.
Finally, since no real guideline has been provided as for the choice of p with any real data set, we provide a fully data-driven algorithm, which exhibits automatic adaptation to the complexity of the model collection.

1.5 Change-points detection

The main concern of change-points detection is to detect abrupt changes in the distribution of the observations (t_i, Y_i), where Y_i denotes the value of the signal Y observed at the deterministic point t_i. The purpose is then to recover the location of these changes.
This problem has a wide range of applications, from time-series in the financial area ([32]) to biological data [41]. Other instances can be found in [7, 16].

Unlike the *on-line* viewpoint where the observations are sequentially given, we rather consider the *off-line* approach in which the data are all observed at the same time and the problem is to recover any potential change *a posteriori*.
There are numerous asymptotic results [40, 31] which correspond to the consistency viewpoint, attempting to recover the "true segmentation". On the contrary, our approach is fully non-asymptotic, in the same line as that of Lebarbier [33] and Baraud *et al.* [5] to name but a few.

Let us assume we observe n points $(t_1, Y_1), \ldots, (t_n, Y_n) \in [0, 1] \times \mathbb{R}$ satisfying

$$Y_i = s(t_i) + \sigma_i \epsilon_i, \quad (\epsilon_i)_i \ \ i.i.d.\,, \ \mathbb{E}\left[\, \epsilon_i \,\right] = 0, \ \mathrm{Var}(\epsilon_i) = 1 \quad \text{and} \quad (\sigma_i)_i \in (\mathbb{R}_+)^n \,,$$

where s denotes the unknown piecewise constant regression function. In this framework, the complexity of the collection of models is exponential [5] and C_p-like penalties are completely misleading [13]. Birgé and Massart [12] derived a penalized criterion, which has been proved to enjoy some optimality properties in a gaussian homoscedastic setting. Besides to the best of our knowledge, nothing has been done in the fully heteroscedastic setting with rich model collections.
Our purpose is to detect change-points in the mean of the signal under heteroscedasticity, and without any assumption about the shape of the distribution of the residuals.

Chapter 7 is dedicated to an extensive simulation study, based on theoretical considerations. A first conclusion is that although Birgé and Massart's penalty is an effective measure of the complexity of rich collections of models in the homoscedastic setting, it is no longer the case under heteroscedasticity. We then propose a resampling-based strategy as a reliable measure of complexity both in the homoscedastic and in the heteroscedastic frameworks. This procedure exhibits better performances in the choice of the number of change-points than Birgé and Massart's criterion.
Furthermore, we propose to replace the empirical risk minimization in the latter procedure by the Lpo algorithm, which prevents us from overfitting and results in better segmentations in the heteroscedastic framework.

Bibliography

[1] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

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

[3] Sylvain Arlot. *Resampling and Model Selection.* PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[4] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[5] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with unknown variance. *The Annals of Statistics*, 00:00, 2008.

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

[7] M. Basseville and N. Nikiforov. *The Detection of Abrupt Changes - Theory and Applications.* Prentice-Hall: Information and System Sciences Series, 1993.

[8] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Staitstical Society. Series B*, 57(1):289–300, 1995.

[9] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive Linear Step-up Procedures that control the False Discovery Rate. *Biometrika*, 93(3):491–507, 2006.

[10] Y. Benjamini and L. Wei. A distribution-free multiple-test procedure that controls the false discovery rate. Technical Report RP-SOR-99-3, Tel Aviv University, Department of Statistics and O.R., 1999.

[11] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multipe testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

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

[13] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 2006.

[14] G. Blanchard and E. Roquain. Self-consistent multiple testing procedures. Technical report, ArXiV, 2008.

[15] L. Breiman. The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. *J. Amer. Statist. Assoc.*, 87(419):738–754, 1992.

[16] B. Brodsky and B. Darkhovsky. *Methods in Change-point problems.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[17] P. Burman. Comparative study of Ordinary Cross-Validation, v-Fold Cross-Validation and the repeated Learning-Testing Methods. *Biometrika*, 76(3):503–514, 1989.

[18] P. Burman. Estimation of optimal transformation using v-fold cross-validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3):314–245, 1990.

[19] G. Castellan. Density estimation via exponential model selection. *IEEE transactions on information theory*, 49(8):2052–2060, 2003.

[20] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

[21] R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Springer, 1993.

[22] B. Efron. Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[23] B. Efron. The jackknife, the bootstrap and other resampling plans. volume 38 of *CBMS-NSFRegional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

[24] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[25] S. Geisser. A predictive approach to the random effect model. *Biometrika*, 61(1):101–107, 1974.

[26] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[27] C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of Statistical Royal Society. Series B*, 64(3):499–517, 2002.

[28] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

[29] I. Ibragimov and R. Khas'minskij. *Statistical Estimation. Asymptotic Theory*. Springer-Verlag, Berlin, 1981.

[30] A. Juditsky and S. Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. *Bernoulli*, 10(2):187–220, 2004.

[31] M. Lavielle and E. Moulines. Détection de ruptures multiples dans la moyenne d'un processus aléatoire. In *C.R. Acad. Sci. Paris*, volume t. 324 of *Série 1*, pages 239–243, 1997.

[32] M. Lavielle and G. Teyssière. Detection of Multiple Change-Points in Multivariate Time Series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.

[33] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

[34] E. Lebarbier and T. Mary-Huard. Une introduction au critère BIC : Fondements théoriques et interprétation. *Journ. de la Société française de statistique*, 147(1):39–57, 2006.

[35] K.-C. Li. Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975, 1987.

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

[37] T. Mary-Huard, S. Robin, and J.-J. Daudin. A penalized criterion for variable selection in classification. *J. Mult. Anal.*, 98(4):695–705, 2007.

[38] P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2007.

[39] A. D. R. McQuarrie and C.-L. Tsai. *Regression and Time-Series Model Selection*. World Scientific, Singapore, 1998.

[40] B. Q. Mia and L. C. Zhao. On detection of change points when the number is unknown. *Chinese J. Appl. Probab. Statist.*, 9(2):138–145, 1993.

[41] F. Picard. *Process segmentation/clustering Application to the analysis of array CGH data*. PhD thesis, Université Paris-Sud 11, 2005.

[42] M. H. Quenouille. Approximate tests of correlation in time series. *J. Royal Statist. Soc. Series B*, 11:68–84, 1949.

[43] D. B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

[44] Marie Sauvé. Histogram selection in non gaussian regression. Technical Report 5911, INRIA, may 2006.

[45] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[46] J. Shao. Model Selection by Cross-Validation. *Journal of the American Statis. Association*, 88(422):486–494, 1993.

[47] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[48] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[49] K. Singh. On the Asymptotic Accuray of Efron's Bootstrap. *The Annals of Statistics*, 9(6):1187–1195, 1981.

[50] C.J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.

[51] 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.

[52] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *JRSS B*, 39(1):44–47, 1977.

[53] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Staitstical Society. Series B*, 64(3):479–498, 2002.

[54] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B*, 66(1):187–205, 2004.

[55] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *PNAS*, 100(16):9440–9445, 2003.

[56] N. Verzelen and F. Villers. Goodness-of-fit Tests for high-dimensional Gaussian linear models. Technical report, 2007.

[57] Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

[58] P. Zhang. Model selection via multifold cross-validation. *The Annals of Statistics*, 21(1):299–313, 1993.

Chapter 2

Model selection

The main purpose of this chapter is to review the large amount of literature about model selection. However we do not attempt to be exhaustive, which is out of the scope of this thesis. our goal is just to collect and summarize what appears to us as some of the main questions in model selection.
We first describe the rationale of model selection and recall the two strategies we may pursue, that is estimation or selection.
We then provide an overview of different model selection frameworks, from penalization to multiple testing *via* the Bayes factor and statistical testing.
Finally, we present a concurrent approach to model selection, that is model aggregation, through different aspects. We mention model aggregation (or mixing), bayesian model averaging as well as some widespread algorithms in the machine learning community such as boosting and bagging.

2.1 Model selection set-up

2.1.1 The model choice

What is a model and what for?

A *model* may be defined as a mathematical structure attempting to describe the mechanism, which generates the observations we have in hand. As an illustration, we give a few examples of what a model may be.
EXAMPLES:

- From a general point of view, a model sometimes refers to a set $\mathcal{P}_\Theta = \{P_\theta \mid \theta \in \Theta\}$ where each P_θ denotes a probability distribution [85]. Provided Θ is in one-to-one mapping with a subset of $\mathbb{R}^d$, the model is said *parametric*, whereas it is *nonparametric* otherwise.

- In a nonparametric setting, we may also define a model as a set of candidate functions (in density estimation or regression), from which we choose an approximation of the unknown target. For instance, a model may be the linear space of piecewise constant functions defined from a given partition of $[0, 1]$.

Thus, a model S is only an *approximation* of the truth. There is no warranty that the target parameter s we try to estimate belongs to S. Following [34], we could say that "truth is infinite dimensional" whereas

S looks rather like a finite dimensional model.

Since the truth is too complex to be understood in its full level of generality, a model provides a simplification of it (thanks to its smaller "dimensionality"), which turns out to provide more insight in the phenomenon of interest. For instance, a variable selection procedure applied to the thousands of genes from a cDNA microarray experiment enables to more easily detect those which are more likely to be connected to the studied mechanism.

Estimation given a model

In the estimation of a parameter s, a model S is a set of candidates from which we choose a reliable estimator. This choice is made thanks to a *contrast function* $\gamma : \mathcal{S} \times \mathcal{Z} \mapsto \mathbb{R}$ such that

$$P\gamma(\cdot) := \mathbb{E}\left[\, \gamma(\cdot, Z)\, \right]$$

is minimized by s over $\mathcal{S}$, where the expectation is taken with respect to a random variable $Z \sim P$. Note that for any estimator $\widehat{s}$, $P\gamma(\widehat{s}) = \mathbb{E}\left[\, \gamma(t, Z)\, \right]_{|t=\widehat{s}}$ is a random variable assessing the performance of $\widehat{s} = \widehat{s}(Z_1, \ldots, Z_n)$ (computed from the data) with respect to the true distribution P. In the regression context, it is named *prediction error* or sometimes the *loss* associated with $\widehat{s}$. This contrast function is related to a *loss function* $\ell : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}_+$ that measures the discrepancy between $\widehat{s}$ and s and defined by

$$\forall t \in \mathcal{S}, \qquad \ell(t, s) = P\gamma(t) - P\gamma(s) \geq 0.$$

In the density framework, an example of loss and contrast functions are the L^2-norm on $[0, 1]$:

$$\ell(t, s) = \|t - s\|^2 \qquad \text{with} \qquad \gamma(t, Z) = \|t\|^2 - 2t(Z)$$

[20, 12]. Its usual alternative is the Kullback-Leibler divergence

$$\ell(t, s) = \int \log\left(\frac{t(x)}{s(x)}\right) s(x)\, dx \qquad \text{with} \qquad \gamma(t, Z) = \log\left[\, t(Z)\, \right]$$

[12, 96, 35].
As for the regression setup, the least-squares contrast is defined for any $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$, by

$$\gamma(t, Z) = \left[\, t(X) - Y\, \right]^2,$$

while the quadratic loss is equal to

$$\ell(t, s) = \int_{\mathbb{R}} \left(t(x) - s(x)\right)^2 P_X(dx),$$

where P_X denotes the unknown distribution of X [12, 8, 11]. The likelihood-based approach may be used as well [81]. From a historical viewpoint ([34]), the quadratic-loss function is more widespread than its log-based counterpart due essentially to computation-time reasons since log-likelihood methods may require intensive computations. Moreover, estimators are the same in the gaussian regression framework. In the clustering framework, we mention that [15] use a different contrast, specially designed to properly suit the clustering purpose in a log-likelihood framework.

Ideally from the choice of such a contrast, we would like to find the minimizer of $P\gamma(\cdot)$ over model S:

$$\overline{s}(P) := \operatorname{Argmin}_{t \in S} P\gamma(t),$$

provided it exists and is uniquely defined. This function is the "best" (unknown) element in the model S, since it is the "closest point" in S to the target s. Note that $\overline{s}(P)$ coincides with s provided s belongs to S.
As we aim at being as close as possible to $\overline{s}(P)$, our strategy consists in introducing the sample mean of the contrast for any $t \in S$:

$$P_n\gamma(t) := 1/n \sum_{i=1}^{n} \gamma(t, Z_i) = \gamma_n(t),$$

which we minimize over S as well. The rationale behind this is simply that provided $P_n\gamma(t)$ is properly concentrated around its expectation, the minimizer of γ_n over S should be close to that of $P\gamma$. The typical mathematical tools to check this concentration requirement are the so-called concentration inequalities such as those of Bernstein and Talagrand... [82, 83, 69]. Following our strategy, we derive an estimator of s associated with the model S, which is the so-called empirical risk minimizer

$$\widehat{s}_S := \mathrm{Argmin}_{t\in S}\gamma_n(t).$$

Parsimony principle

Let us assume that the model S is a finite dimensional vector space with dimension D. Since the target parameter s does not necessarily belong to S, we could be tempted to replace S by a larger model S' such that $S \subsetneq S'$ and S' is closer to s than S, thus reducing the *bias* between the model and the target. The resulting estimator better fits *the data in hand*.
On the other hand, we only have a weak confidence in a model with as many parameters as the number of observations, since on average only a single observation is used to estimate each parameter of the resulting estimator. Such a model is overly overfitted, which means that if we were given new observations, the resulting histogram would unlikely fit the data in a satisfactory way.
With a finite number of observations, the empirical distribution may deviate from the true distribution P of the observations, and this deviation may be all the more large as we have only few data. Thus being too close to the observations (overfitted) may send us away from our target s.
These considerations suggest that a "reliable" model should be the result of a tradeoff between a bias term (which tells us how far from s the model S is) and another term accounting for the overfitting phenomenon.

The loss of the best estimator in S is by definition $\ell\,(s,\,\widehat{s}\,)$, which may be split into two terms

$$\ell(\,\widehat{s}\,, s) = [\,P\gamma(\overline{s}(P)) - P\gamma(s)\,] + [\,P\gamma(\,\widehat{s}\,) - P\gamma(\overline{s}(P))\,]\,. \tag{2.1}$$

By construction, both these terms are nonnegative. The first one in the right-hand side is the *bias term*. It is equal to $\ell\,(s, \overline{s}(P)\,)$ and must be understood as the "distance" between s and S. The second term is called the *variance term* and quantifies the price to pay for estimating $\overline{s}(P)$ in S. Whereas the first term decreases when the model grows, the second one rises since $\widehat{s}$ may all the more deviates as the dimension is large (and the model more flexible). Therefore, the best estimator $\widehat{s}$ reaches a tradeoff between bias and variance.
Remark: Sometimes, the variance term is named *estimation error*. Throughout this manuscript, we will also use the name "model complexity" [88], since it may be related to a complexity measure of the model [86, 87].

2.1.2 Model collection

Model collection and approximation theory

The quality of a model strongly depends on its distance to the unknown parameter s. There is therefore no warranty that a model is actually close to our target without any additional assumption about s. This is all the more true as we only consider a single model. This is the reason why we rather use a list of models, the *model collection*, rather than a single one.

Sometimes it happens that we have some prior knowledge about the problem in hand. With CGH (Comparative Genomic Hybridization) data for instance, the biological meaning of the underlying phenomenon entails that the regression function we try to estimate is piecewise constant, which provides some insight in the choice of a convenient model. Without any such information, we may only make various (more or less realistic) assumptions about the target and choose the model collection according to these assumptions.

This choice is strongly related to the approximation theory. Ideally for each set of assumptions about s, approximation theory helps in selecting the best approximating models (see [12]). In other words, approximation theory makes the bias term in (2.1) as small as possible by an appropriate choice of the collection. We refer the reader to the book of DeVore and Lorentz [38] for a wide range of results in approximation theory.

Uncertainty of the model choice

Due to the lack of information about s, we may expect that the choice of a model within a list of plausible candidates will provide us with a better final model than if we had arbitrarily selected a single model and then made some inference from it. Model selection aims at designing a procedure from the data, which provides a reliable model picked out from a pre-chosen list of candidates. However in the same way as we commit an error when estimating $\overline{s}(P)$ in the model S, the choice of the "best" model within our family of models introduces some additional uncertainty we have to take into account.

As for now and in what follows, we denote by $\mathcal{M}_n$ a countable collection of indices, which is allowed to depend on the sample size n. For each $m \in \mathcal{M}_n$, S_m denotes a finite dimensional vector space of dimension D_m that we call a model. Furthermore, let us define $\mathbb{S}$ by

$$\mathbb{S} := \cup_{m \in \mathcal{M}_n} S_m,$$

while $\mathcal{S}$ denotes the set s belongs to. For each m, the best approximation of s in S_m is denoted by

$$s_m := \mathrm{Argmin}_{t \in S_m} P\gamma(t),$$

and $\widehat{s}_m$ refers to the empirical contrast minimizer over S_m

$$\widehat{s}_m := \mathrm{Argmin}_{t \in S_m} P_n\gamma(t).$$

We are now in position to describe the model selection strategy. With each model S_m an estimator $\widehat{s}_m$ is associated. According to a given criterion crit : $\mathcal{M}_n \mapsto \mathbb{R}$, we define the candidate model $\widehat{m}$ within $\mathcal{M}_n$ as

$$\widehat{m} := \mathrm{Argmin}_{m \in \mathcal{M}_n} \mathrm{crit}(m),$$

provided it exists and is unique. Then by construction, the final estimator is

$$\widetilde{s} = \widehat{s}_{\widehat{m}}.$$

Following the same idea as in (2.1), we may rewrite the loss of $\widetilde{s}$ as follows

$$\ell(s, \widetilde{s}) = [P\gamma(\widetilde{s}) - P\gamma(s_{\widehat{m}})] + \left[P\gamma(s_{\widehat{m}}) - P\gamma(s_{\overline{m}(P)})\right] + \left[P\gamma(s_{\overline{m}(P)}) - P\gamma(s)\right], \tag{2.2}$$

where

$$\overline{m}(P) := \mathrm{Argmin}_{m \in \mathcal{M}_n} P\gamma(s_m)$$

represents the best attainable model over the whole collection, if we knew the true distribution P. The right-most term in (2.2) is the approximation bias of the collection. It is the collection counterpart of the bias term in (2.1). It decreases as the collection grows and even vanishes provided s belongs to $\mathbb{S}$. The second term in the right-hand side is the new term with respect to (2.1). It represents the price for ignoring which model is the best one. The remaining term is the variance term in the estimation of $s_{\widehat{m}} \in S_{\widehat{m}}$.

Cardinality and collection complexity issues

A naive idea is to compare as many models as possible in order to reduce the collection bias in (2.1), while avoiding in the same time high dimensional models in order to prevent from overfitting. However, two simple ideas turn out to suggest how misleading such a strategy could be.

Let us define $\widehat{m}^*$ such that $\widehat{m}^* := \mathrm{Argmin}_{m\in\mathcal{M}_n} P\gamma(\widehat{s}_m)$ exists and is unique. Then for any $\epsilon > 0$, we observe that the probability

$$\mathbb{P}\left[\min_{m\in\mathcal{M}_n} \mathrm{crit}(m) + \epsilon \leq \mathrm{crit}(\widehat{m}^*)\right]$$

is increasing with the cardinality of $\mathcal{M}_n$. Thus, considering a larger number of models would increase the probability of committing a mistake. Hence, the cardinality of the model collection is an important issue to take into account.
In a similar way for a given cardinality of the collection $\mathcal{M}_n$, it will be all the more difficult to distinguish between models as these are strongly alike. Thus, we see that increasing the number of models with the same dimension in the neighbourhood of the dimension of $S_{\widehat{m}^*}$ may induce the same phenomenon.
A conclusion can be drawn from the two previous points: the cardinality of the model collection as well as the named *collection complexity* [9] should be taken into account.

Collection complexity (sometimes called *richness*) characterizes how alike models in the collection are, that is the structure of the collection of models. The dimension of a model is a measure of "similarity" between models, which has been proposed in [20, 12]. Besides, an example of collection complexity measure is given by the cardinality of $\mathcal{M}_n(D) := \{m \in \mathcal{M}_n \mid D_m = D\}$ for each dimension D ([20, 22, 9]). For another quantification of the complexity see Yang [88].
Very often, two regimes of collection complexity are distinguished: the *polynomial* and the *exponential* ones.

Definition 2.1.1.

1. *The collection complexity is said polynomial if there exists positive constant $\delta > 0$ such that*

$$\forall 1 \leq D \leq n, \qquad \mathrm{Card}\left(\mathcal{M}_n(D)\right) \leq D^\delta.$$

2. *Otherwise, the collection is said to have an exponential complexity.*

Examples:

- A prototypical example of polynomial collection of models is the *nested collection*, which means that $m \to D_m$ is a one-to-one mapping and $Dm < D_{m'} \Rightarrow S_m \subset S_{m'}$.
- The problem of variable selection with ordered variables is another illustration of a polynomial complexity, whereas the same question with unordered variables is fully exponential [20, 8, 11, 69, 9].

2.1.3 Estimation and selection

Estimation purpose

A first possible assessment of the quality of a model selection procedure can be made through the *estimation/prediction* viewpoint [94, 22]. In this state of mind, we look for the best estimator of s among $(\widehat{s}_m)_m$ in terms of risk. The ideal (unachievable) estimator is therefore defined as $\widehat{s}_{\widehat{m}^*(P)}$, where $\widehat{m}^*(P)$ comes from

$$\widehat{m}^*(P) := \mathrm{Argmin}_{m\in\mathcal{M}_n} P\gamma(\widehat{s}_m).$$

For each sample, $\widehat{m}^*(P)$ refers to the optimal model and is moreover random. $\widehat{s}_{\widehat{m}^*(P)}$ is called the *trajectorial-oracle* estimator since it is the best attainable estimator if we knew the true distribution of the observations P.
In the same way, we introduce the *oracle* estimator as $\widehat{s}_{m^*(P)}$ with

$$m^*(P) := \mathrm{Argmin}_{m\in\mathcal{M}_n} \mathbb{E}\left[\, P\gamma(\widehat{s}_m)\,\right],$$

where the expectation is taken with respect to $\widehat{s}_m$. In this case, $m^*(P)$ is no longer a random variable. Both these oracle estimators depend on the sample size n and may subsequently differ from a value of n

to another.
Besides, neither the trajectorial-oracle nor the oracle model does necessary contain s even if the latter belongs to $\mathbb{S} = \cup_m S_m$. Indeed when a model is used to generate the data (in simulations for instance), there is no warranty that both the few observations at our disposal and the noise level (in a regression framework for instance) provide enough evidences in favour of the true model [12]. In the following, the oracle models are simply denoted by $\widehat{m}^*$ and respectively m^*.
The performance of the final estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$ may therefore be compared with either $\mathbb{E}\left[\, \ell\left(s, \widehat{s}_{\widehat{m}^*}\right) \right] = \mathbb{E}\left[\, \inf_{m \in \mathcal{M}_n} \ell\left(s, \widehat{s}_m\right) \right]$ or $\mathbb{E}\left[\, \ell\left(s, \widehat{s}_{m^*}\right) \right] = \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\, \ell\left(s, \widehat{s}_m\right) \right]$. The former criterion is used for instance in [7, 5], while in what follows we use the second one as our reference [88, 21, 11, 87, 22].

This comparison can be carried out through asymptotic or non-asymptotic arguments, which respectively leads to *(asymptotic) efficiency* or *oracle inequalities* as detailed in the following.
The *efficiency property* consists in taking s as fixed and assessing the asymptotic behaviour ($n \to +\infty$) of $\mathbb{E}\left[\, \ell(\widetilde{s}, s) \right]$ with respect to $\mathbb{E}\left[\, \ell(\widehat{s}_{m^*}, s) \right]$. The goal is to design a model selection procedure providing an estimator $\widetilde{s}$ the risk of which is asymptotically the same as that of $\widehat{s}_{m^*}$ [79, 80, 65, 78].
Remark: The efficiency property is a pointwise criterion, since the assessment is made with respect to a fixed s. Thus, Yang ([94]) argues that this type of pointwise result may give an over-optimistic view of the actual performances of the given procedure.
An *oracle inequality* is a non-asymptotic criterion quantifying how large the risk of $\widetilde{s} = \widehat{s}_{\widehat{m}}$ may be with respect to $\inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\, \ell(\widehat{s}_m, s) \right]$ up to some constant and remainder terms [20, 12, 21] to name but a few. In fact, an ideal model selection procedure would provide an estimator $\widetilde{s}$ such that $\mathbb{E}\left[\, \ell(\widetilde{s}, s) \right] = \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\, \ell(\widehat{s}_m, s) \right]$. Unfortunately, this goal cannot be reached and we rather have to content ourselves with:

$$\mathbb{E}\left[\, \ell(\widetilde{s}, s) \right] \leq C \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\, \ell(\widehat{s}_m, s) \right], \tag{2.3}$$

where $C > 1$ is a constant independent from s. According to this oracle inequality, $\widetilde{s}$ works almost as well as the best possible estimator among $(\widehat{s}_m)_m$.
Remark: Such a non-asymptotic result enables us to let the model collection depend on n. Indeed, an interesting question is the growth rate of $\mathcal{M}_n$ at which such an oracle inequality may be obtained. This kind of problem cannot be addressed from the usual asymptotic viewpoint where the model collection is kept fixed [78].

The previous criterion is also pointwise and may suffer the same troubles as those mentioned by Yang [94]. This is the reason why we are very interested in an optimality property uniform with respect to s named *adaptivity*. This notion relates to model selection through a close relationship between approximation theory and oracle inequalities.
If we assume that s belongs to a well known space of smooth functions (for instance s belongs to a Hölder ball with radius $L > 0$ and exponent $\alpha \in (0, 1]$), some families of models (Grenander [52]) are known to "optimally" approximate the target s [38]. "Optimally" means that the risk of the estimators built from these models reaches the *minimax rate* [42, 96]. We recall that the minimax risk is defined with respect to a set of parameters $\mathcal{T}$ to which s belongs:

$$R_n(\mathcal{T}) := \inf_{\widehat{s}} \sup_{s \in \mathcal{T}} \mathbb{E}\left[\, \ell(\widehat{s}, s) \right],$$

where $\widehat{s}$ denotes any estimator of s. The minimax risk provides the best possible performance of an estimator of s in the least favourable case, that is when s is as far as possible from it.
From this, we first introduce *adaptivity with respect to* θ, for s belonging to a functional space $\mathcal{T}_\theta$ and $\theta \in \Theta$. We say that a model selection procedure is adaptive with respect to θ if the risk of the resulting estimator reaches the minimax rate in the estimation of $s \in \mathcal{T}_\theta$, without knowing in advance the true parameter θ. Subsequently, an estimator is said to be *adaptive in the minimax sense with respect to* $(\mathcal{T}_\theta)_{\theta \in \Theta}$ if it is adaptive with respect to θ, uniformly in $\theta \in \Theta$. For an extensive and unified presentation of adaptivity, we refer the interested reader to [12], as well as to [20] about various results in this field.
More formally for a given family of sets $(\mathcal{T}_\theta)_{\theta \in \Theta}$ (typically some balls with respect to a given semi-norm), we compare the risk $\mathbb{E}\left[\, \ell(\widetilde{s}, s) \right]$ of an estimator $\widetilde{s}$ to the minimax risks $(R_n(\mathcal{T}_\theta))_\theta$.

Definition 2.1.2. *We say that a sequence of estimators* $(\widehat{s}_n)_n$ *is adaptive with respect to* $(\mathcal{T}_\theta)_{\theta\in\Theta}$ *in the minimax sense if for any* $\theta \in \Theta$, *there exists* $C(\theta) > 0$ *(independent from* n *and the target* s*) such that*

$$\sup_{s\in\mathcal{T}_\theta} \mathbb{E}\left[\,\ell(\widehat{s}_n, s)\,\right] \leq C(\theta) R_n(\mathcal{T}_\theta).$$

It means that the risk of $\widetilde{s}$ is at most of the same order as the minimax risk (with respect to n) or in other words, that $\widetilde{s}$ reaches (up to a constant) the best possible performance of an estimator of s with respect to $(\mathcal{T}_\theta)_{\theta\in\Theta}$.

If we now assume that our model selection procedure results in a final estimator $\widehat{s}_{\widehat{m}}$ satisfying an oracle inequality such as (2.3), then for a given assumption on the smoothness of s, approximation theory provides us with a collection of models depending on n ("sieves" [52]), such that $(\widehat{s}_{\widehat{m}_n})_{n\in\mathbb{N}}$ will be adaptive with respect to some balls in the minimax sense. Such a procedure performs as well as if we had known the smoothness of s in advance.

Selection strategy

In the estimation viewpoint, we look for the best possible estimation (or prediction) in terms of risk minimization. The ideal model is the oracle, defined as the minimizer of the true risk and we aim at selecting a model which is as close as possible to the oracle.
Unlike the previous strategy, the gold standard in *selection* is the "true model" [22], which we try to recover. This "true model" is denoted by S_{m_0} and defined as follows. $s \in S_{m_0}$ and S_{m_0} is the smallest model containing s. Let us assume that a true mode exists in the sequel for the sake of simplicity.
In other words,

$$m_0 := \operatorname{Argmin}_{m\in\mathcal{M}_n} P\gamma(s_m),$$

and $s = s_{m_0}$ (see [77, 78, 22]).
When s does not belong to our model collection, we may try to recover s_{m_0} instead of s itself, which is the closest approximation to s in $S = \cup_m S_m$ ([63]).
Remark: This point of view is that of BIC (Bayesian Information Criterion).

The assessment of the quality of $\widehat{m}$ is made through the *consistency* property. It is an asymptotic criterion according to which $\widehat{m}$ is said to be consistent if $\mathbb{P}\left(\widehat{m} = m_0\right) \xrightarrow[n\to+\infty]{} 1$ [65, 77] We also refer to [78] for a review about efficiency and consistency in linear models.
Remark: The consistency property may be defined in a slight different context, following the idea of Yang [95]. Indeed, the target function is no longer the unachievable limit of a sequence of estimators like s_{m_0}, but is rather defined for each n provided n is large enough.

2.2 Several strategies of model selection

2.2.1 Model selection *via* penalization

Deterministic penalties

Given a family $\mathcal{M}_n$ of models and the associated family of estimators $(\widehat{s}_m)_m$, model selection by *penalized criterion* consists in the choice of the best model as the minimizer of a criterion *crit*$(\cdot)$. This quantity is defined by

$$\operatorname{crit}(m) = \gamma_n(\widehat{s}_m) + \operatorname{pen}(m),$$

where $\gamma_n(\widehat{s}_m)$ assesses how well $\widehat{s}_m$ fits the data, while the penalty pen$(\cdot)$ accounts for the model complexity and possibly the collection complexity as well (see [22]).
This idea dates back to the early 70's with the seminal papers of both Akaike about FPE ([3]) and AIC ([4]), and Mallows about C_p ([68]). As illustrated by Mallows' heuristics in the regression framework,

these approaches intend to design an unbiased estimator of the risk of $\widehat{s}_m$ (see [69]), which we minimize over $\mathcal{M}_n$. To name but a few, other examples of penalized criteria are the Bayesian Information Criterion (BIC) introduced by Schwarz [76], the Minimum Description Length (MDL) of Rissanen [73] and GIC_{λ_n} proposed by Nishii [70]. Whereas FPE, AIC and GIC_{λ_n} are defined with the quadratic loss, AIC, BIC and MDL are derived in the log-likelihood context. Except both MDL and GIC_{λ_n}, Baraud *et al.* [9] point out that the above penalties result from an asymptotic heuristics and may suffer some strong dependence on the sample size as well as the type of model collection in hand.
This is the reason why a lot of work has been devoted to derive penalties in the non-asymptotic setting (see [12] for an introduction and [69] for a wide study of such penalties in various frameworks). As for the density estimation, we mention Barron and Cover [13] (with MDL), Birgé and Massart [20] for the quadratic loss and Castellan [35] for the Kullback-Leibler divergence. In the regression framework, Birgé and Massart [21] provide penalties that are generalizations on the C_p criterion in a gaussian setting when the variance is known. Baraud [8] extends these penalties to the case of unknown variance, which is also one of the purposes of Baraud *et al.* [9], where authors derive new penalties in a wide range of settings from variable selection to change-points detection for instance.
REMARK: All these penalized criteria are justified in a homoscedastic framework due to the high level of technicalities in the heteroscedastic setting.

Deterministic penalties are derived from bounds on the uniform deviation of $\ell(s, \widehat{s}_m)$ around its expectation over the whole collection of models. This approach relies on some concentration inequalities like that of Talagrand [83] or its Bousquet's version [25]. We refer to Massart [69] for a review on this topic and to Boucheron *et al.* [24] for a link with moment inequalities.
Unlike the usual bias-variance tradeoff, the need for some uniform control of the fluctuations of $\ell(s, \widehat{s}_m)$ over the whole collection of models raises the collection complexity issue with rich collections. Thus provided there are lots of models with the same dimension or not, we have to distinguish between penalties.

- C_p-like penalties are valid when the collection of models is not too rich and may be expressed as

$$\mathrm{pen}(m) = CD_m,$$

  where $C > 1$ denotes an unknown universal constant. Justifications for them may be found in Baraud [8] for regression and Castellan [35] in a density estimation setup for instance.

- As for richer collections of models, Birgé and Massart [20] in density estimation and Barron [12] for the regression have shown that the appropriate penalty structure is of the type

$$\mathrm{pen}(m) = c_1 D_m + c_2 D_m \log\left(\frac{N}{D_m}\right),$$

  where c_1 and c_2 are universal positive constants, and N denotes the total number of variables in the unordered variable selection problem for instance. From a theoretical viewpoint, they argued that the $\log-$term is unavoidable (see Birgé and Massart [21] and Donoho and Johnstone [40]) and should be understood as the price to pay for ignoring the oracle.

Note that concentration inequalities provide tight bounds up to some constants, which induces some further work in order to get the best possible values for these constants. Lebarbier [62] has successfully performed an intensive simulation study and found that optimal constants in the gaussian change-points setup are $c_1 = 5$ and $c_2 = 2$. Actually, these constants may only be determined up to a multiplicative constant. In a recent paper, Birgé and Massart [22] discuss this point and describe a strategy named the "slope heuristics", which is theoretically proven to yield an adaptive choice of this unknown multiplicative constant. For not too rich collections of models, Arlot and Massart [6] provide a justification of this heuristics in the random design regression framework for a heteroscedastic noise, without any distributional assumption on the residuals.
REMARK: Lavielle [61] suggests to use a C_p-like penalty whatever the collection complexity. Since this penalty is determined up to an unknown constant, he developed another heuristics which provides an adaptive choice for this constant as well. Actually, this procedure exhibits rather good performance in the change-points detection at least [71]. From a theoretical viewpoint, this procedure relates to the above Birgé and Massart criterion in that when the dimension of the

oracle model $D^* \ll n$ in the change-points detection framework, the penalty may be written as $\mathrm{pen}(m) = c_1 D_m + c_2 D_m \log\left(\frac{n}{D_m}\right) \approx D_m\,(c_1 + c_2 \log n) = C D_m$, where C denotes a constant. The conclusion we may draw is that this heuristics should work as long as $D^* \ll n$, but may encounter some troubles otherwise.

Random penalties

Since they are derived from concentration inequalities, penalized criteria based on deterministic penalties may be seen as upper bounds of the ideal criterion $\ell\,(s, \widehat{s}_m\,)$. Arlot [7] introduces the *ideal penalty*, which is defined for any $m \in \mathcal{M}_n$ by

$$\ell\,(s, \widehat{s}_m\,) = P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{id}(m) \qquad \text{with} \qquad \mathrm{pen}_{id}(m) := (P - P_n)\,\gamma(\widehat{s}_m).$$

Hence we conclude that any (meaningful) deterministic penalty is an upper bound of the ideal penalty (up to some multiplicative unknown constants).

Although concentration inequalities are sharp tools, we may try to estimate the risk. To this aim, another strategy is to use an approximation of $\mathrm{pen}_{id}(m)$ rather than an upper bound, which results in random penalties.
Efron [44, 45] introduced the bootstrap, which provides an approximation of the true distribution and applied this resampling algorithm to model selection ([46]). Fromont [50] formalized this idea and successfully developed bootstrap penalties in classification. Arlot [7, 5] generalized this idea to other resampling schemes.

Despite the lack of theoretical results about resampling algorithms due to the high technicality of the proofs, random penalties are useful and really attractive, since they do not depend on any distributional assumption. Whereas some deterministic penalties are only justified in a gaussian framework for instance, resampling ones are working no matter what is the distribution of the data or even if the noise is homoscedastic or heteroscedastic. In particular when the model collection is not too rich, Arlot [7] (Chap.4) showed that some penalties designed and justified in the homoscedastic case are suboptimal with respect to random penalties under heteroscedasticity.

However, a major drawback of these penalties is that they may turn out to be computation time consuming. Some explicit formulas may be derived in some particular settings like exchangeable weights combined with regressograms (see [7] Chap.6). Nevertheless in more general situations, such penalties could sill be hard to compute. There is therefore a tradeoff between the gain we may expect from the use of resampling penalties and the time we are willing to spend on their computation.

Interest of overpenalization

The first step in model selection *via* penalization was made by Mallows [68] and Akaike [3]. Their strategy relies on the unbiased estimation of the risk corresponding to each model. Indeed as we look for the minimizer of $\ell\,(s, \widehat{s}_m\,)$ over $\mathcal{M}_n$, a plausible solution may be brought by the minimization of any reliable estimator of the latter quantity. Although this strategy seems effective when applied within each model S_m, it is no longer true as soon as we are concerned with the whole collection. Actually, the unbiased estimation of the risk is not a sufficient condition to recover the best model as we can check considering Mallows' C_p in the homoscedastic change-points detection framework if the true variance is known.

$$C_p(m) = \gamma_n(\widehat{s}_m) + 2\sigma^2 \frac{D_m}{n}$$

is an unbiased estimator of the risk $R(m) = \ell\,(s, s_m\,) + \sigma^2 D_m/n$. Yet, the following inequality holds

$$\mathbb{E}\left[\inf_{m\in\mathcal{M}_n} C_p(m)\right] \leq \inf_{m\in\mathcal{M}_n} \mathbb{E}\,[\,C_p(m)\,] = \inf_{m\in\mathcal{M}_n} \mathbb{E}\,[\,\ell\,(s, \widehat{s}_m\,)\,]$$

where the discrepancy between the two sides widens with the collection complexity. This does not prove anything since the resulting $\widehat{m}$ could behave very well. However, it enlightens that minimizing an estimator of the risk does not necessarily provide the minimizer of the latter (even an approximation of it). For instance, Birgé and Massart [22] argue that C_p is not suited in some circumstances due to the high collection complexity.

"Overpenalizing" (resp. "underpenalizing") means we use a penalty that is larger (resp. lower) with high probability than pen_{id}. Note that either over- or under-penalization induce a bias of the resulting criterion with respect to the ideal one. Any underpenalization may lead to the choice of a larger model than the best one and to very poor performance [22], which we would like to avoid.
The ideal penalty introduced by Arlot [7] is a random variable that fluctuates around its expectation. Some concentration inequalities may provide an inequality of the following type for each model m with high probability

$$\mathbb{E}\left[\,\mathrm{pen}_{id}(m)\,\right] - \varphi(m) \leq \mathrm{pen}_{id}(m) \leq \mathbb{E}\left[\,\mathrm{pen}_{id}(m)\,\right] + \psi(m),$$

where φ and ψ are both positive functions, increasing with the complexity of model m. Thus, overpenalizing consists in designing a penalty so that its expectation upper bounds $\mathbb{E}\left[\,\mathrm{pen}_{id}(m)\,\right] + \psi(m)$ uniformly over $\mathcal{M}_n$. Thus the richer the collection of models, the stronger the amount of overpenalization.

From an asymptotic viewpoint, another interesting aspect of overpenalization may lie in the ability of the resulting procedure to recover the true model, provided it belongs to the collection in hand. Shao [78] studies various penalized criteria in the context of linear regression, with respect to essentially two properties: asymptotic efficiency and consistency. It turns out that several penalized criteria such as C_p, FPE_λ or BIC may be understood through the analysis of the GIC_{λ_n} criterion:

$$GIC_{\lambda_n}(m) = \frac{S_n(m)}{n} + \lambda_n \frac{\widehat{\sigma}^2 D_m}{n},$$

where $S_n(m) = \sum_{i=1}^n \left(Y_i - \widehat{s}_m(t_i)\right)^2$, $\lambda_n > 0$ and $\widehat{\sigma}^2$ may be thought of as a consistent estimate of the variance. We set $\mathcal{M}$ a nested collection of models. Let assume that $m_0 \in \mathcal{M}$ is the true model. We define $\mathcal{M}_b = \{m \in \mathcal{M} \,|\, m \subsetneq m_0\}$ the set of biased models and $\mathcal{M}_b^c = \{m \in \mathcal{M} \,|\, m_0 \subset m\}$. Indeed, Shao explains that $\lambda_n \equiv 2$ corresponds to C_p (and AIC at least asymptotically), while $\lambda_n = \log n$ has the same asymptotic behaviour as BIC. This criterion ($\lambda_n \equiv 2$) cannot properly distinguish between several models in $\mathcal{M}_b^c$ ([78] Th.1) and tends to overfit. On the contrary, Shao proves that provided $\lambda_n \to +\infty$ and $\lambda_n/n \to 0$, GIC_{λ_n} is consistent and so is BIC as well ([78] Th.2).

About a possible conciliation between estimation and selection

In his work, Shao [78] summarizes results about AIC and BIC which are both optimal, but in apparently different ways. Indeed whereas AIC is asymptotically efficient, BIC selects the true model (provided it belongs to the considered model collection) with probability growing to 1 as $n \to +\infty$ (consistency). However, we may then wonder whether it is possible to design a new criterion which would combine these two optimality properties. Several attempts have been made to design such a model selection criterion, for instance from MDL. Thus under some conditions, Barron *et al.* [14] show that the latter behaves alternatively like BIC or AIC, depending on whether data come from a parametric or nonparametric model. Besides in a bayesian framework, Hansen and Yu [53, 54] propose a modification of MDL named gMDL, which relies on a F-statistic and appears as a compromise between BIC and a AIC-like penalty. By mixing AIC-based and BIC-based estimators, Yang [93] provides a predictor capturing the behaviour of the best one of them.

Although some optimality properties may be shared by both AIC and BIC (see [78] for a review of various penalized criteria with asymptotic properties), it turns out that these properties are not the right ones we should look for. Indeed, Brown *et al.* [33] emphasize that asymptotic efficiency may be misleading as a pointwise optimality criterion that strongly depends on the estimated function. Thus, we should prefer the adaptivity property satisfied by AIC for instance. It provides some information about the uniform ability of a procedure to provide a reliable estimation.

Yang[93] answers by the negative to the main question and gives a counter example in the gaussian linear regression framework, which illustrates the essential discrepancy between AIC and BIC. The conclusion he draws is that if a model selection procedure is designed to be consistent, then it cannot be adaptive in the minimax sense, that is optimal from an estimation viewpoint.

2.2.2 Statistical tests

There seem to be a relationship between model selection and statistical tests. For example, let us consider a linear regression model in which a F-test is performed for nested models. The null hypothesis is " μ, the parameter of interest belongs to a given finite dimensional vector space W", while its alternative is "μ belongs to $V \setminus W$", where V also denotes a finite dimensional vector space containing W. Even if the rejection of the null hypothesis only means that there is no strong enough evidence in favour of $\mu \in W$ (and not that the alternative hypothesis is true), it turns out that such a test enables to successively discard some candidate models. In a gaussian setting, Baraud *et al.* [10] propose a test procedure, related to model selection through the rewriting of its rejection region and provided the true model exists and corresponds to the null hypothesis.

In a quite recent paper, Birgé [19] describes a framework relying on the statistical test theory and attempts to perform model selection. He starts drawing the picture of different troubles suffered by empirical contrast minimizers such as the MLE (*i.e.* Maximum Likelihood Estimator). These troubles are a lack of robustness, of regularity of the likelihood and possibly a too high massiveness of the parameter space. He suggests the use of statistical test theory in order to overcome these deficiencies. From a set of parameters from which we are looking for the best estimator, and given a set of tests (corresponding to the contrast function in the usual approach), our estimator is the parameter for which both the null hypothesis is not rejected and the rejection region is as small as possible (according to a certain distance). However, the practical use of the whole procedure is almost impossible due to prohibitive computational costs.

2.2.3 Bayes factor and posterior predictive loss

As far as we know about bayesian ideas, model selection is essentially related to statistical tests and decision theory. To distinguish between models of a given finite family, we may use the Bayes factor. Let us consider two models m_0 and m_1 respectively associated with the null and the alternative hypothesis. Kass and Raftery [60] define the Bayes factor B_{10} as the ratio of the predictive distribution of m_1 to that of m_0, that is

$$B_{10} = \frac{P(Obs \,|\, m_1)}{P(Obs \,|\, m_0)} \,,$$

where Obs denotes the data. The Bayes factor does not depend on any prior distribution of either m_0 or m_1, contrary to the ratio of posterior distributions it is related to by the straightforward relation

$$\frac{P(m_1 \,|\, Obs)}{P(m_0 \,|\, Obs)} = \frac{P(Obs \,|\, m_1)}{P(Obs \,|\, m_0)} \frac{\pi(m_1)}{\pi(m_0)} \,,$$

where $\pi(m_k)$ and $P(m_k \,|\, Obs)$ respectively denote the prior and the posterior distributions of model m_k. While Bayes factor expresses how much each model among $\{m_0, m_1\}$ may be actually used to predict the observed data, the ratio of posterior distributions quantifies the amount of support of each model by the observations.

A comparison between p-value and Bayes factor can be pursued [72].
In statistical test theory provided the null hypothesis is rejected, we can only conclude that there was not enough evidence in favour of this hypothesis. In no way we can interpret this rejection as an acceptance of the alternative. Moreover for usual statistical tests (like the likelihood ratio for instance), m_0 and m_1 are nested.
From the bayesian viewpoint, the decision of rejecting any hypothesis is taken according to the Bayes factor value with respect to 1, which provides both some evidence in favour of one model and conversely against the other one.

Besides when only two models are compared, both p-value and Bayes factor may be used to choose a model. It turns out to be more intricate with several models, due to the multiple testing issue (Chapter 5).
Raftery [72] argues against p-values that rejection or acceptance of some hypotheses depend on a given cutoff, which entails some troubles when a large number of models are compared since more and more p-values are below this cutoff. On the contrary, the Bayes factor induces an order between models, which is unsensitive to the number of compared models. The final model, that is the first rank one, remains reliable.
Actually, this point of view about the multiple testing issue (Chapter 5) is somewhat questionable. Rather than p-values themselves, it is the standard practice of people with p-values that may be argued against. Moreover when a large number of p-values are considered, multiple testing procedures [16, 43, 32] have been specially designed to cope with the kind of trouble mentioned in [72].

Since it depends on integrals, for which closed-form expression are hardly obtained, the Bayes factor may involve some intensive numerical calculations, or stochastic approximation algorithms. Moreover, all these methods may induce a large gap between the true value and its (numerical) approximation, which creates some difficulties in truly rejecting or accepting any hypothesis or model. To this end, Kass and Raftery [60] provide some tables that are guidelines in order to quantify the amount of evidence in favour of which a given model may be rejected for instance. The BIC criterion is related to the Bayes factor and provides a reliable approximation to it, considerably reducing computational costs.

Another bayesian approach to model selection relies on what is called *the posterior predictive loss* by Gelfand and Ghosh [51]. They justify their approach arguing against Bayes factor that it is hard to compute with large data sets and suffers some deficiencies in the case of noninformative prior distributions, which are widespread in complex hierarchical models. Moreover, Bayes factor turns out to be an optimal criterion in the particular case of 0-1 loss, when model selection is viewed through hypothesis testing [59]. Nonetheless, Gelfand and Ghosh [51] emphasize that it may be worth considering more general losses. They present a loss-predictive-based strategy which amounts to successively use each model to predict a new set of observations and then quantify through a loss function the deviation of these new data from the original observed ones. The loss function enables to reach a kind of tradeoff between a good fit of the model to the preliminary observations and a reasonable variation of the predicted data around the model prediction (small model complexity). Notice that this approach in a gaussian regression framework results in a AIC-like penalized criterion for instance [51].
REMARK: The bayesian view of model selection is somehow different from the "frequentist" one. Indeed the latter takes into account the randomness of the observations, whereas the former takes them for granted, *i.e.* deterministic. Following this idea, the bayesian analysis of this problem does not pay attention to the collection complexity and only cares about risk estimation for each model, while taking into account some randomness in the models.

2.2.4 Model selection and multiple testing

There exist a relationship between model selection by penalized criteria and the multiple testing framework described in Chapter 5. This link is described in a recent paper of Abramovich *et al.* [2], which addresses the problem of recovering the non-zero components in the mean of a gaussian vector with a constant known noise. This question has already been widely studied, essentially motivated by applications in the signal processing area by wavelet thresholding with sparse signals [41, 1] for instance. In this set-up, the model is

$$\forall i = 1, \ldots, n \qquad Y_i = \mu_i + \sigma\epsilon_i, \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

where $\mu \in \mathbb{R}^n$ and the standard deviation $\sigma > 0$ is known.
In [2], authors follow the hard thresholding strategy which consists in estimating the $i-$th component of the mean by the $i-$th coordinate of the signal provided it is strong enough with respect to a given threshold. In other words for $t > 0$,

$$\forall i, \qquad \widehat{\mu}_i(t) = \left| \begin{array}{ll} Y_i & , \text{ if } |Y_i| \geq t \\ 0 & , \text{ otherwise} \end{array} \right. .$$

Two different optimal thresholds t are compared to each other. The first approach relies on the named Benjamini-Hochberg multiple testing procedure (Section 1.3.2), which aims at controlling the FDR (False Discovery Rate), while the second one consists in minimizing a penalized criterion. On the one hand, the parallel of the FDR-based procedure with penalized criteria enables to derive an adaptivity property with respect to some ℓ_p-balls with a given level of sparsity. The main interests of this results lie in its uniformity over a wide range of parameter sets and in the adaptation of the procedure to the unknown sparsity of the signal. An extension to the case of sparse exponential variables has been made by Donoho and Jin [39], where numerous asymptotic minimax results are proved. [2] stress that no other estimation procedure do allow us to achieve such an adaptation result. On the other hand, the use of the FDR procedure underlines the potential interest of penalties based on the quantiles of some random variables related to the problem in hand. Such penalties could be seen as reliable alternatives to the usual ones, based on the dimension. Note that above strategy may be related to a recent work of Baraud *et al.* [9], where penalties are expressed as functions of some F-statistic quantiles. Moreover, the authors illustrate the interest of this promising strategy in a wide range of simulation experiments.

2.3 Aggregation as an alternative to model selection

Model selection turns out to be a useful tool in order to get more insight in the phenomenon in hand. In the variable selection problem, the choice of one subset of covariates may give us more insight in the understanding of the considered problem. We gain in interpretability by focusing on the more relevant variables to explain what has generated the observations. Besides provided a true model exists, using model selection in order to recover it, or at least to choose a reliable approximation to it seems appropriate.

Nevertheless, several models may be very close to one another according to our criterion, each one being reliable as a final model. Moreover since we do not have the true distribution in hand but only a finite sample from it, discarding all these models except one may seem questionable. On top of that, we may be more interested in the estimation of some parameters rather than in the interpretation of the resulting model. This reasoning leads us to an other strategy named *model aggregation* or *model averaging* for instance. It relies on the idea of combining the estimators associated with models in $\mathcal{M}_n$ through an appropriate choice of weights, which determines the amount of influence of each estimator in the final estimation.

Model aggregation has become very popular in the last decade and is now involved in various domains such as density estimation [89], regression [90, 64], statistical learning [84], individual sequences prediction [36], boosting [74],... Following Yang [92], we may distinguish two different strategies of aggregation: combining *for adaptation* and combining *for improvement*.

2.3.1 Aggregation for adaptation

Aggregation for adaptation aims at combining estimators so that the mixing one automatically reaches the best performance among the family of estimators.
In the density estimation setting, Yang [89] points out that the notion of aggregation for adaptivity (with respect to the initial procedures) leads to adaptivity in the minimax sense, provided some estimators among our collection satisfy this property. He first proposes information-theoretic mixing strategies for both density estimation [89] and regression [90], and then builds another algorithm named ARM [91], which convexly combines regression procedures thanks to cross-validation. Applied with AIC and BIC as initial procedures, Yang [93] empirically shows that the mixing procedure seems to share the optimality properties of the two criteria. He also exhibits some counter-example in which mixing strategies systematically outperform upon selection ones. The price to pay for ignoring which one of the initial procedures is the best one is settled to be of order $\mathcal{O}(\log(p)/n)$, where p is the total number of combined procedures, and n denotes the cardinality of the data. As long as p remains at most polynomial in n, this penalty term remains acceptable. Similar bounds are derived by Leung and Barron [64] for the problem of estimation of the mean of a gaussian vector with known homoscedastic noise. An estimator is proposed that is the weighted average of all the projection estimators in the family. The corresponding risk is assessed thanks to Stein's unbiased risk estimator, which provides oracle inequalities with constant 1, while Yang derives results in a more general framework at the price of larger constants.

As for prediction of individual sequences, Cesa-Bianchi and Lugosi [36] develop another strategy relying on the assessment of a given procedure through the use of a cumulative loss and regret minimization. From a beforehand given set of experts, a mixing strategy is proposed that convexly combines each expert proposal of prediction. For a set of p any experts, the proposed strategy performs as well as the best expert up to an additive penalty term of order $\sqrt{\log(p)/n}$, which may be extended to the convex hull of a finite set of p experts. With further specifications on the set of experts, they manage to get some smaller bounds derived from an accurate study of the minimax regret, which essentially depends on the geometrical properties of the set of experts and relates to empirical process theory. This strategy is fruitfully applied [37] to infinite classes of experts and log-likelihood based losses for instance.

2.3.2 Aggregation for improvement

Another goal we may pursue in model aggregation is averaging for improvement. In particular, we now compare an averaging strategy to the best possible one, instead of the best performance of any initial procedure, which is a stronger requirement. Juditsky and Nemirovski [58] describe a gaussian regression setting with bounded regression function and standard deviation in which they propose a linear combination (with constraints coefficients) of p preliminary (random) functions, which performs as well as the best possible such combination up to a penalty term of order $\mathcal{O}\left(\sqrt{\log(p)/n}\right)$, where denotes the sample size. Moreover, they show that this penalty term cannot be improved when $p \geq n/\log(n)$.
Remark: This term decreases slower than in the aggregation for adaptation framework [91, 93], highlighting that this purpose is more difficult to reach.
Nonetheless, this bound may be improved at least when $p \leq \sqrt{n}$ [92]. Indeed in the same context, Yang uses the ARM algorithm in order to show that provided $p \geq \sqrt{n}$, the bound of [58] is the best possible one, whereas it may be reduced to $\log n/n^{1-\alpha}$ as long as $0 < \alpha < 0.5$. Note that the latter bound turns out to be optimal up to a logarithmic factor and is anyway faster than that of [58].

Bayesian model averaging When a model has been chosen by model selection, it is taken for granted and estimators of some quantities of interest are computed without taking into account the uncertainty lying in the preliminary model choice. This point is already mentioned in Leung and Barron [64]. BMA (*i.e.* Bayesian Model Averaging) intends to overcome this matter and builds estimators defined as a weighted average of estimators based on each model. Note that since inferring the importance of a given variable from only one model is somewhat troublesome [56], the goal of BMA is averaging for improvement.
However even if such an approach seems meaningful, it combines several essential issues such as summations over an overly large number of models or computation of several implicit integrals. Hoeting [55] and Hoeting *et al.* [56] describe some possible ways to circumvent these two difficulties.
The first one comes from an idea of Madigan and Raftery [67] named "Occam's window", which originates in graphical models. Instead of summing over all possible models, which is definitely prohibitive, they try to discard as many models as possible that does not seem to be supported by the data. First, they exclude models for which the integrated likelihood is too far from the optimal one according to a user-specified cutoff. The second rule consists in rejecting the most complex one from a pair of nested models when there is evidence in favour of the simplest one. Although authors explain that these rules enables a drastic reduction of the number of models to consider, these rules may seem *ad-hoc* and subsequently unsatisfactory. Another strategy relies on the use of MC^3 (*i.e.* Markov Chain Monte-Carlo Model Composition), which yields a stochastic approximation of these problematic sums.
As for numeric approximation of implicit integrals, some powerful algorithms may apply, but nevertheless they are computation time consuming. That is the reason why some approximations like Laplace's one may be preferred, which may lead to the BIC criterion in certain circumstances [72].

Overcome instability In the sequel, we focus on the classification problem in which given N observations $(X_1, Y_1), \ldots, (X_N, Y_N)$ drawn from an unknown distribution P on $\mathcal{X} \times \{0, 1\}$, we would like to "learn" the relationship between instances X_i in the feature space and the corresponding labels Y_i in order to accurately predict the label Y of any new instance X. We refer the interested reader to Boucheron *et al.* [23] for a review of some recent developments in statistical and machine learning.

In classification and more generally in model selection, some estimation or prediction procedures are known to suffer some "instability" (Section 3.2.1) and [26]), increasing the risk of the resulting estimator [29]. For instance, subset selection [29] and CART [31] are famous instances belonging to this class of procedures.
Bagging (*i.e.* **b**ootstrap **agg**regat**ing**) has been proposed by Breiman [27] as an attempt to reduce this troublesome instability. It consists in a mixing strategy relying on the bootstrap resampling of the original data (Section 3.1.2). At each step of the B bootstrap resamplings, a new estimator is computed and the final one is defined as the average of the B resulting estimators. Breiman [28] argues that while reducing the variance of the initial procedure, bagging does not increase its bias. The bagging estimator subsequently improves upon each bootstrap one. However, bagging turns out to be effective only when applied to instable procedures, otherwise it may induce some slightly worse performances [27, 17].
REMARK: We could say that instability may appear when several models or procedures are reliable to describe observations. Even a small perturbation of the sample may induce a change in the procedure result. Although the bayesian viewpoint takes the observations for granted and rather put randomness in the model, we could nevertheless say that bagging intends to take into account the uncertainty in the model choice due to the finiteness of the sample just as bayesian model averaging.

Boosting weak algorithms An interesting question in the early 90s was to know whether it was possible to design a *strong learning* algorithm from a *weak learning* one, that is to *boost* any weak learner. This problem has been positively solved by Schapire [74], who raised in the same time a wide range of algorithms called *boosting* (algorithms).
Given a training set of N points $(X_i, Y_i) \in \mathcal{X} \times \{0, 1\}$, each instance X_i is given a probability distribution mass p_i ($p_i = 1/N$ in the uniform case for example). A weak learner (or weak learning algorithm) takes in input both the training points and the associated probability distribution on instances X_i. A learning algorithm is a "weak" learner if it returns a classifier which labels the training points only slightly better than a random guess. Then, this classifier is called weak classifier. The "weak" terminology aims at contrasting with what we call "(strong) learners and/or classifiers". Roughly, we may define a (strong) learning algorithm as an algorithm that may perform as well as we want with high probability and a polynomial complexity [74].
A boosting algorithm is based on a weak learner it will boost through a re-weighting mechanism of the instances in the training set. Thus when a weak classifier has been output by the weak learner, a larger distribution mass is allocated to misclassified points, the whole process being iterated T times. At each round t, updating the probability distribution of the instances enforces the algorithm to concentrate on the "hardest" points to classify. The final (strong) classifier is then defined as a convex combination of classifiers obtained at each step, with weights depending on the amount of error observed at each round of the algorithm.

Although boosting algorithms such as AdaBoost [47] are empirically known to perform very well, there is still things to understand about the way they behave. Indeed for instance, the precise reasons why the generalization error of boosting still decreases while the empirical risk on the training set has already vanished for some time are not completely established yet. Thus whereas Schapire *et al.* [75] propose an explanation relying on the margin maximization, Breiman [30] shows that margin maximization may lead to overfitting and hence to an increase in generalization error. Another (not completely satisfactory) view about boosting has been described in [49, 48], where it is presented as a gradient-descent algorithm, which enables several variants of the original boosting algorithm depending on the pseudo-loss function (LogitBoost [49], L^2Boost [18],...).
Another important issue of boosting is overfitting. In the initial boosting algorithm, the weak learner aims at minimizing the empirical risk without any regularization step. Not only this minimization problem is non-convex, but it leads to systematic overfitting without any regularization step [75, 30], provided the number T of iterations is large enough. Quite recently, several potential solutions have been studied such as the choice of a "simple" set of candidate weak classifiers, the limitation of the number of iterations

[57], the choice of an alternative pseudo-loss function [18, 66, 97]. From a computational viewpoint, this pseudo-loss is chosen to be convex (efficient optimization algorithms). Moreover, [66, 97] proved that an appropriate choice of this loss function enables to regularize the minimization procedure and yields a consistency result of the boosting classifier towards the Bayes risk, as $N \to +\infty$.

Remark: To some extent, Bagging and Boosting algorithms share some similarities, which justifies several comparisons [28, 75]:

- Instances in the training set are associated with a probability mass, which is updated at each step,
- On each round, a new (weak) classifier is computed from the re-weighted sample,
- The final classifier is a convex combination of the different weak classifiers.

Nevertheless, it is worth underlying some differences:

- In the boosting strategy, the allocated distribution mass is larger for misclassified points, which enables the boosting algorithm to concentrate on the hardest points to classify. On the contrary, the purpose of bootstrap is to explore the set of all re-samples without focusing on some of them.
- Whereas the weights of the boosting classifier depend on the performance of each weak classifier, the weights of its bagging counterpart are all equal to $1/T$.

Bibliography

[1] F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Computer Statistical Data Analysis*, 22:351–361, 1996.

[2] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to Unknown Sparsity by controlling the False Discovery Rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[3] H. Akaike. Statistical predictor identification. *Ann. Inst. Statisti. Math.*, 22:203–217, 1969.

[4] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[5] S. Arlot. Model selection by resampling penalization. *Electronic journal of Statistics*, 00:00, 2008.

[6] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning*, page submitted, 2008.

[7] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[8] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[9] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with unknown variance. *The Annals of Statistics*, 00:00, 2008.

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

[11] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

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

[13] A. Barron and T. M. Cover. Minimum Complexity Density Estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

[14] A. Barron, Y. Yang, and B. Yu. Asymptotically optimal function estimation by minimum complexity criteria. In *In Proc. IEEE Int. Symp. Information Theory*, 1994.

[15] J.-P. Baudry, G. Celeux, and J.-M. Marin. Selecting Models Focussing on the Modeller's Purpose. In *Proceedings in Computational Statistics, 18th COMPSTAT symposium*, 2008.

[16] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Staitstical Society. Series B*, 57(1):289–300, 1995.

[17] P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[18] P. Bühlmann and B. Yu. Boosting with $L2$loss: Regression and classification. *J. Amer. Statistic. Assoc.*, 98(462):324–339, 2003.

[19] L. Birgé. Model selection via testing: an alternative to (penalized) maximum lielihood estimators. *Annales de l'Insitut Henri Poincaré*, 42:273–325, 2006.

[20] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgensen, and G. Yang, editors, *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

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

[22] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 2006.

[23] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM Probab. Statist.*, 9:323–375, 2005.

[24] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment Inequalities for Functions of Independent Random Variables. *Annals of Probability*, 33:514–560, 2005.

[25] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris*, 1:495–500, 2002.

[26] O. Bousquet and A. Elisseff. Stability and Generalization. *J. Machine Learning Research*, 2:499–526, 2002.

[27] L. Breiman. Bagging predictors. Technical Report 421, Department of Statistics, University of California ai Berkeley, 1994.

[28] L. Breiman. Bias, Variance, and arcing classifiers. 1996.

[29] L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2382, 1996.

[30] L. Breiman. Prediction Games and Arcing Algorithms. *Neural Computation*, 11(7):1493–1517, 1999.

[31] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Chapman & Hall, 1984.

[32] P. Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6:199, 2005.

[33] L.D. Brown, M.G. Low, and L.H. Zhao. Superefficiency in nonparametric function estimation. *The Annals of Statistics*, 25:2607–2625, 1997.

[34] K. P. Burnham and D. R. Anderson. *Model selection and inference: A Practical Information-Theoretic Approach*. Springer, 1998.

[35] G. Castellan. Density estimation via exponential model selection. *IEEE transactions on information theory*, 49(8):2052–2060, 2003.

[36] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *The Annals of Statistics*, 2(6):1865–1895, 1999.

[37] N. Cesa-Bianchi and G. Lugosi. Worst-Case Bounds for the Logarithmic Loss of Predictors. *Machine Learning*, 43:247–264, 2001.

[38] R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Springer, 1993.

[39] D. Donoho and J. Jin. Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *The Annals of Statistics*, 34(6):2980–3018, 2006.

[40] D. Donoho and I. Johnstone. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, 81:425–455, 1994.

[41] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

[42] D.L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26:879–921, 1998.

[43] S. Dudoit, J. Popper Shaffer, and J. C. Boldrick. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18(1):71–103, 2003.

[44] B. Efron. Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[45] B. Efron. The jackknife, the bootstrap and other resampling plans. volume 38 of *CBMS-NSFRegional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

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

[47] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[48] J. Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[49] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[50] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66(2–3):165–207, 2006.

[51] A. E. Gelfand and S. K. Gosh. Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11, 1998.

[52] U. Grenander. On the theory of mortality measurement. *Skandinavisk Aktuarietidskrift*, 39(2):125–153, 1956.

[53] M. Hansen and B. Yu. Bridging AIC and BIC: an MDL model selection criterion. In *In Proc. IEEE Information Theo. Workshop on Detection, Estim., Classif. and Imaging*, page 63, 1999.

[54] M. Hansen and B. Yu. Model selection and principle of mimum description length. *J. Amer.Statist. Assoc.*, 96:746–774, 2001.

[55] J. Hoeting. Methodology for Bayesian Model Averaging: An Update. International Biometrics Conference Proceedings, 2002.

[56] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian Model Averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[57] W. Jiang. Process consistency for AdaBoost. *The Annals of Statistics*, 32(1):13–29, 2004.

[58] A. Juditsky and A. Nemirovski. Functional agregation for nonparametric regression. *The Annals of Statistics*, 28(3):681–712, 2000.

[59] J.B. Kadane and J.M. Dickey. *Bayesian decision theory and the simplification of models*. In Evaluations of Econometric Models. New York, Academic Press, 1980.

[60] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statstical Association*, 90(430):773–795, 1995.

[61] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85:1501–1510, 2005.

[62] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

[63] E. Lebarbier and T. Mary-Huard. Une introduction au critère BIC : Fondements théoriques et interprétation. *Journ. de la Société française de statistique*, 147(1):39–57, 2006.

[64] G. Leung and A.R. Barron. Information Theory and Mixing Least-Squares Regression. *IEEE transactions on information theory*, 52(8):3396–3410, 2006.

[65] K.-C. Li. Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975, 1987.

[66] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of boosting methods. *The Annals of Statistics*, 32(1):30–55, 2004.

[67] D. Madigan and A. Raftery. Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window. *J. Amer. Statistical Assoc.*, 89(428):1535–1546, 1994.

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

[69] P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2007.

[70] R. Nishii. Asymptotic properties of criteria for selection of variables in multipe regression. *The Annals of Statistics*, 12(2):758–765, 1984.

[71] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 27(6):electronic access, 2005.

[72] A. E. Raftery. Bayesian Model Selection in Social Research. *Siociological Methodology*, 25:111–163, 1995.

[73] J. Rissanen. Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431, 1983.

[74] R. Schapire. The Strength of Weak Learnability. *Machine Learning*, 5:197–227, 1990.

[75] R. Schapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[76] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[77] J. Shao. Model Selection by Cross-Validation. *Journal of the American Statis. Association*, 88(422):486–494, 1993.

[78] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[79] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[80] R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statisti. Math.*, 35:415–423, 1983.

[81] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *JRSS B*, 39(1):44–47, 1977.

[82] M. Talagrand. Concentration of measure and isoperimetric inequqlities in product spaces. *Publications mathématiques de l'I.H.É.S.*, 81:73–205, 1995.

[83] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996.

[84] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[85] A. B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Mathématiques et Applications. Springer-Verlag, 2003.

[86] S. van de Geer. M-estimation using penalties or sieves. *J. Statis. Plan. and Infer.*, 108:55–69, 2002.

[87] M. Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003.

[88] Y. Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499, 1999.

[89] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Analysis*, 74(1):135–161, 2000.

[90] Y. Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87, 2000.

[91] Y. Yang. Nonparametric Regression and prediction with Dependent Errors. *Bernoulli*, 7(4):633–655, 2001.

[92] Y. Yang. Aggregating regression procedures for a better peformance. 2003.

[93] Y. Yang. Regression with multiple candidate model: selection or mixing? *Statist. Sinica*, 13:783–809, 2003.

[94] Y. Yang. Can the strength of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92:937–950, 2005.

[95] Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

[96] Y. Yang and A. Barron. Information-Theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599, 1999.

[97] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.

Chapter 3

Cross-validation

This chapter is devoted to a presentation of the family of cross-validation algorithms, and may be split into two parts.
The first one is the state-of-the art about cross-validation. We describe cross-validation algorithms as some elements of the broader class of resampling algorithms. The main focus is given to the leave-p-out cross-validation. We then recall the cross-validation heuristics and briefly draw the comparison with the well-known bootstrap. We also review the tradeoff between bias, variance and computational complexity, which occurs with the main issue of cross-validation, that is the choice of the splitting ratio.
The second half is dedicated to the closed-form expressions we can derive for the leave-p-out based estimator of the risk in both density estimation and regression. Actually whereas this estimator is usually computationally infeasible, we are able to derive closed-form expressions with a wide range of estimators such as projection estimators, kernel estimators, and regressograms.
We conclude by showing that the leave-p-out risk estimator can be interpreted as a penalized criterion which overpenalizes. Moreover, the amount of overpenalization increases with the cardinality of the test set.

3.1 CV and resampling

3.1.1 Historical viewpoint

Substitute for the resubstitution error

From a historical viewpoint, practitioners used to assess the performance of an estimator from the same data that have been previously used to compute it. Nowadays, it is commonly admitted that such an assessment provides an overly optimistic view of the estimator in hand [29]. At the early 30s, some practitioners in the Psychology area introduced the *cross-validation* (CV) strategy as a reliable alternative to this *resubstitution error* in the assessment of the performance of a predictor [53, 46, 44].
Some preliminary formulations of the CV algorithm may be found in Hills [45] and in the seminal work of Lachenbruch and Mickey [52]. However, the first precise statement of the CV algorithm named *leave-one-out* (Section 3.2.1) (Loo) only comes with Mosteller and Tukey [61], who straightforwardly inspired the first theoretical studies of CV in a regression context [74].
Mainly, the first papers dedicated to the study of CV are actually devoted to the description of some

CV algorithms such as the Loo [74] or the *V-fold cross-validation* (VFCV) [32, 33]. Some closed-form expressions are derived and extensive simulation studies are provided.

Various purposes and frameworks

The main attractive aspects of CV are the intuitive rationale behind it as well as its ability to be applied without any restrictive assumption on the data distribution for instance. Indeed although it cannot be claimed that CV is the universally best estimation algorithm, it is usually thought that it provides reliable estimators in a wide rage of circumstances. That is the reason why it is widely used in very different frameworks.

Regression In regression, Stone [74] and Geisser [32, 33] try to accurately estimate the prediction error of some predictors in order to choose one of them. Still in the regression setting, Burman [17] addresses the problem of risk estimation, while Picard and Cook [62] pursue model selection via subset selection. Lugosi and Nobel [57] as well as Wegkamp [83] also tackle the problem of model selection in which they enlighten CV as a measure of the model complexity. Some other purposes may be reached by use of CV. For instance, we may point out the determination of the regularization parameter for SVM [10], the Lasso and the ridge regression [77]. In the context of aggregation, Breiman [16] and Yang [85] also propose a mixing strategy relying on CV-based weights.

Density Simultaneously, Rudemo [68] (for histograms and kernel estimators) and Bowman [13] (for kernel estimators only) propose similar CV heuristics and provide closed-form expressions for the resulting risk estimators. Their performances are monitored in wide simulation studies. Hall [37] shows that these heuristics lead to consistent and asymptotically optimal kernel estimates in the quadratic loss case. Stone [73] presents a similar procedure to the Loo. Although it is applied in the multivariate case under mild assumptions on marginal distributions, this procedure still exhibits asymptotic optimality properties as well. From a different viewpoint but still with kernel estimators, Hall [39] explore the influence of the interaction between the target density and the chosen kernel on the behaviour of the CV procedure based on the Kullback-Leibler loss. In front of such a large number of reliable alternatives, Sain *et al.* [69] carry out a wide comparison study in the multivariate setup, where CV is compared to bootstrap and other usual asymptotic approximations. From this study, an unexpected conclusion is drawn: CV all the more outperforms upon other competing strategies as the problem dimensionality grows. More recently in a model selection perspective, van der Laan *et al.* [80] study various types of CV technics in order to assess their accuracy in a general density estimation setting. Some essentially asymptotic results are derived about the convergence of risk estimators.

Classification In a set of two papers, Efron [28, 29] studies the classification problem and more precisely the discrepancy (*i.e. optimism*) between the true prediction error and the empirical risk. The optimism expectation is estimated through two different strategies, one based on Loo and the other one on bootstrap. After deriving some closed-form expressions for the optimism, Efron [28, 29] makes the comparison between CV and bootstrap and also to penalized criteria in some simulation experiments, noticing some evidence in favour of bootstrap and CV.

A theoretical comparison has been carried out from a general viewpoint by Blanchard and Massart [11] who derive an oracle inequality for the *hold-out* procedure (Section 3.2.1). This result emphasizes the adaptivity of CV to the classification noise condition. Indeed although CV-based approaches seem to be rather crude in comparison with the refinement of some penalties like local Rademacher complexities, it turns out to outperform upon such penalties both from some theoretical and practical aspects.

Some further comparisons between the hold-out and some penalized criteria such as SRM (Vapnik [82]) and MDL (Rissanen [66]) have been made by Kearns [47] and Kearns *et al.* [48], where the discrepancy between the generalization error of the CV-based estimate and that of the target is upper bounded. Two main conclusions are drawn about the hold-out: It may lead to a larger error than that of any tested penalized criterion, the amount of which depends on the proportion of data devoted to the test set. Nonetheless,

there are situations in which CV may outperform upon penalized criteria.
Following Kearns' papers, Bartlett *et al.* [7] extended the comparison between hold-out and penalized criteria to the data-dependent penalties. Unlike the hold-out which estimates the excess loss for each candidate classifier, they point out that data-dependent penalties intend to estimate the complexity of each class $\mathcal{F}_k$ through: $\sup_{f \in \mathcal{F}_k} L(f) - \widehat{L}(f)$. This yields more accurate estimators when the noise level is high, but rather loose ones when the noise level drops.
Another aspect of CV in the classification literature concerns the derivation of some named *worst-case bounds* for the generalization error of the Loo estimate in a given model. Such bounds are of interest since they warranty the performance of the Loo-based estimator independently from the input distribution, the target function and even from the estimation algorithm. Indeed, Devroye and Wagner [24] provide one of the first bounds of this type under some *stability* assumptions on the algorithm (Section 3.2.1). Kearns and Ron [49] weaken this stability assumption and further describe the behaviour of the Loo through the comparison with the empirical risk. In particular, they conclude that for a very wide family of algorithms on a model of finite VC dimension, the worst-case bound on the Loo generalization error is the same (or at least of the same order) as that resulting from the empirical risk minimization.

Asymptotic results

As underlined by Blanchard and Massart [11] in the classification context, there is quite a wide gap between theory about CV and its practical use. More precisely, most of the theoretical results about CV are asymptotic.
The first ones are due to Stone [75] in the regression setup with the log-likelihood loss function. His main result settles the asymptotic equivalence between Loo and the AIC criterion. Li [55] follows Stone's strategy of comparing Loo with several penalized criteria such as Mallows' C_p [58] and also the called GCV (Generalized Cross-Validation) of Craven and Wahba [21], which should be seen more as a penalized criterion than as a cross-validation technique. He also proves their asymptotic optimality through their asymptotic equivalence. However unlike Stone for whom the index set of his predictors can be $[0, 1]$, Li only deals with discrete ones.
Zhang [88] prefers focusing on the linear regression model, in which he studies the asymptotic behaviour of several *multifold* cross-validation procedures (Section 3.2.1). Some closed-form expressions as well as asymptotic expansions are derived and Zhang concludes that a multifold CV procedure named *leave-p-out* (Section 3.2.1) and the FPE_α criterion [72] are asymptotically equivalent for a proper choice of the regularization parameter.

Rather than making comparisons between criteria, Stone [73] straightforwardly proves the asymptotic optimality of the Loo in the multivariate density estimation framework via kernel estimators. The large amount of literature about the asymptotic properties of CV justifies the impressive review paper of Shao [71] in which numerous results are embedded in a more general viewpoint and thus clarified. For instance, he distinguishes the criteria according to their asymptotic efficiency or consistency properties.
From a consistency viewpoint, Yang [86] focuses on the optimal choice of the splitting ratio in CV procedures (Section 3.2.2). Unlike the previous work of Shao [70] who simply stated that in a parametric setting, this splitting ratio must converge to 1 in order to reach consistency, Yang proves a more general result which encompasses both parametric and nonparametric settings. Indeed in view of consistency, he succeeds in relating the magnitude of the splitting ratio to the convergence rate of each estimator in the finite set of candidates.

In a model selection framework, Burman [17, 18] aims at designing an accurate estimator of the risk of a given estimator. To do so, he studies some multifold CV algorithms. From the expansions of the two first moments of risk estimators, he proposes a correction, which is supposed to partially reduce the bias of the evaluated algorithms. He concludes by arguing that as an estimator of the risk, Loo and the corrected Loo provide the best estimators among CV-based risk ones.
Remark: This correction presents some similarities with the recent and more general work of Arlot [3].

3.1.2 Resampling

The resampling background

The word *resampling* means drawing one (or several) new sample(s) from the original one, hence "re-sampling". Let $Z_{1,n} := Z_1, \ldots, Z_n$ be a sample of n observations from an unknown distribution P, with empirical distribution P_n. Then, resampling from $Z_{1,n}$ consists in drawing a new sample $Z^*_{1,m}$ for $m \in \mathbb{N}^*$, such that the conditional distribution of $Z^*_{1,m}$ given $Z_{1,n}$ is known. If $m < n$, we rather speak about *subsampling*.
Among resampling algorithms, two different schemes may be distinguished:

- The *parametric resampling* is used when the unknown distribution P is assumed to be parametric, that is $P \in \{P_\theta \,|\, \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$. Thus, the initial sample yields an estimator for the optimal θ: $\widehat{\theta}$, which we use to draw new observations $Z^*_{1,m} \sim P_{\widehat{\theta}} =: Q_n$ given $Z_{1,n}$.
- The *nonparametric resampling* consists in randomly picking out new observations $Z^*_{1,m}$ among $\{Z_1, \ldots, Z_n\}$ according to a known conditional distribution given $Z_{1,n}$ denoted by Q_n.

In the sequel, we focus on the nonparametric resampling scheme.

Let us assume we now attempt to estimate the unknown distribution (or a given parameter related to it) of a statistic $L(P, Z_{1,n})$, which may be written as a function of both P and the original sample (think about the generalization error of any predictor for example). Then provided the common distribution Q_n of $Z^*_{1,n}$ is close to P (in the sense of the convergence in law), the resampling heuristics says that we can *plug* Q_n and $Z^*_{1,n}$ *in* $L(\cdot, \cdot)$ in place of respectively P and $Z_{1,n}$. Thus under some reasonable assumptions, the conditional distribution $\mathcal{L}^*_n$ of $L(Q_n, Z^*_{1,n})$ given $Z_{1,n}$ estimates that of $L(P, Z_{1,n})$, $\mathcal{L}_n$:

$$\mathcal{L}\left(L(Q_n, Z^*_{1,n}) \,|\, Z_{1,n}\right) \approx \mathcal{L}\left(L(P, Z_{1,n})\right).$$

Whereas in the *real world* we cannot access any parameter of interest due to our ignorance of P, the "resampling world" provides us with a statistic from which we are able to compute everything, therefore yielding some information about the original unknown distribution.

Besides, numerous estimators and statistics only depend on the sample data $Z_{1,n}$ through their empirical distribution $P_n = 1/n \sum_{i=1}^n \delta_{X_i}$. We may then rewrite the previous heuristics as

$$\mathcal{L}\left(L(Q_n, P^*_n) \,|\, Z_{1,n}\right) \approx \mathcal{L}\left(L(P, P_n)\right),$$

where $P^*_n := 1/n \sum_{i=1}^n \delta_{Z^*_i}$ denotes the empirical distribution of the resample. This points out that another way to define resampling strategies is through the random empirical measure P^*_n. Indeed, we may rewrite P^*_n as a weighted average of the dirac measures of the original observations

$$P^*_n = P^W_n := \frac{1}{n}\sum_{i=1}^n W_i \delta_{Z_i},$$

where the weights $W_{1,n} = W_1, \ldots, W_n$ are exchangeable random variables [76], independent from the Z_is and constrained by the choice of the resampling scheme [59, 6, 64]. This generalized view of resampling is often named *weighted bootstrap*. For instance, Efron's bootstrap [26, 28] corresponds to the choice $(W_1, \ldots, W_n)$, following a multinomial distribution $\mathcal{M}(1/n, \ldots, 1/n; n)$ (see Section 3.1.2 for more examples). Note that the weights are not always chosen to be either positive or integers. However, a common requirement is that $\mathbb{E}W_i = 1$. We now illustrate the kind of result we may expect in the exchangeable bootstrap in the mean framework [59, 35]:

Theorem 3.1.1. *Let $W_n = (W_{n,1}, \ldots, W_{n,n})$ be a vector of n exchangeable random variables independent from the sequence $Z_{1,n}$, satisfying the following conditions:*
A1. *For any n and j, $W_{n,j} \geq 0$ and $\sum_j W_{n,j} = 1$* ***A2.*** $\mathrm{Var}(W_{n,1}) = \mathcal{O}\left(n^{-2}\right)$*,*
A3. $\max_{1\leq j\leq n} \sqrt{n}\,|W_{n,j} - 1/n| \xrightarrow{P} 0$ ***A4.*** $n^{-1}\sum_{j=1}^n (W_{n,j} - 1/n)^2 \xrightarrow{P} c^2 > 0$.
Further assume that $\mathbb{E}Z_i^2 = \sigma^2 < +\infty$, then

$$\mathcal{L}\left(\sqrt{n}\left(\overline{Z}^*_n - \overline{Z}_n\right) \,|\, Z_{1,n}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, c^2\sigma^2) \qquad a.s.\ ,$$

where $\overline{Z}_n^* = 1/n \sum_{j=1}^n W_{n,j} Z_j$ *and* $\overline{Z}_n = 1/n \sum_{j=1}^n Z_j$.

A straightforward application of the CLT theorem provides the convergence in distribution of the centered empirical process towards a gaussian distribution with mean 0 and variance σ^2

$$\mathcal{L}\left(\sqrt{n}\left(\overline{Z_n} - \mathbb{E}Z_1\right)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Since the mean $\mathbb{E}Z_1$ as well as the variance σ^2 are unknown, we do not actually know the limit distribution. However, Theorem 3.1.1 asserts that we may use the limit distribution of the resampling empirical process normalized by c as an approximation to that of the original process.

Resampling botany

In a few words, CV is a resampling method consisting in (randomly or not) splitting the data into two sets [74, 33]. The first one is named the *training set* and is devoted to compute the considered estimator, while the second one is called the *test set*, which is used to assess the accuracy of the latter estimator. Writing the generalization error of a given predictor as $r = L(P, P_n)$ with the same notations as before, a CV estimator of r is obtained by replacing (P, P_n) by (P^{train}, P^{test}), where P^{train} denotes the empirical distribution of the data in the training set and P^{test}, that of the data in the test set. Since the involved resamples are actually subsamples, CV may be seen as a subsampling procedure (see Section 3.2.1 for a further more detailed presentation of CV).
Depending on the splitting scheme of the data, several CV techniques are widely used, but the cornerstone of the process remains the independence property between the test and the training set. This requirement is usually met through the usual assumption that $Z_{1,n}$ is an *i.i.d.* vector of observations. Nevertheless, this sufficient condition may be relaxed and some attempts have been done to deal with the time-series dependence type [19]. The idea is simply to remove from the training set all the data which are dependent with respect to those in the test set (see also the book of Györfi *et al.* [36]).
Another subsampling procedure is the *jackknife*, introduced by Quenouille [65] and also Tukey [79], which intends to yield bias and variance estimations for a given estimator. However, Efron [26] explains reasons why it is outperformed upon by the bootstrap, which is more widely applicable and now widespread.

Strictly speaking even if resampling already appears in a primitive form with the introduction of CV techniques, it has only emerged as a worthwhile matter of study following the work of Efron [26, 28] about bootstrap. The latter presents some similarities with the jackknife, at least in the opportunity to derive bias and variance estimates for a given estimator as well as confidence intervals. However from the comparison of both of them, Efron [26, 27] concludes that bootstrap outperforms upon jackknife. Moreover, he shows how jackknife may be understood through a linear expansion of the bootstrap estimator.
Due to some deficiencies of the bootstrap with several respects [42, 14], different variants of it have arisen, which may be understood through the choice of various weights in the resample empirical distribution.

- Rubin [67] introduces a choice of weights based on the Dirichlet distribution, named *bayesian bootstrap*.
- The *m out of n bootstrap* may be found in Giné [35]. He exhibits several situations in which classical bootstrap fails and proves that subsampling with replacement asymptotically works. Politis and Romano [63] have previously studied this methodology without replacement.
- Another failure of classical bootstrap stems from the non-*i.i.d.* case as noted by Liu [56], for instance in the heteroscedastic regression setup [84, 8]. Thus, Wu [84] suggests *the wild bootstrap*, which is a resampling scheme relying on *i.i.d.* weights $W_n = W_{n,1}, \ldots, W_{n,n}$ satisfying for any i, $\mathbb{E}W_{n,i} = 1$, $\mathbb{E}\left(W_{n,i} - 1\right)^2 = 1$ and $\mathbb{E}\left(W_{n,i} - 1\right)^3 = 1$. This strategy has been further studied by Härdle and Mammen [42] in a statistical test framework of a parametric hypothesis against a nonparametric one.

Note that the introduction of *the Edgeworth expansion* [40] turns out to be a useful tool in the study (and towards a more accurate understanding) of bootstrap procedures. Actually, bootstrap is known to automatically provide a right first order term [38], but may suffer some troubles due to worse higher order terms. To overcome this, Hall [38] proposes an “auto-correction” of the bootstrap, obtained by iteratively applying bootstrap. However, he recognizes that this procedure cannot be applied an arbitrarily large number of times due to some overfitting. In order to prevent from some troubles with the second order term, Wu [84], Liu [56] use the wild bootstrap. As for the generalized bootstrap (with various weights), Mason and Newton [59] study the first order term of such procedures with exchangeable weights, while Hall and Mammen [41] perform such an analysis in a more general framework for the second order term.

Resampling penalties

Resampling has also been used in the model selection *via* penalization, as a means to design some new data-driven penalties. Indeed, concentration inequalities [60] are usually involved in the derivation of deterministic penalties, hence the latter may be seen as upper bounds of the ideal penalty (see Section 2.2.1 and [3]). A natural question is to know how rough is this upper bound. Moreover except penalties of an asymptotic flavour, deterministic penalties are determined up to some unknown constants which have to be further fixed, for instance by an intensive simulation step [54]. These are some reasons why estimates (rather than upper bounds) of the ideal penalty have been designed, using resampling strategies. However, there is a price to pay for such estimators which is the computational cost: Drawing B bootstrap resamples for each model is more time-consuming than computing once the AIC criterion for example.

Efron [28, 29] computes the optimism, defined as the bias of the empirical risk with respect to the generalization error in a classification setting. He formulates this discrepancy in terms of covariance and hence introduces *penalties by covariance*. Bootstrap- and CV-based estimators of this bias are studied. Another aspect of penalty by covariance is explored by Daudin and Mary-Huard [22] who address the problem of variable selection in classification by introducing the swapping criterion as a data-driven complexity measure. Breiman [14] adopts a similar decomposition of the generalization error as Efron. In the subset selection framework, he makes the comparison between the classical bootstrap, AIC and what he calls *little bootstrap*.
Still in classification, Koltchinskii [50] presents and studies *Rademacher complexities*, which are data-driven penalties. Due to their good performances, they are further analyzed and compared with bootstrap penalties by Fromont [31]. In a wide review, Bartlett *et al.* [7] study several resampling-based complexity measures such as Rademacher, hold-out and maximum discrepancy penalties to name but a few.

In the general regression setup, Lugosi and Nobel [57] describe a procedure based on a half splitting of the data. This is similar in the spirit to hold-out and exhibits very good results, but it turns out to be computationally demanding. Wegkamp [83] exploits the hold-out idea to propose a computationally feasible procedure for which he derives an oracle inequality with explicit constants. Koltchinskii [51] studies and proves the accuracy of *local Rademacher penalties*, which differ from the previous ones by the fact we do not consider the maximal deviation of the process over the entire class, but only over a subset determined according to a given parameter. More recently, Arlot [3, 2] has developed a *random penalty* based strategy to perform model selection, which encompasses a large part of the previous viewpoints.

3.2 What is CV?

3.2.1 CV heuristics

Rationale in the *i.i.d.* setting

Beforehand, we introduce some notations that will be repeatedly used throughout this chapter.
Let $Z_{1,n} = Z_1, \ldots, Z_n \in \mathcal{Z}$ be a n sample drawn from an unknown distribution P, with empirical distribution denoted by P_n. For a given resampling scheme, let define $W_n = (W_{n,1}, \ldots, W_{n,n})$ a vector of random variables *independent from* $Z_{1,n}$. Furthermore, we set $Z_{1,n}^W = Z_1^W, \ldots, Z_n^W$, a resample and $P_n^W = 1/n \sum_{i=1}^n W_{n,i} Z_i$, its empirical distribution. In the sequel, $\widehat{s} = \widehat{s}(Z_{1,n})$ denotes any estimator of a parameter of interest $s \in S$, and $\gamma : \ S \times \mathcal{Z} \to \mathbb{R}$ is a contrast function. We recall that for any $t \in S$, $P\gamma(t) := \mathbb{E}\left[\, \gamma(t, Z) \,\right]$, where $Z \sim P$. Here, we aim at describing the CV heuristics from a very general viewpoint, including both density estimation and random-design regression.

CV has essentially been designed to estimate the risk of any estimator $\widehat{s}$ of s. It attempts to estimate $r = \mathbb{E}\left[\, P\gamma(\, \widehat{s}\,)\,\right]$, which may be rewritten as

$$r = \mathbb{E}_{Z_{1,n}}\left[\mathbb{E}_Z\left[\gamma\left(\, \widehat{s}\,(Z_{1,n}, Z)\,\right)\right]\right],$$

where $\mathbb{E}_Z$ and $\mathbb{E}_{Z_{1,n}}$ both represent expectations with respect to Z and $Z_{1,n}$.
REMARK: We point out that there are two levels of randomness in the above expression since Z and $Z_{1,n}$ are *independent*.
This is the cornerstone of the CV strategy, which intends to exploit these two randomness levels by splitting the data into a *training set* and a *test set*. Roughly speaking, the idea is simply to use the data in the training set to build the estimator, while the test set is devoted to the assessment of the estimator performance. Provided $Z_{1,n}$ is made of independent data, training and test sets are independent by construction as well. Independence is actually the main requirement of CV and even in the non *i.i.d.* case studied by Burman *et al.* [19], authors exploit the order of the dependence structure in removing some data from the sample so that they recover independence.

In the bootstrap idea, we would estimate the distribution of $\widehat{r} = \mathbb{E}_Z\left[\gamma\left(\, \widehat{s}\,(Z_{1,n}), Z\,\right)\right]$. For the sake of simplicity, assume $\widehat{r}$ only depends on $Z_{1,n}$ through its empirical distribution, that is $\widehat{r} = \widehat{r}\,(P_n,\, P)$. Resamples $Z_{1,n}^W$ would be drawn with exchangeable weights W_n such that $W_n \sim \mathcal{M}(1/n, \ldots, 1/n;\, n)$, so that

$$\mathcal{L}\left(\widehat{r}(P_n,\, P)\right) \approx \mathcal{L}\left(\widehat{r}(P_n^W,\, P_n) \mid Z_{1,n}\right).$$

Then, $r = \mathbb{E}_{Z_{1,n}}\widehat{r}$ is estimated as a by-product thanks to the conditional expectation of $\widehat{r}(P_n^W,\, P_n)$ given $Z_{1,n}$:

$$r \approx \mathbb{E}_W\left[\,\widehat{r}(P_n^W,\, P_n)\,\right] = \frac{1}{B}\sum_{b=1}^{B}\widehat{r}(P_n^b,\, P_n),$$

where E_W denotes the expectation with respect to the weights W_n, B is the total number of bootstrap resamples and P_n^b is the empirical distribution of the b-th resample.
REMARK: At each round of the bootstrap procedure, there is no independence between $Z_{1,n}^W$ and $Z_{1,n}$. Besides, the estimation of r results from a kind of asymptotic viewpoint, that is when B is large enough.

Unlike the bootstrap scheme, CV focuses on the independence property of $Z_{1,n}$ and Z. Let us denote by W_n a binary vector corresponding to one CV scheme and associated with observations in the training set (which will be detailed in the following section), while $\overline{W}_n$ denotes its natural counterpart representing the training set data. Then,

$$\widehat{r}\,(P_n,\, P) = \mathbb{E}_Z\left[\gamma\left(\, \widehat{s}\,(Z_{1,n}), Z\,\right)\right] \approx \widehat{r}\left(P_n^{\overline{W}},\, P_n^W\right),$$

and

$$r = \mathbb{E}_{Z_{1,n}}\left[\,\widehat{r}\,(P_n,\, P)\,\right] \approx \mathbb{E}_W\left[\,\widehat{r}\left(P_n^{\overline{W}},\, P_n^W\right)\right],$$

where the expectation $\mathbb{E}_W$ is taken with respect to the vector W_n. Note that at each step of the process, $Z_{1,n}^W$ and $Z_{1,n}^{\overline{W}}$ remain independent, unlike what happens in the bootstrap which takes P_n as fixed and resamples from it.

Description of CV methods

In model selection, another interest of resampling methods, especially of CV, is their ability to work with any estimator [86] in a wide range of frameworks [11], contrary to (deterministic) penalized criteria which require a preliminary study of this estimator to design the appropriate penalty. Indeed if we think about AIC-like penalties, there is no immediate warranty for them to be suited to any other estimator than the empirical contrast minimizer. However, the price for such a generality level is essentially the computation cost, which may be very high.
These two remarks as well as the high technicalities of the proofs all motivate the numerous variants of CV algorithms (see also [25] for a extensive review about CV):

- From a historical viewpoint, the *leave-one-out* (Loo) was the first CV scheme that appeared in a quite formalized version in Mosteller and Tukey [61], and then in [74]. It consists in successively removing each observation from the original data and computing the estimator from the $n-1$ remaining ones. The performance of the resulting estimator is then assessed thanks to the removed point. The final Loo risk estimator is defined as the average over the n possible test sets. In order to stick to the resampling formalism, Loo corresponds to the choice of a random vector W_n, such that $W_{n,j} \in \{0, n\}$, $\mathbb{P}(W_{n,j} > 0) = 1/n$ for any j, and $\sum_{j=1}^{n} W_{n,j}/n = 1$. The Loo risk estimator is expressed as

$$\widehat{R}_1(A) = \frac{1}{n}\sum_{i=1}^{n} \gamma(A(Z_{1,n}^{(i)}), Z_i),$$

where $Z_{1,n}^{(i)}$ represents $Z_{1,n}$ from which Z_i has been removed and A denotes an estimation algorithm, that is an application that takes as input some data and outputs an estimator. In a nutshell, $A(Z_{1,n})$ is the estimator provided by algorithm A, computed from $Z_{1,n}$.

- The *leave-p-out* (Lpo), with $p \in \{1, \ldots, n-1\}$, may be seen as a generalization of the Loo to which it amounts when $p = 1$. It appears in a general framework in Burman [17], and in a linear regression setup in [70, 87]. In the density estimation setting, Celisse and Robin [20] derive a closed-form expression for the Lpo estimator in the estimation of the risk of histograms. It consists in the same procedure as that of the Loo, except that at each of the $\binom{n}{p}$ rounds we remove p observations (instead of only one). The corresponding weights are defined by $W_{n,i} \in \{0, n/p\}$ for any i, $\sum_{i=1}^{n} p/nW_{n,i} = p$ and the probability of any such vector is $\binom{n}{p}^{-1}$. Thus with the same notations as before, the Lpo risk estimator (also named Lpo estimator or Lpo risk) is finally

$$\widehat{R}_p(A) = \binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_p} \left[\frac{1}{p} \sum_{i \in e} \gamma\left(A(Z_{1,n}^{\overline{e}}), Z_i\right) \right],$$

where $\mathcal{E}_p = \{(i_1, \ldots, i_p) \,|\, i_1 < \ldots < i_p,\ i_j \in \{1, \ldots, n\}\}$, $e \in \mathcal{E}_p$ and $\overline{e} = \{1, \ldots, n\} \setminus e$.

- Due to the high computational burden of the previous procedures, Geisser [32, 33] introduces an alternative algorithm named *V-fold cross-validation* (VFCV). The VFCV has been studied in Burman [17, 18] who suggests a correction in order to remove some bias. It relies on a preliminary random (or not) choice of a partition of the data in V subsets of approximately equal size n/V. Each subset is successively left out, and the $V-1$ remaining ones are used to compute the estimator, while the last one is dedicated to its performance assessment. The V-fold risk estimator is the average of the V resulting estimators. For a given random partition of the data, the above description results in V weight vectors W_n of respective probability $1/V$, satisfying $W_{n,i} \in \{0, V\}$ for any i, and $\sum_{i=1}^{n} W_{n,i} = n/V$. This leads us to the following VFCV estimator:

$$\widehat{R}_{\text{VFCV},V}(A) = \frac{1}{V}\sum_{v=1}^{V} \left[\frac{V}{n} \sum_{i \in e_v} \gamma\left(A(Z_{1,n}^{\overline{e_v}}), Z_i\right) \right],$$

where e_v denotes the n/V indices of the V-th subset.

- *The Hold(-p)-out* (Hpo), with $p \in \{1, \ldots, n-1\}$, is one of the simplest (to analyze) CV procedures, consisting in randomly partitioning the data in a training and a test sets. But unlike the preceding procedures, the estimator computation and its assessment are only performed once. Since there is no averaging on several resamples, this simple procedure has been often studied (see [7, 11] in classification, [57, 83] in regression). For a randomly chosen $e \in \mathcal{E}_p$, its simple expression is

$$\widehat{R}_{\text{Hpo},p}(A) = \frac{1}{p}\sum_{i \in e} \gamma\left(A(Z_{1,n}^{\overline{e}}), Z_i\right).$$

Which one to use?

Nowadays, the always increasing amount of data results in very large sample sizes ($n \gg 1$). Since the Loo [74] requires the computation of one estimator for each successively removed observation, it may be too computationally demanding. To overcome this problem, Geisser [33] proposed the VFCV algorithm, which only requires the computation of V estimators (as many as we have subsets of data). Thus provided $V \ll n$, it is less expensive to use VFCV than Loo. However, the latter relies on a preliminary random partitioning of the data in V subsets. This additional randomness may induce some unwanted variability [20]. A similar remark applies to Hpo, since the common intuition about it is that choosing only a subset of the data may be misleading if unfortunately these data are not fully representative of the underlying phenomenon.

Keeping this additional randomness issue in mind, Lpo [70] may appear as the "gold standard". Indeed, it does not introduce any additional variability, since all the $\binom{n}{p}$ resamples are taken into account. To go further, VFCV may be understood as an approximation of the "ideal" Lpo up to some fluctuations due to the additional randomness the former introduces. Note that other attempts to approximate the Lpo have been proposed such that the repeated learning-testing method [17] for instance.
Nevertheless, the price to pay for such an "optimality" is once more the computational issue. The Lpo computation requires to explore the $\binom{n}{p}$ resamples, which is intractable even for relatively small n and p. In some specific settings, closed-form expressions may be derived for the Lpo estimator [20].
Throughout this manuscript, we will focus on the study of the Lpo. In fact, we think that this resampling scheme may provide some more insight in the behaviour of CV techniques for which, some more work may be done towards a deeper understanding.

Stability and CV

For an algorithm, the notion of *stability* characterizes the way the algorithm is "sensitive" to any change in the input data. Intuitively, if a slight perturbation in the data induces a strong "variation" of the result of the algorithm, the latter is said to be unstable. In order to explain what a "variation" may be, we may think about classification. For a given algorithm, the resulting classifier may be different if we only remove one observation from the initial sample, which is a first instance of variation. If we rather consider the generalization error associated to such an algorithm. The same change in the data is likely to induce a different classifier, but the resulting error may remain the same as that of the previous one. Anyway, another type of possible variation is a change in the generalization error.

The notion of stability is studied by Breiman [15] in the regression setup when describing the bagging strategy. For a training set $\mathcal{T}$ and a predictor A such that $A(X, \mathcal{T})$ is the prediction of an instance X from the training set $\mathcal{T}$, Breiman quantifies the instability as the discrepancy between $\left(\mathbb{E}_{\mathcal{T}}\left[\, A(X,\mathcal{T})\,\right]\right)^2$ and $\mathbb{E}_{\mathcal{T}}\left[\, A(X,\mathcal{T})\,\right]^2$, where $\mathbb{E}_{\mathcal{T}}$ denotes the expectation with respect to the training set. In other words, the amount of instability of A is given by the following mean variance of $A(X, \mathcal{T})$:

$$\mathbb{E}_{(X,Y)}\left(\mathrm{Var}_{\mathcal{T}}\left[\, A(X,\mathcal{T})\,\right]\right),$$

where $\mathbb{E}_{(X,Y)}$ is the expectation with respect to (X, Y) and $\mathrm{Var}_{\mathcal{T}}$ denotes the variance with respect to $\mathcal{T}$.

The different levels at which we may check the stability of an algorithm (change of classifier or in the generalization error and so on...) have led to many definitions of stability, lots of them being given in the classification context. For instance, Kearns and Ron [49] distinguish between changes in the classification rule and changes in the error of this rule to formulate the subsequent definitions of stability. From this, they derive what they call "sanity-check bounds" for the Loo error, highlighting that some bad performance of Loo, in terms of a large gap with respect to the true error, may be accounted for by a lack of stability of the algorithm in hand. In turn, Bousquet and Elisseff [12] extend this type of analysis to the regression framework. They detail several notions of stability and compare bounds for the empirical risk to bounds for Loo. From the closeness of these bounds, they conclude that the stability approach essentially concerns stable algorithms for which these two error measures may be close to one another. The relationship between Loo and stability has been further explored by Evgeniou *et al.* [30] who consider several classification algorithms, which they combine. The Loo error of each algorithm as well as that of the mixing procedure are compared. Thus, the stability concept enables them to explain the better behaviour of the Loo when applied to the mixing strategy, which aims at reducing instability.

3.2.2 How to split the data?

Bias-Variance-Computation tradeoff

An appreciable property we may require for our CV estimator is that it reaches the best *bias-variance tradeoff*, as an estimator of the risk.
Remark: We do not speak about the bias and the variance of the estimator the risk of which is estimated by CV, but rather of the risk estimator itself.

It turns out that both the bias and the variance of the CV estimator of the risk are related to the size p of the test set (see [43] and Section 3.4.1).
In a regression context, Burman [17] expands bias and variance of the VFCV estimator as functions of the empirical distribution of the data. He provides first and second order terms of these functions, thus proving their strong dependence on both p and the sample size n. In a density estimation setup, explicit formulas for the bias and variance of the Lpo estimator may be found in [20], which highlight the dependence on p. Thus for instance, removing or adding some data from or to the training set will influence the bias-variance tradeoff achieved by the corresponding CV estimator.

The common intuition about that may be described by the following two extreme situations [43]. A small training set yields poorly fitted (and subsequently biased) estimators, since they have been designed from much less data. However, the corresponding test set is large and the resulting accuracy assessment is very sharp. Conversely, a large training set provides nearly unbiased estimators (see Bousquet and Elisseff [12] in classification), but the small test set only yields a rough performance evaluation, with possibly large variance. Similar considerations may be found in Yang [86] embedded in a selection strategy.
Remark: The above reasoning should not be taken for granted as enlighten by Burman's asymptotic result [17]. Thus, the first order terms of his expansions indicate that both the bias and the variance decrease when p goes down, which is supported by his simulation results. The conclusion we may draw from this point is that such a general result does not exist and that the behaviour of the CV estimate as a function of p will be setup dependent.

On top of the previous bias-variance tradeoff, we should add a common issue to all CV algorithms: the computational burden. Indeed, the computational feasibility must be taken into account in the choice of p. Burman [17, 18] mentions this point to argue in favour of VFCV since the "optimal" Loo sometimes turns out to be computationally intractable. A reasonable strategy could therefore consist in using the VFCV estimator with V as large as we can afford. Moreover, since small Vs induce a bias, Burman propose a correction of it by adding a term preventing from too strong deviations. An appropriate choice of p is very intricate due to the coupling of several often contradictory aspects: bias, variance and computation time. Note that in a recent work, Arlot [3] developed a general strategy based on resampling penalties. His idea is to decouple the aforementioned quantities in order to facilitate the choice of p [2].

Asymptotic results about splitting

At the core of any CV algorithm (Lpo, Hpo or VFCV) is the choice of p, the size of the test set. There is already a large amount of theoretical results on the matter, most of them being of asymptotic nature. Among all of them, two broad classes may be distinguished according to the adopted viewpoint (estimation or selection).

Indeed, a large number of results concern the comparison between CV procedures and penalized criteria, known to reach asymptotic efficiency. Thus, Stone [75] establishes the asymptotic equivalence between Loo and AIC in a parametric regression setting. Li [55] considers a linear regression setup and establishes the same result for the Loo and the C_p, while in the same setting Zhang [87] tackles the (asymptotic) equivalence between the Lpo and the FPE criterion [1] and more precisely describes the appropriate choice of p so that Lpo is asymptotically equivalent to FPE_α. Optimality properties or at least a better understanding of the Lpo may be deduced from these comparisons. Since AIC or Mallows' C_p pursue (and enjoy) asymptotic efficiency, we may expect that the Loo, and to some extent the Lpo for small ps, intend to do the same. Some direct studies of CV estimators provide the same kind of information as in Stone [73] in the density estimation context.

Through an appropriate choice of p, Lpo may also be used in an attempt to recover the "true" model, that is in a consistency viewpoint. In the linear regression context, Shao [70] establishes that the Lpo estimator reaches consistency if and only if we require that $p/n \to 1$, which means that asymptotically all the data should be placed in the test set.
Due to the wide range of settings in which all these results are derived, Shao [71] has intended to somewhat clarify the landscape on the matter. Besides, interested in recovering the best estimator among a finite family of competitors, Yang [86] gives a more qualified statement than his predecessors, relating the optimal choice of p to some key properties of the estimators in hand like their convergence rates.

Risk estimation or model selection

If our goal is the risk estimation of a given estimator, then taking into account the bias and the variance (with computational considerations) alone turn out to be sufficient to provide a reliable risk estimator, provided a good estimator is that one reaching the best bias-variance tradeoff. Burman [17] furnishes some simulation experiments to argue in this direction for instance.

Nonetheless if the actual purpose is model selection (and not only risk estimation), then it may be necessary to take into account the structure of the model collection, in order to prevent from overfitting. If we think about the closeness of the Lpo procedure (with an appropriate choice of p) to some well-known penalized criteria, this is likely to be true. For instance, Stone [75] proves the asymptotic equivalence between Loo and AIC. Since we know some situations in which the use of AIC is misleading (complete subset selection for instance [14]), it seems reliable that Loo may suffer similar troubles.
Furthermore, Zhang [87] explicitly relates Lpo for a suitable choice of p to the FPE criterion in the following way. For a given model m of dimension D_m, we recall that the FPE criterion is defined as

$$FPE(m) = RSS(m) + \alpha \widehat{\sigma}^2 D_m,$$

where $\alpha \in \mathbb{R}_+^*$, $RSS(m)$ denotes the sum of squares for model m and $\widehat{\sigma}^2$ is any consistent estimator of σ^2 for instance. Then, Zhang explains that asymptotic equivalence between Lpo and FPE applies provided

$$\alpha = 1 + \frac{1}{1-\lambda},$$

with $p/n = \lambda + o(1)$. Thus, if p is kept independent from n, $\lambda = 0$ and $\alpha = 2$, which is quite similar to the C_p. Furthermore since α is an increasing function of λ, the regularization parameter is larger than 2 and increases as $\lambda > 0$ grows towards 1, which is synonymous with overpenalization (Section 2.2.1).

Obviously, the preceding heuristics is only asymptotic and intends to provide more insight on the problem. Actually, most of the existing results about the optimal determination of p are asymptotic [71], whereas model selection *via* penalization from the estimation viewpoint as well as applications require finite-sample results.

3.3 Closed-form expressions

3.3.1 Preliminary calculations

Purpose and strategy

In the present section, we aim at deriving some closed-form expressions for Lpo estimators in the density estimation and the regression contexts.
In density estimation, we observe n realizations of random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, where s denotes

the unknown density of P with respect to the Lebesgue measure on $\mathcal{X}$. In what follows, we will alternatively address the problem of density estimation on $\mathcal{X} = [0,1]$ when we use general projection estimators, or $\mathcal{X} = \mathbb{R}$ with kernel estimators.
As for the regression problem, we assume that we are given a set of *independent* observations $Z_1, \ldots, Z_n$ where for any i, $Z_i = (X_i, Y_i) \in [0,1] \times \mathcal{Y}$ satisfies

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i, \quad \mathbb{E}\left[\epsilon_i \mid X_i\right] = 0, \quad \mathbb{E}\left[\epsilon_i^2 \mid X_i\right] = 1,$$

where $s \in L^2([0,1])$ is the unknown regression function and $\sigma : \ [0,1] \to \mathbb{R}_+$. For each i, $X_i \sim \mu_i$ so that the X_is are *not necessarily identically distributed*. We have in mind the fixed-design regression case for instance [5].

Our purpose is to derive closed-form formulas with as few restrictions as possible on the loss function as well as on the type of estimators we will consider. Actually, we have in mind two broad collections of estimators which are projection estimators and kernel-based estimators.

Projection estimators and empirical risk minimizer

In the following, we describe the relationship between projection estimators and empirical risk minimizer, with respect to the statistical framework.
Let us define $\{\varphi_\lambda\}_{\lambda \in \Lambda_n}$ a family of functions in $L^2([0,1], \nu)$, where ν denotes the Lebesgue measure on $[0,1]$ and Λ_n is a countable set of indices. For any $m \in \mathcal{M}_n$, set $\Lambda(m) \subset \Lambda_n$ such that $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$ is an orthonormal family of functions. Let S_m denote the linear space of dimension D_m spanned by $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$.
We call $\widehat{s}$ a *projection estimator* any estimator of s such that

$$\widehat{s}_m = \sum_{\lambda \in \Lambda(m)} \widehat{\beta}_\lambda \varphi_\lambda \qquad \text{with} \qquad \widehat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n \theta_\lambda(Z_i),$$

where $Z_1, \ldots, Z_n$ denote some observations (density or regression) and θ_λ is a function independent from the observations for any λ [78].
In the density estimation framework, $\theta_\lambda = \varphi_\lambda$ for every λ and

$$\widehat{s}_m = \sum_{\lambda \in \Lambda(m)} \widehat{\beta}_\lambda \varphi_\lambda \qquad \text{with} \qquad \widehat{\beta}_\lambda = P_n \varphi_\lambda := \frac{1}{n} \sum_{i=1}^n \varphi_\lambda(Z_i). \tag{3.1}$$

A typical example of estimator is the histogram for which $\varphi_\lambda = \mathbb{1}_{I_\lambda} / \sqrt{|I_\lambda|}$, where $\{I_\lambda\}_{\lambda \in \Lambda(m)}$ denotes a partition of $[0,1]$ and $|I_\lambda|$ represents the length of the interval I_λ.
In the regression setting, $\theta_\lambda\left[(x,y)\right] = y\varphi_\lambda(x)$ and

$$\widehat{s}_m = \sum_{\lambda \in \Lambda(m)} \widehat{\beta}_\lambda \varphi_\lambda \qquad \text{with} \qquad \widehat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n Y_i \varphi_\lambda(X_i).$$

Remark: The latter estimator is essentially used for uniformly distributed X_is, due to the approximation it results from.

These two estimators belong to a broader class of estimators which may be expressed as

$$\widetilde{s}_m = \sum_{\lambda \in \Lambda(m)} \widetilde{\beta}_\lambda \varphi_\lambda,$$

where $\widetilde{\beta}_\lambda$ is to be chosen. At this step, we have to distinguish the density setup from the regression one. Indeed in density estimation, looking for the $\widetilde{\beta}_\lambda$s which minimize the empirical risk exactly results in the same estimator as in (3.1): projection estimator and empirical risk minimizer (ERM) coincide.
On the contrary, this is not as such in the regression setup since there, we have

$$\widetilde{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n Y_i \frac{\varphi_\lambda(X_i)}{1/n \sum_{j=1}^n \varphi_\lambda^2(X_j)}.$$

We conclude that the ERM deserve a particular treatment.
REMARKS:

- Projection estimators are mainly used in fixed-design regression with deterministic $X_i = i/n$ and some further explanations with can be found in Tsybakov [78]. In the signal processing area, we refer to Genovese and Wasserman [34] for an approach based on wavelets.
- Regressograms are an example of ERM. They have been extensively used in change-points detection for instance [54, 3, 4].
- ERM and projection estimator coincide for the particular choice of the trigonometric basis with a fixed design i/n [78].

In conclusion, two cases must be distinguished in the sequel. We first derive closed-form expressions for projection estimators both in the density estimation and the regression frameworks. Then, we address the same question with the ERM in the regression context.

Preliminaries

From the above comments, we see that two different situations have to be distinguished.
It is easy to check that every projection estimators $\widehat{s}_m$ may be written as

$$\forall 1 \leq i \leq n, \qquad \widehat{s}_m(Z_i) = \frac{1}{n}\sum_{j=1}^{n} H_m(Z_j, Z_i), \tag{3.2}$$

where $H_m(\cdot,\,\cdot)$ is a function which may be expressed in terms of the basis vectors. Furthermore, note that this expression is also strongly connected to kernel density estimators as well. Indeed, provided K denotes a kernel and $h > 0$ a smoothing parameter, we may define

$$\forall 1 \leq i \leq n, \qquad \widehat{s}_h(X_i) = \frac{1}{n}\sum_{j=1}^{n} K_h\left(X_j - X_i\right).$$

This general shape will be useful in the sequel to carry out our calculations.

In comparison, we observe that we cannot hope for any formula like (3.2) with the ERM or kernel-based estimators in regression. The only one we can expect is respectively

$$\forall 1 \leq i \leq n, \qquad \widehat{s}_m(Z_i) = \frac{1}{n}\sum_{j=1}^{n} \frac{H_m(Z_j, Z_i)}{G_m(Z_{1,n};\, Z_i)} \quad \text{or} \quad \widehat{s}_m(Z_i) = \frac{1}{n}\sum_{j=1}^{n}\sum_{\lambda\in\Lambda(m)} \frac{H_m^{\lambda}(Z_j, Z_i)}{G_m^{\lambda}(Z_{1,n};\, Z_i)}, \tag{3.3}$$

where we point out the dependence of the denominator on the whole sample used to build the estimator.

An instructive calculation

As a justification of the following, we point out that most of the loss functions that are usually used may be expanded through their Taylor expansions. Thus for the sake of simplicity, we will carry out a preliminary calculation with a monomial loss of degree $k \in \mathbb{N}^*$, which is sufficient to check the minimal requirements we need to reach our goal.
In the density setting, this loss is

$$\ell\,(s, \widehat{s}\,) = \|s - \widehat{s}\,\|_k^k = \int_{[0,1]} (s(t) - \widehat{s}\,(t))^k \, dt,$$

while the corresponding contrast in regression is $\gamma\,[\,\widehat{s}\,, (X_i, Y_i)\,] = (Y_i - \widehat{s}\,(X_i))^k$. We recall the general formula of the Lpo estimator:

$$\widehat{R}_p(m) = \binom{n}{p}^{-1} \sum_{e\in\mathcal{E}_p} \frac{1}{p}\sum_{i\in e} \gamma\left[\,\widehat{s}_m(Z_{1,n}^{\overline{e}}), Z_i\,\right].$$

Whatever the context, we must be able to compute $\sum_{e\in\mathcal{E}_p} \left[\, \widehat{s}_m(Z_{1,n}^{\overline{e}};\, t)\,\right]^j$ for any $t \in [0,1]$, with $1 \leq j \leq k$. The general expression (3.2) leads us to

$$\sum_{e\in\mathcal{E}_p} \left[\, \widehat{s}_m(Z_{1,n}^{\overline{e}};\, t)\,\right]^j = \frac{1}{n^j} \sum_{1\leq r_1,\ldots,r_j\leq n} \prod_{\ell=1}^{j} H_m(Z_{r_l}, t) \left(\sum_{e\in\mathcal{E}_p} \prod_{s=1}^{j} \mathbb{1}_{\{r_s\in\overline{e}\}} \right). \tag{3.4}$$

Remarks:

- Since the expression in brackets can be computed by hand, any potential restriction comes from another part of the expression.
- The summation in brackets is made over less and less terms as j increases. This term even vanishes when j is larger than the cardinality of the training set $\overline{e}$.
- A tradeoff raises between the size of the exponent j and what is computable by hand or at least reasonably computable by a computer, that is $\sum_{1\leq r_1,\ldots,r_j\leq n} \prod_{\ell=1}^{j} H_m(Z_{r_l}, t)$.
- Other losses may be considered, but only through their approximations by their first expansion terms, until a "reasonable" order. But this is out of the scope of the present manuscript.

Thus except if we have an efficient way to compute the latter expression, otherwise we will content ourselves with $k = 2$, that is with the quadratic loss function.

Further specifications

Expression (3.3) suggests that both ERM and kernel-based estimators in regression will not behave like projection estimators. We therefore give further details about the derivation of closed-from expressions for them.
From (3.3), we notice that the denominator also depends on $Z_{1,n}^{\overline{e}}$ (points in the training set), which prevents us from simply inverting the sums. We then obtain

$$\sum_{e\in\mathcal{E}_p} \left[\, \widehat{s}_m(Z_{1,n}^{\overline{e}};\, t)\,\right]^j = \frac{1}{n^j} \sum_{1\leq r_1,\ldots,r_j\leq n} \prod_{\ell=1}^{j} H_m(Z_{r_l}, t) \left(\sum_{e\in\mathcal{E}_p} \frac{\prod_{s=1}^{j} \mathbb{1}_{\{r_s\in\overline{e}\}}}{G_m\left(Z_{1,n}^{\overline{e}}, t\right)} \right), \tag{3.5}$$

which suggests we need to further specify the way $G_m(Z_{1,n}^{\overline{e}}, t)$ depends on $Z_{1,n}^{\overline{e}}$ since the expression in brackets cannot be straightforwardly computed by hand anymore, unlike (3.4).
Either $G_m(Z_{1,n}^{\overline{e}}, t)$ is a function of the number of observations that enter into the calculation of $G_m(Z_{1,n}^{\overline{e}}, t)$ for a given t (which happens when we use some piecewise constant estimators for instance), or $G_m(Z_{1,n}^{\overline{e}}, t)$ is a more complex function of these observations. In the latter case, we cannot compute any closed-form formula since the algorithmic complexity is almost the same as that of an exhaustive exploration of the $\binom{n}{p}$ resamples.
On the other hand provided the former situation holds, we are able to compute the quantity in brackets in (3.5). For instance, this occurs when we use regressograms and also with the specific kernel satisfying $K_h(Z_i - Z_j) = \mathbb{1}_{(|Z_i - Z_j|\leq h)}$ for $h > 0$.

3.3.2 Closed-form Lpo estimators

Due to the specificity of regressogram and kernel estimator in regression, these deserve a separate treatment. In the next section, we first deal with the general *projection estimators in both density estimation and regression* as well as with *kernel density estimators*, which all share the same expression as (3.2).

Projection estimators and kernel density estimators

Before providing formulas both in density estimation and regression, we need the following lemma which provides the key quantities.

Lemma 3.3.1. *Let $\widehat{s}_m(Z_{1,n}^{\overline{e}})$ be a generic projection estimator based on model S_m and computed from the training data $Z_{1,n}^{\overline{e}}$. Then,*

$$\sum_{e\in\mathcal{E}_p} \left\|\widehat{s}_m(Z_{1,n}^{\overline{e}})\right\|_2^2 = \frac{1}{(n-p)^2}\left[\binom{n-1}{p}\sum_{k=1}^n \|H_m(Z_k,\cdot)\|_2^2 + \binom{n-2}{p}\sum_{k\neq\ell}\langle H_m(Z_k,\cdot), H_m(Z_\ell,\cdot)\rangle_2\right], \tag{3.6}$$

$$\sum_{e\in\mathcal{E}_p}\sum_{i\in e} \widehat{s}(X_{1,n}^{\overline{e}})(X_i) = \frac{1}{n-p}\binom{n-2}{p-1}\sum_{i\neq j} H_m(Z_j,Z_i), \tag{3.7}$$

$$\sum_{e\in\mathcal{E}_p}\sum_{i\in e}\left[\,\widehat{s}_m(Z_{1,n}^{\overline{e}})(Z_i)\right]^2 = \frac{1}{(n-p)^2}\left[\binom{n-2}{p-1}\sum_{i\neq k}\left[\,H_m(Z_k,Z_i)\right]^2 + \binom{n-3}{p-1}\sum_{i\neq k\neq \ell} H_m(Z_k,Z_i)H_m(Z_\ell,Z_i)\right], \tag{3.8}$$

$$\sum_{e\in\mathcal{E}_p}\sum_{i\in e} Y_i\widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i) = \frac{1}{n-p}\binom{n-2}{p-1}\sum_{i\neq j} Y_i H_m(Z_j,Z_i), \tag{3.9}$$

where $\langle\cdot,\cdot\rangle_2$ denotes the inner product in $L^2(\mathcal{X})$.

The proof of Lemma 3.3.1 is deferred to Section 3.5.

Density estimation From the preceding result, we deduce the general expressions for projection and kernel estimators in density estimation.

Proposition 3.3.1. *For any density $t:\ [0,1]\to\mathbb{R}_+$, the contrast function associated with the L^2-loss is defined by $\gamma(t,X) = \|t\|^2 - 2t(X)$.*
Let $\widehat{s}_m$ denote the projection estimator onto the model S_m. Then,

$$H_m(X_i,X_j) = \sum_{\lambda\in\Lambda(m)} \varphi_\lambda(X_j)\varphi_\lambda(X_i)$$

and for any $p\in\{1,\ldots,n-1\}$,

$$\widehat{R}_p(m) = \frac{1}{n(n-p)}\sum_{\lambda\in\Lambda(m)}\left[\sum_j \varphi_\lambda^2(X_j) - \frac{n-p+1}{n-1}\sum_{j\neq k}\varphi_\lambda(X_j)\varphi_\lambda(X_k)\right]. \tag{3.10}$$

Besides, let $\widehat{s}_m$ denote the kernel density estimator based on a symmetric kernel K, with smoothing parameter $m>0$. Then,

$$H_m(X_j,X_i) = K_m(X_j-X_i) := K\left[\,(X_j-X_i)/m\,\right].$$

Moreover for any $p\in\{1,\ldots,n-1\}$,

$$\widehat{R}_p(m) = \frac{1}{n-p}\|K_m\|^2 + \frac{(n-p-1)}{n(n-1)(n-p)}\sum_{k\neq\ell} K_m^*(X_k-X_\ell) - \frac{2}{n(n-1)}\sum_{k\neq\ell} K_m(X_k,X_\ell), \tag{3.11}$$

where $K_m^* := (K\star K)_m$ and $\star$ denotes the convolution product.

Note that the computation of (3.10) and (3.11) is not more expensive than the usual computation of the Gram matrix with the usual kernel estimators, that is $(K_h(X_i,X_j))_{1\leq i,j\leq n}$.

Proof. In the density estimation framework, the contrast associated with the L^2-loss is $\gamma(t,X) = \|t\|^2 - 2t(X)$. Subsequently, the Lpo estimator is

$$\widehat{R}_p(m) = \binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\left\|\widehat{s}_m(Z_{1,n}^{\overline{e}})\right\|^2 - \frac{2}{p}\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\sum_{i\in e}\widehat{s}_m(Z_{1,n}^{\overline{e}})(Z_i).$$

Besides, the general projection estimator is

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda\, \varphi_\lambda,$$

which provides $H_m(Z_j, Z_i) = \sum_{\lambda\in m} \varphi_\lambda(Z_j)\,\varphi_\lambda(Z_i)$. The simple application of (3.6) and (3.7) provides the expected conclusion.
As for the kernel estimator, the result is straightforward since $H_m(Z_j, Z_i) = K_m(Z_j - Z_i)$. □

We are now in position to apply Proposition 3.3.1 to peculiar examples such as histograms and kernels.

Corollary 3.3.1. *With the same notations as before, assume that $\widehat{s}_m$ denotes the histogram estimator built from the partition $I(m) = (I_1, \ldots, I_{D_m})$ of $[0,1]$ in D_m intervals of length $\omega_\lambda = |I_\lambda|$. Then for $p \in \{1, \ldots, n-1\}$,*

$$\widehat{R}_p(m) = \frac{1}{(n-1)(n-p)} \sum_{\lambda=1}^{D_m} \frac{1}{\omega_\lambda}\left[(2n-p)\,\frac{n_\lambda}{n} - n(n-p+1)\left(\frac{n_\lambda}{n}\right)^2\right], \tag{3.12}$$

where $n_\lambda = \sharp\{i \mid X_i \in I_\lambda\}$.
Assume now that K is the gaussian kernel. Then for any smoothing parameter $h > 0$,

$$\widehat{R}_p(h) = \frac{1}{2\sqrt{\pi}(n-p)h}\left[1 + \frac{n-p-1}{n(n-1)}\sum_{i\neq j} e^{-\frac{1}{4}\left(\frac{X_i-X_j}{h}\right)^2}\right] - \sqrt{\frac{2}{\pi}}\frac{1}{n(n-1)h}\sum_{i\neq j} e^{-\frac{1}{2}\left(\frac{X_i-X_j}{h}\right)^2} \tag{3.13}$$

Proof. (3.12) comes simply from the application of (3.10) with $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{\omega_\lambda}$, while (3.13) results from (3.11) with $K:\ x \mapsto 1/\sqrt{2\pi}\exp\left(-1/2x^2\right), \quad \forall x \in \mathbb{R}$. □

Regression The case of projection estimator in regression is handled in the same way, but with a different contrast function. Now, we give the Proposition 3.3.1 counterpart.

Proposition 3.3.2. *For any observations $Z = (X, Y)$ and any function $t:\ [0,1] \to \mathbb{R}$, the contrast function associated with the L^2-loss is defined by $\gamma(t, Z) = (Y - t(X))^2$.*
Let $\widehat{s}_m$ denote the projection estimator onto the model S_m. Then,

$$H_m(Z_j, Z_i) = \sum_{\lambda\in\Lambda(m)} Y_j\varphi_\lambda(X_j)\varphi_\lambda(X_i).$$

Moreover for any $p \in \{1, \ldots, n-1\}$,

$$\widehat{R}_p(m) = \frac{1}{n(n-1)}\left[\frac{1}{n-p}\sum_{i\neq j} H_m^2(Z_j, Z_i) + \frac{n-p-1}{(n-p)(n-2)}\sum_{i\neq j\neq k} H_m(Z_j, Z_i)H_m(Z_k, Z_i) - 2\sum_{i\neq j} Y_i H_m(X_j, X_i)\right] + \frac{1}{n}\sum_{i=1}^n Y_i^2. \tag{3.14}$$

Furthermore,

$$\sum_{i\neq k}(H_m(Z_k, Z_i))^2 = \sum_\lambda [\,S_\lambda(0,2)S_\lambda(2,2) - S_\lambda(2,4)\,] + \sum_{\lambda\neq\lambda'}[\,S_{\lambda,\lambda'}(0,1,1)S_\lambda(2,1,1) - S_{\lambda,\lambda'}(2,2,2)\,], \tag{3.15}$$

$$\begin{aligned}\sum_{i\neq k\neq \ell} H_m(Z_k, Z_i)H_m(Z_\ell, Z_i) = {} & \sum_\lambda [S_\lambda(0,2)S_\lambda^2(1,1) - 2S_\lambda(1,3)S_\lambda(1,1) - S_\lambda(0,2)S_\lambda(1,1) + S_\lambda(1,3) + \\ & S_\lambda(2,4)] + \\ & \sum_{\lambda\neq\lambda'} [S_{\lambda,\lambda'}(1,0,1)S_{\lambda,\lambda'}(1,1,0)S_{\lambda,\lambda'}(0,1,1) - S_{\lambda,\lambda'}(1,0,1)S_{\lambda,\lambda'}(1,2,1) - S_{\lambda,\lambda'}(0,1,1)S_{\lambda,\lambda'}(2,1,1) + \\ & 2S_{\lambda,\lambda'}(2,2,2) - S_{\lambda,\lambda'}(1,1,2)S_{\lambda,\lambda'}(1,1,0)],\end{aligned} \tag{3.16}$$

$$\sum_{i\neq j} Y_i H_m(Z_j, Z_i) = \sum_\lambda [\,S_\lambda^2(1,1) - S_\lambda(2,2)\,], \tag{3.17}$$

where $S_\lambda(a,b) = \sum_{i=1}^n Y_i^a \varphi_\lambda^b(X_i)$ and $S_{\lambda,\lambda'}(a,b,c) = \sum_{i=1}^n Y_i^a\varphi_\lambda^b(X_i)\varphi_{\lambda'}^c(X_i)$.

Proof. The contrast function is now $\gamma(t,Z) = [Y - t(X)]^2$. The general Lpo risk estimator may be expressed as the sum of three terms

$$\begin{aligned}\widehat{R}_p(m) &= \left[\frac{1}{p}\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\sum_{i\in e} Y_i^2\right] + \left[\frac{1}{p}\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\sum_{i\in e}\left[\widehat{s}_m(Z_{1,n}^{\overline{e}})(Z_i)\right]^2\right] - \\ &\quad 2\left[\frac{1}{p}\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\sum_{i\in e} Y_i\widehat{s}_m(Z_{1,n}^{\overline{e}})(Z_i)\right].\end{aligned}$$

The first term is dealt with thanks to Lemma 3.5.1 in Section 3.5. For the two last ones, we apply respectively (3.8) and (3.9) to conclude. □

Regressograms and kernel-based estimators in regression

From (3.3), we have seen that the denominator dependence on the sample is a crucial issue for us to derive closed-form formulas. If this dependence is too complex, we cannot provide any effective calculation of (3.3). In the sequel, we first justify the shape of expression (3.3) in the case of both regressograms and general kernels. Then, we illustrate the previous assertion with a kernel estimator.

Let $I(m) = (I_1, \ldots, I_{D_m})$ be a partition of $[0,1]$ in D_m intervals indexed by m and let S_m denote the vector space of all piecewise constant functions on $I(m)$. Then, the regressogram associated with S_m is

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda}, \quad \text{where} \quad \widehat{\beta}_\lambda = \frac{\sum_{j=1}^n Y_j \mathbb{1}_{I_\lambda}(X_j)}{\sum_{k=1}^n \mathbb{1}_{I_\lambda}(X_k)}\cdot$$

Note that this estimator is uniquely defined only if there is at least one observation in each interval I_λ of the partition. In the sequel, we only consider models for which this requirement holds.
By inverting the sums, we can rewrite $\widehat{s}_m$ as

$$\forall 1 \le i \le n, \quad \widehat{s}_m(X_i) = \frac{1}{n}\sum_{j=1}^n \left(\sum_{\lambda\in\Lambda(m)} \frac{H_m^\lambda(Z_j, Z_i)}{G_m^\lambda(Z_{1,n}, Z_i)}\right),$$

where $H_m^\lambda(Z_j,Z_i) = Y_j \mathbb{1}_{I_\lambda}(X_j)\mathbb{1}_{I_\lambda}(X_i)$ and $G_m^\lambda(Z_{1,n},Z_i) = 1/n\sum_{k=1}^n \mathbb{1}_{I_\lambda}(X_k)$.
With a kernel-based estimator, we immediately notice that

$$\forall 1 \le i, \le n, \quad \widehat{s}_m(X_i) = \frac{1}{n}\sum_{j=1}^n \frac{H_m(Z_j,Z_i)}{G_m(Z_{1,n},Z_i)},$$

with $H_m(Z_j,Z_i) = Y_j K_m(X_j - X_i)$ and $G_m(Z_{1,n},Z_i) = 1/n\sum_{k=1}^n K_m(X_k - X_i)$, where K denotes a kernel and m its bandwidth.

Let us now illustrate the necessity of further requirements in order to derive closed-form expressions with a general kernel. As before, we must be able to compute expressions like

$$\sum_{e\in\mathcal{E}_p} \mathbb{1}_{(i\in e)}\widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i) = \sum_{j=1}^n H_m(Z_j,Z_i)\sum_{e\in\mathcal{E}_p}\frac{\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})}}{\sum_{k\in\overline{e}} K_m(X_k - X_i)}\cdot$$

Assume that $\sum_{k\in\overline{e}} K_m(X_k - X_i)$ takes $Q_i \le n-p$ known values $A_i(q),\ q = 1,\ldots,Q_i$, then we can rewrite the above expression as

$$\sum_{e\in\mathcal{E}_p} \mathbb{1}_{(i\in e)}\widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i) = \sum_{j=1}^n H_m(Z_j,Z_i)\sum_{q=1}^{Q_i} A_i(q)^{-1}\sum_{e\in\mathcal{E}_p}\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})}\mathbb{1}_{\left(\sum_{k\in\overline{e}} K_m(X_k - X_i) = A_i(q)\right)} \quad (3.18)$$

From (3.18), we see that provided $\mathrm{Card}\left(\left\{e\in\mathcal{E}_p \mid A_i(q) = \sum_{k\in\overline{e}} K_m(X_k - X_i)\right\}\right)$ is known for each q, a closed-form expression may be derived, whereas no such formula can be expected otherwise. Fortunately,

this requirement is fulfilled by the kernel $K(t) = \mathbb{1}_{(|t|\leq 1)}$ as well as regressograms, which we will consider in the sequel.
REMARKS:

- On the contrary if the denominator continuously depends on the observations, the knowledge of $\operatorname{Card}\left(\left\{e \in \mathcal{E}_p \mid A_i(q) = \sum_{k\in\overline{e}} K_m(X_k - X_i)\right\}\right)$ amounts to explore the $\binom{n}{p}$ resamples.

- An important issue is to make sure that the kernel estimators and regressograms are uniquely defined. Indeed, the denominator could vanish, provided m is too small (kernel estimator) or if there is no observation in at least one interval of the partition (regressogram). In the sequel, we assume that either the smoothing parameter m is large enough, or that there is at least one in each interval of the considered partitions.

Once these estimators are well defined, another issue occurs when applying the Lpo. Since this resampling scheme consists in removing p observations from the n original ones, if p is larger than the number of points in a given interval (for the regressogram) and since the resampling is made exhaustively, some intervals are emptied. Obviously, the same phenomenon raises with the kernel estimator. We now describe a possible solution to this problem.
First notice that when working with regressograms,

$$\forall 1 \leq i \leq n, \quad \widehat{s}_m(X_i) = \frac{1}{n}\sum_{j=1}^{n} \frac{H_m^{\lambda_i}(Z_j, Z_i)}{G_m^{\lambda_i}(Z_{1,n}, Z_i)},$$

where λ_i denotes the index λ such that $X_i \in I_\lambda$ ($I = (I_1, \ldots, I_D)$ is a partition of $[0,1]$). Since the Lpo estimator may be written as

$$\widehat{R}_p(m) = \frac{1}{p}\left[\operatorname{Card}\left(\mathcal{E}_p\right)\right]^{-1} \sum_{i=1}^{n} \left[\sum_{e\in\mathcal{E}_p} \mathbb{1}_{(i\in e)} \left(Y_i - \widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i)\right)^2\right],$$

the question amounts to warranty that $\widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i)$ is well defined for each point X_i. Our idea is to replace $\mathcal{E}_p$ by a subset denoted by $\mathcal{E}_p'(i)$ *depending on* i, such that $G_m^{\lambda_i}(Z_{1,n}^{\overline{e}}, Z_i)$ is non null for every $e \in \mathcal{E}_p'(i)$. In other words for each i, we are approximating the global expectation with respect to the resampling

$$\mathbb{E}_W\left[\gamma(\widehat{s}_m(P_n^{\overline{W}}), Z_i)\right]$$

by a conditional version of it

$$\mathbb{E}_W\left[\gamma(\widehat{s}_m(P_n^{\overline{W}}), Z_i) \mid G_m^{\lambda_i}(P_n^{\overline{W}}, Z_i) \neq 0\right].$$

Thus, all the forthcoming closed-form expressions follow this convention.
We now provide the results for both regressogram and kernel estimators. For the sake of simplicity, we denote the current estimator by $\widehat{s}_m(X_i) = 1/n\sum_{j=1}^{n} H_m(Z_j, Z_i)/G_m(Z_{1,n}, Z_i)$.

Proposition 3.3.3. *With the above notation, let $\widehat{s}_m$ denote either the regressogram in S_m built from the partition $\{I_\lambda\}$, or a kernel-based estimator with smoothing parameter $m > 0$. Then, we have*

$$H_m(Z_j, Z_i) = Y_j \mathbb{1}_{I_{\lambda_i}}(X_j)\mathbb{1}_{I_{\lambda_i}}(X_i) \quad \text{and} \quad G_m(Z_{1,n}, Z_i) = 1/n\sum_{k=1}^{n} \mathbb{1}_{I_{\lambda_i}}(X_k)\mathbb{1}_{I_{\lambda_i}}(X_i)$$

for the regressogram and

$$H_m(Z_j, Z_i) = Y_j K_m(X_j - X_i) \quad \text{and} \quad G_m(Z_{1,n}, Z_i) = 1/n\sum_{k=1}^{n} K_m(X_k - X_i)$$

for the kernel-based estimator. Moreover for any $1 \leq p \leq n-1$, the Lpo estimator is

$$\widehat{R}_p(m) = \frac{1}{p}\sum_{i=1}^{n} \frac{1}{N_i}\left[\mathbb{1}_{(n_i=1)}\{+\infty\} + \mathbb{1}_{(n_i\geq 2)}\left\{ Y_i^2 A_i - 2\sum_{j\neq i} Y_i H_m(Z_j, Z_i) B_i + \sum_{j\neq i}\left(H_m(Z_j,Z_i)\right)^2 C_i \right\} + \right.$$
$$\left. \mathbb{1}_{(n_i\geq 3)} \sum_{j\neq i}\sum_{k\neq j, k\neq i} H_m(Z_j,Z_i)H_m(Z_k,Z_i)D_i \right],$$

where $n_i = n\, G_m(Z_{1,n}, Z_i)$ and

$$\begin{aligned} N_i &= 1 - \mathbb{1}_{(p\geq n_i)}\binom{n-n_i}{p-n_i}/\binom{n}{p}, \\ A_i &= V_i(0) - \frac{V_i(1)}{n_i} \quad \text{and} \quad B_i = \frac{A_i}{n_i - 1}, \\ C_i &= \frac{V_i(-1)}{n_i-1} - \frac{V_i(0)}{n_i(n_i-1)}, \\ D_i &= \frac{(n_i+1)V_i(0) - V_i(1) - n_i V_i(-1)}{n_i(n_i-1)(n_i-2)}. \end{aligned}$$

For any $k \in \{-1, 0, 1\}$,

$$V_i(k) = \sum_{r=1\vee(p-n_i)}^{n_i\wedge(n-p)} r^k \frac{\binom{n-p}{r}\binom{p}{n_i-r}}{\binom{n}{n_i}}.$$

The proof is essentially based on some intricate algebra and is given in Section 3.5.
Remark: With regressograms, N_i, A_i, B_i, C_i, D_i and V_i only depends on i through the interval X_i belongs to. These quantities may therefore be calculated beforehand, whereas it is no longer the case with the kernel estimator.

These expressions may be further developed, which yields the following corollary:

Corollary 3.3.2. *With the same notations as in Proposition 3.3.3, we get*

$$\widehat{R}_p(m) = \sum_{\lambda\in m} \frac{1}{pN_\lambda}\left[\mathbb{1}_{(n_\lambda=1)}\{+\infty\} + \mathbb{1}_{(n_\lambda\geq 2)}\left\{ S_{\lambda,2}\left(A_\lambda + C_\lambda(n_\lambda - 1)\right) - 2B_\lambda\left(S_{\lambda,1}^2 - S_{\lambda,2}\right)\right\} + \right.$$
$$\left. \mathbb{1}_{(n_\lambda\geq 3)}\left\{ D_\lambda(n_\lambda-2)\left[S_{\lambda,1}^2 - S_{\lambda,2}\right]\right\}\right],$$

for regressograms, with $N_\lambda, A_\lambda, \ldots, D_\lambda$ being the same as in Proposition 3.3.3 when $X_i \in I_\lambda$. $S_{\lambda,1} = \sum_{j=1}^n Y_j \mathbb{1}_{I_\lambda}(X_j)$ and $S_{\lambda,2} = \sum_{j=1}^n Y_j^2 \mathbb{1}_{I_\lambda}(X_j)$.
For kernel estimator, we derive

$$\widehat{R}_p(m) = \sum_{i=1}^n \frac{\mathbb{1}_{(n_i\geq 2)}}{pN_i}\left\{ Y_i^2\left[A_i + \left(2(B_i + \mathbb{1}_{(n_i\geq 3)}D_i) - C_i\right)K_m(0)\right] - 2Y_i\sum_{j=1}^n Y_j K_m(X_j - X_i)\left[B_i + \right.\right.$$
$$\left.\left. \mathbb{1}_{(n_i\geq 3)} D_i K_m(0)\right] + \sum_{j=1}^n Y_j^2 K_m^2(X_j - X_i)\left[C_i - \mathbb{1}_{(n_i\geq 3)}D_i\right] + \left(\sum_{j=1}^n Y_j K_m(X_j - X_i)\right)^2 \mathbb{1}_{(n_i\geq 3)} D_i \right\} +$$
$$\sum_{i=1}^n \mathbb{1}_{(n_i=1)}\{+\infty\}.$$

3.4 CV and penalized criteria

3.4.1 Moment calculations

In this section, our goal is to provide explicit expressions for the two first moments (at least for the density estimation) of the Lpo risk estimator, and then of its bias. By doing so, we are looking for the weakest requirements on the distribution of the observations, which warranty the existence of these quantities.

Density

We first deal with general projection estimators for which we provide explicit expectation and variance.

Proposition 3.4.1. *With the same notations as in Proposition 3.3.1, we have for any* $1 \leq p \leq n-1$,

$$\begin{aligned}
\mathbb{E}\widehat{R}_p(m) &= \frac{1}{n-p}\sum_{\lambda\in\Lambda(m)}\left[\mathbb{E}\varphi_\lambda^2(X) - (\mathbb{E}\varphi_\lambda(X))^2\right] - \sum_{\lambda\in\Lambda(m)}(\mathbb{E}\varphi_\lambda(X))^2,\\
\mathrm{Var}\left[\widehat{R}_p(m)\right] &= (n(n-1)(n-p))^{-2}\left[2\beta^2 t_1\sum_\lambda\left(P\varphi_\lambda^2\right)^2 + 4\alpha\beta t_1\sum_\lambda P\varphi_\lambda^3 P\varphi_\lambda + n\alpha^2\mathbb{E}\left(\sum_\lambda\varphi_\lambda^2\right)^2 - \right.\\
&\quad n\alpha^2\left(\sum_\lambda P\varphi_\lambda^2\right)^2 + 2\beta^2 t_1\sum_{\lambda\neq\lambda'}(P\varphi_\lambda\varphi_{\lambda'})^2 + 4\beta^2 t_2\mathbb{E}\left(\sum_\lambda\varphi_\lambda P\varphi_\lambda\right)^2 + \\
&\quad \left.(-4n+6)t_1\beta^2\left(\sum_\lambda(P\varphi_\lambda)^2\right)^2 + 4\alpha\beta t_1\sum_{\lambda\neq\lambda'}P\varphi_\lambda^2\varphi_{\lambda'}P\varphi_{\lambda'} - 4t_1\alpha\beta\sum_\lambda P\varphi_\lambda^2\sum_{\lambda'}(P\varphi_{\lambda'})^2\right],
\end{aligned}$$

where $P\varphi_\lambda = \mathbb{E}\varphi_\lambda(X)$, *and*

$$\begin{aligned}
\alpha &= n-1 & \beta &= n-p+1,\\
t_1 &= n(n-1) & t_2 &= t_1(n-2),
\end{aligned}$$

The technical proof is given in Section 3.5. Note that these formulas may be derived provided $P\left|\varphi_\lambda\right|^3 < +\infty$ for any $\lambda \in \Lambda(m)$, which is satisfied if s is assumed to be bounded and $\int\left|\varphi_\lambda\right|^3 < +\infty$ (φ_λ continuous and compact supported for instance).

If we apply Proposition 3.4.1 to histograms estimators, we obtain the following expressions for the expectation and the variance of the Lpo risk estimator:

Corollary 3.4.1. *For any* $\lambda \in \Lambda(m)$, *set* $\alpha_\lambda = \mathbb{P}(X_i \in I_\lambda)$. *Then,*

$$\begin{aligned}
\mathbb{E}\left[\widehat{R}_p(m)\right] &= \frac{1}{n-p}\sum_{\lambda\in m}\frac{1}{\omega_\lambda}\alpha_\lambda(1-\alpha_\lambda) - \sum_{\lambda\in m}\frac{1}{\omega_\lambda}\alpha_\lambda^2,\\
\mathrm{Var}\left[\widehat{R}_p(m)\right] &= \frac{p^2 q_2(n,\alpha,\omega) + p\, q_1(n,\alpha,\omega) + q_0(n,\alpha,\omega)}{[n(n-1)(n-p)]^2},
\end{aligned}$$

where

$$\begin{aligned}
\forall(i,j) &\in \{1,\ldots,3\}\times\{1,2\}, \quad s_{i,j} = \sum_{k=1}^{D}\alpha_k^i/\omega_k^j,\\
q_2(n,\alpha,\omega) &= n(n-1)\left[2s_{2,2} + 4s_{3,2}(n-2) + s_{2,1}^2(-4n+6)\right],\\
q_1(n,\alpha,\omega) &= n(n-1)\left[-8s_{2,2} - 8s_{3,2}(n-2)(n+1) - 4s_{1,1}s_{2,1}(n-1) - \right.\\
&\quad \left. 2s_{2,1}^2(-4n^2+2n+6)\right],\\
q_0(n,\alpha,\omega) &= n(n-1)\left[s_{1,2}(n-1) - 2s_{2,2}(n^2-2n-3) + \right.\\
&\quad 4s_{3,2}(n-2)(n+1)^2 - s_{1,1}^2(n-1) + \\
&\quad \left. 4s_{1,1}s_{2,1}(n^2-1) + s_{2,1}^2(-4n+6)(n+1)^2\right].
\end{aligned}$$

The bias of the Lpo risk estimator may be a more interesting quantity to work with. From Proposition 3.4.1, we derive its expression.

Corollary 3.4.2. *With the same notations as in Proposition 3.4.1 and for any projection estimator, the bias of the Lpo estimator equals*

$$\begin{aligned}\mathbb{B}\left[\widehat{R}_p(m)\right] := \mathbb{E}\widehat{R}_p(m) - R_n(m) &= \frac{p}{n(n-p)}\sum_{\lambda\in m}\left[\mathbb{E}\varphi_\lambda^2(X) - (\mathbb{E}\varphi_\lambda(X))^2\right],\\ &= \frac{p}{n(n-p)}\sum_{\lambda\in\Lambda(m)}\mathrm{Var}\left[\varphi_\lambda(X)\right] \geq 0,\end{aligned}$$

where $R_n(m) = \mathbb{E}\left[\|\widehat{s}_m\|^2 - 2\int_{[0,1]} s\,\widehat{s}_m\right]$.

Similar results may be obtained for kernel estimators as exposed in the following proposition.

Proposition 3.4.2. *With the same notations as in Proposition 3.3.1 and provided K is a symmetric kernel such that $K \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and $s \in L^1(\mathbb{R})$, the expectation, bias and variance are respectively given by*

$$\begin{aligned}\mathbb{E}\left[\widehat{R}_p(m)\right] &= \frac{1}{n-p}\left(\|K_m\|^2 + (n-p-1)\int_{\mathbb{R}}\left[\mathbb{E}K_m(X-t)\right]^2 dt\right) - 2\mathbb{E}\left[K_m(X-Y)\right],\\ \mathbb{B}\left[\widehat{R}_p(m)\right] &= \frac{p}{n(n-p)}\left(\|K_m\|^2 - \mathbb{E}\left[K^*(X-Y)\right]\right),\\ \mathrm{Var}\left[\widehat{R}_p(m)\right] &= \frac{p^2\widetilde{q}_2 + p\widetilde{q}_1 + \widetilde{q}_0}{n(n-1)(n-p)^2},\end{aligned}$$

where X and Y are independent random variables, drawn from P, and

$$\begin{aligned}\widetilde{q}_2 &= A_2,\\ \widetilde{q}_1 &= -2nA_2 - A_1,\\ \widetilde{q}_0 &= n^2A_2 + nA_1 + A_0,\end{aligned}$$

and

$$\begin{aligned}A_2 &= 4(n-2)\left(\mathbb{E}_X\left(\mathbb{E}_Y h(X,Y)\right)^2 - \left(\mathbb{E}h(X,Y)\right)^2\right) + 2\mathrm{Var}\left(h(X,Y)\right),\\ A_1 &= 8(n-2)\left(\mathbb{E}h(X,Y)\mathbb{E}K_m^*(X-Y) - \mathbb{E}_X\left[\mathbb{E}_Y h(X,Y)\mathbb{E}_Y K_m^*(X-Y)\right]\right) +\\ &\quad 4\left(\mathbb{E}h(X,Y)\mathbb{E}K_m^*(X-Y) - \mathbb{E}h(X,Y)K_m^*(X-Y)\right),\\ A_0 &= 4(n-2)\left(\mathbb{E}_X\left(\mathbb{E}_Y K_m^*(X-Y)\right)^2 - \left(\mathbb{E}K_m^*(X-Y)\right)^2\right) + 2\mathrm{Var}\left(K_m^*(X-Y)\right),\\ h(X,Y) &= K_m^*(X-Y) - 2K_m(X-Y),\end{aligned}$$

with $\mathbb{E}_X$ and $\mathbb{E}_Y$ denoting expectations with respect to X and Y.

Remarks:

- Note that a sufficient condition for the bias of the Lpo risk to be nonnegative is that $K(0) \geq K(t)$ for any $t \in \mathbb{R}$, which holds for the gaussian kernel for instance.
- The variance expression of kernel estimator comes from the $\mathbb{U}$-statistic nature of the Lpo risk estimator. We refer to [81] for an introduction to $\mathbb{U}$-statistics and to [23] for an extensive study in the empirical process framework.

Regression

In the regression context, we will content ourselves with closed-form expressions for expectations, since any computation of the variance would involve the existence of every moments until the order four. Note that up to this requirement and more complicate algebra, the variance of the Lpo estimator may be explicitly computed as well.

We first provide the expectation of the projection estimator for which simple calculations lead to:

Proposition 3.4.3. *With the same notations as Proposition 3.3.2, for a given sample of observations* $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$*, we have*

$$\mathbb{E}\left(\widehat{R}_p(m)\right) = \frac{1}{n-p}\left[\sum_{\lambda} P\varphi_\lambda^2 \mathbb{E}\left(Y^2\varphi_\lambda^2(X)\right) + \sum_{\lambda\neq\lambda'} \mathbb{E}\left(Y^2\varphi_\lambda(X)\varphi_{\lambda'}(X)\right) P\varphi_\lambda\varphi_{\lambda'}\right] +$$
$$\frac{n-p+1}{n-p}\left[\sum_{\lambda} P\varphi_\lambda^2 \left(\mathbb{E}\left[Y\varphi_\lambda(X)\right]\right)^2 + \sum_{\lambda\neq\lambda'} P\varphi_\lambda\varphi_{\lambda'}\mathbb{E}\left(Y\varphi_\lambda(X)\right)\mathbb{E}\left(Y\varphi_{\lambda'}(X)\right)\right] -$$
$$2\sum_{\lambda}\left[\mathbb{E}\left(Y\varphi_\lambda(X)\right)\right]^2.$$

Remark: In the above result, the expectations are taken with respect to the joint distribution of (X, Y).

We now provide expectations of the Lpo risk estimator for the regressograms and kernels. We point out that these expectations are actually conditional expectations given the design points. Indeed, we are not able to derive closed-form expressions in the random design setting. The best we could do in such a case is an approximation of these expectations as in the recent work of Arlot [3]. The proof of this proposition is given in Section 3.5.

Proposition 3.4.4. *In the fixed-design setting, let us assume we observe n random variables* $Y_i = s(X_i) + \sigma_i\epsilon_i$*, where* $(\epsilon_i)_i$ *are i.i.d. centered random variables such that* $\mathbb{E}\epsilon_i^2 = 1$*. With the notations of Corollary 3.3.2, the Lpo risk expectation for regressograms equals*

$$\mathbb{E}\left[\widehat{R}_p(m)\right] = \sum_{\lambda\in m}\frac{1}{pN_\lambda}\Big[\mathbb{1}_{(n_\lambda=1)}\{+\infty\} + \mathbb{1}_{(n_\lambda\geq 2)}\Big\{\Big[F_{\lambda,2} + n_\lambda\left(\sigma_\lambda^r\right)^2\Big]\left(A_\lambda + C_\lambda(n_\lambda-1)\right) -$$
$$2B_\lambda\left(F_{\lambda,1}^2 - F_{\lambda,2}\right)\Big\} + \mathbb{1}_{(n_\lambda\geq 3)}\left\{D_\lambda(n_\lambda-2)\left[F_{\lambda,1}^2 - F_{\lambda,2}\right]\right\}\Big],$$

where $F_{\lambda,2} = \sum_{i=1}^n s^2(X_i)\mathbb{1}_{I_\lambda}(X_i)$*,* $F_{\lambda,1} = \sum_{i=1}^n s(X_i)\mathbb{1}_{I_\lambda}(X_i)$ *and* $\left(\sigma_\lambda^r\right)^2 = \sum_{i=1}^n \sigma_i^2\mathbb{1}_{I_\lambda}(X_i)/n_\lambda$*.*
In the same way for kernels, we get

$$\mathbb{E}\left[\widehat{R}_p(m)\right] = \sum_{i=1}^n \frac{\mathbb{1}_{(n_i\geq 2)}}{pN_i}\Bigg\{\left(s^2(X_i)+\sigma_i^2\right)A_i - 2s(X_i)\sum_{j\neq i}s(X_j)K_m(X_j-X_i)B_i +$$
$$\sum_{j\neq i}\left(s^2(X_j)+\sigma_j^2\right)K_m^2(X_j-X_i)C_i +$$
$$\mathbb{1}_{(n_i\geq 3)}\left\{\sum_{j\neq i}\sum_{k\neq i,k\neq j} s(X_j)s(X_k)K_m(X_j-X_i)K_m(X_k-X_i)\right\}\Bigg\} + \sum_{i=1}^n \mathbb{1}_{(n_i=1)}\{+\infty\}.$$

3.4.2 CV as a random penalty

Ideal and Lpo penalties

In model selection, our goal is to find the "best" model among $(S_m)_m$, in terms of a given criterion from a family of estimators $(\widehat{s}_m)_m$. A very common way to reach this goal is the minimization of a penalized criterion $\mathrm{crit}(\cdot)$ defined as

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m),$$

where γ is the contrast function that measures the quality of an estimator and $\mathrm{pen}(\cdot) : \mathcal{M}_n \to \mathbb{R}_+$ denotes the penalty term, which takes into account the complexity of the model m.
Ideally, the optimal criterion we would like to minimize over $\mathcal{M}_n$ is the random quantity

$$\mathrm{crit}_{id}(m) = P\gamma(\widehat{s}_m) := \mathbb{E}\gamma\left(\widehat{s}_m, Z\right),$$

with the expectation taken with respect to $Z \sim P$. The interpretation of crit_{id} as a penalized criterion comes from

$$\mathrm{crit}_{id}(m) = P_n\gamma(\widehat{s}_m) + [\, P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m)\,] .$$

The quantity in square brackets is named the ideal penalty

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{pen}_{id}(m) := P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m).$$

Following the CV strategy, we perform model selection by minimizing the Lpo risk estimator over $\mathcal{M}_n$, provided $\mathcal{M}_n$ is not too large. Thus for a given $1 \leq p \leq n-1$, the candidate $\widehat{m}$ is

$$\widehat{m} = \mathrm{Argmin}_{m\in\mathcal{M}_n} \widehat{R}_p(m).$$

The idea that there is a strong relationship between penalized criteria and CV is strongly supported by the large amount of literature about the comparison between these two aspects [75, 55, 87]. Therefore, we may try to include the CV strategy into the wider scope of penalized criteria minimization. Thus,

$$\widehat{m} = \mathrm{Argmin}_{m\in\mathcal{M}_n} P_n\gamma(\widehat{s}_m) + \left[\widehat{R}_p(m) - P_n\gamma(\widehat{s}_m) \right] .$$

In the above expression, the quantity in square brackets is called the Lpo penalty:

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{pen}_p(m) := \widehat{R}_p(m) - P_n\gamma(\widehat{s}_m).$$

This Lpo penalty may be subsequently understood as a random penalty. It is an alternative to C_p-like penalties as a measure of the model complexity.
Remark: A similar approach applied to the Loo can be found in Birgé and Massart [9].

Thanks to this parallel between CV and penalized criteria, we attempt to get more insight in the behaviour of CV techniques, for instance with respect to the parameter p.

Lpo overpenalization

In this section, we aim at making comparison between pen_{id} and pen_p, so that we would like to characterize some features in the behaviour of pen_p with respect to p. This comparison is carried out through the expectations of these criteria, which are both random variables.
Since there is no model structure in the kernel-based approach, we do not apply the following to kernel estimators. In the sequel, we only consider general projection estimators in density estimation, especially histograms, and also regressograms.

The following result we provide concerns density estimation. Here, the question is the assessment of the gap (in expectation) between the Lpo penalty and the ideal one. We first address this question with general projection estimators.
We start with a lemma which introduces the forthcoming proposition.

Lemma 3.4.1. *With the same notations as before with any projection estimator $\widehat{s}_m$ onto S_m, we obtain*

$$\begin{aligned}
\mathbb{E}\left[\, \mathrm{pen}_{id}(m) \,\right] &= \frac{2}{n} \sum_{\lambda\in m} \mathrm{Var}\left(\varphi_\lambda(X)\right), \\
\mathbb{E}\left[\, \mathrm{pen}_{p}(m) \,\right] &= \frac{2n-p}{n(n-p)} \sum_{\lambda\in m} \mathrm{Var}\left(\varphi_\lambda(X)\right).
\end{aligned}$$

We now state the main assertion about the Lpo penalty associated with projection estimators in density estimation.

Proposition 3.4.5. *For any $m \in \mathcal{M}_n$, let $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$ denote an orthonormal basis of S_m and $\widehat{s}_m$, the projection estimator onto S_m. Following Proposition 3.4.1, we get*

$$\forall m \in \mathcal{M}_n,\ \forall 1 \leq p \leq n-1, \qquad \mathbb{E}\left[\, \mathrm{pen}_p(m) - \mathrm{pen}_{id}(m) \,\right] = \frac{p}{n(n-p)} \sum_{\lambda \in m} \mathrm{Var}\left(\varphi_\lambda(X)\right) \geq 0.$$

Remark: As this quantity remains positive whatever p, we conclude that the Lpo penalty always overpenalizes, which remains true for any orthonormal basis. Moreover, the amount of overpenalization increases with p. Thus, the Loo provides the weakest overpenalization of order $\mathcal{O}\left(1/n^2\right)$, whereas the Lpo with $p \simeq n/2$ (which is similar to the 2-fold CV) corresponds to an overpenalization of the same amount as the expectation of the ideal penalty, that is $\mathcal{O}\left(1/n\right)$. Note that this overpenalization phenomenon may be related to the "intuitive" bias increase, when less and less data are used in the training set (see Corollary 3.4.2).

In comparison with the previous expressions, calculations involving the regressogram are by far more intricate and come from Proposition 3.4.4 and the following lemmas.

Lemma 3.4.2. *With the same notations as Proposition 3.4.4, we have*

$$\mathbb{E}\left[\, \mathrm{pen}_{id}(m) \,\right] = \frac{2}{n} \sum_{\lambda \in m} \left(\sigma_\lambda^r\right)^2 \qquad \text{and} \qquad \mathbb{E}\left[\, P_n \gamma\left(\widehat{s}_m\right) \,\right] = \sum_\lambda \frac{n_\lambda - 1}{n} \left[\left(\sigma_\lambda^r\right)^2 + \frac{n_\lambda}{n_\lambda - 1} \left(\sigma_\lambda^b\right)^2 \right],$$

$$\mathbb{E} S_{\lambda,2} = n_\lambda \left[\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2 + \beta_\lambda^2 \right] \qquad \text{and} \qquad \mathbb{E} S_{\lambda,1}^2 = n_\lambda \left(\sigma_\lambda^r\right)^2 + n_\lambda^2 \beta_\lambda^2\,,$$

where $\left(\sigma_\lambda^r\right)^2 := 1/n_\lambda \sum_{i=1}^n \sigma_i^2 \mathbb{1}_{I_\lambda}(X_i)$ and $\left(\sigma_\lambda^b\right)^2 := 1/n_\lambda \sum_{i=1}^n \left[\, s(X_i) - \beta_\lambda \,\right]^2 \mathbb{1}_{I_\lambda}(X_i)$.

We also need some further details about coefficients $V_\lambda(k)$, which is given by the following result.

Lemma 3.4.3. *For any $\lambda \in \Lambda(m)$, let X denote a random variable following a hypergeometric distribution $X \sim \mathcal{H}(n_\lambda, n-p, n)$. Then,*

$$\begin{aligned} V_\lambda(0) &= \mathbb{P}\left[\, X \in \{1 \vee p - n_\lambda, \ldots, n_\lambda \wedge n-p\} \,\right] = 1 - \mathbb{1}_{(p \geq n_\lambda)} \binom{p}{n_\lambda} / \binom{n}{n_\lambda}, \\ V_\lambda(1) &= \mathbb{E}\left[\, X \,\right] = \frac{n_\lambda(n-p)}{n} \qquad \text{and} \qquad V_\lambda(-1) = \mathbb{E}\left[\, X^{-1} \mathbb{1}_{(X>0)} \,\right], \\ V_\lambda(1) V_\lambda(-1) &\geq 1. \end{aligned}$$

with $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$.

Due to the high complexity of expressions in hand, the following result is not given in full generality, but we make a few assumptions in order to enlighten the underlying phenomenon as clearly as possible.

Proposition 3.4.6. *With the notations of Proposition 3.4.4, let us assume that for any $\lambda \in \Lambda(m)$, $n_\lambda \geq 3$. Then if $p < n_\lambda$ for every λ, we have*

$$\begin{aligned} \mathbb{E}\left[\, \mathrm{pen}_p(m) \,\right] &= \sum_{\lambda \in m} \left\{ \left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2 \frac{n_\lambda}{n_\lambda - 1} \right\} \left(\frac{p-n}{np} + \frac{n}{p(n-p)} V_\lambda(1) V_\lambda(-1) \right), \\ &\geq \frac{n - p/2}{n-p}\, \mathbb{E}\left[\, \mathrm{pen}_{id}(m) \,\right]. \end{aligned}$$

Moreover, assume that $p \geq \left(n_\lambda \vee \left[\, \sqrt{a_\lambda} - 1 \,\right] n / \sqrt{a_\lambda} \right)$ for any $\lambda \in \Lambda(m)$, where $a_\lambda = n_\lambda / V_\lambda(1) V_\lambda(-1)$. Then,

$$\begin{aligned} \mathbb{E}\left[\, \mathrm{pen}_p(m) \,\right] &= \sum_{\lambda \in \Lambda(m)} \left\{ \left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2 \frac{n_\lambda}{n_\lambda - 1} \right\} \left[(n_\lambda - 1) \frac{n-p}{np} + \frac{1}{N_\lambda} \left(\frac{n V_\lambda(1) V_\lambda(-1)}{p(n-p)} - \frac{n_\lambda(n-p)}{np} \right) \right], \\ &\geq \frac{n-p/2}{n-p}\, \mathbb{E}\left[\, \mathrm{pen}_{id}(m) \,\right]. \end{aligned}$$

The proof is provided in Section 3.5.
Thanks to the above proposition, we have characterized the behaviour of the Lpo penalty as an overpenalizing criterion both for small and large values of p. But as previously described (Section 2.2.1), this overpenalization does not necessary appear as a drawback of the approach. Indeed whereas a biased criterion may be misleading when trying to accurately estimate the risk of an estimator, this bias may prevent some well known troubles in the model selection framework.

3.5 Proofs

3.5.1 Closed-form Lpo estimator

Proof of Lemma 3.3.1

The first remark is that for each $e \in \mathcal{E}_p$, we have

$$\begin{aligned}
\forall t \in [0,1], \quad \widehat{s}_m(Z_{1,n}^{\overline{e}})(t) &= \frac{1}{n-p}\sum_{j\in\overline{e}} H_m(Z_j,t) = \frac{1}{n-p}\sum_{j=1}^{n} H_m(Z_j,t)\mathbb{1}_{(j\in\overline{e})}, \\
\text{and} \quad \sum_{i\in e}\widehat{s}_m(Z_{1,n}^{\overline{e}})(Z_i) &= \frac{1}{n-p}\sum_{i=1}^{n}\sum_{j\in\overline{e}} H_m(Z_j,Z_i)\mathbb{1}_{(i\in e)} = \frac{1}{n-p}\sum_{i\neq j} H_m(Z_j,Z_i)\mathbb{1}_{(j\in\overline{e})}\mathbb{1}_{(i\in e)}.
\end{aligned}$$

Then, the Lemma follows from the following combinatorial results

Lemma 3.5.1. *For any $i \neq j \neq k \in \{1,\ldots,n\}$,*

$$\begin{aligned}
\sum_{e\in\mathcal{E}_p}\mathbb{1}_{(j\in\overline{e})} &= \binom{n-1}{p} \quad \text{and} \quad \sum_{e\in\mathcal{E}_p}\mathbb{1}_{(j\in\overline{e})}\mathbb{1}_{(k\in\overline{e})} = \binom{n-2}{p-1}, \\
\sum_{e\in\mathcal{E}_p}\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})}\mathbb{1}_{(k\in\overline{e})} &= \binom{n-3}{p-1} \quad \text{and} \quad \sum_{e\in\mathcal{E}_p}\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})} = \binom{n-2}{p-1},
\end{aligned}$$

where we stress that the sum is made over the resamples i, j and k are kept fixed.

Proof. $\sum_{e\in\mathcal{E}_p}\mathbb{1}_{(j\in\overline{e})}$ may be interpreted as the number of subsets of $\{1,\ldots,n\}$ of size p (denoted by e) which do not contain j, since $j \in \overline{e}$. Thus, it is the number of possible choices of p non ordered and different elements among $n-1$.
The other equalities follow from a similar argument. □

Proof Proposition 3.3.3

We begin with a lemma which will be useful in the proof.

Lemma 3.5.2.

$$\begin{aligned}
\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\mathbb{1}_{\left((n-p)G_m(Z_{1,n}^{\overline{e}},Z_i)=r\right)}\mathbb{1}_{(i\in e)} &= \left(1-\frac{r}{n_i}\right)\mathbb{P}\left(X=r\right), \qquad (3.19)\\
\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\mathbb{1}_{\left((n-p)G_m(Z_{1,n}^{\overline{e}},Z_i)=r\right)}\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})} &= \frac{r(n_i-r)}{n_i(n_i-1)}\mathbb{P}\left(X=r\right), \\
\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\mathbb{1}_{\left((n-p)G_m(Z_{1,n}^{\overline{e}},Z_i)=r\right)}\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})}\mathbb{1}_{(k\in\overline{e})} &= \frac{r(r-1)(n_i-r)}{n_i(n_i-1)(n_i-2)}\mathbb{P}\left(X=r\right),
\end{aligned}$$

where X is random variable which follows a hypergeometric distribution $\mathcal{H}(n_i, n-p, n)$.

Proof. (Lemma 3.5.2)
We have to remember that the choice of X_i determines an interval in which n_i observations lie. Besides $(n-p)G_m(Z_{1,n}^{\overline{e}}, Z_i)$ denotes the number of instances of the training set among these n_i points, given

that X_i is in the test set and belongs to these n_i observations as well. Therefore, (3.19) represents the number of subsets of $\{X_1, \ldots, X_n\} \setminus \{X_i\}$ of cardinality $n-p$ such that r points are among the n_i former observations. Thus, it equals $\binom{n_i-1}{r}\binom{n-n_i}{n-p-r}$. Divided by $\binom{n}{p}$ and after recombination, it gives

$$\binom{n_i-1}{r}\binom{n-n_i}{n-p-r}/\binom{n}{p} = (1 - r/n_i)\,\mathbb{P}\,(X = r)\,,$$

where $\mathbb{P}\,(X=r) = \binom{n-p}{r}\binom{p}{n_i-r}/\binom{n_i}{n}$.
The other equalities are obtained following the same reasoning. □

The first step in the proof of Proposition 3.3.3 consists in writing the Lpo risk estimator with the convention discussed earlier, which results in replacing $\mathcal{E}_p$ by $\mathcal{E}'_p(i)$ for each i in the expressions. Besides, let assume that m satisfies $\forall i,\ n_i \geq 2$.

$$\begin{aligned}
\widehat{R}_p(m) &= \frac{1}{p}\sum_{i=1}^n \left[\,\mathrm{Card}(\mathcal{E}'_p(i))\,\right]^{-1} \sum_{e\in\mathcal{E}'_p} \mathbb{1}_{(i\in e)} \left(Y_i - \widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i)\right)^2, \\
&= \frac{1}{p}\sum_{i=1}^n \left[\,\mathrm{Card}(\mathcal{E}'_p(i))\,\right]^{-1} \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} \mathbb{1}_{(i\in e)} \left(Y_i - \widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i)\right)^2, \\
&= \sum_{i=1}^n \frac{1}{pN_i}\binom{n}{p}^{-1} \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} \mathbb{1}_{(i\in e)} \left\{Y_i^2 - 2Y_i\widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i) + \widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i)^2\right\}, \qquad (3.20)
\end{aligned}$$

where $N_i := \mathrm{Card}(\mathcal{E}'_p)/\,\mathrm{Card}(\mathcal{E}_p)$.
Splitting the previous expression into three parts, we see that

$$\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} Y_i^2 \mathbb{1}_{(i\in e)} = Y_i^2 \left[\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} \mathbb{1}_{(i\in e)}\right], \qquad (3.21)$$

$$\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} Y_i\widehat{s}_m(Z_{1,n}^{\overline{e}})(Z_i)\mathbb{1}_{(i\in e)} = \frac{1}{n-p}\sum_{j\neq i} H_m(Z_j,Z_i)Y_i \left[\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} \frac{\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})}}{G_m(Z_{1,n}^{\overline{e}},Z_i)}\right], \qquad (3.22)$$

$$\begin{aligned}
\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} \left[\,\widehat{s}_m(Z_{1,n}^{\overline{e}})(Z_i)\,\right]^2 \mathbb{1}_{(i\in e)} = \frac{1}{(n-p)^2} \sum_{j\neq i,k\neq i} H_m(Z_j,Z_i)H_m(Z_k,Z_i) \Bigg[\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z_{1,n}^{\overline{e}},Z_i)\neq 0\right)} \\
\frac{\mathbb{1}_{(i\in e)}\mathbb{1}_{(j\in\overline{e})}\mathbb{1}_{(k\in\overline{e})}}{G_m^2(Z_{1,n}^{\overline{e}},Z_i)}\Bigg]. \qquad (3.23)
\end{aligned}$$

The main point is the computation of the expressions in square brackets, which may be achieved by first remembering that for each i, either $G_m(Z_{1,n}^{\overline{e}}, Z_i) = 1/(n-p)\sum_{j\in\overline{e}} \mathbb{1}_{I_{\lambda_i}}(X_j)$ for regressograms, or $G_m(Z_{1,n}^{\overline{e}}, Z_i) = 1/(n-p)\sum_{j\in\overline{e}} \mathbb{1}_{(|X_i-X_j|\leq m)}$ for kernel-based estimators. Thus, $G_m(Z_{1,n}^{\overline{e}}, Z_i)$ only depends on $Z_{1,n}^{\overline{e}}$ through the cardinality of the set of observations belonging to a given interval which itself depends on X_i. Subsequently notice that for each i,

$$(n-p)G_m(Z_{1,n}^{\overline{e}}, Z_i) \in \{1 \vee (p-n_i), \ldots, (n-p)\wedge n_i\}\,.$$

Then, (3.21), (3.22) and (3.23) give

$$\begin{aligned}
\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z^{\overline{e}}_{1,n},Z_i)\neq 0\right)} Y_i^2 \mathbb{1}_{(i\in e)} &= \mathbb{1}_{(n_i\geq 2)} Y_i^2 \left[\sum_r \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left((n-p)G_m(Z^{\overline{e}}_{1,n},Z_i)=r\right)} \mathbb{1}_{(i\in e)} \right], \\
\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z^{\overline{e}}_{1,n},Z_i)\neq 0\right)} Y_i \widehat{s}_m(Z^{\overline{e}}_{1,n})(Z_i) \mathbb{1}_{(i\in e)} &= \mathbb{1}_{(n_i\geq 2)} \sum_{j\neq i} H_m(Z_j,Z_i) Y_i \\
&\quad \left[\sum_r \frac{1}{r} \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left((n-p)G_m(Z^{\overline{e}}_{1,n},Z_i)=r\right)} \mathbb{1}_{(i\in e)} \mathbb{1}_{(j\in\overline{e})} \right], \\
\sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left(G_m(Z^{\overline{e}}_{1,n},Z_i)\neq 0\right)} \left[\widehat{s}_m(Z^{\overline{e}}_{1,n})(Z_i) \right]^2 \mathbb{1}_{(i\in e)} &= \mathbb{1}_{(n_i\geq 3)} \sum_{j\neq i, k\neq i} H_m(Z_j,Z_i) H_m(Z_k,Z_i) \\
&\quad \left[\sum_r \frac{1}{r^2} \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\left((n-p)G_m(Z^{\overline{e}}_{1,n},Z_i)=r\right)} \mathbb{1}_{(i\in e)} \mathbb{1}_{(j\in\overline{e})} \mathbb{1}_{(k\in\overline{e})} \right],
\end{aligned}$$

where $\sum_r$ is taken over $\{1 \vee (p-n_i), \ldots, (n-p)\wedge n_i\}$.
Then, plugging the results of Lemma 3.5.2 in the above expressions as well as in (3.20) enables to get the right expression for the Lpo estimator.

Besides, note that $\mathrm{Card}(\mathcal{E}'_p)$ equals the number of subsets of size $n-p$ of $\{X_1,\ldots,X_n\}$, such that at least one element of this subset belongs to either I_{λ_i} or to $\{X_j \mid |X_i - X_j| \leq m\}$. Thus, $\mathrm{Card}(\mathcal{E}'_p) = \binom{n}{p} - \binom{n-n_i}{n-p}\mathbb{1}_{p\geq n_i}$ and hence

$$N_i = 1 - \binom{n-n_i}{n-p} \Big/ \binom{n}{p} \mathbb{1}_{p\geq n_i},$$

which concludes the proof.

3.5.2 Moments calculations

Proof of Proposition 3.4.1

The expectation is a straightforward consequence of (3.10).
The variance calculation is not difficult, but very technical. We only give the main step of this proof.
First, let define $A_\lambda = \sum_{j=1}^n \varphi_\lambda^2(X_j)$ and $B_\lambda = \sum_{j\neq k} \varphi_\lambda(X_j)\varphi_\lambda(X_k)$. Set $\alpha = n-1$ and $\beta = n-p+1$, such that $n(n-1)(n-p)\widehat{R}_p(m) = \sum_\lambda (\alpha A_\lambda + \beta B_\lambda)$. Then, $\left[\sum_\lambda (\alpha A_\lambda + \beta B_\lambda)\right]^2 = \sum_\lambda \left(\alpha^2 A_\lambda^2 + \beta^2 B_\lambda^2 + 2\alpha\beta A_\lambda B_\lambda\right) + \sum_{\lambda\neq\lambda'} \left(\alpha^2 A_\lambda A_{\lambda'} + \beta^2 B_\lambda B_{\lambda'} + 2\alpha\beta A_\lambda B_{\lambda'}\right)$. After some calculation,

the different terms are respectively equal to

$$
\begin{aligned}
\mathbb{E}\sum_\lambda A_\lambda^2 &= \sum_\lambda \left[n\, P\varphi_\lambda^4 + t_1\left(P\varphi_\lambda^2\right)^2\right],\\
\mathbb{E}\sum_\lambda B_\lambda^2 &= \sum_\lambda \left[4t_2 P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2 + 2t_1\left(P\varphi_\lambda^2\right)^2 + t_3\left(P\varphi_\lambda\right)^4\right],\\
\mathbb{E}\sum_\lambda A_\lambda B_\lambda &= \sum_\lambda \left[2t_1 P\varphi_\lambda^3 P\varphi_\lambda + t_2 P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2\right],\\
\mathbb{E}\sum_{\lambda\neq\lambda'} A_\lambda A_{\lambda'} &= n\left[\mathbb{E}\left(\sum_\lambda \varphi_\lambda^2(X)\right)^2 - \sum_\lambda P\varphi_\lambda^4\right] + t_1\left[\left(\sum_\lambda P\varphi_\lambda^2\right)^2 - \sum_\lambda\left(P\varphi_\lambda^2\right)^2\right],\\
\mathbb{E}\sum_{\lambda\neq\lambda'} B_\lambda B_{\lambda'} &= 2t_1\sum_{\lambda\neq\lambda'}\left(P\varphi_\lambda\varphi_{\lambda'}\right)^2 + 4t_2\left[\mathbb{E}\left(\sum_\lambda \varphi_\lambda(X)P\varphi_\lambda\right)^2 - \sum_\lambda P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2\right] +\\
&\qquad t_3\left[\left(\sum_\lambda\left(P\varphi_\lambda\right)^2\right)^2 - \sum_\lambda\left(P\varphi_\lambda\right)^4\right],\\
\mathbb{E}\sum_{\lambda\neq\lambda'} A_\lambda B_{\lambda'} &= 2t_1\sum_{\lambda\neq\lambda'} P\varphi_\lambda^2\varphi_{\lambda'}P\varphi_{\lambda'} + t_2\left[\mathbb{E}\left(\sum_\lambda \varphi_\lambda^2(X)\right)\sum_{\lambda'}\left(P\varphi_{\lambda'}\right)^2 - \mathbb{E}\left(\sum_\lambda \varphi_\lambda^2(X)\left(P\varphi_\lambda\right)^2\right)\right].
\end{aligned}
$$

On the other hand,

$$
\left(n(n-1)(n-p)\mathbb{E}\left[\widehat{R}_p(m)\right]\right)^2 = n^2\alpha^2\left(\sum_\lambda P\varphi_\lambda^2\right)^2 + t_1^2\beta^2\left(\sum_\lambda\left[P\varphi_\lambda\right]^2\right)^2 + 2n\alpha\beta t_1\left(\sum_\lambda P\varphi_\lambda^2\right)\sum_{\lambda'}\left(P\varphi_{\lambda'}\right)^2.
$$

Combining these two expressions yields the variance after some simplifications.

Proof of Corollary 3.4.2

We have to compute $R_n(m)$ for any model m.

$$
\begin{aligned}
R_n(m) &:= \mathbb{E}\left[\|\widehat{s}_m\|^2\right] - 2\mathbb{E}\left[\int_{[0,1]} s\widehat{s}_m\right],\\
&= \sum_\lambda \mathbb{E}\left(P_n\varphi_\lambda\right)^2 - 2\sum_\lambda\left(P\varphi_\lambda\right)^2,\\
&= \frac{1}{n}\sum_\lambda \operatorname{Var}\left(\varphi_\lambda(X)\right) - \sum_\lambda\left(P\varphi_\lambda\right)^2.
\end{aligned}
$$

Proof of Proposition 3.4.2

The expectation immediately is a consequence of Proposition 3.3.1.
The conclusion for the bias comes from the calculation of $\mathbb{E}\left[\|\widehat{s}_m\|^2\right] - 2\int_{\mathbb{R}} s\,\mathbb{E}\left[\widehat{s}_m\right]$, since we see that

$$
\begin{aligned}
\mathbb{E}\left[\|\widehat{s}_m\|^2\right] &= \frac{1}{n}\|K_m\|^2 + (1-1/n)\,\mathbb{E}\left[K_m^*(X-Y)\right],\\
\int_{\mathbb{R}} s\,\mathbb{E}\left[\widehat{s}_m\right] &= \mathbb{E}\left[K_m(X-Y)\right].
\end{aligned}
$$

The variance expression relies on the $\mathbb{U}$-statistic nature of the Lpo estimator. Indeed,

$$
\operatorname{Var}\left[\widehat{R}_p(m)\right] = \operatorname{Var}\left[\frac{1}{n(n-1)}\sum_{i\neq j} g\left(X_i, X_j\right)\right],
$$

where $g(X_i, X_j) = h(X_i, X_j) - 1/(n-p)K_m^*(X_i - X_j)$ and $h(X_i, X_j) = K_m^*(X_i - X_j) - 2K_m(X_i - X_j)$. By setting

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} g(X_i, X_j),$$

we notice that U_n is a U-statistic of order 2. Thanks to a known result about the variance of $\mathbb{U}$-statistics [81], it entails that

$$\mathrm{Var}\left[U_n \right] = \frac{4(n-2)}{n(n-1)} \zeta_1 + \frac{2}{n(n-1)} \zeta_2,$$

where

$$\zeta_1 = \mathrm{Var}\left[\mathbb{E}\left(g(X,Y) \mid X \right) \right] \qquad \text{and} \qquad \zeta_2 = \mathrm{Var}\left[g(X,Y) \right].$$

Then, we develop ζ_1 and ζ_2:

$$\zeta_1 = \mathbb{E}_X \left[\mathbb{E}_Y h(X,Y) \right]^2 + \frac{1}{(n-p)^2} \mathbb{E}_X \left[\mathbb{E}_Y K_m^*(X-Y) \right]^2 - \frac{2}{n-p} \mathbb{E}_X \left[\mathbb{E}_Y h(X,Y) \mathbb{E}_Y K_m^*(X-Y) \right],$$

$$\zeta_2 = \mathrm{Var}\left[h(X,Y) \right] + \frac{1}{(n-p)^2} \mathrm{Var}\left[K_m^*(X-Y) \right] - \frac{2}{n-p} \left\{ \mathbb{E}\left[h(X,Y) K_m^*(X-Y) \right] - \mathbb{E} h(X,Y) \mathbb{E} K_m^*(X-Y) \right\}.$$

Gathering the terms according to their $1/(n-p)$ exponent, we recover the desired expression.

Proof Proposition 3.4.4

This proposition comes from the following equalities:

$$\mathbb{E}\left[\sum_{i=1}^n Y_i^2 \mathbb{1}_{I_\lambda}(X) \right] = \sum_{i=1}^n \sigma_i^2 \mathbb{1}_{I_\lambda}(X_i) + \sum_{i=1}^n s^2(X_i) \mathbb{1}_{I_\lambda}(X_i),$$

$$\mathbb{E}\left[\sum_{i=1}^n Y_i \mathbb{1}_{I_\lambda}(X) \right]^2 = \sum_{i=1}^n \sigma_i^2 \mathbb{1}_{I_\lambda}(X_i) + \left(\sum_{i=1}^n s(X_i) \mathbb{1}_{I_\lambda}(X_i) \right)^2,$$

which enables to conclude for regressograms.
Similar calculations lead to the result for kernel-based estimator.

3.5.3 Lpo penalty

Proofs of Lemmas 3.4.2 and 3.4.3

Proof. (Lemma 3.4.2) The elementary calculations in the proof of Proposition 3.4.4 lead to

$$\mathbb{E} S_{\lambda,2} = n_\lambda \left[\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2 + \beta_\lambda^2 \right], \qquad \text{and} \qquad \mathbb{E} S_{\lambda,2} = n_\lambda \left(\sigma_\lambda^r\right)^2 + n_\lambda^2 \beta_\lambda^2,$$

with

$$\left(\sigma_\lambda^r\right)^2 = 1/n_\lambda \sum_{i=1}^n \sigma_i^2 \mathbb{1}_{I_\lambda}(X_i), \qquad \text{and} \qquad \left(\sigma_\lambda^b\right)^2 = 1/n_\lambda \sum_{i=1}^n \left(s(X_i) - s_m(X_i) \right)^2 \mathbb{1}_{I_\lambda}(X_i),$$

$$\beta_\lambda = 1/n_\lambda \sum_{i=1}^n s(X_i) \mathbb{1}_{I_\lambda}(X_i).$$

We recall that s_m denotes the orthogonal projection of s onto S_m and that $s_m = \sum_\lambda \beta_\lambda \mathbb{1}_{I_\lambda}$.
Now, let compute $\mathbb{E}\left[\mathrm{pen}_{id}(m) \right]$. First, we notice that

$$\mathbb{E}\left[P\gamma(\widehat{s}_m) \right] = \sum_\lambda \left\{ \frac{1}{n} \sum_i \left(\sigma_i^2 + \mathbb{E}\left[\left(s(X_i) - \widehat{\beta}_\lambda \right)^2 \right] \right) \mathbb{1}_{I_\lambda}(X_i) \right\},$$

$$\mathbb{E}\left[P_n\gamma(\widehat{s}_m) \right] = \sum_\lambda \left\{ \frac{1}{n} \sum_i \left(\sigma_i^2 + \mathbb{E}\left[\left(s(X_i) - \widehat{\beta}_\lambda \right)^2 \right] \right) \mathbb{1}_{I_\lambda}(X_i) + \frac{1}{n} \sum_i \mathbb{E}\left[2\epsilon_i \left(s(X_i) - \widehat{\beta}_\lambda \right) \right] \mathbb{1}_{I_\lambda}(X_i) \right\}.$$

Hence,

$$\mathbb{E}\left[\,\mathrm{pen}_{id}(m)\,\right] = \frac{2}{n}\sum_{\lambda}\left(\sigma_\lambda^r\right)^2 .$$

Furthermore,

$$\mathbb{E}\left[\,P_n\gamma(\widehat{s}_m)\,\right] = \sum_{\lambda}\left\{\frac{1}{n}\left(n_\lambda\left(\sigma_\lambda^r\right)^2 + \sum_i \mathbb{E}\left[\left(s(X_i)-\widehat{\beta}_\lambda\right)^2\right]\mathbb{1}_{I_\lambda}(X_i)\right) - \frac{2}{n}\left(\sigma_\lambda^r\right)^2\right\},$$

and

$$\mathbb{E}\left(s(X_i)-\widehat{\beta}_\lambda\right)^2 = \mathbb{E}\left(s(X_i)-s_m(X_i)\right)^2 + \frac{1}{n_\lambda^2}\sum_{j=1}^n \sigma_j^2\mathbb{1}_{I_\lambda}(X_j).$$

Hence, we recover the result

$$\mathbb{E}\left[\,P_n\gamma(\widehat{s}_m)\,\right] = \frac{1}{n}\sum_{\lambda}(n_\lambda-1)\left(\left(\sigma_\lambda^r\right)^2 + \frac{n_\lambda}{n_\lambda-1}\left(\sigma_\lambda^b\right)^2\right).$$

$\square$

Proof. (Lemma 3.4.3)
By definition, $V_\lambda(0) = \mathbb{P}\left[\,X \in \{1\vee p-n_\lambda,\dots,n_\lambda\wedge n-p\}\,\right]$, where $X \sim \mathcal{H}(n_\lambda, n-p, n)$. Moreover, $\mathbb{P}\left[\,X \in \{0\vee p-n_\lambda,\dots,n_\lambda\wedge n-p\}\,\right] = 1$ and $\mathbb{P}\left[\,X \in \{1\vee p-n_\lambda,\dots,n_\lambda\wedge n-p\}\,\right] = \sum_{r=1\vee(p-n_\lambda)}^{n_\lambda\wedge(n-p)}\mathbb{P}(X=r)$, where 0 has been excluded, whence $V_\lambda(0) = 1-\mathbb{P}(X=0)\,\mathbb{1}_{(n_\lambda\le p)}$. Other equalities follow from the definition of $V_\lambda(k)$.
On the other hand, $g:\mathbb{R}_+\to\mathbb{R}_+^*$ such that $g(t)=\mathbb{1}_{\mathbb{R}_+^*}(t)/t$ is convex a.e.. A simple application of Jensen inequality to g provides that $g(\mathbb{E}X)\le\mathbb{E}g(X)$, which implies $1\le V_\lambda(1)V_\lambda(-1)$.

$\square$

Proof of Proposition 3.4.6

Assume that model m satisfies $\forall\lambda\in\Lambda(m)$, $n_\lambda\le 2$. Then, a slight modification of Proposition 3.4.6 provides

$$\begin{aligned}\widehat{R}_p(m) = \sum_{\lambda\in m}\frac{1}{pN_\lambda}\Bigg\{&\mathbb{1}_{(n_\lambda\ge3)}\left(S_{\lambda,2}-\frac{S_{\lambda,1}^2}{n_\lambda}\right)\left[V_\lambda(0)-\frac{V_\lambda(1)}{n_\lambda-1}+\frac{n_\lambda}{n_\lambda-1}V_\lambda(-1)\right]+\\ &\mathbb{1}_{(n_\lambda=2)}\left(S_{\lambda,2}\left[\frac{n_\lambda^2+1}{n_\lambda(n_\lambda-1)}V_\lambda(0)-\frac{n_\lambda+1}{n_\lambda(n_\lambda-1)}V_\lambda(1)+V_\lambda(-1)\right]-2S_{\lambda,1}^2\left[\frac{V_\lambda(0)}{n_\lambda-1}-\frac{V_\lambda(1)}{n_\lambda(n_\lambda-1)}\right]\right)\Bigg\}.\end{aligned}$$

Thus, we see that assuming that every $n_\lambda\ge3$ simplify the above expression, which provides

$$\mathbb{E}\left[\widehat{R}_p(m)\right] = \sum_{\lambda\in m}\frac{1}{pN_\lambda}\left(\mathbb{E}S_{\lambda,2}-\frac{\mathbb{E}S_{\lambda,1}^2}{n_\lambda}\right)\left[V_\lambda(0)-\frac{V_\lambda(1)}{n_\lambda-1}+\frac{n_\lambda}{n_\lambda-1}V_\lambda(-1)\right].$$

We then use Lemma 3.4.2 to get

$$\mathbb{E}\left[\widehat{R}_p(m)\right] = \sum_{\lambda\in m}\frac{1}{pN_\lambda}\left(\left(\sigma_\lambda^r\right)^2+\frac{n_\lambda}{n_\lambda-1}\left(\sigma_\lambda^b\right)^2\right)\left[\,(n_\lambda-1)V_\lambda(0)-V_\lambda(1)+n_\lambda V_\lambda(-1)\,\right].$$

From now on, we distinguish two cases: $p<n_\lambda$ and $p\ge n_\lambda$ for every $\lambda\in\Lambda(m)$.

1- Let assume that $p < n_\lambda$ for every λ. Then, $N_\lambda = 1 = V_\lambda(0)$ and we have

$$\begin{aligned}\mathbb{E}\left[\widehat{R}_p(m)\right] &= \sum_{\lambda\in m} \frac{1}{p}\left((\sigma_\lambda^r)^2 + \frac{n_\lambda}{n_\lambda - 1}\left(\sigma_\lambda^b\right)^2\right)\left[n_\lambda - 1 - \frac{n_\lambda(n-p)}{n} + \frac{n}{(n-p)}V_\lambda(1)V_\lambda(-1)\right],\\ &= \sum_{\lambda\in\Lambda(m)} \frac{1}{p}\left((\sigma_\lambda^r)^2 + \frac{n_\lambda}{n_\lambda - 1}\left(\sigma_\lambda^b\right)^2\right)\left[n_\lambda - 1 - \frac{n_\lambda(n-p)}{n} + \frac{n}{(n-p)}V_\lambda(1)V_\lambda(-1)\right].\end{aligned}$$

Another use of Lemma 3.4.2 provides us with

$$\begin{aligned}\mathbb{E}\left[\operatorname{pen}_p(m)\right] &:= \mathbb{E}\left[\widehat{R}_p(m)\right] - \mathbb{E}\left[P_n\gamma(\widehat{s}_m)\right],\\ &= \sum_{\lambda\in m}\left((\sigma_\lambda^r)^2 + \frac{n_\lambda}{n_\lambda - 1}\left(\sigma_\lambda^b\right)^2\right)\times\\ &\qquad\qquad\left[\frac{n_\lambda - 1}{p} - \frac{n_\lambda - 1}{n} - \frac{n_\lambda(n-p)}{np} + \frac{n}{p(n-p)}V_\lambda(1)V_\lambda(-1)\right],\\ &= \sum_{\lambda\in\Lambda(m)}\left((\sigma_\lambda^r)^2 + \frac{n_\lambda}{n_\lambda - 1}\left(\sigma_\lambda^b\right)^2\right)\left[\frac{p-n}{np} + \frac{n}{p(n-p)}V_\lambda(1)V_\lambda(-1)\right].\end{aligned}$$

Finally, both Lemma 3.4.2 and Lemma 3.4.3 give the expected result.

2- If we now assume $p \geq n_\lambda$, $N_\lambda < 1$, set

$$\begin{aligned}A_\lambda &= \frac{n_\lambda - 1}{p}\left((\sigma_\lambda^r)^2 + \frac{n_\lambda}{n_\lambda - 1}\left(\sigma_\lambda^b\right)^2\right),\\ B_\lambda &= -\frac{V_\lambda(1)}{n_\lambda - 1} + \frac{n_\lambda}{n_\lambda - 1}V_\lambda(-1).\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}\left[\widehat{R}_p(m)\right] &= \sum_\lambda \frac{A_\lambda}{N_\lambda}\left[N_\lambda + B_\lambda\right],\\ &= \sum_\lambda A_\lambda\left[1 + B_\lambda\right] + \left(\sum_\lambda \frac{A_\lambda}{N_\lambda}\left[N_\lambda + B_\lambda\right] - \sum_\lambda A_\lambda\left[1 + B_\lambda\right]\right),\\ &= \sum_\lambda A_\lambda\left[1 + B_\lambda\right] + \sum_\lambda A_\lambda B_\lambda\left(\frac{1}{N_\lambda} - 1\right)\cdot\end{aligned}$$

Since $N_\lambda < 1$, a sufficient condition for the preceding result to remain true is that $B_\lambda \geq 0$, which means

$$-\frac{n_\lambda(n-p)}{n} + \frac{n}{(n-p)}V_\lambda(1)V_\lambda(-1) \quad\geq\quad 0.$$

This holds in the particular case when

$$p \quad\geq\quad \frac{(\sqrt{a_\lambda} - 1)}{\sqrt{a_\lambda}}n,$$

where $a_\lambda = n_\lambda / V_\lambda(1)V_\lambda(-1)$.

Bibliography

[1] H. Akaike. Statistical predictor identification. *Ann. Inst. Statisti. Math.*, 22:203–217, 1969.

[2] S. Arlot. Model selection by resampling penalization. *Electronic journal of Statistics*, 00:00, 2008.

[3] Sylvain Arlot. *Resampling and Model Selection.* PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[4] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with unknown variance. *The Annals of Statistics*, 00:00, 2008.

[5] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[6] P. Barbe and P. Bertail. *The Weighted Bootstrap.* Lecture Notes in Statistics. Springer-Verlag, New York, 1995.

[7] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1–3):85–113, 2002.

[8] R. Beran. Discussion of "Jackknife, bootstrap and other resampling methods in regression analysis" by C. F. Wu. *The Annals of Statistics*, 14:1295–1298, 1986.

[9] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgensen, and G. Yang, editors, *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

[10] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531, 2008.

[11] G. Blanchard and P. Massart. Discussion: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2664–2671, 2006.

[12] O. Bousquet and A. Elisseff. Stability and Generalization. *J. Machine Learning Research*, 2:499–526, 2002.

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

[14] L. Breiman. The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. *J. Amer. Statist. Assoc.*, 87(419):738–754, 1992.

[15] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[16] L. Breiman. Stacked Regression. *Machine Learning*, 24:49–64, 1996.

[17] P. Burman. Comparative study of Ordinary Cross-Validation, v-Fold Cross-Validation and the repeated Learning-Testing Methods. *Biometrika*, 76(3):503–514, 1989.

[18] P. Burman. Estimation of optimal transformation using v-fold cross-validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3):314–245, 1990.

[19] P. Burman, E. Chow, and D. Nolan. A Cross-validatory method for dependent data. *Biometrika*, 81(2):351–358, 1994.

[20] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

[21] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[22] J.-J. Daudin and T. Mary-Huard. Estimation of the conditional risk in classification: The swapping method. *Comput. Stat. Data Anal.*, 52(6):3220–3232, 2008.

[23] V. H. de la Peña and E. Giné. *Decoupling: From Dependence to independence*. Springer-Verlag, New York, 1999.

[24] L. Devroye and T. J. Wagner. Distribution-Free performance Bounds for Potential Function Rules. *IEEE Transaction in Information Theory*, 25(5):601–604, 1979.

[25] S. Dudoit and M. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

[26] B. Efron. Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[27] B. Efron. The jackknife, the bootstrap and other resampling plans. volume 38 of *CBMS-NSFRegional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

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

[29] B. Efron. How biased is the Apparent Error Rate of a Prediction Rule? *J. Amer. Statist. Assoc.*, 81(394):461–470, 1986.

[30] T. Evgeniou, M. Pontil, and A. Elisseff. Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers. *Machine Learning*, 55:71–97, 2004.

[31] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66(2–3):165–207, 2006.

[32] S. Geisser. A predictive approach to the random effect model. *Biometrika*, 61(1):101–107, 1974.

[33] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[34] C. Genovese and L. Wasserman. Confidence sets for nonparametricwavelet regression. *The Annals of Statistics*, 33(2):698–729, 2005.

[35] E. Giné. *Lectures on some aspects of the bootstrap. In Lectures on Probability and Statistics: Ecole d'été de Probabilité de Saint-Flour, XXVI-1996*, volume 1665 of *Lecture Notes in Math.* Springer-Verlag, Berlin, 1997.

[36] L. Györfi, W. Härdle, W. Sarda, and P. Vieu. *Nonparametric Curve Estimation from Time Series.* Springer-Verlag, New York, 1989.

[37] P. Hall. Large Sample Optimality of Least Squares Cross-Validation in Density Estimation. *The Annals of Statistics*, 11(4):1156–1174, 1983.

[38] P. Hall. On the bootstrap and confidence intervals. *The Annals of Statistics*, 14(4):1431–1452, 1986.

[39] P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15(4):1491–1519, 1987.

[40] P. Hall. *The Booststrap and Edgeworth Expansion.* Springer-Verlag, New York, 1992.

[41] P. Hall and E. Mammen. On general resampling algorithms and their performance in distribution estimation. *The Annals of Statistics*, 22(4):2011–2030, 1994.

[42] W. Härdle and E. Mammen. Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21(4):1926–1947, 1993.

[43] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, 2001.

[44] P. A. Herzberg. The parameters of cross-validation. *Psychometrika*, 34:Monograph Supplement, 1969.

[45] M. Hills. Allocation Rules and their Error Rates. *J. Royal Statist. Soc. Series B*, 28(1):1–31, 1966.

[46] P. Horst. *The Prediction of Persoanl Adjustment.* New York Social Science Research Council, 1941.

[47] M. Kearns. A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split. *Neural Comput.*, 9(5):1143–1161, 1997.

[48] M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning*, 27:7–50, 1997.

[49] M. Kearns and D. Ron. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, 11:1427–1453, 1999.

[50] V. Koltchinkii. Penalties and structural risk minimization. *IEEE Tans. Inf. Theory*, 47:1902–1914, 2001.

[51] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[52] P. A. Lachenbruch and M. R. Mickey. Estimation of Error Rates in Discriminant Analysis. *Technometrics*, 10(1):1–11, 1968.

[53] S. C. Larson. The shrinkage of the coefficient of multiple correlation. *J. Edic. Psychol.*, 22:45–55, 1931.

[54] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

[55] K.-C. Li. Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975, 1987.

[56] Regina Y. Liu. Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.*, 16(4):1696–1708, 1988.

[57] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864, 1999.

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

[59] D. M. Mason and M. A. Newton. A rank statistics approach to the consistency of a general boostrap. *The Annals of Statistics*, 20(3):1611–1624, 1992.

[60] P. Massart. *Concentration Inequalities and Model Selection.* Lecture Notes in Mathematics. Springer, 2007.

[61] F. Mosteller and J. W. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology, Vol. 2.* Addison-Wesley, 1968.

[62] R. R. Picard and R. D. Cook. Cross-Validation of Regression Models. *J. Amer. Statist. Assoc.*, 79(387):575–583, 1984.

[63] D. N. Politis and J. P. Romano. Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions. *The Annals of Statistics*, 22(4):2031–2050, 1994.

[64] J. Praestgaard and J. A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21:2053–2086, 1993.

[65] M. H. Quenouille. Approximate tests of correlation in time series. *J. Royal Statist. Soc. Series B*, 11:68–84, 1949.

[66] J. Rissanen. Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431, 1983.

[67] D. B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

[68] M. Rudemo. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.

[69] S. R. Sain, K. A. Baggerly, and D. W. Scott. Cross-Validation of Multivariate Densities. *Journal of the American Statistical Association*, 89(427):807–817, 1994.

[70] J. Shao. Model Selection by Cross-Validation. *Journal of the American Statis. Association*, 88(422):486–494, 1993.

[71] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[72] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[73] C.J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.

[74] 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.

[75] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *JRSS B*, 39(1):44–47, 1977.

[76] R. L. Taylor, P. Daffer, and R. F. Patterson. *Limit theorems for sums of exchangeable random variables*. Amer. Math. Soc., 1988.

[77] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. Royal Statist. Soc. Series B*, 58(1):267–288, 1996.

[78] A. B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Mathématiques et Applications. Springer-Verlag, 2003.

[79] J. W. Tukey. Bias and confidence in not quite large samples (Abstract). *Ann. Math. Statist.*, 29:614, 1958.

[80] M. van der Laan, S. Dudoit, and S. Keles. Asymptotic Optimality of Likelihood Based Cross-Validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 4, 2004.

[81] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[82] V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[83] M. Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003.

[84] Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1350, 1986. With discussion and a rejoinder by the author.

[85] Y. Yang. Adaptive Regression by Mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588, 2001.

[86] Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

[87] P. Zhang. Model selection via multifold cross-validation. *The Annals of Statistics*, 21(1):299–313, 1993.

[88] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.

Chapter 4

Optimal risk estimation *via* cross-validation in density estimation

Here is a paper that has been published in Comput. Statist. and Data Analysis in 2008. The underlying idea is to provide an "optimal" estimator of the risk of a given family of histograms in order to perform model selection.
We first derive a closed-form expression for the leave-p-out estimator of the risk in the particular case of kernel estimator and histogram. Plug-in estimators of its bias and variance are deduced so that we define the best $\widehat{p}$ of the leave-p-out estimator among $1 \leq p \leq n-1$ as that one which provides the smallest estimated mean square error.
It turns out that we are able to derive closed-form expressions for all the involved quantities, which results in a computationally reasonable data-driven procedure.
Some simulations are carried out to highlight the better behaviour of the leave-p-out with respect to the V-fold cross-validation as well as to the leave-one-out in the multiple testing framework. Following the strong requirement of a referee, we have also performed a simulation with completely irregular histograms (of exponential complexity), which results in the better behaviour of the proposed procedure with respect to the leave-one-out.

4.1 Abstract

The problem of density estimation is addressed by minimization of the L^2-risk for both histogram and kernel estimators. This quadratic risk is estimated by leave-p-out cross-validation (Lpo), which is made possible thanks to closed-form formulas, contrary to common belief. The potential gain in the use of Lpo with respect to V-fold cross-validation (V-fold) in terms of the bias-variance tradeoff is highlighted. An exact quantification of this extra variability, induced by the preliminary random partition of the data in the V-fold, is proposed. Furthermore exact expressions are derived for both the bias and the variance of the risk estimator with histograms. Plug-in estimates of these quantities are provided, while their accuracy is assessed thanks to concentration inequalities. An adaptive selection procedure for p in the case of histograms is subsequently presented. This relies on minimization of the mean square error of the Lpo risk estimator. Finally a simulation study is carried out which first illustrates the higher reliability of

the Lpo with respect to the V-fold, and then assesses the behavior of the selection procedure. For instance optimality of leave-one-out (Loo) is shown, at least empirically, in the context of regular histograms.

4.2 Introduction

Histograms are a widespread tool in the context of nonparametric density estimation thanks to their simple interpretability. For regular histograms it is known that the asymptotic optimal partition width in terms of risk minimization is of the order $n^{-1/3}$, especially in the case of L^2-loss [5]. Some work has been done for other loss functions with similar results (see [3] for L^1-loss for example). [15] presented a selection procedure based on leave-one-out (Loo) and essentially regular histograms. He gave analytical expression for his Loo estimators but did not assess the performance of his procedure except by simulations. When such closed-form formulas no longer exist, some efficient algorithms may sometimes be proposed (see [8]).
As their approximation properties are clearly better than those of histograms, kernel estimators have already been widely studied in the context of density estimation. Their convergence rates outperform those of histograms and they are especially well suited for smooth densities. In the aforementioned article, [15] already gives exact expressions of kernel based risk estimates, but only in the Loo case. A large part of the literature is devoted to the search for an "optimal" bandwidth. Nevertheless, the lack of a closed-form and tractable formula for the risk estimator usually leads only to asymptotic considerations which are sometimes unsatisfactory (see [12]).
The present paper has two purposes. In Section 2, we provide closed-form formulas for the leave-p-out (Lpo) estimator of the L^2-risk, in the case of both histogram and kernel estimators. These formulas and the following results can be easily extended to the multiple dimension case. In the specific case of histograms, analytical expressions of both the bias and the variance of the Lpo risk are given with an assessment of their accuracy. Secondly, we see in Section 3 how much better the Lpo estimator may be with respect to the V-fold by means of exact calculations as well. We therefore furnish a quantification of the unwanted variability induced by the preliminary partition of the data in the V-fold. Trying to hit the best bias-variance tradeoff in the specific case of histograms, we present a selection procedure for p that relies on the minimization of the mean square error (MSE) of the risk estimator. In Section 4, we show an illustration of how unduly variable the V-fold kernel estimator may be. We also discuss simulation results about density estimation. As for regular histograms, an empirical optimality result is shown for the Loo, whereas better results are obtained for Lpo ($p \neq 1$) in the case of non-regular histograms for instance.

4.3 Closed-form expressions for the Lpo risk estimator

4.3.1 Framework

We consider n independent identically distributed random variables $X_1, \ldots, X_n$ with values in the probability space $(\mathcal{X}, \mathcal{B}, \mu)$, where μ is the Lebesgue measure. Let us assume that their common distribution P is absolutely continuous with respect to μ and denote by s its density w.r.t μ. As it depends on the context, $\mathcal{X}$ is to be precisely defined later.
Set $\widehat{s}$ any estimator belonging to a given class $\mathcal{S}$. We aim at estimating s, that is at finding the "closest" estimator to s among $\mathcal{S}$ in terms of minimization of the L^2-risk:

$$R(\widehat{s}) = \mathbb{E}_s \left[\| s - \widehat{s} \|_2^2 \right].$$

Ideally, we would like to find s^* such that

$$\begin{aligned} s^* = \mathrm{Argmin}_{\widehat{s} \in \mathcal{S}} R(\widehat{s}) &= \mathrm{Argmin}_{\widehat{s} \in \mathcal{S}}\, \mathbb{E}_s \left[\| s \|_2^2 + \| \widehat{s} \|_2^2 - 2 \int_{\mathcal{X}} s(x)\, \widehat{s}(x)\, dx \right], \\ &= \mathrm{Argmin}_{\widehat{s} \in \mathcal{S}} L(\widehat{s}), \end{aligned} \tag{4.1}$$

where

$$L(\widehat{s}) = \mathbb{E}_s \left[\| \widehat{s} \|_2^2 - 2 \int_{\mathcal{X}} s(x)\, \widehat{s}(x)\, dx \right]. \tag{4.2}$$

Note that s^* is obviously unreachable due to its dependence on s.

4.3.2 Cross-validation estimators

For example in the kernel density estimation context, looking for the optimal bandwidth commonly involves the leave-one-out cross-validation (Loo) in order to estimate the risk function. This procedure consists in splitting the observations into two parts. The $n-1$ first observations are used to compute a well chosen estimator, whereas the last observation is devoted to the assessment of the accuracy of this estimator. We run that process n times, corresponding to the n possible choices of a subsample of size $n-1$ among n observations. We finally average the n resulting quantities, thus obtaining the Loo-risk estimate [6]. This method provides computation-time tractable estimators that are nearly or totally unbiased (see respectively (3.4) and (2.8) in [15]). Nevertheless the resulting estimation is known to suffer some variability which could attain undesirable levels (see [4]). An alternative method is the leave-p-out cross-validation (Lpo). Similarly to the Loo, observations are split into two subsets of cardinality $n-p$ and p, that are used respectively for computing the estimator and measuring its performance. The average of the $\binom{n}{p}$ obtained quantities provides an estimator referred to as the Lpo estimate.

Definition 4.3.1 (Lpo risk estimator)**.** *Set $X_1, \ldots, X_n$ n independent identically distributed random variables. For $p \in \{1, \ldots, n-1\}$, let $\mathcal{E}_p$ be the set of all possible p-subsets of $\{1, \ldots, n\}$. For any $e \in \mathcal{E}_p$, $\overline{e} = \{1, \ldots, n\} \setminus e$ and $X^{\overline{e}} = \{X_i / \, i \in \overline{e}\}$. With the same notations as before,*

$$\widehat{L}_p(\widehat{s}\,) = \frac{1}{\binom{n}{p}} \sum_{e \in \mathcal{E}_p} \left[\| \, \widehat{s}^{\,\overline{e}} \, \|_2^2 - 2 \, \frac{1}{p} \sum_{i \in e} \widehat{s}^{\,\overline{e}}(X_i) \right],$$

where $\widehat{s}^{\,\overline{e}}$ denotes any estimator of $\mathcal{S}$ built from $X^{\overline{e}}$.

Generally, this estimator is computation-time prohibitive due to the $\binom{n}{p}$ subsamples we have to visit. Moreover, we have to keep in mind the widespread intuition that the larger p is, the less variable but the more biased the Lpo estimator should be.
Lpo estimation being currently held as untractable, the $V-$fold cross-validation (V-fold) is the usual adopted compromise: preventing excessive variability and being weakly biased (see [4] and [10]). It consists in randomly partitioning the data into V subsets of equal size ($V = n/p$). One of them being fixed, the estimator is computed on the others and its performance is evaluated on the fixed one. Averaging the V resulting quantities (one for each fixed element of the partition), we get the named V-fold estimator. We underline that unlike the Lpo case, the V subsets are disjoint. In addition, the essential drawback of this method is its dependence on a preliminary random splitting of the data into V groups (V being determined by the practitioner), which introduces unwanted randomness.

4.3.3 Explicit expression for the Lpo risk estimator

Histograms and kernels

In the sequel when considering histograms, let $\mathcal{M} = \mathcal{M}_D$ be the set of all possible partitions of $\mathcal{X} = [0,1]$ in D intervals. For any $m \in \mathcal{M}$, $m = (I_k)_{k=1,\ldots,D}$, where the intervals I_k are ordered from left to right, and for any $k \in \{1, \ldots, D\}$, $\omega_k = |I_k|$ denotes the length of I_k. We set

$$\Omega = \Omega_D = \{\omega \in [0,1]^D / \, \forall k = 1, \ldots, D, \; \omega_k = |I_k|, \; I_k \in m, \; m \in \mathcal{M}\}.$$

The associated histogram is defined as

$$\widehat{s}_\omega = \sum_{k=1}^{D} \frac{n_k}{n\omega_k} \mathbb{1}_{I_k}, \tag{4.3}$$

where n_k denotes the number of observations in I_k.
For sake of simplicity, we only consider $\mathcal{X} = [0,1]$. However, the following results can be extended to $\mathcal{X} = [0,1]^d$. Note that we do not require histograms to be regular. Although from a computational point of view we cannot work with completely irregular ones, the collection of partitions we visit may be made of irregular ones of a special type, as we will see in Section 4.5.2 for instance.
In short, ω will abusively be referred to as the partition to which it corresponds.
In the following, we refer to kernels as those introduced by [14] and [13] defined on $\mathcal{X} = \mathbb{R}$. As above for

histograms, results about kernels can be easily generalized to $\mathcal{X} = \mathbb{R}^d$. For any such kernel K, h is called the bandwidth and $K_h(\cdot) = 1/h\, K(\cdot/h)$. The corresponding density estimator is for any $h > 0$,

$$\widehat{s}_h(\cdot) = \frac{1}{n}\sum_{i=1}^{n} K_h(\cdot - X_i)\,. \tag{4.4}$$

$\mathcal{S}$ denotes the set of all possible estimators, *i.e.* either histograms built from $\mathcal{M}$ or kernels, depending on what we are talking about.

Leave-p-out and variability control

We now point out the two possible ways of applying Loo to the estimation of (4.2). According to the above Loo description, we estimate $\int_{[0,1]} s(x)\widehat{s}(x)\,dx$ by $\frac{1}{n}\sum_{i=1}^{n}\widehat{s}^{\,(i)}(X_i)$, where $\widehat{s}^{\,(i)}$ denotes the estimate ($\widehat{s}^{\,(i)} \in \mathcal{S}$) computed from the $n-1$ data $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$. However two estimators of $\mathbb{E}_s||\widehat{s}||_2^2$ may be proposed, so that we can define two Loo estimators of $L(\widehat{s}\,)$:

$$\widehat{L}_1^{(1)}(\widehat{s}) = \quad \parallel \widehat{s} \parallel_2^2 \quad - \quad 2\frac{1}{n}\sum_{i=1}^{n}\widehat{s}^{\,(i)}(X_i), \tag{4.5}$$

$$\widehat{L}_1^{(2)}(\widehat{s}) = \frac{1}{n}\sum_{i=1}^{n}\parallel \widehat{s}^{\,(i)} \parallel_2^2 \quad - \quad 2\frac{1}{n}\sum_{i=1}^{n}\widehat{s}^{\,(i)}(X_i). \tag{4.6}$$

To quantify the relationship between the two possible estimators, [15] gave the following

Lemma 4.3.1. *In the case of a regular D-piece histogram ($\omega = 1/D$) with the above notations,*

$$\widehat{L}_1^{(1)}(\omega) \quad = \quad \frac{2}{(n-1)\omega} - \frac{n+1}{(n-1)\omega}\sum_{k=1}^{D}\left(\frac{n_k}{n}\right)^2, \tag{4.7}$$

$$\widehat{L}_1^{(2)}(\omega) \quad = \quad \frac{2n-1}{(n-1)^2\omega} - \frac{n^2}{(n-1)^2\omega}\sum_{k=1}^{D}\left(\frac{n_k}{n}\right)^2. \tag{4.8}$$

Then,

$$\widehat{L}_1^{(1)}(\omega) - \widehat{L}_1^{(2)}(\omega) = \mathcal{O}\left(\frac{1}{n^2}\right).$$

Provided n is large enough, both these estimators should give similar results.
The Loo estimator is known for its possible high-level of variability. To prevent this drawback, we consider Lpo estimation which enables us to control the variance of our estimator by means of the choice of p. As noticed by [**?**] but only in the case of the Loo, applying cross-validation to histograms provides us with an explicit formula for the L^2-risk estimator. We extend this remark in the case of the Lpo estimation in which we get explicit theoretical expression for the risk estimate. First, we give some notations and the definition of the two possible Lpo estimators of the risk.

Definition 4.3.2 (Lpo risk estimators)**.** *Set $X_1, \ldots, X_n$ n i.i.d. random variables. For $p \in \{1, \ldots, n-1\}$, let $\mathcal{E}_p$ be the set of all possible p-subsets of $\{1, \ldots, n\}$. For any $e \in \mathcal{E}_p$, $\overline{e} = \{1, \ldots, n\} \setminus e$ and $X^{\overline{e}} = \{X_i / \, i \in \overline{e}\}$. With the same notations as before,*

$$\widehat{L}_p^{(1)}(\widehat{s}) \quad = \quad \parallel \widehat{s} \parallel_2^2 -2\binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\quad \frac{1}{p}\sum_{i\in e}\widehat{s}^{\,\overline{e}}(X_i), \tag{4.9}$$

$$\widehat{L}_p^{(2)}(\widehat{s}) \quad = \quad \binom{n}{p}^{-1}\sum_{e\in\mathcal{E}_p}\left[\parallel \widehat{s}^{\,\overline{e}} \parallel_2^2 -2\,\frac{1}{p}\sum_{i\in e}\widehat{s}^{\,\overline{e}}(X_i)\right], \tag{4.10}$$

where $\widehat{s}^{\,\overline{e}}$ denotes any estimator of $\mathcal{S}$ built from $X^{\overline{e}}$.

We clearly see the analogy with the expressions (4.5) and (4.6) in the case of the Loo and like [15], we get explicit theoretical formulas for those estimates.

Exact Lpo risk estimator for histograms and kernels

First of all, we give the following combinatorial lemma.

Lemma 4.3.2. *Set a $D-$partition (I_k), $e \in \mathcal{E}_p$ and $X^e = \{X_i / \, i \in \overline{e}\}$. For any $k \in \{1, \ldots, N\}$, define $n_k^e = |\{i / \, X_i \in X^e \cap I_k\}|$ and $n_k^{\overline{e}} = n_k - n_k^e$.*
Then for $p \in \{1, \ldots, n-1\}$,

$$\begin{aligned} \forall k \in \{1, \ldots, N\}, \qquad \sum_{e \in \mathcal{E}_p} n_k^e \, n_k^{\overline{e}} &= n_k(n_k-1)\binom{n-2}{p-1} , \\ \sum_{e \in \mathcal{E}_p} (\, n_k^{\overline{e}}\,)^2 &= n_k \binom{n-1}{p} + n_k(n_k-1)\binom{n-2}{p} . \end{aligned}$$

From these elementary relations, inverting the sums yields

Theorem 4.3.1 (Lpo risk estimators for histograms)**.** *For any $p \in [\![1, n-1]\!]$, and any $\omega \in \Omega$,*

$$\widehat{L}_p^{(1)}(\omega) = \widehat{L}_1^{(1)} = \frac{2}{n-1} \sum_{k=1}^{N} \frac{n_k}{n\omega_k} - \frac{n+1}{n-1} \sum_{k=1}^{N} \frac{1}{\omega_k} \left(\frac{n_k}{n}\right)^2 , \tag{4.11}$$

$$\widehat{L}_p^{(2)}(\omega) = \frac{2n-p}{(n-1)(n-p)} \sum_{k=1}^{N} \frac{n_k}{n\omega_k} - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_{k=1}^{N} \frac{1}{\omega_k} \left(\frac{n_k}{n}\right)^2 . \tag{4.12}$$

In (4.11), the independence with respect to p prevents us from controlling the variance of our estimator as we were expecting to. Using Lpo only on the integral term in (4.2) ends up being the same as applying Loo. Such a behavior is essentially the consequence of linearity in (4.9). Introducing non-linear term in (4.12) with the L^2 norm, we obtain some dependence on p that will be useful to control the variability. Unlike the Loo case, large values of p could induce a significant discrepancy between $\widehat{L}_p^{(2)}(\omega)$ and $\widehat{L}_p^{(1)}(\omega)$. From now on as we want to control the variance of the risk estimator by an appropriate choice of p, we consider $\widehat{L}_p^{(2)}(\omega)$ as the Lpo risk estimator for histograms for a given partition ω. In the following, $\widehat{L}_p(\omega) = \widehat{L}_p^{(2)}(\omega)$. Note that such a closed-form formula is all the more interesting as it avoids a large amount of computations: we do not have to compute an estimator from the training set and then assess its performance on the remaining data at each step.
As for kernels, similar calculations lead to

Theorem 4.3.2 (Lpo risk estimators for kernels)**.** *Following the definition (4.4), set K a kernel and $h > 0$, its bandwidth. Then for any $p \in \{1, \ldots, n-1\}$,*

$$\begin{aligned} \widehat{L}_p(h) = \widehat{L}_p(K,h) &= \frac{1}{(n-p)} \, ||K_h||_2^2 + \frac{n-p-1}{n(n-1)(n-p)} \sum_{i \neq j} K^*_{i,j} \\ &\quad - \frac{2}{n(n-1)} \sum_{i \neq j} K_{i,j} , \end{aligned} \tag{4.13}$$

where $K^*_{i,j} = (K * K)_h \, (X_i - X_j)$, with denoting by $*$ the convolution product and $K_{i,j} = K_h(X_i - X_j)$.

Remarks:

- The computation of each term relies only on the calculation of the matrix $(K_h(X_i - X_j))_{1 \leq i,j \leq n}$, which is the elementary piece you have to deal with whenever you use kernel-based estimates.

- In the sequel, $\widehat{L}_p(w)$ denotes the histogram risk estimator built from ω whereas $\widehat{L}_p(h)$ is its kernel counterpart, with bandwidth h.

In order to show how easy the Lpo risk estimator is to compute, we deal with the case of the widespread Gaussian kernel $K^G(.) = 1/\sqrt{2\pi}\, e^{-1/2(.)^2}$.

Lemma 4.3.3. *For any $p \in [\![1, n-1]\!] = \{1, \ldots, n-1\}$ and $h > 0$, we get*

$$
\begin{aligned}
\widehat{L}_p^G(h) &= \frac{1}{2\sqrt{\pi}(n-p)h}\left[1 + \frac{n-p-1}{n(n-1)}\sum_{i\neq j} e^{-\frac{1}{4}\left(\frac{X_i - X_j}{h}\right)^2}\right] \\
&\quad - \sqrt{\frac{2}{\pi}}\frac{1}{n(n-1)h}\sum_{i\neq j} e^{-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2}.
\end{aligned}
$$

Note that the computation of this quantity only relies on the matrix with general term $\left(\exp\left[-\left(\left(X_i - X_j\right)/h\right)^2\right]\right)_{1\leq i,j\leq n}$. At the beginning of the fourth part, we show an example of density estimation using this risk estimate.

4.3.4 Closed formula of the bias and the variance for histograms

Theoretical expressions

Thanks to the aforementioned exact expressions of the Lpo risk in the case of histograms, we are now in position to give theoretical expressions for both the bias and the variance of $\widehat{L}_p(\omega)$. The following proposition is proven in the appendix.

Proposition 4.3.1 (Exact bias and variance expressions)**.** *Let ω correspond to a $D-$partition $(I_k)_k$ of $[0,1]$ and for any $k \in \{1, \ldots, D\}$, $\alpha_k = \Pr(X_1 \in I_k)$ such that $\alpha = (\alpha_1, \ldots, \alpha_D)$. For any $(i,j) \in \{1, \ldots, 3\} \times \{1, 2\}$, $\quad s_{i,j} = \sum_{k=1}^{D} \alpha_k^i / \omega_k^j$. Then,*

$$
\begin{aligned}
B_p(\omega) &= B_p(\alpha, \omega) = \frac{p}{n(n-p)}\sum_{k=1}^{D(\omega)} \frac{\alpha_k(1-\alpha_k)}{\omega_k}, && (4.14) \\
V_p(\omega) &= V_p(\alpha, \omega) = \frac{p^2\varphi_2(n,\alpha,\omega) + p\,\varphi_1(n,\alpha,\omega) + \varphi_0(n,\alpha,\omega)}{[n(n-1)(n-p)]^2}, && (4.15)
\end{aligned}
$$

where

$$
\begin{aligned}
\varphi_2(n,\alpha,\omega) &= 2n(n-1)\left[(n-2)(s_{2,1} + s_{1,1} - s_{3,2}) - ns_{2,2} - (2n-3)s_{2,1}^2\right], \\
\varphi_1(n,\alpha,\omega) &= -2n(n-1)(3n+1)\left[(n-2)(s_{2,1} - s_{3,2}) - ns_{2,2}\right] + \\
&\quad 2n(n-1)\left[2(n+1)(2n-3)s_{2,1}^2 + (-3n^2 + 3n + 4)s_{1,1}\right], \\
\varphi_0(n,\alpha,\omega) &= 4n(n-1)(n+1)\left[(n-2)(s_{2,1} - s_{3,2}) - ns_{2,2}\right] - \\
&\quad 2n(n-1)\left[(n^2 + 2n + 1)(2n-3)s_{2,1}^2 + (2n^3 - 4n - 2)s_{1,1}\right] + \\
&\quad n(n-1)^2(s_{1,2} - s_{1,1}^2).
\end{aligned}
$$

Plug-in estimates

From expressions (4.14) and (4.15), replacing the unknown α_k by $\widehat{\alpha}_k = n_k/n$ for each k provides us with plug-in estimates.

Definition 4.3.3. *Set $\widehat{\alpha} = (\frac{n_1}{n}, \ldots, \frac{n_D}{n})$, where $\forall k \in \{1, \ldots, D\}$, $n_k = |\{i / X_i \in I_k\}|$. With the same notations as in Proposition 4.3.1, we define*

$$
\begin{aligned}
\widehat{B}_p(\omega) = B_p\left(\widehat{\alpha}, \omega\right) &= \frac{p}{n(n-p)}\sum_{k=1}^{D(\omega)} \frac{1}{\omega_k}\frac{n_k}{n}\left(1 - \frac{n_k}{n}\right), && (4.16) \\
\widehat{V}_p(\omega) = V_p\left(\widehat{\alpha}, \omega\right) &= \frac{p^2\varphi_2(n,\widehat{\alpha},\omega) + p\,\varphi_1(n,\widehat{\alpha},\omega) + \varphi_0(n,\widehat{\alpha},\omega)}{[n(n-1)(n-p)]^2}. && (4.17)
\end{aligned}
$$

We want to assess their accuracy by use of a concentration inequality for sums of independent random variables, which we just recall for the sake of sufficiency.
Note that the only requirement is the independence of the variables.

Theorem 4.3.3 (Bernstein's tail inequality). *Let $X_1, \ldots, X_n$ be independent real-valued random variables with zero mean, such that $\forall i,\ |X_i| \leq c$ a.s.. Defining $\sigma^2 = n^{-1} \sum_{i=1}^n \mathrm{Var} X_i$ and $S_n = \sum_{i=1}^D X_i$, we obtain for any $\epsilon > 0$,*

$$\Pr\left(\frac{1}{n} S_n > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2(\sigma^2 + \frac{c}{3}\,\epsilon)}\right). \tag{4.18}$$

Unlike Hoeffding's tail inequality, which does not really take the variance into account, Bernstein's bound is of the same order as $e^{-n\epsilon}$ if $\sigma^2 \ll \epsilon$, while it is of the order $e^{-n\epsilon^2}$ with Hoeffding's inequality. The former is consequently tighter when the variance of the observations may be weak, which could be the case in our statement.
Both the mean-value theorem and the above Bernstein inequality yield the following proposition. A brief proof is given in the appendix.

Proposition 4.3.2 (Accuracy of the plug-in bias estimator). *With the same notations as before, for any $p \in [\![1, n-1]\!]$ and any given $D-$partition $m \in \mathcal{M}$ associated with ω, we have*

$$\forall \epsilon > 0, \quad \Pr(|\widehat{B}_p(\omega) - B_p(\omega)| \geq \epsilon) \leq 2D \exp\left[-\frac{n^3(n-p)^2}{2p^2 M(\omega)^2 (1 + \frac{n(n-p)}{3pM(\omega)}\,\epsilon)}\,\epsilon^2\right], \tag{4.19}$$

where $M(\omega) = \sum_{k=1}^D \frac{1}{\omega_k}$.

In the worst case we have $M(\omega) \leq n^2$, as the number of elements of the expected partition does not exceed the number of observations. Nonetheless according to the Bernstein inequality, if the partition bandwidth (*i.e.* $\min_k \omega_k$) is not too small, the bound decreases like $\exp\left[-\frac{3n^2(n-p)}{2pM(\omega)}\,\epsilon\right]$.
Furthermore, p makes the bound worse as it grows: according to (4.14), a rise in p makes the scale of variation of the bias around its expectation higher, hence causing bigger fluctuations. Typically a large value of p may be chosen when s is not smooth enough, which generates larger variations. This bound may be consequently quite realistic. Using the binomial distribution of n_k and the Cramer-Chernov method, it is possible to get a similar bound.
In the same way, we can obtain such a type of bound for the plug-in variance estimator thanks to concentration inequality and the mean-value theorem. Thus, we give

Proposition 4.3.3 (Accuracy of the plug-in variance estimator). *For any $p \in [\![1, n-1]\!]$, and $\omega \in \Omega \cap [0,1]^D$, consider $\mathcal{L}([0,1]^D, \mathbb{R})$ endowed with $\|.\|$. Set $r_{n,p} = [n(n-1)(n-p)]^2$ and $\varphi(x) = r_{n,p} V_p(x, \omega)$. Defining $M(n,p,\omega) = \sup_{x \in [0,1]^D} \|D_x \varphi\|$, ($D_x\varphi \in \mathcal{L}([0,1]^D, \mathbb{R})\ \forall x$), we get $\forall \epsilon > 0$,*

$$\Pr(|\widehat{V}_p(\omega) - V_p(\omega)| \geq \epsilon) \leq 2D \exp\left[-\frac{n}{2}\left(\frac{r_{n,p}}{M(n,p,\omega)}\,\epsilon\right)^2 \left(1 + \frac{r_{n,p}}{3M(n,p,\omega)}\,\epsilon\right)^{-1}\right]. \tag{4.20}$$

The same interpretation applies to this upper bound which is derived from Bernstein's inequality. Notice that the smoothness condition is expressed here through the term $M(n,p,\omega)$, which is well defined because of the continuity of $x \in [0,1]^D \mapsto D_x\varphi \in \mathcal{L}([0,1]^D, \mathbb{R})$. It therefore means that the smoother the density, the more accurate the plug-in variance estimator. This proof is deferred to the appendix as well.

It is easily seen that the aforementioned plug-in estimators are both biased. As for the bias, a simple calculation provides the following unbiased estimator:

Lemma 4.3.4. *For any ω,*

$$\widetilde{B}_p(\omega) = \frac{p}{(n-1)(n-p)} \sum_{k=1}^{D(\omega)} \frac{\widehat{\alpha}_k(1-\widehat{\alpha}_k)}{\omega_k}.$$

Note that as the bias of the plug-in variance estimator is more intricate and involves the unknown density s, an unbiased estimate of the variance is not achievable.

4.4 Reliability of estimators

4.4.1 Exact calculation of the gap between variances of Lpo and V-fold estimators

The V-fold procedure relies on a preliminary random partitioning of the data in V subsets of cardinality $p = n/V$. Not only does it suffers from the lack of a criterion for choosing which V is preferable (this is also the case for the Lpo risk), but this supplementary randomness also induces unwanted variability. The latter might be non negligible, especially in the choice of the bandwidth h, which highlights the interest in the use of the Lpo risk estimator rather than V-fold one.
In order to give evidence of this phenomenon, we first need the following notations. Let $\{E_1, \ldots, E_V\}$ denote a random partition of $\{1, \ldots, n\}$ in V disjoint subsets of size p. For any $e \in \mathcal{E}_p$, we define the following random variables which are not independent

$$\begin{aligned} Z_e &= 1, \quad \text{if } e \in \{E_1, \ldots, E_V\}, \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

Then the V-fold risk estimator can be expressed as:

Definition 4.4.1 (V-fold risk estimator)**.** *For any kernel K and $h > 0$, we define*

$$\widehat{L}_V(h) := \frac{1}{V} \sum_{k=1}^{V} T_{E_k} = \frac{1}{V} \sum_{e \in \mathcal{E}_p} Z_e \, T_e, \tag{4.21}$$

with $T_e = ||\widehat{s}_h^{\,\overline{e}}||_2^2 - \frac{2}{p} \sum_{i \in e} \widehat{s}_h^{\,\overline{e}}(X_i), \ \forall e \in \mathcal{E}_p$.

In order to quantify the variability due to the random partition, we denote by $\mathbb{E}_Z[\,.\,]$, the expectation with respect to the Z_e variables. Thanks to the following proposition, we compare the variance of the Lpo risk estimator with that of the V-fold one:

Proposition 4.4.1. *For any $h > 0$,*

$$\begin{aligned} \mathbb{E}_Z\left[\widehat{L}_V(h)\right] &= \widehat{L}_p(h), && (4.22)\\ \mathbb{V}_{X,Z}\left(\widehat{L}_V(h)\right) &= \mathbb{E}_X\left[\mathbb{V}_Z\left(\widehat{L}_V(h)\right)\right] + \mathbb{V}_X\left[\widehat{L}_p(h)\right], && (4.23) \end{aligned}$$

and

$$\begin{aligned} \mathbb{V}_Z\left(\widehat{L}_V(h)\right) = &\left(\frac{1}{V} - \frac{1}{\binom{n}{p}}\right) \sum_{e \in \mathcal{E}_p} \frac{T_e^2}{\binom{n}{p}} + \\ &\left(1 - \frac{1}{V}\right) \sum_{e \cap e' = \emptyset} \frac{T_e \, T_{e'}}{\binom{n}{p}\binom{n-p}{p}} - \sum_{e \neq e' \in \mathcal{E}_p} \frac{T_e \, T_{e'}}{\binom{n}{p}^2}, \end{aligned} \tag{4.24}$$

where $\mathbb{V}_Z$ is the variance with respect to Z_e, $\mathbb{E}_X$ and $\mathbb{V}_X$ denote the expectation and variance with respect to the data, and $\mathbb{V}_{X,Z}$ represents the variance with respect to both the data and the Z_e's.

As both terms on the right-hand side of the equality are positive, (4.23) states that there is an increase in the variability of the risk estimator due to the random partition: $\mathbb{V}_{X,Z}\left(\widehat{L}_V(h)\right) > \mathbb{V}_X\left[\widehat{L}_p(h)\right]$. Combined with $\mathbb{E}_Z\left[\widehat{L}_V(h)\right] = \widehat{L}_p(h)$, this entails that both these estimators share the same bias whereas the V-fold is more variable than its Lpo counterpart: Lpo risk seems to be more reliable an estimator than V-fold (see Section 4.5.1 and more precisely Figure 4.2).
We are interested in assessing how large the discrepancy between variances of the two estimators is. To do this, we calculate $\mathbb{V}_Z\left(\widehat{L}_V(h)\right)$. We simplify (4.24) in the same way as the Lpo risk estimator. Due to its intricate expression, this exact and computable formula is not shown here. However it is used in Section 4 to compute the exact value of $\mathbb{V}_Z\left(\widehat{L}_V(h)\right)$.

4.4.2 Choice of the parameter p

As in the V-fold where V is a user-specified parameter, we have to choose the value of p from which we apply the leave-p-out. The criterion we use in this purpose is the mean square error (MSE) of the Lpo risk defined by

$$MSE(\omega, p) = MSE\left(\widehat{L}_p(\omega)\right) = (B_p(\omega))^2 + V_p(\omega).$$

In the following we work with a fixed $\omega \in [0,1]^D$. Consequently, we omit ω in the notation and consider the MSE as a function of p: $MSE(\omega, p) = MSE(p)$ that we want to minimize. The theoretical minimum location p^* offers the best tradeoff between the square bias and the variance of the estimator.
Exploiting the above plug-in estimators, we propose the following $\widehat{MSE}(\widehat{L}_p(\omega)) = \widehat{MSE}(p)$.

Definition 4.4.2 (Plug-in MSE estimator)**.** *With the same notations as in (4.16) and (4.17), we define*

$$\widehat{MSE}(p) = \left(\widehat{B}_p(\omega)\right)^2 + \widehat{V}_p(\omega). \tag{4.25}$$

Obviously, the accuracy of this estimator can be straightforwardly derived from that of $\widehat{B}_p$ and $\widehat{V}_p$ (see (4.19) and (4.20)), and is not explicitly mentioned.

Our present goal is to find $\widehat{p} \in [\![1, n-1]\!]$ such that

$$\widehat{p} = \text{Argmin}_{p \in [\![1,n-1]\!]} \widehat{MSE}(p). \tag{4.26}$$

The first step of our strategy is to determine the global minimum location $\widehat{p}_{abs}$ of that criterion in $\mathbb{R}$:

$$\widehat{p}_{abs} = \text{Argmin}_{p \in \mathbb{R}} \widehat{MSE}(p). \tag{4.27}$$

Then we shall build $\widehat{p}$ from this $\widehat{p}_{abs}$. Differentiating $x \in \mathbb{R} \mapsto \widehat{MSE}(x)$ provides an explicit expression for $\widehat{p}_{abs}$. Thus, we get

Theorem 4.4.1 (Exact $\widehat{p}_{abs}$)**.** *Set for any $x \in \mathbb{R}$, $\psi(x) = \widehat{MSE}(x)$. With the notations of proposition 4.3.1, we obtain*

$$\psi(x) = \frac{x^2[\varphi_3(n,\widehat{\alpha},\omega) + \varphi_2(n,\widehat{\alpha},\omega)] + x\,\varphi_1(n,\widehat{\alpha},\omega) + \varphi_0(n,\widehat{\alpha},\omega)}{[n(n-1)(n-x)]^2},$$

where $\varphi_3(n,\widehat{\alpha},\omega) = (n-1)^2(s_{1,1} - s_{2,1})^2$.
In the sequel, we drop arguments $(n, \widehat{\alpha}, \omega)$. Moreover,

- *if $\varphi_0 = 0$, then $x \in [1, n-1] \mapsto \psi(x)$ increases.*
- *otherwise, $\varphi_0 > 0$ and*
  - *(a) if*
  $$2n[\varphi_3 + \varphi_2] + \varphi_1 \neq 0, \tag{4.28}$$
  then $\widehat{p}_{abs}$ exists and $\widehat{p}_{abs} = -(n\varphi_1 + 2\varphi_0)/(2n[\varphi_2 + \varphi_3] + \varphi_1)$. Moreover, if $\widehat{p}_{abs} \notin [1, n-1]$, then $x \in [1, n-1] \mapsto \psi(x)$ increases.
  - *(b) Otherwise, $2n[\varphi_2 + \varphi_3] + \varphi_1 = 0$ implies that ψ increases in $[1, n-1]$.*

The elementary proof of this theorem is also deferred to the appendix.
Condition (4.28) should not to be too restrictive, which is satisfied in our simulations (see Section 4.5.2).
Note that Theorem 4.4.1 does not assert that $\widehat{p}_{abs} \in [1, n-1]$.
Thanks to Theorem 4.4.1, we now define $\widehat{p}$ as

$$\widehat{p} = \left| \begin{array}{ll} k\left(\widehat{p}_{abs}\right), & \text{if } \widehat{p}_{abs} \in [1, n-1] \\ 1, & \text{otherwise} \end{array} \right., \tag{4.29}$$

where $k(x)$ denotes the closest integer to x.
Note that $\widehat{p} = \widehat{p}(\omega)$ as we have worked with a given $\omega \in [0,1]^D$.

4.4.3 Adaptive selection procedure for histograms

Although we are unable to provide an adaptive procedure for kernel estimators, it is possible at least for histograms. Ideally we would like first to find ω^* defined as

$$\omega^* = \mathrm{Argmin}_{\omega\in\Omega} L(\widehat{s}_\omega) = \mathrm{Argmin}_{\omega\in\Omega} \mathbb{E}_s \left[\| \widehat{s}_\omega \|_2^2 - 2 \int_{[0,1]} s(x)\, \widehat{s}_\omega(x)\, dx \right],$$

which corresponds to the best histogram. As $L(\widehat{s}_\omega)$ depends on the unknown s, we use $\widehat{L}_p(\omega)$. A natural idea would be to look for

$$\widehat{\omega} = \mathrm{Argmin}_{\omega\in\Omega} \widehat{L}_p(\omega)\,.$$

The resulting $\widehat{\omega}$ depends on the value p of the leave-p-out procedure. As a criterion for choosing the parameter p, we introduce the mean square error (MSE) for which we get an explicit expression. As above, we use a plug-in estimator of the MSE that we want to minimize as a function of p. The resulting optimal $\widehat{p}$ realizes the best tradeoff between both the bias and the variance of $\widehat{L}_p(\omega)$.
Procedure

1. For each $\omega \in \Omega$, determine the optimal $\widehat{p}(\omega) = \mathrm{Argmin}_{p\in [\![1,n-1]\!]} \widehat{MSE}(\widehat{L}_p(\omega))$.
2. For each $\omega \in \Omega$, compute $\widehat{L}_{\widehat{p}\,(\omega)}(\omega)$.
3. Set $\widehat{\omega} = \mathrm{Argmin}_{\omega\in\Omega} \left\{ \widehat{L}_{\widehat{p}\,(\omega)}(\omega) \right\}$.
4. Define the final histogram by $\widetilde{s} = \widehat{s}_{\widehat{\omega}}$.

This procedure can be applied to any collection of partitions we want to explore.

4.5 Simulations and discussion

4.5.1 Influence of V on the choice of the optimal bandwidth

A large part of the literature about kernel estimators is devoted to the choice of an optimal bandwidth h^* defined as

$$h^* = \mathrm{Argmin}_{h>0} \mathbb{E}\left[||s - \widehat{s}_h||_2^2 \right].$$

In the following illustration, we are interested in both providing evidence of the only weak confidence we can have in the V-fold with respect to the Lpo risk estimator and in showing how influential p could be on the choice of $\widehat{h}$ for kernels. In the following example, we estimate $s = 0.5{*}\mathcal{N}(0, 9.10^{-2}) + 0.5{*}\mathcal{N}(1, 9.10^{-2})$. A sample of size $n = 500$ is drawn and we use the Gaussian kernel risk estimator given in Lemma 4.3.3.

Reliability of the V-fold

In order to understand how the V-fold is related to the Lpo, we can think of the V-fold as a way to approximate the ideal Lpo, since the latter used to be considered as unreachable. Except for small fluctuations due to random partitioning of the data, the V-fold with $V = n/p$ is expected to behave nearly as the Lpo estimator does. We provide here a kind of confidence region for the V-fold estimator (Figure 4.1). The width of this region expresses how wrong the V-fold estimate may be. With that objective, we use a closed-form formula for (4.24) as already mentioned. In Figure 4.1, the exact computation of $\mathbb{V}_Z\left(\widehat{L}_V(h)\right) = \widehat{\sigma}(h)^2$, for $p = 100$ ($V = 5$) enables us to plot $h \mapsto \widehat{L}_p(h) \pm 2\widehat{\sigma}(h)$. We observe the quite wide range of variation for the selected bandwidth $\widehat{h}$ as $\widehat{h}_{Lpo} = 0.1$, whereas $\widehat{h}_{Lpo+2\widehat{\sigma}(h)} = 0.115$ and $\widehat{h}_{Lpo-2\widehat{\sigma}(h)} = 0.08$.
As it introduces no unwanted variability, the Lpo risk estimator is more reliable than the V-fold one, all the more as this supplementary variability could be non negligible, as is observed in Figure 4.2. From 100 samples of size $n = 100$ drawn from $\mathcal{N}(0, 0.25)$, we compute both the V-fold and Lpo risk estimates, which provides us with an estimation of their variances. Whereas the surface corresponding to the Lpo risk estimator is rather smooth, that of V-fold is much more irregular, with rather large fluctuations around Lpo estimation values.

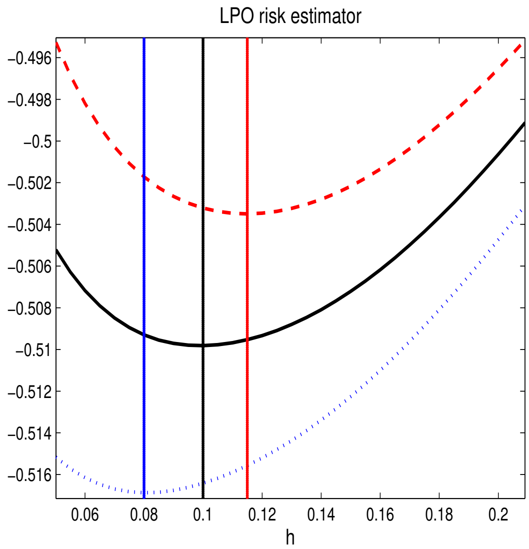

Figure 4.1: Graph of the Lpo risk versus h for $p = 100$. Plain line depicts the Lpo risk, dashed line denotes $\widehat{L}_p(h) + 2\widehat{\sigma}(h)$, dotted line represents $\widehat{L}_p(h) - 2\widehat{\sigma}(h)$. Vertical lines denotes minimum location for each criterion

Dependence of the Lpo risk estimator on p

An assessment of the dependence of $\widehat{h}$ on p is possible thanks to the left panel in Figure 4.3, which shows how influential the parameter p is on the values of the selected bandwidth $\widehat{h} = \mathrm{Argmin}_{h>0}\widehat{L}_p(h)$, chosen by leave-p-out. In this simulation, we use a sample of size $n = 500$ drawn from the following Gaussian mixture: $2/5\,\mathcal{N}(0, 0.09) + 3/5\,\mathcal{N}(0, 0.25)$. We notice the large variation of $\widehat{h}$ as $\widehat{h}$ varies from 0.22 for $p = 1$ to 0.34 for $p = 450$. There is therefore a real potential effect of the choice of p. The influence of p on the Lpo risk estimator itself is observable on the right panel in Figure 4.3, where the Lpo risk estimator is plotted versus p and h.

4.5.2 Density estimation by regular histogram

In the following simulations, we consider the classical density estimation framework. This study was carried out with five different densities: the uniform distribution on [0,1], a two-step density $s(x) = 3e/5\mathbb{1}_{[0,e^{-1}]}(x) + 2/[5(1-1/e)]\mathbb{1}_{[e^{-1},1]}(x)$, the two Beta densities $\beta(1,5)$ and $\beta(5,10)$, and two Beta-mixture densities. The successive sample sizes are $n = 100$, 500, 1000 and 2000. We use regular histograms with D bins, $D \in [\![1, K_{max}]\!]$, with $K_{max} = 25$, 50, 100, 150 and 200. Each condition was repeated 500 times for $n \leq 500$ and only 250 times otherwise.
The main observation for our purpose is that the procedure described in Section 4.4.3 always selects $\widehat{p} = 1$. The reason may be that regular histograms are not sensitive enough to catch the slight variation in the risk estimation, enabled by the change of p. As expected, the Loo risk estimator is all the more accurate as the smoothness of the density increases. Thus, the best results are obtained for the uniform density.
Contrary to what is usually thought, the leave-one-out procedure applied on regular histograms seems to achieve the best tradeoff between a small bias and a variance that is not too high with respect to other cross-validation procedures. Note that condition (4.28) was never violated in our simulations.

4.5.3 Multiple testing context and non-regular histograms

We want to carry out the simultaneous test of a very large number n of hypotheses from which an unknown number $n_0 > 0$ are true. It is well known that a key quantity to estimate in this context is the proportion $\pi_0 = n_0/n$ (see [11]). It appears that a way to obtain an estimator of π_0 is to estimate the density of the p-values corresponding to each hypothesis we are testing. These p-values are assumed to be *i.i.d.* and

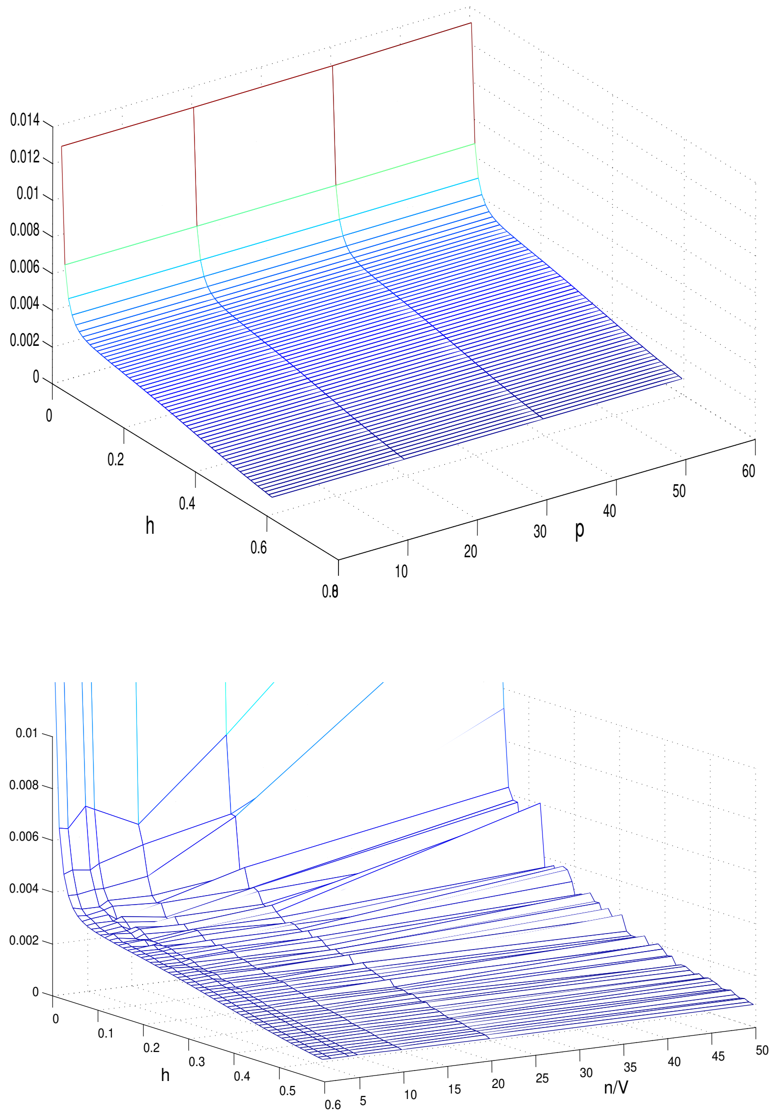

Figure 4.2: **Top panel:** Graph of the empirical variance of the Lpo risk estimator versus the bandwidth h and the parameter p. **Bottom panel:** Graph of the empirical variance of the V-fold estimator versus the bandwidth h and the parameter n/V (to enable comparisons).

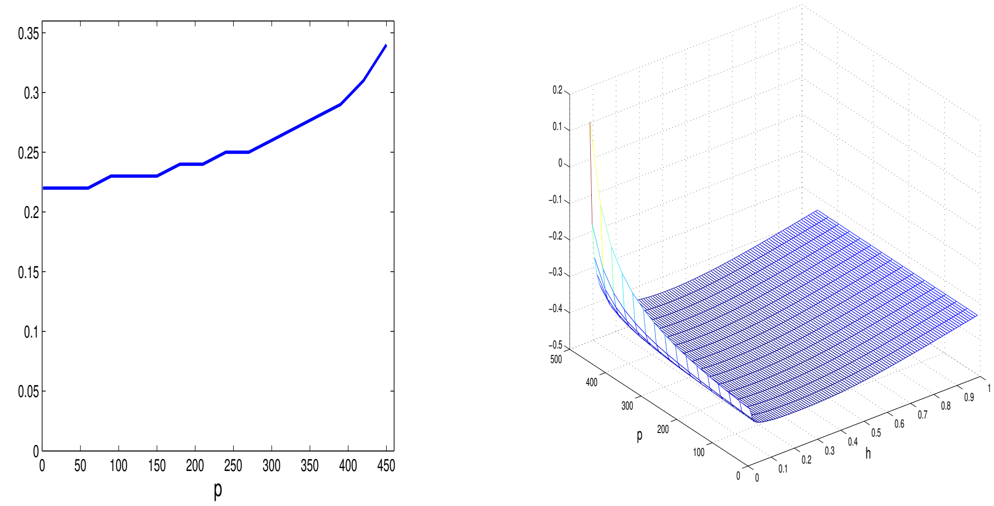

Figure 4.3: **Left panel:** Graph of the selected bandwidth $\widehat{h}$ with respect to the different values of p, using the Lpo risk estimator **Right panel:** Graph of the Lpo risk estimator versus the bandwidth h and parameter p.

drawn from the density

$$s = \pi_0 \mathbb{1}_{[0,1]} + (1 - \pi_0) f,$$

where f is an unknown density on $[0, 1]$, that concentrates around 0, decreases and vanishes when far enough from 0. Figure 4.4 depicts an example of a histogram obtained from the Hedenfalk data set ([7]). We consider three sample sizes $n = 500$, 1000 and 2000. Due to computational reasons, each condition was repeated 500 times for $n = 500$, whereas only 250 times for $n = 1000$ and 2000. For f, we therefore use a Beta density $\beta(1, A)$, the parameter of which is fixed successively at levels $A = 10$, 25, 50, 100 and 1000. The real proportion π_0 successively equals 0.5, 0.75, 0.9 and 0.95. Our prior knowledge of s leads us to consider irregular histograms, in the same way as [2], obtained as follows. For each $K \in [\![1, K_{max}]\!]$, that denotes the number of bins of a preliminary regular histogram, we consider all histograms with $D \in [\![1, K]\!]$ bins, built from the latter by agglomerating its columns from 1 to $\lambda = (D-1)/K$. An example of such a histogram is given in Figure 4.5. Eventually the procedure depends on another parameter K_{max} that is the maximal size of the considered regular histograms. It is set at levels 60, 80, 100 and 120. Unlike what happens for regular histograms, the proposed procedure most of the time chooses $\widehat{p} \neq 1$, and even much larger values than 1. The considered irregular histograms are more flexible and seem quite well suited to this kind of problem. Nevertheless, the still poor approximation properties of histograms often result in the final choice of the same histogram as the leave-one-out. For instance, we get different histograms in only 7.4% of the simulations for $\pi_0 = 0.95$ and $A = 5$, but up to 37.8% when $\pi_0 = 0.5$ and $A = 50$. In the multiple testing framework, since f vanishes when far enough from 0, a plausible estimator of π_0 may be

$$\widehat{\pi}_0 = \frac{\sharp \left\{ i / \, P_i \in \left[\widehat{\lambda}, 1 \right] \right\}}{m(1 - \widehat{\lambda}\,)},$$

where $[\widehat{\lambda}, 1]$ denotes the interval corresponding to the widest column of the selected histogram and $P_1, \ldots, P_n$ are *i.i.d.* random variables drawn from s. Thus in Table 4.1, we observe that both Loo and Lpo estimators overestimate π_0. However, the Lpo based estimation is less biased than the Loo. Note that there is no observable dependence of the selected histogram on the parameter K_{max}.

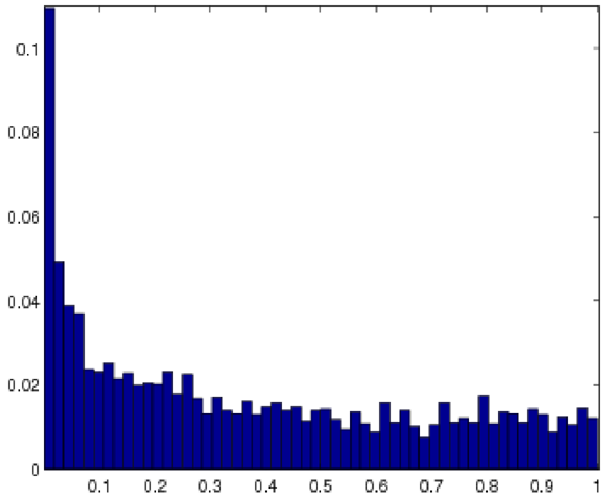

Figure 4.4: Regular histogram of p-values from Hedenfalk data with 56 bins.

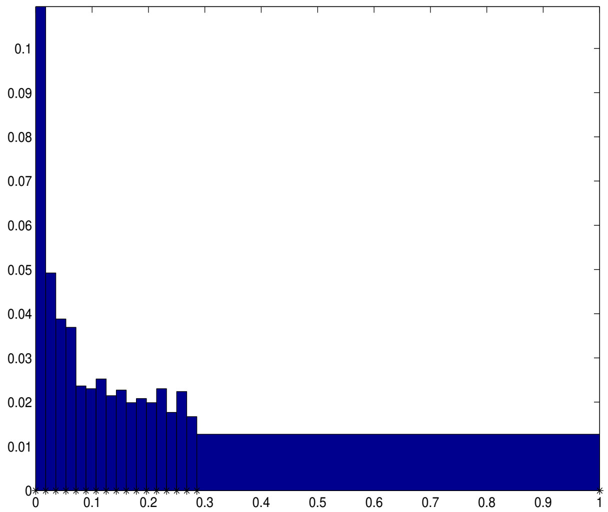

Figure 4.5: Example of agglomerated histogram obtained by agglomeration of bins from 1 to 0 of the Hedenfalk histogram.

π_0	$\widehat{\pi}_0^{Lpo} - \pi_0$	$\widehat{\pi}_0^{Loo} - \pi_0$
0.5	0.014	0.016
0.9	0.008	0.010
0.95	0.005	0.007

Table 4.1: Example of the empirical bias of the Lpo based estimator of π_0 ($\widehat{\pi}_0^{Lpo}$) and of its Loo counterpart ($\widehat{\pi}_0^{Loo}$) for $\pi_0 \in \{0.5, 0.9, 0.95\}$, $n = 1000$, $A = 10$ and 250 simulations.

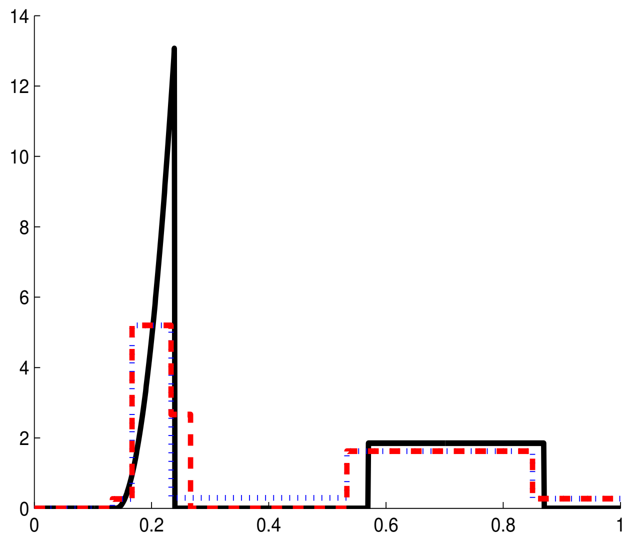

Figure 4.6: **Left panel:** Example of Lpo (resp. Loo) histogram in dashed line (resp. dotted line) with respect to the true mixture density (plain line).

4.5.4 Density estimation by irregular histograms

As the lack of flexibility of regular histograms is in question, we consider irregular histograms. They are obtained from a preliminary thin and regular grid of $[0, 1]$ with $K \in \mathbb{N}^*$ pieces. A histogram of dimension D corresponds to a partition of $[0, 1]$ in D elements, each one being the union of elementary pieces of the preliminary grid. Our goal is to choose the best histogram among all possible histograms obtained from the given preliminary grid. The dimension D of such a histogram satisfies $D \in [\![1, D_{max}]\!]$, with $D_{max} \leq K$.
We simulate 500 samples of size $n = 900$ drawn from s, where $\forall x \in [0, 1], \quad s(x) = 4/3\ (x - a)^2/(b - a)^3 \mathbb{1}_{[a,b]}(x) + 5/9\ \mathbb{1}_{[c,d]}(x)/(d - c)$, with $a = 0.14$, $b = 0.24$, $c = 0.57$ and $d = 0.87$. Here, $K = 60$ and $D_{max} = 30$. The selection of the best histogram is made by use of dynamic programming, which was introduced in [1] and essentially consists of a fast algorithm that provides the optimal solution.
Figure 4.6 shows a noticeable difference between histograms selected either by Lpo or by leave-one-out: between the two distribution supports, the Loo estimation is larger than 0 whereas its Lpo counterpart is 0. Both histograms differ from each other in more than 90% of the simulations. The empirical risk of the Lpo histogram estimator, computed from the 500 repetitions, is systematically lower than that of the Loo one.

4.5.5 Discussion

First, this work provides some tools to better understand the behaviour of cross-validation and more specifically, that of the leave-p-out cross-validation. To our knowledge, it is the first one that gives such results in the finite sample framework. Thanks to the closed-form expressions we derived, we are in position to give a kind of recipe for choosing the parameter p and inferring the behaviour of the V-fold cross-validation as well. For instance, our results suggest that a V-fold based bandwidth may be less reliable than a Lpo based one.
As for regular histograms, we get an empirical optimality result that identifies the leave-one-out risk estimator as the best one in terms of the bias-variance tradeoff, among procedures based on cross-validation. The change in the behavior of the proposed procedure when using non-regular histograms of a special type may be due to rather poor approximation properties of regular histograms, especially in case of spatial inhomogeneity. In Section 4.5.4, the compact-support densities used to build the mixture density s exploit that idea. Combined with irregular histograms, noticeable differences are obtained between leave-p-out

and leave-one-out estimators, Lpo with $p \neq 1$ leading to both an observable better spatial adaptation and a lower true risk.
As for kernels, our simulation study shows a strong dependence of the Lpo risk estimator and the resulting bandwidth $\widehat{h}$ on the parameter p. A future work will develop a similar criterion to that of histograms, in order to choose the parameter p when kernels are used. To this end, exact calculations or at least asymptotic tools may be explored. However in the context of kernels, simulations have shown how unreliable the V-fold risk estimator may be. Unlike the Lpo, its preliminary random partitioning of the data induces unwanted variability and may lead to a strong over- or under-smoothing.

4.6 Appendix

4.6.1 Sketch of proof of Theorem 4.3.2

Theorem 4.3.2 relies on the following

Lemma 4.6.1. *With the same notations as above, set e a subset of $\{1,\ldots,n\}$ of size p. Then,*

$$\begin{array}{rl} \forall i \in [\![1,n]\!], & \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\{i\in\overline{e}\}} = \binom{n-1}{p}, \\ \forall i,\, j \in [\![1,n]\!],\ i\neq j, & \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\{i\in\overline{e}\}}\mathbb{1}_{\{j\in\overline{e}\}} = \binom{n-2}{p}, \\ & \sum_{e\in\mathcal{E}_p} \mathbb{1}_{\{i\in e\}}\mathbb{1}_{\{j\in\overline{e}\}} = \binom{n-2}{p-1}. \end{array}$$

■

4.6.2 Proof of Proposition 4.3.1

For the bias, a simple calculation of the expectation of the estimator provides (4.14). The variance is more complex to obtain although it is not essentially different. The only things we can use are that for each k, $\mathbb{1}_{I_k}(X_i),\ i=1\ldots,n$ are *i.i.d.* and consequently, $n_k \sim \mathcal{B}in(n,\alpha_k)$. So we give the following

Lemma 4.6.2. *For any $k\neq l$,*

$$\begin{array}{rcl} \mathbb{E}_s[n_k n_l] & = & n(n-1)\alpha_k\alpha_l. \\ \mathbb{E}_s[n_k^2 n_l^2] & = & n(n-1)\alpha_k\alpha_l + n(n-1)(n-2)[\alpha_k\alpha_l^2+\alpha_k^2\alpha_l+(n-3)\alpha_k^2\alpha_l^2]. \\ \mathbb{E}_s[n_k n_l^2] & = & n(n-1)\alpha_k\alpha_l + n(n-1)(n-2)\alpha_k^2. \end{array}$$

For the first moments of the binomial law, we refer to [9]. After some calculation, we get

$$\begin{array}{rcl} (n-p)^2\mathbb{E}_s\widehat{L}_{n-p}^2 & = & \dfrac{(n-p+1)(n-2)}{n(n-1)}\left[2(n-p+1)s_{2,1}s_{1,1}-\right. \\ & & \left.2(p-2)s_{3,2}+(n-3)(n-p+1)s_{2,1}^2+2(2n-p)s_{2,1}\right]+ \\ & & \dfrac{s_{1,2}}{n}-2s_{2,2}\dfrac{n-p+1}{n-1}(2n-p)+s_{1,1}^2\dfrac{n-1}{n}\cdot \end{array}$$

And combining with

$$(n-p)^2\left(\mathbb{E}_s\widehat{L}_{n-p}\right)^2 = s_{1,1}^2+(n-p+1)^2s_{2,1}^2-2(n-p+1)s_{1,1}s_{2,1},$$

we obtain the desired expression for the variance.

■

4.6.3 Proof of Proposition 4.3.2

Set $B_p(\alpha, \omega) = \varphi(\alpha) = p/\left(n(n-p)\right)\sum_{k=1}^{D}\varphi_k(\alpha_k)$, with $\forall x \in [0,1],\ \varphi_k(x) = x(1-x)/\omega_k$. Using that $\sup_{x\in[0,1]^D} ||D_x\varphi|| \leq \sum_{k=1}^{D} 1/\omega_k = M(\omega)$, we get

$$\begin{aligned}
\Pr(|\widehat{B}_p(\omega) - B_p(\omega)| \geq \epsilon) &\leq \Pr\left(M(\omega)\, ||\alpha - \widehat{\alpha}||_\infty \geq \frac{n(n-p)}{p}\,\epsilon\right) , \\
&\leq \Pr\left(\max_k \left|\frac{n_k}{n} - \alpha_k\right| \geq \frac{n(n-p)}{p\, M(\omega)}\,\epsilon\right) , \\
&\leq \sum_k \Pr\left(\left|\frac{n_k}{n} - \alpha_k\right| \geq \frac{n(n-p)}{p\, M(\omega)}\,\epsilon\right) ,
\end{aligned}$$

which enables us to conclude by applying Bernstein's inequality.

■

4.6.4 Proof of proposition 4.3.3

We differentiate φ defined by $\varphi(x) = r_{n,p}V_p(x,\omega),\ \forall x \in [0,1]$ which is continuously differentiable (polynomial). Thanks to the continuity of $x \mapsto \|D_x\varphi\|$ and the compactness of $[0,1]^D$, $\sup_{x\in[0,1]^D} \|D_x\varphi\| < \infty$. Denoting this quantity by $M(n,p,\omega)$, we then have

$$\forall x, y \in [0,1]^D, \quad |\varphi(x) - \varphi(y)| \leq M(n,p,\omega)||x-y||_\infty .$$

This induces that

$$\begin{aligned}
\Pr(|\widehat{V}_p - V_p| \geq \epsilon) &= \Pr\left(|\varphi(\alpha) - \varphi\left(\widehat{\alpha}\right)| \geq \epsilon\, r_{n,p}\right) , \\
&\leq \Pr\left(M(n,p,\omega)\, \|\alpha - \widehat{\alpha}\|_\infty \geq \epsilon\, r_{n,p}\right) , \\
&\leq \sum_k \Pr\left(\left|\frac{n_k}{n} - \alpha_k\right| \geq \frac{r_{n,p}}{M(n,p,\omega)}\,\epsilon\right) , \\
\text{(Bernstein)} \quad &\leq 2D\exp\left[-\frac{n}{2}\left(\frac{r_{n,p}}{M(n,p,\omega)}\,\epsilon\right)^2 \frac{1}{1 + \frac{r_{n,p}}{3M(n,p,\omega)}\,\epsilon}\right] .
\end{aligned}$$

■

4.6.5 Sketch of proof of Proposition 4.4.1

The main argument in the proof is this combinatorial lemma:

Lemma 4.6.3. *For any* $p \in [\![1, n-1]\!]$,

1. $\forall e \in \mathcal{E}_p, \quad \mathcal{L}\left(Z_e\right) = \mathcal{B}\left(V/\binom{n}{p}\right)$,
2. $\forall e \neq e' \in \mathcal{E}_p, \quad \mathbb{E}_Z\left[Z_e\, Z_{e'}\right] = V(V-1)/\binom{n}{p}\binom{n-p}{p}$.

■

4.6.6 Proof of theorem 4.4.1

Excluding cases where there exists k_0 such that $n_{k_0} = n$,

$$\widehat{B}_x = \frac{x}{n(n-x)}\sum_k \frac{n_k/n(1-n_k/n)}{\omega_k} > 0 \text{ in } \mathbb{R}^* \setminus \{n\}$$

implies that $\psi > 0$ in $\mathbb{R}^* \setminus \{n\}$ ($\psi(x) \xrightarrow[x\to n]{} +\infty$ as is the case for $\widehat{B}_x^2$).
Furthermore, $\psi(0) = \frac{\varphi_0}{n^2}$ ·

1. If $\varphi_0 = 0$, then $\forall x \in \mathbb{R} \setminus \{n\},\ \psi(x) = x\frac{(\varphi_2+\varphi_3)x+\varphi_1}{(n-x)^2}$ · If $n((\varphi_2 + \varphi_3)n + \varphi_1) = 0$, there exists $\gamma \in \mathbb{R}$ such that $(\varphi_2 + \varphi_3)x + \varphi_1 = \gamma(n - x)$, whence $\forall x \in \mathbb{R} \setminus \{n\},\ \psi(x) = \frac{\gamma x}{n-x}$, which is excluded by $\psi > 0$ in $\mathbb{R}^* \setminus \{n\}$. So, $n((\varphi_2 + \varphi_3)n + \varphi_1) > 0$ and $\forall x \in \mathbb{R} \setminus \{n\},\ x(\varphi_2 + \varphi_3 x + \varphi_1) > 0$.
   Suppose $\varphi_2 + \varphi_3 = 0$, then as above $\varphi_1 = 0$ and $\psi \equiv 0$, which contradicts $\psi > 0$ in $\mathbb{R} \setminus \{n\}$.
   If $\varphi_2 + \varphi_3 \neq 0$, $\varphi_1 = 0$ and $\psi(x) = \frac{(\varphi_2+\varphi_3)x^2}{(n-x)^2}$ with $\varphi_2 + \varphi_3 > 0$, ψ strictly increases in $[1, n - 1]$.

2. Otherwise, $\varphi_0 \neq 0$ and $\psi > 0$ in $\mathbb{R} \setminus \{n\}$ as the continuity of ψ at 0 implies that $\varphi_0 > 0$.

In the following, we consider the latter case.
$\psi > 0$ in $\mathbb{R} \setminus \{n\}$, so that $x \mapsto (\varphi_2+\varphi_3)x^2+\varphi_1 x+\varphi_0 > 0$ in $\mathbb{R} \setminus \{n\}$. At $x = n$, $(\varphi_2+\varphi_3)n^2+\varphi_1 n+\varphi_0 \neq 0$. Otherwise, we would write either $(\varphi_2 + \varphi_3)x^2 + \varphi_1 x + \varphi_0 = c(n - x)^2$ or $(\varphi_2 + \varphi_3)x^2 + \varphi_1 x + \varphi_0 = (n - x)(ax+b)$, $a, b, c \in \mathbb{R}$. Thus, either respectively $\psi(x) \not\xrightarrow[x\to n]{} +\infty$, or $\psi(x) \xrightarrow[x\to n^-]{} +\infty$ and $\psi(x) \xrightarrow[x\to n^+]{} -\infty$ for example, which is excluded. Then,

$$\varphi_2 + \varphi_3 n^2 + \varphi_1 n + \varphi_0 > 0, \qquad \text{(continuity)} \tag{4.30}$$

and we obtain that the discriminant $\Delta = \varphi_1^2 - 4\varphi_0(\varphi_2 + \varphi_3) < 0$, whence $\varphi_2 + \varphi_3 > 0$.
The derivative ψ' in $\mathbb{R} \setminus \{n\}$ satisfies

$$\psi'(x) = \frac{(2n(\varphi_2 + \varphi_3) + \varphi_1)x + n\varphi_1 + 2\varphi_0}{(n - x)^3} \,.$$

1. If $2n(\varphi_2 + \varphi_3) + \varphi_1 = 0$, (4.30) implies $n\varphi_1 + 2\varphi_0 > 0$, that is, ψ grows in $[0, n[$.

2. Otherwise, there exists an extremum location $p_0 = -\frac{n\varphi_1+2\varphi_0}{2n(\varphi_2+\varphi_3)+\varphi_1}$ for ψ.

   (a) If $2n(\varphi_2+\varphi_3)+\varphi_1 < 0$, then (4.30) yields both that $n\varphi_1+2\varphi_0 > 0$ and $p_0 = \frac{n\varphi_1+2\varphi_0}{-(2n(\varphi_2+\varphi_3)+\varphi_1)} > n$. The sign of the derivative indicates that ψ increases in $[0, n[$.

   (b) If $2n(\varphi_2+\varphi_3)+\varphi_1 > 0$, in a similar way we get $p_0 < n$. So, either p_0 belongs to $[0, n[$ or $p_0 < 0$ in which case ψ also increases in $[0, n[$.

■

Bibliography

[1] R. E. Bellman and S. E. Dreyfus. *Applied Dynamic Programming*. Princeton, 1962.

[2] G. Castellan. Modified Akaike's criterion for histogram density estimation. Technical Report 99.61, Université de Paris-Sud, 1999.

[3] L. Devroye and L. Diaconis. *Nonparametric density estimation: the $L1$ view*. New York: Wiley, 1985.

[4] A. Elisseff and M. Pontil. Leave-one-out error and stability of learning algorithms with applications. In S. Basu C. Michelli In J. Suykens, G. Horvath and J. Vandewalle, editors, *Learning Theory and Practice*, ASI Series. IOS Press, Amsterdam; Washington, DC, 2002.

[5] D. Freedman and P. Diaconis. On the histogram as a density estimator: $L2$ Theory. *Z. Wahrscheinlischkeittheorie Verw. Gebiete*, 57:453–476, 1981.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2001.

[7] I. Hedenfalk, D. Duggan, Y.D. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O.P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene expression profiles in hereditary breast cancer. *New Engl. Jour. Medicine*, 344:539–548, 2001.

[8] M. Hubert and S. Engelen. Fast cross-validation of high-breakdown resampling methods for PCA. *Comput. Stat. Data Anal.*, 51(10):5013–5024, 2007.

[9] N. Johnson, S. Kotz, and A. Kemp. *Univariate Discrete Distributions*. General Probability and Mathematical Statistics. Wiley, 2005.

[10] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In N. Lavrac and S. Wrobel, editors, *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.

[11] M. Langaas, B. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Staitstical Society. Series B*, 67(4):555–572, 2005.

[12] J.S. Marron and M.P. Wand. Exact mean intergrated squared errors. *The Annals of Statistics*, 20:712–736, 1992.

[13] E. Parzen. On estimation of a probability density and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[14] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *Annals of Mathematical Statistics*, 27:642–669, 1956.

[15] M. Rudemo. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.

Chapter 5

A leave-p-out estimator of the proportion of true null hypotheses

The main concern of this chapter is the estimation of the unknown proportion of true null hypotheses in the multiple testing framework. This work is tightly connected to the previous chapter since the same approach is applied in this specific context.
However, we derive a new estimator of the unknown proportion π_0, which enjoys some asymptotic theoretical properties and provides a tight asymptotic control of the false discovery rate.
A simulation study enables the comparison between the proposed estimator and some widely used competitors, which results in the better performance of our proposal.
Furthermore, this new estimator allows to cope with some troubles referred to as the "U-shape" condition, where other estimators dramatically fail.

5.1 Abstract

In the multiple testing context, a challenging problem is the estimation of the proportion π_0 of true-null hypotheses. A large number of estimators of this quantity rely on identifiability assumptions that either appear to be violated on real data, or may be at least relaxed. Under independence, we propose an estimator $\widehat{\pi}_0$ based on density estimation using both histograms and cross-validation.
Due to the strong connection between the false discovery rate (FDR) and π_0, many multiple testing procedures (MTP) designed to control the FDR may be improved by introducing an estimator of π_0. We provide an example of such an improvement (plug-in MTP) based on the procedure of Benjamini and Hochberg. Asymptotic optimality results may be derived for both $\widehat{\pi}_0$ and the resulting plug-in procedure. The latter ensures the desired asymptotic control of the FDR, while it is more powerful than the BH-procedure.
Finally, we compare our estimator of π_0 with other widespread estimators in a wide range of simulations. We obtain better results than other tested methods in terms of mean square error (MSE) of the proposed estimator. Finally, both asymptotic optimality results and the interest in tightly estimating π_0 are confirmed (empirically) by results obtained with the plug-in MTP.

5.2 Introduction

Multiple testing problems arise as soon as several hypotheses are tested simultaneously. Like in test theory, we are concerned with the control of type-I errors we may commit in falsely rejecting any tested hypothesis. Post-genomics, astrophysics or neuroimaging are typical areas in which multiple testing problems are encountered. For all these domains, the number of tests may be of the order of several thousands. Suppose we are testing each of m hypotheses at level $0 < \alpha < 1$, the probability of at least one false positive (*e.g.* false rejection) may equal $m\alpha$ in the worst case. A possible way to cope with this is to use the Bonferroni procedure ([8]), which consists in testing each hypothesis at level α/m. However, this method is known to be drastically conservative.

Since we may be more interested in controlling the proportion of false positives among rejections rather than the total number of false positives itself, Benjamini and Hochberg [3] introduced the false discovery rate (FDR), defined by

$$FDR = \mathbb{E}\left[\frac{FP}{1 \vee R}\right],$$

where $a \vee b = \max(a, b)$, FP denotes the number of false positives and R is the total number of rejections. A large part of the literature is devoted to the building of multiple testing procedures (MTP) that upper bound FDR as tightly as possible ([4, 5]). For instance, that of Benjamini and Hochberg (BH-procedure) [3] ensures the following inequality under independence

$$FDR \leq \pi_0 \alpha \leq \alpha,$$

where π_0 denotes the unknown proportion of true null hypotheses, while α is the actual level at which we want to control the FDR. Since π_0 is unknown, the BH-procedure suffers some loss in power, which is all the more deep as π_0 is small. A natural idea to overcome this drawback is the computation of an accurate π_0 estimator, which would be plugged in the procedure. Thus π_0 appears as a crucial quantity that is to be estimated, hence the large amount of existing estimators. We refer to [15, 6] for reviews on this topic. The randomness of this estimation needs to be taken into account in the assessment of the procedure performance ([11, 24]).

In many of quite recent papers about multiple testing (see [6, 9, 10, 11]), a two-component mixture density is used to describe the behaviour of p-values associated with the m tested hypotheses. As usual for mixture models, we need an assumption that ensures the identifiability of the model parameters. Thus, most of π_0 estimators rely on the strong assumption that there are only p-values following a uniform distribution on $[0, 1]$ in a neighbourhood of 1. However, Pounds *et al.* [17] recently observed the violation of this key assumption. They pointed out that some p-values associated with induced genes may be artificially sent near to 1, for example when a one-sided test is performed while the non-tested alternative is true. To overcome this difficulty, we propose to estimate the density of p-values by some non-regular histograms, providing a new estimator of π_0 that remains reliable in the Pounds' framework thanks to a relaxed "identifiability assumption".

In the context of density estimation with the quadratic loss and histograms, asymptotic considerations have been used by Scott ([22]) for instance. A drawback of this approach relies on regularity assumptions made on the unknown distribution. Some AIC-type penalized criteria as in Barron *et al.* [1] could be applied as well. However, such an approach depends on some unknown constants that have to be calibrated at the price of an intensive simulation step (see [16] in the regression framework). As it is both regularity-assumption free and computationally cheap, we address the problem by means of cross-validation, first introduced in this context by Rudemo ([18]). More precisely, the leave-p-out cross-validation (LPO) is successfully applied following a strategy exposed in Celisse *et al.* [7]. Unlike Schweder and Spjøtvoll's estimator of π_0 ([21]), which depends on a user-specified parameter λ, ours is fully adaptive thanks to the LPO-based approach.

The paper is organized as follows. In Section 1, we present a cross-validation based estimator of π_0 (denoted by $\widehat{\pi}_0$) derived from a strategy previously exposed in [7], from which we only remind few necessary notations (see Section 1.2), but no result. This estimator relies on several assumptions that are fully specified and a description of the whole π_0 estimation procedure is given. Section 2 is devoted to asymptotic results such as consistency of $\widehat{\pi}_0$. Then we propose a plug-in multiple testing procedure (plug-in MTP), based on the same idea as that of Genovese *et al.* [11]. It is compared to the BH-procedure in terms of power and its asymptotic control of the FDR is derived. Section 3 is mainly concerned with the performance assessment of our π_0 estimation procedure in a wide range of simulations. A comparison with other existing and widespread methods is carried out.The influence of the π_0 estimation on the power of the plug-in MTP is inferred as well. Experimental evidences for almost overall improvements are obtained with the proposed method.

5.3 Estimation of the proportion of true null hypotheses

5.3.1 Mixture model

Let $P_1, \ldots, P_m$ be m *i.i.d.* random variables following a density g on $[0,1]$. $P_1, \ldots, P_m$ denote the p-values associated with the m tested hypotheses. Taking into account the two populations of ($\mathbf{H}_0$ and $\mathbf{H}_1$) hypotheses, we assume ([6, 9, 11]) that g may be written as

$$\forall x \in [0,1], \qquad g(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x),$$

where f_0 (resp. f_1) denotes the density of $\mathbf{H}_0$ (resp. $\mathbf{H}_1$) p-values, that is p-values corresponding to true null (resp. false null) hypotheses. π_0 is the unknown proportion of true null hypotheses. Moreover, we assume that f_0 is continuous, which ensures that $f_0 = 1$: $\mathbf{H}_0$ p-values follow the uniforme distribution $\mathcal{U}([0,1])$. Subsequently, the above mixture becomes

$$\forall x \in [0,1], \qquad g(x) = \pi_0 + (1-\pi_0) f_1(x), \tag{5.1}$$

where both π_0 and f_1 remain to be estimated.
Most of existing π_0 estimators rely on a sufficient condition which ensures the identifiability of π_0. This assumption may be expressed as follows

$$\exists \lambda^* \in]0,1] / \quad \forall i \in \{1, \ldots, m\},\ P_i \in [\lambda^*, 1] \Rightarrow P_i \sim \mathcal{U}([\lambda^*, 1]). \qquad (\mathbf{A})$$

($\mathbf{A}$) is therefore at the origin of Schweder and Spjøtvoll's estimator ([21]), further studied by Storey ([24, 25]). It depends on a cut-off $\lambda \in [0,1]$ from which only $\mathbf{H}_0$ p-values are observed. This estimation procedure is further detailed in Section 5.5. The same idea underlies the adaptive Benjamini and Hochberg step-up procedure described in [4], based on the slope of the cumulative distribution function of p-values. If we assume $\lambda^* = 1$ (that is $f_1(1) = 0$), Grenander [12] and Storey *et al.* [26] choose $\widehat{g}(1)$ to estimate π_0, where $\widehat{g}$ denotes the estimator of g. Genovese *et al.* [11] use $(1 - G(t))/(1 - t),\ t \in (0,1)$ as an upper bound of π_0, which becomes (for t large enough) an estimator as soon as ($\mathbf{A}$) is true.
However, this assumption may be strongly violated as noticed by Pounds *et al.* [17]. This point is detailed in Section 5.5.2. Following this remark, we propose the milder assumption ($\mathbf{A'}$):

$$\exists \Lambda^* = [\lambda^*, \mu^*] \subset (0,1] / \quad \forall i \in \{1, \ldots, m\},\ P_i \in \Lambda^* \Rightarrow P_i \sim \mathcal{U}(\Lambda^*). \qquad (\mathbf{A'})$$

While it is a generalization of ($\mathbf{A}$), this assumption remains true in Pounds' framework as we will see in Section 5.5.2 . Scheid *et al.* [19] proposed a procedure named *Twilight*, which consists in a penalized criterion and provides, as a by-product, an estimation of π_0. Since this procedure does not rely on assumption ($\mathbf{A}$), it should be taken as a reference competitor in the simulation study (Section 5.5) with respect to our proposed estimators.

5.3.2 A leave-p-out based density estimator

If g satisfies (**A'**), any "good estimator" of this density on Λ^* would provide an estimate of π_0. Since g is constant on the whole interval Λ^*, we adopt histogram estimators. Note that we do not really care about the rather poor approximation properties of histograms outside of Λ^* as our goal is essentially the estimation of Λ^* and of the restriction of g to Λ^*, denoted by $g_{|\Lambda^*}$ in the sequel.
For a given sample of observations $P_1, \ldots, P_m$ and a partition of $[0,1]$ in $D \in \mathbb{N}^*$ intervals $I = (I_k)_{k=1,\ldots,D}$ of respective length $\omega_k = |I_k|$, the histogram $\widehat{s}_\omega$ is defined by

$$\forall x \in [0,1], \qquad \widehat{s}_\omega(x) = \sum_{k=1}^{D} \frac{m_k}{m\,\omega_k} \mathbb{1}_{I_k}(x),$$

where $m_k = \sharp\{i \in [\![1,m]\!] : P_i \in I_k\}$.
If we denote by $\mathcal{S}$ the collection of histograms we consider, the "best estimator" among $\mathcal{S}$ is defined in terms of the quadratic risk:

$$\begin{aligned}\widetilde{s} &= \mathrm{Argmin}_{s\in\mathcal{S}}\mathbb{E}_g\left[||g-s||_2^2\right], \\ &= \mathrm{Argmin}_{s\in\mathcal{S}}\left\{\mathbb{E}_g\left[||s||_2^2\right] - 2\int_{[0,1]} s(x)g(x)\,dx\right\},\end{aligned} \tag{5.2}$$

where the expectation is taken with respect to the unknown g. According to (5.2), we define R by

$$R(s) = \mathbb{E}_g\left[||s||_2^2\right] - 2\int_{[0,1]} s(x)g(x)\,dx. \tag{5.3}$$

In (5.3) we notice that R still depends on g that is unknown. To get rid of this, we use a cross-validation estimator of R that will achieve the best trade-off between bias and variance. Following ([13]), we know that leave-one-out (LOO) estimators may suffer from some high level variability. For this reason we prefer the use of leave-p-out (LPO), keeping in mind that the choice of the parameter p will enable the control of the bias-variance trade-off.
At this stage, we refer to Celisse *et al.* [7] for an exhaustive presentation the leave-p-out (LPO) based strategy. Hereafter, we remind the reader what LPO cross-validation consists in and then, give the main steps of the reasoning. First of all, it is based on the same idea as the well-known leave-one-out (see [13] for an introduction) to which it reduces for $p = 1$. For a given $p \in [\![1, m-1]\!]$, let split the sample $P_1, \ldots, P_m$ into two subsets of respective size $m-p$ and p. The first one, called training set, is devoted to the computation of the histogram estimator whereas the second one (the test set) is used to assess the behaviour of the preceding estimator. These two steps have to be repeated $\binom{m}{p}$ times, which is the number of different subsets of cardinality p among $\{P_1, \ldots, P_m\}$.

Closed formula of the LPO risk This outlined description of the LPO leads to the following closed formula for the LPO risk estimator of $R(\widehat{s}_\omega)$ (see [7]): For any partition $I = (I_k)_{k=1,\ldots,D}$ of $[0,1]$ in D intervals of length $\omega_k = |I_k|$ and $p \in [\![1, m-1]\!]$,

$$\widehat{R}_p(\omega) = \frac{2m-p}{(m-1)(m-p)}\sum_{k=1}^{D}\frac{m_k}{m\omega_k} - \frac{m(m-p+1)}{(m-1)(m-p)}\sum_{k=1}^{D}\frac{1}{\omega_k}\left(\frac{m_k}{m}\right)^2, \tag{5.4}$$

where $m_k = \sharp\{i \in [\![1,m]\!] : \ P_i \in I_k\}$, $k = 1,\ldots,D$. As it may be evaluated with a computational complexity of only $O\left(m\log m\right)$, (5.4) means that we have a very efficient estimator of the quadratic risk $R(\widehat{s}_\omega)$. Now, we propose a strategy for the choice of p that relies on the minimization of the mean square error criterion (MSE) of our LPO estimator of the risk. Indeed among $\{\widehat{R}_p(\widehat{s}_\omega) : \ p \in [\![1, m-1]\!]\}$, we would like to choose the estimator that achieves the best bias-variance trade-off. This goal is reached by means of the MSE criterion, defined as the sum of the square bias and the variance of the LPO risk estimator. Thanks to (5.4), closed formulas for both the bias (5.5) and the variance (5.6) of LPO risk estimator may be derived. We recall here these expressions that come from [7].

Bias and variance of the LPO risk estimator Let ω correspond to a $D-$partition $(I_k)_k$ of $[0,1]$ and for any $k \in \{1,\ldots,D\}$, $\alpha_k = \Pr[P_1 \in I_k]$ such that $\alpha = (\alpha_1,\ldots,\alpha_D) \in [0,1]^D$.
Then for any $p \in [\![1, m-1]\!]$,

$$B_p(\omega) = B_p(\alpha,\omega) = \frac{p}{m(m-p)}\sum_{k=1}^{D}\frac{\alpha_k(1-\alpha_k)}{\omega_k}, \tag{5.5}$$

$$V_p(\omega) = V_p(\alpha,\omega) = \frac{p^2\varphi_2(m,\alpha,\omega) + p\,\varphi_1(m,\alpha,\omega) + \varphi_0(m,\alpha,\omega)}{[m(m-1)(m-p)]^2}, \tag{5.6}$$

where

$$
\begin{aligned}
\forall (i,j) &\in \{1,\ldots,3\} \times \{1,2\}, \quad s_{i,j} = \sum_{k=1}^{D} \alpha_k^i / \omega_k^j, \\
\varphi_2(m,\alpha,\omega) &= m(m-1)\left[2s_{2,2} + 4s_{3,2}(m-2) + s_{2,1}^2(-4m+6)\right], \\
\varphi_1(m,\alpha,\omega) &= m(m-1)\left[-8s_{2,2} - 8s_{3,2}(m-2)(m+1) - 4s_{1,1}s_{2,1}(m-1)-\right. \\
&\quad \left. 2s_{2,1}^2(-4m^2+2m+6)\right], \\
\varphi_0(m,\alpha,\omega) &= m(m-1)\left[s_{1,2}(m-1) - 2s_{2,2}(m^2-2m-3)+\right. \\
&\quad 4s_{3,2}(m-2)(m+1)^2 - s_{1,1}^2(m-1)+ \\
&\quad \left. 4s_{1,1}s_{2,1}(m^2-1) + s_{2,1}^2(-4m+6)(m+1)^2\right].
\end{aligned}
$$

Plug-in estimators may be obtained from the preceding quantities by just replacing α_k with $\widehat{\alpha}_k = m_k/m$ in the expressions. Following our idea about the choice of p, we define for each (partition) ω the best theoretical value p^* as the minimum location of the MSE criterion:

$$
p^* = \text{Argmin}_{p \in [\![1,m-1]\!]} MSE(p) = \text{Argmin}_p \left\{ B_p(\omega)^2 + V_p(\omega) \right\}. \tag{5.7}
$$

The main point is that this minimization problem has an explicit solution named $p^*_{\mathbb{R}}$, as stated by Theorem 3.1 in [7]. For the sake of clarity, we recall the MSE expression:

Minimum location expression With the same notations as for the bias and the variance, we obtain for any $x \in \mathbb{R}$,

$$
MSE(x) = \frac{x^2[\varphi_3(m,\alpha,\omega) + \varphi_2(m,\alpha,\omega)] + x\,\varphi_1(m,\alpha,\omega) + \varphi_0(m,\alpha,\omega)}{[m(m-1)(m-x)]^2},
$$

where $\varphi_3(m,\alpha,\omega) = (m-1)^2(s_{1,1} - s_{2,1})^2$.
Thus, we define our best choice $\widehat{p}$ for the parameter p by

$$
\widehat{p} = \left| \begin{array}{ll} k\left(\widehat{p}_{\mathbb{R}}\right), & \text{if } \widehat{p}_{\mathbb{R}} \in [1, m-1] \\ 1, & \text{otherwise} \end{array} \right., \tag{5.8}
$$

where $k(x)$ denotes the closest integer near to x and $\widehat{p}_{\mathbb{R}}$ has the same definition as $p^*_{\mathbb{R}}$, but with $\widehat{\alpha}$ instead of α in the expression.
Remark: There may be a real interest in choosing adaptively the parameter p, rather than fixing $p = 1$. Indeed in the regression framework for instance, Shao [23] and Yang [28] underline that the simple and widespread LOO may be sub-optimal with respect to LPO with a larger p. In the linear regression set-up, Shao even shows that $p/m \to 1$ as $m \to +\infty$ is necessary to get consistency in selection.

5.3.3 Estimation procedure of π_0

Collection of non-regular histograms

We now precise the specific collection of histograms we will consider. For given integers $N_{min} < N_{max}$, we build a regular grid of $[0,1]$ in N intervals (of length $1/N$) with $N \in [\![N_{min}, N_{max}]\!]$. For a couple of integers $0 \le k < \ell \le N$, we define a unique histogram made of first k regular columns of width $1/N$, then a wide central column of length $(\ell - k)/N$ and finally $N - \ell$ thin regular columns of width $1/N$. An example of such an histogram is given in Figure 5.1. The collection $\mathcal{S}$ of the histograms we consider is defined by

$$
\mathcal{S} = \bigcup_{N \in [\![N_{min}, N_{max}]\!]} \mathcal{S}_N,
$$

where

$$
\forall N, \qquad \mathcal{S}_N = \left\{ \widehat{s}_\omega \;:\; w_{k+1} = (\ell-k)/N,\; w_i = 1/N \; for \; i \neq k+1,\; 0 \le k < \ell \le N \right\}.
$$

Provided $(\mathbf{A}')$ is fulfilled, we expect for each N a selected histogram with its wide central interval $[\lambda, \mu]$ close to Λ^*. The comparison of all these histograms (one per value of N) enables to relax the dependence of each selected histogram on the grid width $1/N$. Note that $N_{\min}$ may always be chosen equal to 1.

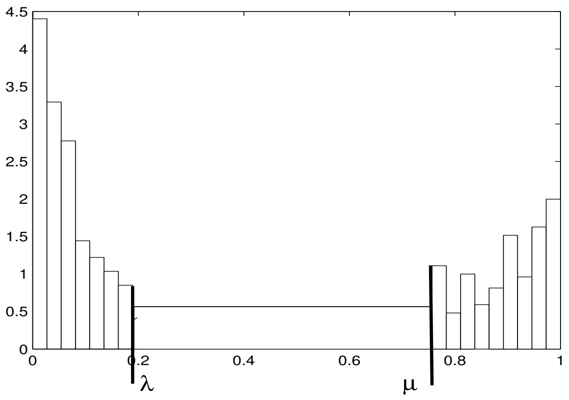

Figure 5.1: Example of non-regular histogram in $\mathcal{S}$. There are $k = 7$ regular columns from 0 to $\lambda = k/N$, a wide central column from λ to $\mu = \ell/N$, and $N - \ell = 7$ regular column of width $1/N$ from μ to 1.

Estimation procedure

Following the idea at the beginning of Section 5.3.2, $\widehat{\pi}_0$ will consist of the height of the selected histogram on its central interval $[\lambda, \mu]$. More precisely, we propose the following estimation procedure for π_0. For each partition (represented here by the vector ω), we compute $\widehat{p}\,(\omega) = \mathrm{Argmin}_p \widehat{MSE}(p, \omega)$, where $\widehat{MSE}$ denotes the MSE estimator obtained by plugging m_k/m in place of α_k in expressions of (5.7). The best (in terms of the bias-variance trade-off) LPO estimator of the quadratic risk $R(\widehat{s}_\omega)$ is therefore $\widehat{R}_{\widehat{p}(\omega)}(\omega)$. Then we choose the histogram that reaches the minimum of the latter criterion over $\mathcal{S}$. From this histogram, we finally get both the interval $[\widehat{\lambda}, \widehat{\mu}]$, which estimates Λ^*, and

$$\widehat{\pi}_0 = \widehat{\pi}_0(\widehat{\lambda}, \widehat{\mu}) \stackrel{def}{=} \frac{\sharp\left\{i:\, P_i \in \left[\widehat{\lambda}, \widehat{\mu}\right]\right\}}{m(\widehat{\mu} - \widehat{\lambda}\,)} \cdot$$

These steps are outlined hereafter
Procedure:

1. For each partition denoted by ω, define $\widehat{p}\,(\omega) = \mathrm{Argmin}_p \widehat{MSE}(p, \omega)$.
2. Find the best partition $\widehat{\omega} = \mathrm{Argmin}_\omega \widehat{R}_{\widehat{p}(\omega)}(\omega)$.
3. From $\widehat{\omega}$, get $(\widehat{\lambda}, \widehat{\mu})$.
4. Compute the estimator $\widehat{\pi}_0 = \frac{\sharp\left\{i:\, P_i \in \left[\widehat{\lambda}, \widehat{\mu}\right]\right\}}{m(\widehat{\mu} - \widehat{\lambda}\,)} \cdot$

5.4 Asymptotic results

5.4.1 Pointwise convergence of LPO risk estimator

Lemma 5.4.1. *Following the notations in Section 5.3.2, for any $p \in \{1, m-1\}$ and ω, we have*

$$MSE(p, \omega) = \mathcal{O}_{m \to +\infty}\left(1/m\right),$$

Moreover if $8 s_{1,1} s_{2,1} - 2 s_{1,1}^2 + 8 s_{3,2} - 10 s_{2,1}^2 - 4 s_{2,2} \neq 0$ *, then*

$$\widehat{p}(\omega)/m \xrightarrow[m \to +\infty]{a.s.} \ell_\infty(\omega),$$

where $\ell_\infty(\omega) \in [0, 1]$.

PROOF

1. We see that

$$\begin{aligned}\varphi_3 + \varphi_2 &= 4m^3\left[s_{3,2} - s_{2,1}^2\right] + o(m^3),\\ \varphi_1 &= 8m^4\left[-s_{3,2} + s_{2,1}^2\right] + o(m^4),\\ \varphi_0 &= 4m^5\left[s_{3,2} - s_{2,1}^2\right] + o(m^5).\end{aligned}$$

   Thus for any $p \in \{1, \ldots, m-1\}$ and partition of size vector $\omega \in [0,1]^D$ we have

$$MSE(p, \omega) = \mathcal{O}_{m\to+\infty}\left(\frac{1}{m}\right).$$

2. Simple calculations lead to

$$\frac{p^*_{\mathbb{R}}(\omega)}{m} \xrightarrow[m\to+\infty]{} \frac{8s_{3,2} - 4s_{2,2} + 4s_{1,1}s_{2,1} - 8s_{2,1}^2}{8s_{1,1}s_{2,1} - 2s_{1,1}^2 + 8s_{3,2} - 10s_{2,1}^2 - 4s_{2,2}} =: \ell(\alpha, \omega).$$

   For any k $\widehat{\alpha}_k \xrightarrow[m\to+\infty]{a.s.} \alpha_k$, the continuous mapping theorem implies the almost sure convergence. Finally, the result follows by setting $\ell_\infty(\omega) = \mathbb{1}_{\{\ell(\alpha,\omega)\in[0,1]\}}\ell(\alpha, \omega)$. ■

Proposition 5.4.1. *For any given ω, define $\widehat{p}(\omega)$ as in Section 5.3.2 and $\widehat{L}_p(\omega) = \widehat{R}_p(\omega) + ||g||_2^2$. If $\ell_\infty(\omega) \neq 1$, we have*

$$\widehat{L}(\omega) \stackrel{def}{=} \widehat{L}_{\widehat{p}}(\omega) \xrightarrow[m\to+\infty]{P} L(\omega) \stackrel{def}{=} ||g - s_\omega||_2^2.$$

Remark: Note that the assumption on ℓ_∞ does seem rather natural. It means that the test set must be (at most) of the same size as the training set ($\widehat{p}/(n-\widehat{p}) = \mathcal{O}_P(1)$). Moreover, $\ell_\infty(\omega) = 1$ if and only if $s_{2,1} - s_{2,2} - s_{3,2} = -3\, s_{1,1}$, which holds for very specific densities.

PROOF
The first part of Lemma 5.4.1 implies that $\widehat{R}_p(\omega) - R(\widehat{s}_\omega) \xrightarrow[m\to+\infty]{P} 0$. Combined with $R(\widehat{s}_\omega) \xrightarrow[m\to\infty]{} L(\omega) - ||g||_2^2$, it yields that for any fixed p,

$$\widehat{L}_p(\omega) \xrightarrow[m\to+\infty]{P} L(\omega).$$

Finally, the result follows from both the continuous mapping theorem and the assumption on ℓ_∞.■

5.4.2 Consistency of $\widehat{\pi}_0$

We first emphasize that for a given $N \in \{N_{\min}, \ldots, N_{\max}\}$ any histogram in $\mathcal{S}_N$ is associated with a given partition of $[0,1]$ that may be uniquely represented by (N, λ, μ). We give now the first lemma of the consistency proof.

Lemma 5.4.2. *For $\lambda^* \neq \mu^* \in [0,1]$, let s be a constant density on $[\lambda^*, \mu^*]$. Suppose $N_{\min}$ such that for any $N_{\min} \leq N$, it exists a partition (N, λ, μ) satisfying $0 < \mu - \lambda \leq \mu^* - \lambda^*$. For a given N, let ω_N represent the partition (N, λ_N, μ_N) with $\lambda_N = \lceil N\lambda^* \rceil / N$ and $\mu_N = \lfloor N\mu^* \rfloor / N$. Define s_ω as the orthogonal projection of s onto piecewise constant functions built from the partition associated with ω. If the dimension of a partition is its number of pieces, then ω_N is the partition with the smallest dimension satisfying*

$$\omega_N \in \operatorname{Argmin}_\omega ||s - s_\omega||_2^2.$$

Proof

For symmetry reasons, we deal with partitions, for a given N, made of regular columns of width $1/N$ from 0 to λ and only one column from λ to 1 (*e.g.* we set $\mu = 1$). In the sequel, $I^{(N)}$ denotes the partition associated with ω_N.

1. Suppose that it exists ω_0 such that $s = s_{\omega_0}$. Then $||s - s_{\omega_N}||_2^2 = 0$ and $\omega_N \in \text{Argmin}_\omega ||s - s_\omega||_2^2$.
2. Otherwise, s does not equal to any s_ω.
   (a) If $\lambda^* = k/N$, then $\lambda_N = \lambda^*$. Any subdivision I of $I^{(N)}$ satisfies $||s-s_\omega||_2^2 = ||s-s_{\omega_N}||_2^2$, where ω corresponds to I. Now, let $\mathcal{F}_I$ be the set of piecewise constant functions built from a partition I. For any partition $I = (I_k)_k$ such that $\forall k,\ I_\ell^{(N)} \subset I_k$ for a given ℓ, then $\mathcal{F}_I \subset \mathcal{F}_{I^{(N)}}$. Thus $||s-s_\omega||_2^2 = ||s-s_{\omega_N}||_2^2 + ||s_{\omega_N} - s_\omega||_2^2$, since $s_{\omega_N} - s_\omega \in \mathcal{F}_{I^{(N)}}$. Therefore, $\omega_N \in \text{Argmin}_\omega ||s-s_\omega||_2^2$.
   (b) If $\lambda^* \notin \{1/N, \dots, 1\}$. As before, any subdivision of $I^{(N)}$ will have the same bias, whereas it is larger for any partition containing $I^{(N)}$. So, $\omega_N \in \text{Argmin}_\omega ||s - s_\omega||_2^2$.■

Lemma 5.4.3. *With the same notations as before, we define $L(\omega) = ||s-s_\omega||_2^2$. Let $\widehat{L}$ be a random process indexed by the set of partitions Ω such that $\widehat{L}(\omega') \xrightarrow[m\to+\infty]{P} L(\omega')$, for any $\omega' \in \Omega$. If $\widehat{\omega} \in \text{Argmin}_\omega \widehat{L}(\omega)$, then*

$$\widehat{L}(\widehat{\omega}) \xrightarrow[m\to+\infty]{P} \min\{L(\omega) : \omega \in \Omega\}.$$

Proof

Set $\Gamma \subset \Omega$ such that $\forall \omega \in \Gamma,\ L(\omega) = \min_{\omega' \in \Omega} L(\omega')$ and define $\delta = \min_{\omega \neq \omega' \in \Gamma} |L(\omega) - L(\omega')|/2$. For $|\Omega| = k$ and $|\Gamma| = \ell$, we have the ordered quantities $L(\omega^1) = \dots = L(\omega^k) < L(\omega^{k+1}) \le \dots \le L(\omega^\ell)$. Set $\epsilon > 0$. For each ω^i, it exists m_i (large enough) such that for $m \ge m_i$, $|\widehat{L}(\omega^i) - L(\omega^i)| < \epsilon$, with high probability. For $m_{\max} = \max_i m_i$, we get $\max_{\omega \in \Omega} |\widehat{L}(\omega) - L(\omega)| < \epsilon$ in probability. Thanks to the latter inequality and by definition of $\widehat{\omega}$,

$$L(\widehat{\omega}) < \widehat{L}(\widehat{\omega}) + \epsilon \le \widehat{L}(\omega) + \epsilon < L(\omega) + 2\epsilon, \text{ in Probability}$$

for any $\omega \in \Omega \setminus \Gamma$. Hence, we obtain

$$L(\widehat{\omega}) < \min_{\omega \in \Omega\setminus\Gamma} L(\omega) = L(\omega^{k+1}), \text{ in Probability.}$$

Thus, $\widehat{\omega} \in \Gamma$ with high probability and the result follows.■

Theorem 5.4.1. *For $0 \le \lambda^* < \mu^* \le 1$, let $s :\ [0,1] \mapsto [0,1]$ be a constant function on $[\lambda^*, \mu^*]$ such that s is not constant on any interval I with $[\lambda^*, \mu^*] \subsetneq I$ (if it exists). Suppose $N_{\min}$ such that for any $N_{\min} \le N \le N_{\max}$, it exists a partition (N, λ, μ) satisfying $0 < \mu - \lambda \le \mu^* - \lambda^*$. Set $\Omega = \cup_N \Omega_N$, where Ω_N denotes the partitions associated with S_N. If $\widehat{\pi}_0$ is the estimator described in Section 5.3.3 selected from Ω, then*

$$\widehat{\pi}_0 \xrightarrow[m\to+\infty]{P} \pi_0.$$

Proof

For $\epsilon > 0$ and $N_{\min} \le N \le N_{\max}$,

$$\begin{aligned}
\Pr\left[\,|\pi_0 - \widehat{\pi}_0| > \epsilon\right] &= \Pr\left[\left| s\left(\frac{\lambda^*+\mu^*}{2}\right) - \widehat{s}_{\widehat{\omega}}\left(\frac{\widehat{\lambda}+\widehat{\mu}}{2}\right)\right| > \epsilon\right], \\
&\le \Pr\left[\,[\widehat{\lambda},\widehat{\mu}] \not\subset [\lambda^*,\mu^*]\right] + \Pr\left[||s_{\widehat{\omega}} - \widehat{s}_{\widehat{\omega}}||^2_{2,[\widehat{\lambda},\widehat{\mu}]} > \epsilon^2(\widehat{\mu} - \widehat{\lambda})\right], \\
&\le \Pr\left[\,|L(\widehat{\omega}) - L(\omega_N)| > \delta\right] + \Pr\left[\sup_\omega ||s_\omega - \widehat{s}_\omega||_2^2 > \epsilon^2/N_{\max}\right],
\end{aligned}$$

for some $\delta > 0$ ($||\cdot||_{2,[\widehat{\lambda},\widehat{\mu}]}$ denotes the quadratic norm restricted to $[\widehat{\lambda},\widehat{\mu}]$). As the cardinality of the set of partitions is finite ($N_{\max}$ does not depend on m),

$$\Pr\left[\sup_\omega ||s_\omega - \widehat{s}_\omega||_2^2 > \epsilon^2/N_{\max}\right] \xrightarrow[m\to+\infty]{} 0.$$

We use the following inequality $|L(\widehat{\omega}) - L(\omega_N)| - |L(\widehat{\omega}) - \widehat{L}(\widehat{\omega})| \leq |\widehat{L}(\widehat{\omega}) - L(\omega_N)|$ and the uniform convergence in probability of $\widehat{L} - L$ over Ω ($|\Omega| < +\infty$) to get

$$\Pr\left[\,|L(\widehat{\omega}) - L(\omega_N)| > \delta\right] \quad \leq \quad \Pr\left[\,|\widehat{L}(\widehat{\omega}) - L(\omega_N)| > \delta'\right],$$

for some $\delta' > 0$. The result comes from both Lemma 5.4.2 and Lemma 5.4.3.■

5.4.3 Asymptotic optimality of the plug-in MTP

The following is inspired by both [11] and [25]. In the sequel, we will remind some of their results to state the link. First of all for any $\theta \in [0,1]$, set

$$\forall t \in (0,1], \qquad Q_\theta(t) = \frac{\theta\, t}{G(t)} \quad \text{and} \quad \widehat{Q}_\theta(t) = \frac{\theta\, t}{\widehat{G}(t)},$$

where G (resp. $\widehat{G}$) denotes the (empirical) cumulative distribution function of p-values. Let define the threshold $T_\alpha(\theta) = T(\alpha, \theta, \widehat{G}) = \sup\{t \in (0,1) : \widehat{Q}_\theta(t) \leq \alpha\}$. Now we are in position to define our plug-in procedure:

Definition 5.4.1 (Plug-in MTP)**.** *Reject all hypotheses with p-values less than or equal to the threshold* $T_\alpha(\widehat{\pi}_0)$.

Storey *et al.* [25] established the equivalence between the BH-procedure and the procedure consisting in rejecting hypotheses associated with p-values less than or equal to the threshold $T_\alpha(1)$, named the step-up $T_\alpha(1)$ procedure. We may slightly extend Lemma 1 and Lemma 2 in [25] by using similar proofs, so that they are omitted here.

Lemma 5.4.4. *With the same notations as before, we have*

(i) the step-up procedure $T_\alpha(\widehat{\pi}_0(0,1)) = T_\alpha(1)$ *is equivalent to the BH-procedure in that they both reject the same hypotheses,*

(ii) the step-up procedure $T_\alpha(\widehat{\pi}_0(\widehat{\lambda}, \widehat{\mu}))$ *is equivalent to the BH-procedure with* m *replaced by* $\widehat{\pi}_0(\widehat{\lambda}, \widehat{\mu})$.

Thus, we observe that the introduction of $\widehat{\pi}_0$ (supplementary information) in our procedure entails the rejection of at least as much hypotheses as the BH-procedure (T_α in nonincreasing). Hence our plug-in procedure should be more powerful, provided it controls the FDR at the required level α.
We settle this question now, at least asymptotically, thanks to a slight generalization of Theorem 5.2 in [11] to the case where G is not necessarily concave (see the "U-shape" framework described in Section 5.5.2 for instance). For $t \in [0,1]$, let define $FP(t)$ (resp. $R(t)$) as the number of $\mathbf{H}_0$ (resp. the total number of) p-values lower than or equal to t and set $\Gamma(t) = FP(t)/(R(t) \vee 1)$. Thus,

$$\forall t \in [0,1], \quad FDR(t) = \mathbb{E}\left[\Gamma(t)\right].$$

Theorem 5.4.2. *For any* $\delta > 0$ *and* $\alpha \in [0, \pi_0[$, *define* $\widehat{\pi}_0^\delta = \widehat{\pi}_0 + \delta$. *Assume that the density* f *of* $\mathbf{H}_1$ *p-values is differentiable and is nonincreasing on* $[0, \lambda^*]$, *vanishes on* $[\lambda^*, \mu^*]$ *and is nondecreasing on* $[\mu^*, 1]$. *Then*

(i) Q_{π_0} *is increasing on* $I_\alpha = Q_{\pi_0}^{-1}([0,\alpha])$,

(ii) $\mathbb{E}\left[\Gamma(T_\alpha(\widehat{\pi}_0^\delta))\right] \leq \alpha + o(1)$.

Remarks:
Note that the only interesting choice of α actually lies in $[0, \pi_0)$. If $\alpha \geq \pi_0$, then $FDR(t) \leq \alpha$ is satisfied in the non-desirable case where all hypotheses are rejected.
A sufficient condition on G for the increase of Q_{π_0}, is that G were continuously differentiable and $G'(t) < G(t)/t, \forall t \in (0,1]$. Thus, G may be nondecreasing (not necessarily concave) and Q_{π_0} may increase yet.

To prove Theorem 5.4.2, we first need a useful lemma, the technical proof of which is deferred to Appendix.

Lemma 5.4.5. *With the above notations, for any* $\alpha \in (0,1]$, $T(\alpha,\cdot,\widehat{G}): \ [0,1] \mapsto [0,1]$ *is continuous a.s.. Moreover for any* $\theta \in [0,1]$, $G \mapsto T(\alpha,\theta,G)$ *is continuous on* $\mathcal{B}^+([0,1])$*, the set of positive bounded functions on* $[0,1]$*, endowed with the* $\|\cdot\|_\infty$.

PROOF

(i) As f is differentiable and nonincreasing, G is concave on $[0,\mu^*]$ and Q_{π_0} increases on this interval. Following the above remarks, Q_{π_0} is still increasing provided $G'(t) < G(t)/t$ for $t \in [\mu^*,1]$. Thus provided $G'(t) < G(t)/t, \ \forall t \in [\mu^*,1]$, Q increases on $[\mu^*,1]$. Otherwise, there exists $t_0 \in [\mu^*,1]$ such that $G'(t_0) = G(t_0)/t_0$. Then, the increase of f ensures that $G(x)/x \le G'(x), \ \forall x \ge t_0$. Hence, Q_{π_0} is nonincreasing on $[t_0,1]$. Finally since $Q(\pi_0) = 1$, Q_{π_0} is increasing on I_α.

(ii) Rewrite first the difference

$$
\begin{aligned}
\Gamma\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right) - \alpha &= \Gamma\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right) - Q_{\pi_0}\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right) \\
&\quad + Q_{\pi_0}\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right) - Q_{\pi_0}\left(T(\alpha,\pi_0^\delta,\widehat{G})\right) && (5.9) \\
&\quad + Q_{\pi_0}\left(T(\alpha,\pi_0^\delta,\widehat{G})\right) - Q_{\pi_0}\left(T(\alpha,\pi_0^\delta,G)\right) && (5.10) \\
&\quad + Q_{\pi_0}\left(T(\alpha,\pi_0^\delta,G)\right) - \alpha. && (5.11)
\end{aligned}
$$

Set $\eta > 0$ such that $2\eta < T(\alpha,\pi_0^\delta,G)$. Note that

$$
\Gamma\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right) - Q_{\pi_0}\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right) \le \frac{1}{\sqrt{m}} \left\|\sqrt{m}\,(\Gamma - Q_{\pi_0})\right\|_{\infty,[\eta,1]} + \mathbb{1}_{\{T(\alpha,\widehat{\pi}_0^\delta,\widehat{G}) \le \eta\}}.
$$

Thus thanks to Lemma 5.4.5,

$$
\mathbb{P}\left[T(\alpha,\widehat{\pi}_0^\delta,\widehat{G}) \le \eta\right] \le \mathbb{P}\left[T(\alpha,\pi_0^\delta,G) \le \eta + o_P(1)\right] \xrightarrow[m\to+\infty]{} 0.
$$

Besides, both Theorem 4.4 of [11] and Prohorov's theorem ([27]) imply that

$$
\mathbb{E}\left[\frac{1}{\sqrt{m}} \left\|\sqrt{m}\,(\Gamma - Q_{\pi_0})\right\|_{\infty,[\eta,1]}\right] = o(1).
$$

Hence $\mathbb{E}\left[\Gamma\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right) - Q_{\pi_0}\left(T(\alpha,\widehat{\pi}_0^\delta,\widehat{G})\right)\right] = o(1)$.
Thanks to Lemma 5.4.5, the uniform continuity of Q_{π_0} combined with the convergence in probability of $\widehat{\pi}_0^\delta$ ensure that the expectation of (5.9) is of the order of $o(1)$.
Since $T(\alpha,\pi_0^\delta,G) = \sup\{t: \ Q_{\pi_0}(t) \le \alpha\pi_0/\pi_0^\delta\}$, $\beta = \pi_0/\pi_0^\delta < 1$ and Q_{π_0} is a one-to-one mapping on I, we get $Q_{\pi_0}\left(T(\alpha,\pi_0^\delta,G)\right) = Q_{\pi_0}\left(Q_{\pi_0}^{-1}(\alpha\beta)\right) = \alpha\beta$. Thus,

$$
Q_{\pi_0}\left(T(\alpha,\pi_0^\delta,\widehat{G})\right) - Q_{\pi_0}\left(T(\alpha,\pi_0^\delta,G)\right) = Q_{\pi_0}\left(T(\alpha\beta,\pi_0,\widehat{G})\right) - \alpha\beta\,,
$$

Theorem 5.1 ([11]) applied with $\alpha\beta$ instead of α and $t_0 = Q_{\pi_0}^{-1}(\alpha\beta)$ entails that the expectation of (5.10) is $o(1)$ as well.
Finally, (5.11) is equal to $(\beta-1)\alpha < 0$.■

5.5 Simulations and Discussion

5.5.1 Comparison in the usual framework ($\mu = 1$)

By "usual framework", we mean that the unknown f_1 in the mixture (5.1) is a decreasing density satisfying assumption (**A**): it vanishes on an interval $[\lambda^*,1]$ with λ^* possibly equal to 1. In this framework,

$$
\widehat{\pi}_0 = \frac{\sharp\{i/\,P_i \in [\widehat{\lambda},1]\}}{m\,(1-\widehat{\lambda}\,)}\,.
$$

Table 5.1: Results for the two simulation conditions $(\lambda^*, s) = (0.2, 4)$ and $(\lambda^*, s) = (0.4, 6)$. The LPO and LOO based methods are compared to the Schweder and Spjøtvoll estimator, $\widehat{\pi}_0^{St}$ computed with $\lambda = 0.5$. (All displayed quantities are multiplied by 100.)

$\pi_0 = 0.9$	$\lambda^* = 0.2,\ s = 4$			$\lambda^* = 0.4,\ s = 6$		
Method	Bias	Std	MSE	Bias	Std	MSE
LPO	**0.39**	2.5	**6.41** 10^{-2}	**0.56**	2.8	**8.00** 10^{-2}
LOO	**0.46**	2.3	5.52 10^{-2}	**0.61**	2.7	7.66 10^{-2}
$\widehat{\pi}_0^{St}$	-0.15	3.2	**9.94** 10^{-2}	0.24	3.1	**9.58** 10^{-2}

Except $\widehat{\lambda}$, this general expression was introduced by Schweder *et al.* [21]. Their estimator

$$\widehat{\pi}_0^{SS}(\lambda) = \frac{\sharp\{i/\, P_i \in [\lambda, 1]\}}{m\,(1-\lambda\,)},$$

is based on (**A**) and strongly depends on the parameter $\lambda \in [0, 1]$ that is supposed to be given, but totally unknown in practice. A crucial issue ([15]) is precisely the determination of an 'optimal' λ.

A potential gain in choosing λ

In 2002, Storey [24] studied further this estimator and even proposed ([26]) the systematic value $\lambda = 0.5$ as a quite good choice. In the following, we show that even if assumption (**A**) is satisfied for $\lambda^* = 0.2$ or 0.4, there is a real potential gain in choosing λ in an adaptive way.

In the following simulations, the unknown density f_1 in the mixture (5.1) is a beta density on $[\lambda^*, 1]$ with parameter s:

$$f_1(t) = s/\lambda^*(1 - t/\lambda^*)^{s-1}\mathbb{1}_{[0,\lambda^*]}(t),$$

where $(\lambda^*, s) \in \{(0.2, 4), (0.4, 6)\}$. The beta distribution is all the more sharp in the neighbourhood of 0 as s is large. The proportion π_0 is equal to 0.9, the sample size $m = 1000$ while $n = 500$ repetitions have been made. There does not seem to be any strong sensitivity to the choice of N_{max} (data not shown here), as long as N_{max} is obviously not too small. Until the end of the paper, $N_{min} = 1$ and $N_{max} = 100$.

Table 5.1 shows the simulation results for the leave-p-out (*LPO*) and the leave-one-out (*LOO*) based estimators of π_0, compared to that of Schweder and Spjøtvoll for $\lambda = 0.5$ denoted by $\widehat{\pi}_0^{St}$. We see that in both cases, *LPO* is less biased than *LOO* but slightly more variable, which leads to a higher value for the MSE. This larger variability may be due to the supplementary randomness induced by the choice of $\widehat{\lambda}$. Both *LPO* and *LOO* seem a bit conservative unlike $\widehat{\pi}_0^{St}$, which is however a little less biased. We say that an estimator of π_0 is conservative as soon as it upperbounds π_0 on average. The main conclusion is that the MSE of *LPO* (and *LOO*) is always lower than that of $\widehat{\pi}_0^{St}$, even if the assumption (**A**) is satisfied ($\lambda = 0.5 > \lambda^*$). An adaptive choice of λ may provide a more accurate estimation of π_0, which is all the more important as m grows.

Comparison when $\lambda^* = 1$

We consider now the general (more difficult) case when (**A**) is only satisfied for $\lambda^* = 1$. Thus, f_1 is a beta density of parameter $s : f_1(t) = s(1-t)^{s-1}, \quad t \in [0, 1]$, with $s \in \{5, 10, 25, 50\}$. The sample size $m = 1000$ and $\pi_0 \in \{0.5, 0.7, 0.9, 0.95\}$. Each condition has been repeated $n = 500$ times. We detail below four of the different methods that have been compared in this framework.

Smoother and ***Bootstrap***

In [26], the authors proposed a method consisting in first computing the Schweder and Spjøtvoll estimator on a regular grid of $[0, 1]$ and then adjusting a cubic spline. The final estimator of π_0 is the resulting function evaluated at 1. This procedure is called *Smoother*.

The *Bootstrap* method was introduced in [25]. Authors define the optimal value of λ as the minimizer of the MSE of their π_0 estimator. Since this quantity is unknown, they use an estimation based on bootstrap. They also need to compute $\widehat{\pi}_0(\lambda)$ for values of λ on a preliminary grid of $[0, 1]$.

These methods are available as options of the *qvalue* function in the `R-` package qvalue [26].

Adaptive Benjamini-Hochberg procedure
In the sequel, this procedure is denoted by ABH and we refer to [4] for a detailed description. In outline, the method relies on the idea that the plot of p-values versus their ranks should be (nearly) linear for large enough p-values (likely $\mathbf{H}_0$ p-values). The inverse of the resulting slope provides a plausible estimator based on assumption (**A**).
The ABH procedure may be applied through the function *pval.estimate.eta0* in package fdrtool with the option method= "adaptive" `http://cran.r-project.org/src/contrib/Descriptions/fdrtool.html`.

Twilight
In their article, Scheid *et al.* [19] proposed a penalized criterion based on assumption (**A'**). This is a sum of the Kolmogorov-Smirnov score and a penalty term. The whole criterion is expected to provide the widest possible set of $\mathbf{H}_0$ hypotheses. How the penalty term balances against the Kolmogorov-Smirnov score depends on a constant C that is to be determined. To do so, the authors propose to use bootstrap combined with Wilcoxon tests. Besides, this procedure is iterative and strongly depends on the length of the data, which could be a serious drawback with increasing data sets.
The function *twilight* is available in package twilight [20].

Results
As in the preceding simulation study, LPO and LOO refer to the proposed methods. Figure 5.2 illustrates the performances for all the methods but ABH, for which results are quite poor with respect to other methods (see Table 5.2). We notice that both St_{Sm} and St_{Boot} have systematically larger MSE than the three remaining approaches. Our methods give quite similar results to each other in this framework. $Twilight$, LPO and LOO furnish nearly the same MSE values in the most difficult case $s = 5$, when $\pi_0 > 0.5$. Except for $\pi_0 = 0.5$ and $s = 5$, LPO and LOO all the more outperform upon $Twilight$ as the proportion raises. The better performance of $Twilight$ in this set-up may be due to the classical difference between cross-validation and penalized criteria. Indeed in the context of supervised classification for instance, Kearns *et al.* [14] and Bartlett *et al.* [2] show that cross-validation is used to providing good results, provided the noise level of the signal is not too high. Otherwise, penalized criteria (like $Twilight$) outperform upon cross-validation. In the present context, $s = 5$ means that $\mathbf{H}_1$ p-values are spread on a large part of $[0, 1]$ and not only concentrated in a neighbourhood of 0, while $\pi_0 = 0.5$ indicates a larger number of $\mathbf{H}_1$ p-values in the distribution tail of the Beta density. Thus this situation may be held as the counterpart of the noisy case in supervised classification. Nevertheless, LPO and LOO always outperform $Twilight$ when $\pi_0 > 0.5$. They are even uniformly better than $Twilight$ for $\pi_0 = 0.95$, that is for small proportions of $\mathbf{H}_1$ hypotheses.

5.5.2 Comparison in the U-shape case

The 'U-shape case' refers to the phenomenon underlined by Pounds *et al.* [17] on a real data set made of Affymetrix 'pooled' present-absent p-values (one p-value per probe set). We explore the behaviour of the preceding methods applied to p-values with similar distributions. In our simulation design, the sample is $m = 1000$, while $\pi_0 \in \{0.25, 0.5, 0.7, 0.8, 0.9\}$ and $n = 200$ repetitions of each condition have been made.
Typically, the U-shape case appears when one-sided tests are made whereas the non-tested alternative is true. For example, suppose the test statistics are distributed as a three-component gaussian mixture model

$$\pi_0 \mathcal{N}(0, 2.5\, 10^{-2}) + \frac{1-\pi_0}{2}\left[\mathcal{N}(a, \theta^2) + \mathcal{N}(b, \nu^2)\right], \tag{5.12}$$

where $a < 0$, $b > 0$ and $\theta, \nu > 0$, corresponding to respectively non-induced, under-expressed and over-expressed genes. We want to test whether genes are over-expressed, that is H_0 : '*the mean equals 0*' versus H_1 : '*the mean is positive*'. A test statistic drawn from $\mathcal{N}(a, \theta^2)$ (under-expressed gene) is more likely to have a larger p-value than those under $\mathcal{N}(b, \nu^2)$, which correspond actually to over-expressed genes. This phenomenon is clearly all the more deep as the gap between a and b is high and variances θ^2 and ν^2 are small. Note that a similar shape may be observed when test statistics are ill-chosen.
In order to mimic Pounds' example, we use (5.12) with $-a = b \in \{1, 1.5\}$ and $\theta = \nu \in \{0.5, 0.75\}$. As they were quite similar, results in these different conditions are gathered in Table 5.3. Except LPO and LOO for which this phenomenon is not so strong, any other method all the more overestimates π_0 as the

Table 5.2: Numerical results for different π_0 estimators with $s = 10$ and $\pi_0 \in \{0.5, 0.7, 0.9, 0.95\}$. Four other methods are compared to *LPO* and *LOO*. St_{Sm} denotes *Smoother*, St_{Boot} states for *Bootstrap* and $Twil$ for $Twilight$.(All displayed quantities are multiplied by 100.)

π_0	0.5			0.7		
Method	Bias	Std	MSE	Bias	Std	MSE
LPO	1.4	3.5	14.5 10^{-2}	1.4	3.4	13.6 10^{-2}
LOO	1.6	3.4	13.9 10^{-2}	1.6	3.3	13.4 10^{-2}
St_{Sm}	-0.9	5.1	26.2 10^{-2}	-0.9	6.0	36.2 10^{-2}
St_{Boot}	-2.3	4.0	20.9 10^{-2}	-3.3	4.7	33.3 10^{-2}
$Twil$	-1.0	3.6	14.0 10^{-2}	-1.5	4.2	19.4 10^{-2}
ABH	37.9	8.3	15.0	0.27	2.4	7.6

π_0	0.9			0.95		
Method	Bias	Std	MSE	Bias	Std	MSE
LPO	0.8	3.6	13.7 10^{-2}	0.5	3.1	9.5 10^{-2}
LOO	1.0	3.4	12.5 10^{-2}	0.7	2.9	8.9 10^{-2}
St_{Sm}	-0.5	6.6	43.1 10^{-2}	-1.0	5.5	30.8 10^{-2}
St_{Boot}	-3.7	5.4	43.4 10^{-2}	-3.7	5.1	39.6 10^{-2}
$Twil$	-1.6	4.4	21.8 10^{-2}	-1.6	4.2	20.2 10^{-2}
ABH	9.8	0.4	95.5 10^{-2}	4.9	0.1	24.1 10^{-2}

Table 5.3: Results of the U-shape case for the six compared methods for $\pi_0 \in \{0.25, 0.5, 0.7, 0.8, 0.9\}$.(All displayed quantities are multiplied by 100.)

π_0	0.25			0.5			0.7		
Method	Bias	Std	MSE	Bias	Std	MSE	Bias	Std	MSE
LPO	5.5	6.2	**0.7**	5.5	5.2	**0.6**	5.3	4.4	**0.5**
LOO	6.2	5.7	**0.7**	6.8	5.7	**0.8**	6.6	4.8	**0.7**
St_Sm	75.0	0	56.0	50.0	0	25.0	30.0	0	9.0
St_Bo	43.2	3.2	18.7	28.9	2.2	8.4	17.4	1.6	3.0
$Twil$	73.2	2.5	53.6	47.5	3.0	22.6	27.4	2.3	8.0
ABH	45.5	5.4	21.0	31.4	4.2	10.0	19.8	3.1	4.0

π_0	0.8			0.9		
Method	Bias	Std	MSE	Bias	Std	MSE
LPO	5.3	4.1	**0.4**	4.2	2.7	**0.2**
LOO	6.4	4.1	**0.6**	4.7	2.5	**0.3**
St_Sm	20.0	0	4.0	9.9	0.2	**1.0**
St_Bo	11.6	1.3	1.0	5.4	1.6	**0.3**
$Twil$	17.5	1.8	3.0	8.0	1.3	**0.7**
ABH	13.8	2.3	2.0	7.4	1.3	**0.6**

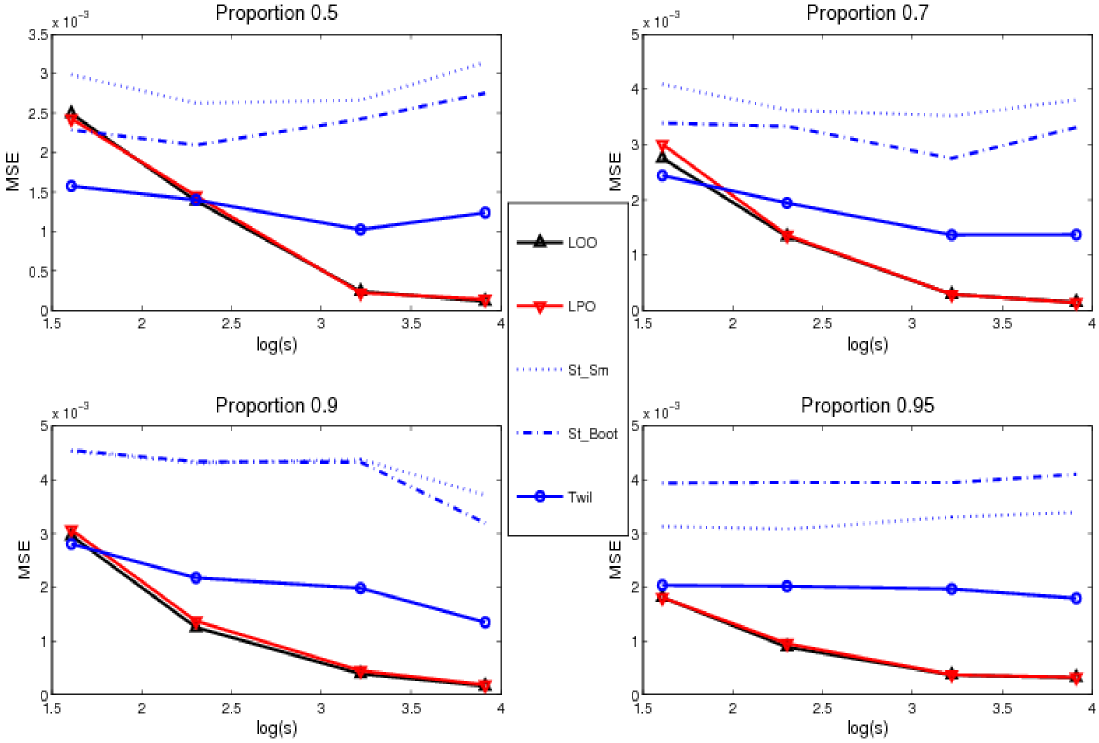

Figure 5.2: Graphs of the MSE of the π_0 estimator versus $\log s$, where s is the parameter of the Beta density. Each graph is devoted to a given proportion, from 0.5 to 0.95. St_{Sm} denotes the MSE obtained for *Smoother*, St_{Boot} that of *Bootstrap* while *Twil* states for *Twilight*.

proportion of p-values under the uniform distribution is small. In our framework, a growth in π_0 entails an increase in the right part of the histogram near 1, which is responsible for the overestimation (violation of assumption (**A**)). On the contrary when $\pi_0 = 0.9$, the violation of assumption (**A**)) is weaker and similar values of MSE are obtained for the competing approaches. In this set-up, *LPO*, *LOO* and St_{Boot} provide systematically the lowest MSE values. In comparison, it is somewhat surprising that *Twilight* overestimates π_0 so much, since it should have remained reliable under assumption (**A'**). Despite the preceding simulation results, we observe a repeated overestimation, which means that the criterion under-penalizes large sets of p-values. The involved penalty may have been designed for the situation before (with only one peak near 0), whereas it may be no longer relevant in this framework. This may be interpreted as a consequence of the higher adaptivity of cross-validation based methods over penalized criteria. Finally it is worth noticing that both the bias and the MSE of *LPO* are systematically lower than those of *LOO*, showing the interest of choosing p in an adaptive way.

5.5.3 Power

Here, we study the influence of the estimation of π_0 on the power of multiple testing procedures obtained as described in Section 5.5.1 for various π_0 estimators. The *Twilight* method is used for comparison, in association with the Benjamini-Hochberg procedure ([3]). Our reference is what we call the Oracle procedure, which consists in plugging the true value of π_0 in the MTP procedure of Section 5.5.1. The same simulations as in Section 5.5.1 are used for this study, which is carried out in two steps. In the first one, we compare procedures in terms of their empirical FDR, in order to assess the expected control for finite samples. Thus, we choose the level $\alpha = 0.15$ at which we want to control the FDR and then compute, for each of the $n = 500$ samples, the corresponding FDP in the terminology of [11], *e.g.* the ratio of the number of falsely rejected hypotheses over the total number of rejections. Finally, we get an estimator of the actual FDR: $\widehat{FDR}$ by averaging the simulation results. Table 5.4 gives results for the LPO and LOO based procedures $\widehat{FDR}_{LPO}$, $\widehat{FDR}_{LOO}$ and also for *Twilight* ($\widehat{FDR}_{Twil}$), Benjamini-Hochberg ($\widehat{FDR}_{BH}$) and

Table 5.4: Values of the empirical estimate of the FDR (%) for the LPO ($\widehat{FDR}_{LPO}$), LOO ($\widehat{FDR}_{LOO}$), *Twilight* ($\widehat{FDR}_{Twil}$), Benjamini-Hochberg ($\widehat{FDR}_{BH}$) and Oracle ($\widehat{FDR}_{Best}$) procedures. s denotes the parameter of the Beta distribution used to generate the data.

s	π_0	$\widehat{FDR}_{LPO}$	$\widehat{FDR}_{LOO}$	$\widehat{FDR}_{Twil}$	$\widehat{FDR}_{BH}$	$\widehat{FDR}_{Best}$
5	0.5	14.15	14.06	14.85	8.35	14.29
	0.7	14.13	14.03	14.85	10.40	14.50
	0.9	15.01	15.01	15.73	14.26	14.81
	0.95	13.23	13.43	13.76	13.13	13.83
10	0.5	14.74	14.69	15.50	6.94	15.02
	0.7	15.14	15.09	15.61	10.29	15.12
	0.9	17.91	17.90	18.08	15.85	17.94
	0.95	14.65	14.65	15.25	14.37	14.95
25	0.5	14.88	14.82	15.51	7.48	15.04
	0.7	14.69	14.64	15.19	10.47	14.84
	0.9	15.50	15.57	16.31	13.56	15.92
	0.95	14.35	14.22	14.51	13.19	14.19
50	0.5	14.76	14.71	15.42	7.40	14.89
	0.7	14.81	14.77	15.23	10.36	14.87
	0.9	13.93	13.82	14.79	13.17	13.98
	0.95	16.12	16.32	16.57	14.65	16.08

Oracle procedures ($\widehat{FDR}_{Best}$). In the second step, we check the potential improvement in power enabled by the LPO-based MTP with respect to the BH-procedure. The assessment of this point is made in terms of the expectation of the proportion of falsely non-rejected hypotheses among true alternatives (named FNR here). This criterion is estimated by the average of the preceding ratio computed from each sample. Table 5.5 displays the empirical FNR values, denoted by $\widehat{FNR}_{LPO}$, $\widehat{FNR}_{LOO}$, $\widehat{FNR}_{Twil}$, $\widehat{FNR}_{BH}$ and $\widehat{FNR}_{Best}$ respectively for the LPO, LOO, *Twilight*, Benjamini-Hochberg and Oracle procedures. In both steps of this study, s denotes the parameter of the Beta distribution that was used to simulate the data.

In comparison to the Oracle procedure (with the true π_0), Table 5.4 shows that the LPO procedure provides an actual value of the FDR that is almost always very close to the best possible one. Moreover in nearly all conditions, LPO outperforms its LOO counterpart and remains a little bit conservative, *e.g.* it furnishes a FDR that is lower or equal to the desired level α. This observation empirically confirms the result stated in Theorem 5.4.2. Besides as expected, the estimation of π_0 entails a tighter control than that of the BH-procedure where $\widehat{\pi}_0 = 1$. Unlike the proposed methods, *Twilight* fails in controlling the FDR at the desired level since $\widehat{FDR}_{Twil}$ is very often larger than $\widehat{FDR}_{Best}$ (the best reachable value), and even larger than α. Subsequently, *Twilight* should not enter in the comparison of methods in terms of power.
Table 5.5 enlightens that proportions of false negatives may be very high in most of the simulation conditions, as shown by the Oracle procedure. Nevertheless, $\widehat{FNR}_{LPO}$ remains very close to the ideal one. As a remark, note that the *Twilight* FNR estimates are also close to the Oracle values, but nearly always lower. As suggested by FDR results, LOO is less powerful that LPO, whereas both of them outperform by far the BH-procedure. Note that the proportion of false negatives strongly decreases when s grows, which means that $\mathbf{H}_1$ p-values are more and more concentrated in the neighbourhood of 0. As the interval on which assumption (**A**) is satisfied is wider, the problem becomes easier. Besides, we observe a fall in power when π_0 grows in general. Indeed for small proportion of true alternatives, the "border" between the two populations of p-values is more difficult to define as a large number of $\mathbf{H}_1$ p-values behave like $\mathbf{H}_0$ ones. Finally note that very often, the LPO procedure shares (nearly) the same power as the Oracle one.

5.5.4 Discussion

In this article, we propose a new estimator of the unknown proportion of true null hypotheses π_0. It relies on first the estimation of the common density of p-values by use of non-regular histograms of a special

Table 5.5: Average proportion of falsely non-rejected hypotheses (%) for the LPO ($\widehat{FNR}_{LPO}$), LOO ($\widehat{FNR}_{LOO}$), *Twilight* ($\widehat{FNR}_{Twil}$), Benjamini-Hochberg ($\widehat{FNR}_{BH}$) and Oracle ($\widehat{FNR}_{Best}$) procedures. s denotes the parameter of the Beta distribution used to generate the data.

s	π_0	$\widehat{FNR}_{LPO}$	$\widehat{FNR}_{LOO}$	$\widehat{FNR}_{Twil}$	$\widehat{FNR}_{BH}$	$\widehat{FNR}_{Best}$
5	0.5	93.94	94.22	91.64	99.78	94.16
	0.7	99.65	99.65	99.59	99.80	99.63
	0.9	99.87	99.87	99.86	99.89	99.86
	0.95	99.91	99.91	99.90	99.92	99.91
10	0.5	25.69	25.91	22.01	96.83	23.22
	0.7	96.36	96.44	95.08	99.16	96.03
	0.9	99.56	99.56	99.54	99.64	99.56
	0.95	99.76	99.76	99.76	99.77	99.74
25	0.5	0.88	0.90	0.70	17.72	0.79
	0.7	22.83	23.04	20.85	61.00	21.93
	0.9	97.89	97.89	97.68	98.49	97.86
	0.95	99.16	99.16	99.06	99.23	99.14
50	0.5	0.96	0.92	0.64	1.58	0.72
	0.7	2.26	2.30	2.01	10.07	2.19
	0.9	82.40	82.47	80.39	88.05	82.08
	0.95	96.74	96.76	96.60	97.15	96.74

type, and secondly on the leave-p-out cross-validation. The resulting estimator enables more flexibility than numerous existing ones, since at least it is still convenient in the "U-shape" case, without any supplementary computational cost.
Our estimator may be linked with that of Schweder and Spjøtvoll for which almost only theoretical results with λ fixed have been obtained by Storey. However unlike the latter, we provide a fully adaptive procedure that does not depend on any user-specified parameter λ. Thus, asymptotic optimality results are here derived with $\lambda = \widehat{\lambda}$. They assert, for instance, that the asymptotic exact control of the FDR with our plug-in MTP is reached.
Eventually, a wide range of simulations enlighten that the proposed π_0 estimator realizes the best bias-variance tradeoff among all tested estimates. Moreover, the proposed plug-in procedure is (empirically) shown to provide the expected control on the FDR (for finite samples), while being a little more powerful than its LOO counterpart. Moreover, the results in Section 5.5.2 confirm the interest in choosing adaptively the parameter p rather than the usual $p = 1$ value. The LPO procedure is very often almost as powerful as the best possible one of this type, obtained when π_0 is known.

5.6 Appendix

PROOF
First, we show that $T(\alpha, \cdot, \widehat{G})$ is right (resp. left) continuous on $(0, 1)$. As it is a similar reasoning, we only deal with right continuity.
Let $(\epsilon_n)_n \in \left(\mathbb{R}_+^*\right)^{\mathbb{N}^*}$ denote a sequence decreasing towards 0. For any $\theta \in (0,1)$, set $\forall n,\ r_n = T(\alpha, \theta + \epsilon_n, \widehat{G})$ *a.s.*. Then $(r_n)_n$ is an almost surely convergent increasing sequence, upper bounded by $T(\alpha, \theta, \widehat{G})$. To prove that $T(\alpha, \theta, \widehat{G})$ is its limit, we show that for any $\delta > 0$, there exists $\epsilon > 0$ satisfying $T(\alpha, \theta + \epsilon, \widehat{G}) \geq T(\alpha, \theta, \widehat{G}) - \delta$. Notice that there exists $\eta > 0$ s.t. $T := T(\alpha, \theta, \widehat{G}) = \sup\{t \in [\eta, 1] : \widehat{Q}_\theta(t) \leq \alpha\}$. Then for $0 < \delta < \eta$, $T - \delta = \sup\left\{u \in [\eta - \delta, 1 - \delta] : \frac{\theta(u+\delta)}{\widehat{G}(u+\delta)} \leq \alpha\right\}$. Provided δ is small enough, $\widehat{G}(u + \delta) = \widehat{G}(u),\ \forall u$. Hence, $T - \delta = \sup\left\{u \in [\eta - \delta, 1 - \delta] : \frac{\theta u}{\widehat{G}(u)} + \frac{\theta\delta}{\widehat{G}(u)} \leq \alpha\right\} \leq T(\alpha, \theta + \epsilon, \widehat{G}) = \sup\left\{t \in [0, 1] : \frac{\theta t}{\widehat{G}(t)} + \frac{\epsilon t}{\widehat{G}(t)} \leq \alpha\right\}$ for any $0 < \epsilon < \delta\theta$, which provides the result.
For the second point, define $G \in \mathcal{B}^+([0, 1])$ and for any sequence $(\epsilon_n)_n \in \left(\mathbb{R}_+^*\right)^{\mathbb{N}}$ decreasing towards 0, let $(H_n)_n \in \left(\mathcal{B}^+([0, 1])\right)^{\mathbb{N}}$ denote a sequence of positive bounded functions satisfying $\forall n,\ ||G - H_n||_\infty \leq \epsilon_n$.

Then for large enough n, we have

$$\frac{\theta t}{G(t) - \epsilon_n} \leq \alpha \quad \Leftrightarrow \quad \frac{\theta t}{G(t)} \leq \alpha \left(1 - \frac{\epsilon_n}{G(t)}\right),$$

and $\alpha(1 - \epsilon_n/||G||_\infty) \leq \alpha$. Thus, $r_n = \sup\{t : \ \theta t/(G(t) + \epsilon_n) \leq \alpha\}$ denotes an increasing sequence that is bounded by $T(\alpha, \theta, G)$. Moreover as $(\epsilon_n)_n$ decreases towards 0, r_n is as close as we want to $T(\alpha, \theta, G)$. The same reasoning may be followed with $r'_n = \sup\{t : \ \theta t/(G(t) - \epsilon_n) \leq \alpha\}$, which concludes the proof. ■

Bibliography

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

[2] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1–3):85–113, 2002.

[3] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Staitstical Society. Series B*, 57(1):289–300, 1995.

[4] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive Linear Step-up Procedures that control the False Discovery Rate. *Biometrika*, 93(3):491–507, 2006.

[5] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multipe testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[6] P. Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6:199, 2005.

[7] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

[8] S. Dudoit, J. Popper Shaffer, and J. C. Boldrick. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18(1):71–103, 2003.

[9] B. Efron. Large-Scale Simultaneous Hypothesis Testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[10] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of American Statistical Association*, 96(456):1151–1160, 2001.

[11] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

[12] U. Grenander. On the theory of mortality measurement. *Skandinavisk Aktuarietidskrift*, 39(2):125–153, 1956.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, 2001.

[14] M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning*, 27:7–50, 1997.

[15] M. Langaas, B. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Staitstical Society. Series B*, 67(4):555–572, 2005.

[16] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

[17] S. Pounds and C. Cheng. Robust estimation of the false discovery rate. *Bioinformatics*, 22(16):1979–1987, 2006.

[18] M. Rudemo. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.

[19] S. Scheid and R. Spang. A Stochastic Downhill Search Algorithm for Estimating the Local False Discovery Rate. *I.E.E.E. Transactions on Computational Biology and Bioinformatics*, 1(3):98–108, 2004.

[20] S. Scheid and R. Spang. Twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics*, 21(12):2921–2922, 2005.

[21] T. Schweder and E. Spjøtvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502, 1982.

[22] D. Scott. On Optimal and Data-Based Histograms. *Biometrika*, 66(3):605–610, 1979.

[23] J. Shao. Model Selection by Cross-Validation. *Journal of the American Statis. Association*, 88(422):486–494, 1993.

[24] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Staitstical Society. Series B*, 64(3):479–498, 2002.

[25] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B*, 66(1):187–205, 2004.

[26] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *PNAS*, 100(16):9440–9445, 2003.

[27] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[28] Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

Chapter 6

Cross-validation in density estimation: A model selection point of view

The present chapter aims at studying the leave-p-out algorithm in various collection complexity settings (polynomial and exponential), but always in the density estimation framework. Indeed, an important question in model selection is that of the complexity of the collection of models in hand. Actually, there are some interactions between the performance of an algorithm and the complexity of the collection.
We first consider the polynomial setup in which we derive an oracle inequality for any projection estimator as well as an adaptivity result in the minimax sense for histograms, with respect to Hölder balls.
In the exponential setting, some simulations illustrate the deficiency of the widely used leave-one-out as well as the leave-p-out with small values of p: both of them suffer some overfitting. We propose to add a penalty term to the leave-p-out risk estimator in order to take into account the high richness of the collection, which is responsible for this overfitting. An oracle inequality is derived in this setting as well.
We also consider the more general problem of overfitting encountered by the leave-one-out even in the polynomial setting. We use some simulation experiments to illustrate this point and show that a larger choice of p in the leave-p-out can balance the overfitting. From a practical point of view, we provide a data-driven penalty which exhibits some automatic adaptation properties to the collection complexity.

6.1 Introduction

Model selection via penalization has been introduced by the seminal works of Mallows and Akaike with respectively C_p [25] and AIC [2], and also by Schwarz [30] who proposed the BIC criterion. AIC and BIC rely on an asymptotic heuristics, which makes them dependent on the model collection in hand and on

the sample size [4]. For instance, the rationale behind Mallows' C_p and AIC is to provide an unbiased estimator of the risk for each model.
More recently, Birgé and Massart [8, 9, 10] have developed a non-asymptotic approach, inspired from the pioneering work of Barron and Cover [6]. From a countable collection of models denoted by $\{S_m\}_{m\in\mathcal{M}_n}$, we choose a family of estimators $(\widehat{s}_m)_{m\in\mathcal{M}_n}$. Their strategy consists in designing an "appropriate" penalized criterion

$$\mathrm{crit}(m) = \gamma_n(\widehat{s}_m) + \mathrm{pen}(m),$$

where $\gamma_n(\widehat{s}_m)$ denotes the empirical contrast at $\widehat{s}_m$ and $\mathrm{pen}: \mathcal{M}_n \mapsto \mathbb{R}_+$ denotes a penalty growing with the model complexity. The quality assessment of this criterion is made through an oracle inequality

$$\mathbb{E}\left[\, \|s - \widehat{s}_{\widehat{m}}\|^2 \,\right] \leq C \inf_{m\in\mathcal{M}_n} \left\{ \mathbb{E}\left[\, \|s - \widehat{s}_m\|^2 \,\right] + R(m,n) \right\},$$

where $C \geq 1$ is a constant independent from the density s and $R(m,n)$ denotes a remainder term with respect to $\mathbb{E}\left[\, \|s - \widehat{s}_m\|^2 \,\right]$. The closer C is to 1, the better the final estimator is. This is the general point of view we adopt throughout this chapter.
In the density estimation framework, Barron *et al.* [5] developed a general approach based on deterministic penalties leading to an oracle inequality with Kullback-Leibler divergence and Hellinger distance. This result has been adapted to the particular case of histograms by Castellan [15, 16] and further studied in [11]. With the quadratic loss, Birgé and Massart [8, 5] also derived some general results with projection estimators, which they apply to get some adaptivity results with respect to the smoothness of s [8].

Here, we address the problem of density estimation *via* cross-validation (CV) algorithms, unlike the aforementioned approaches relying on some deterministic penalties.
CV has been introduced by Stone [34, 35] for the leave-one-out (Loo) and Geisser [19, 20] for the V-fold cross-validation (VFCV) in a regression context and by Stone [33] in density estimation.
Rudemo [28] and Bowman [13] provided some closed-form expressions for the Loo estimator of the risk of histograms or kernel estimators. These results have been recently generalized by Celisse and Robin [17] to the leave-p-out cross-validation (Lpo).
Theoretical results about the effectiveness of CV algorithms are mainly asymptotic and concern the regression framework [24, 31, 32, 40]. As for non-asymptotic results in the density setting, Birgé and Massart [8] obtained an oracle inequality that may be applied to the Loo procedure. But to our knowledge, no result of this type has already been proved for the Lpo algorithm in the density estimation setup.

In the literature about model selection via penalization, an important notion is that of *complexity of the collection of models* [4], named in the sequel collection complexity. This notion already arises with discrete models in the minimum description length of Rissanen [27] and in the work of Barron and Cover [6] about minimum complexity. It is further generalized by Barron *et al.* [5] and Birgé and Massart [8, 9, 10] to the case of continuous models.
Collection complexity refers to the structure of the collection of models we consider. It could be understood as the collection counterpart of the model complexity, which is characterized by the dimension for instance with finite dimensional vector spaces. Following Barron *et al.* [5], the complexity of the collection of models may be quantified by the number of models with the same dimension for instance. In this setting, two wide situations are usually distinguished: the polynomial and the exponential complexity frameworks [10]. Baraud *et al.* [4] as well as Sauvé [29] introduce a complexity index which enables to characterize different complexity levels in the exponential case for instance.
Some connections between resampling algorithms such as Loo and collection complexity can be made through the asymptotic equivalence between Loo and C_p [24] or thanks to the non-asymptotic result of Birgé and Massart [8] in the context of density estimation. Thus in the same way as the AIC performance depends on the collection complexity [10, 14], Loo may be expected to inherit the same properties. For the Lpo, a similar indirect connection comes from the asymptotic equivalence between Lpo and FPE_α [41].

In this chapter, we aim at providing non-asymptotic theoretical results for the Lpo risk estimator in the density estimation framework as well as some simulation results to study the way the "optimal" choice of p and the collection complexity are intertwined.
In Section 6.2, we show that the Lpo algorithm may be understood as a random penalized criterion leading

to an increasing overpenalization as p grows. An oracle inequality is then derived in the polynomial complexity setting, which results in an adaptivity property in the minimax sense with respect to Hölder balls. In the exponential setting, some limitations of the Lpo with small values of p are illustrated in Section 6.3, where we propose the addition of a complexity term to the Lpo risk estimator. The resulting criterion leads to an oracle inequality with a $\log n$ term, which is the price to pay for such "rich" collections. Section 6.4 is devoted to some proofs. Besides through a simulation study in Section 6.5, we show that the Lpo and more generally CV techniques are sensitive to collection complexity even in the polynomial framework. Increasing p to overpenalize then enables to balance this phenomenon as observed in our simulation results. However, there is not yet any reliable guideline to choose p in front of real data. To this end, we propose in Section 6.6 a fully data-dependent algorithm. In some simulations, it is shown to automatically adapt itself to the collection complexity, either in the polynomial or exponential setting, therefore outperforming upon the Lpo strategy, at least for small values of p.

6.2 Polynomial complexity

6.2.1 Overpenalization of the Lpo risk

Ideal and Lpo penalties In model selection, our goal is to find the "best" model among $(S_m)_m$, in terms of a given criterion from a family of estimators $(\widehat{s}_m)_m$. A very common way to reach this goal is the minimization of a penalized criterion $\mathrm{crit}(\cdot)$ defined as

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m),$$

where γ is the contrast function that measures the quality of an estimator and $\mathrm{pen}(\cdot) : \mathcal{M}_n \to \mathbb{R}_+$ denotes the penalty term, which takes into account the complexity of the model S_m. In the density estimation setting, $\gamma(t, X) = \|t\|^2 - 2t(X)$, where $X \sim P$.
Ideally, the optimal criterion we would minimize over $\mathcal{M}_n$ is the random quantity

$$\mathrm{crit}_{id}(m) = P\gamma(\widehat{s}_m) := \mathbb{E}\gamma\left(\widehat{s}_m, Z\right),$$

with the expectation taken with respect to $Z \sim P$, where Z is independent from the original data. The link between these two criteria can be made by rewriting the latter as follows

$$\mathrm{crit}_{id}(m) = P_n\gamma(\widehat{s}_m) + \left[\, P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m)\,\right].$$

The quantity in square brackets is named the ideal penalty

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{pen}_{id}(m) := P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m).$$

Following the CV strategy, we perform model selection by minimizing the Lpo risk estimator over $\mathcal{M}_n$. Thus for a given $1 \leq p \leq n-1$, the candidate $\widehat{m}$ is

$$\widehat{m} = \mathrm{Argmin}_{m \in \mathcal{M}_n} \widehat{R}_p(m).$$

The idea that there is a strong relationship between penalized criteria and CV is strongly supported by the large amount of literature about the comparison of these two aspects [35, 24, 41]. Therefore, we may try to include the CV strategy into the wider scope of penalized criteria minimization. Thus,

$$\widehat{m} = \mathrm{Argmin}_{m \in \mathcal{M}_n} \left\{ P_n\gamma(\widehat{s}_m) + \left[\widehat{R}_p(m) - P_n\gamma(\widehat{s}_m) \right] \right\}.$$

In the above expression, the quantity in square brackets is called the *Lpo penalty*:

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{pen}_p(m) := \widehat{R}_p(m) - P_n\gamma(\widehat{s}_m).$$

This Lpo penalty may be subsequently understood as a random penalty. Note that a similar approach applied to the Loo can be found in Birgé and Massart [8].
Thanks to this parallel between CV and penalized criteria, we attempt to get more insight in the behaviour of CV techniques, for instance with respect to the parameter p.

Lpo overpenalization In this section, we aim at making comparison between pen_{id} and pen_p, so that we characterize some features in the behaviour of pen_p with respect to p. This comparison is carried out through the expectations of these penalties, which are both random variables.

In the sequel, we consider general projection estimators with any orthonormal basis of functions. Let us define $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$ a family of functions in $L^2([0,1],\nu)$, where ν denotes the Lebesgue measure on $[0,1]$ and Λ_n is a countable set of indices. For any $m\in\mathcal{M}_n$, set $\Lambda(m)\subset\Lambda_n$ such that $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ is an orthonormal family of functions. Let S_m denote the linear space of dimension D_m spanned by $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$.
We call $\widehat{s}$ a *projection estimator* any estimator of s such that

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} \widehat{\beta}_\lambda\varphi_\lambda \qquad \text{with} \qquad \widehat{\beta}_\lambda = P_n\varphi_\lambda := \frac{1}{n}\sum_{i=1}^n \varphi_\lambda(Z_i),$$

where $Z_1,\ldots,Z_n$ denote some observations [36]. A typical example of such an estimator is the histogram for which $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$, where $\{I_\lambda\}_{\lambda\in\Lambda(m)}$ denotes a partition of $[0,1]$ and $|I_\lambda|$ represents the length of the interval I_λ.

The main concern of the following result is to assess the behaviour (in expectation) of the Lpo penalty with respect to the ideal one. This question is addressed with general projection estimators.
We start with a preliminary lemma:

Lemma 6.2.1. *With the same notations as before with any projection estimator $\widehat{s}_m$ onto S_m, we obtain*

$$\begin{aligned}\mathbb{E}\left[\,\mathrm{pen}_{id}(m)\,\right] &= \frac{2}{n}\sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left(\varphi_\lambda(X)\right),\\ \mathbb{E}\left[\,\mathrm{pen}_p(m)\,\right] &= \frac{2n-p}{n(n-p)}\sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left(\varphi_\lambda(X)\right).\end{aligned}$$

We now state the main assertion about the Lpo penalty associated with projection estimators in density estimation, which results from the previous lemma.

Proposition 6.2.1. *For any $m\in\mathcal{M}_n$, let $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ denote an orthonormal basis of S_m and $\widehat{s}_m$, the projection estimator onto S_m. Then, we get*

$$\forall m\in\mathcal{M}_n,\ 1\le p\le n-1, \qquad \mathbb{E}\left[\,\mathrm{pen}_p(m)-\mathrm{pen}_{id}(m)\,\right] = \frac{p}{n(n-p)}\sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left(\varphi_\lambda(X)\right)\ge 0.$$

Since this quantity remains nonnegative whatever p, we conclude that the Lpo penalty always overpenalizes, which remains true for any orthonormal basis. Moreover, the amount of overpenalization increases with p. Thus, the Loo provides the weakest overpenalization of order $\mathcal{O}\left(1/n^2\right)$, whereas the Lpo with $p\simeq n/2$ (which is similar to the 2-fold CV) corresponds to an overpenalization of the same order as the expectation of the ideal penalty, that is $\mathcal{O}\left(1/n\right)$.

6.2.2 Oracle inequality

Purpose and strategy In the following, we assess the quality of the Lpo-based model selection procedure through the statement of an oracle inequality. This result is settled in the polynomial complexity framework and holds for any projection estimator. To our knowledge, it is the first non-asymptotic result about the performance of the Lpo in this framework. We point out that unlike the usual approach in model selection via penalization, our purpose is not to design a penalty function since the Lpo estimator itself may be understood as a penalized criterion (Section 6.2.1).

Let us first describe the outlines of our strategy. We start with the definition of $\widehat{s}_{\widehat{m}}$ as the minimizer of the Lpo risk estimator, which leads to an inequality (6.2) written so as we stress the discrepancy between the Lpo estimator and its expectation, for each model in the collection. Then, we show that this discrepancy can be studied on a set of high probability (Lemma 6.2.4) rather than on the whole space . The gap between the Lpo risk and its expectation is evaluated through the use of two concentration inequalities: Bernstein's and a version of Talagrand's inequality (Proposition 6.2.2 and Proposition 6.2.3). By recombination of these different results, we derive the main inequality which holds except on a set of small probability (6.6). The conclusion results from the following lemma:

Lemma 6.2.2. *Let X and Y be two random variables such that* $\forall z > 0, \quad \mathbb{P}\left(X \geq Y + K_1 z + K_2\right) \leq \Sigma e^{-z}$, *where* $K_1,\ K_2,\ \Sigma > 0$. *Then, we have*

$$\mathbb{E}X \leq \mathbb{E}Y + K_1\,\Sigma + K_2.$$

The straightforward proof of this lemma is deferred to Section 6.4.

Main result Our main result relies on several assumptions that we now present and discuss. Set $X \sim s$ and for any m,

$$\phi_m = \sum_{\lambda \in \Lambda(m)} \varphi_\lambda^2 \qquad \text{and} \qquad V_m = \mathbb{E}\,\phi_m(X).$$

Then, we define the following assumptions

(Reg) $\quad \exists \Phi > 0 / \quad \sup_{m \in \mathcal{M}_n} \|\phi_m\|_\infty \leq \Phi\, n / \left(\log n\right)^2,$

($Reg2$) $\quad \exists \Phi > 0 / \quad \sup_{m \in \mathcal{M}_n} \left\{ \sup_{(\alpha_\lambda)_\lambda,\, |a|_\infty = 1} \|\sum_\lambda \alpha_\lambda \varphi_\lambda\|_\infty \right\} \leq \sqrt{\Phi\, n / \left(\log n\right)^2},$

(Ad) $\quad \exists \xi > 0 / \ \forall m \in \mathcal{M}_n$ with $D_m \geq 2, \quad n\mathbb{E}\left[\, \|s_m - \widehat{s}_m\|_2^2 \,\right] \geq \xi D_m,$

(Pol) $\quad \exists \delta > 0 / \ \forall D \geq 1,\ |\{m \in \mathcal{M}_n \mid D_m = D\}| \leq D^\delta.$

Since $\phi_m = \sum_{\lambda \in \Lambda(m)} \varphi_\lambda^2$, $\|\phi_m\|_\infty$ may be understood as a regularity measure of the basis $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$. Thus, ($Reg$) relates the regularity of each considered basis to the amount of data. For instance, let us assume we use histogram estimators based on a partition of $[0,1]$ in D_m intervals $(I_1, \ldots, I_{D_m})$, $\varphi_\lambda = \mathbb{1}_{I_\lambda} / \sqrt{|I_\lambda|}$, where $|I_\lambda|$ is the length of I_λ. Then, (Reg) gives a lower bound for the minimal length of an interval I_λ of the partition with respect to the number of observations. In other words, we cannot consider partitions with an intervals with less than $n/(\log n)^2$ observations.

($Reg2$) is another regularity assumption about $(\varphi_\lambda)_{\lambda \in \Lambda(m)}$. In the specific case of a basis defined from a partition of $[0,1]$, like histograms or piecewise polynomials, (Reg) implies ($Reg2$) . But it is no longer true with general bases of functions. Therefore, the constant Φ is not necessarily the same in both (Reg) and ($Reg2$) . However replacing one of them by the maximum value provides the same constant, which will be assumed in the following.

If we develop $\mathbb{E}\left\|s_m - \widehat{s}_m\right\|^2$, we see that

$$\begin{aligned}
\mathbb{E}\left\|s_m - \widehat{s}_m\right\|^2 &= \sum_{\lambda \in \Lambda(m)} \mathbb{E}\left[\, \nu_n^2\left(\varphi_\lambda\right) \,\right], \\
&= \sum_{\lambda \in \Lambda(m)} \frac{1}{n} \mathrm{Var}\left[\, \varphi_\lambda(X) \,\right].
\end{aligned}$$

For instance if we use histograms, $\mathrm{Var}\left[\, \varphi_\lambda(X) \,\right]$ vanishes if and only if the support of s is included in I_λ. (Ad) therefore requires that for any m, there are always "enough" informative basis vectors, if an informative vector is a vector such that $\mathrm{Var}\left[\, \varphi_\lambda(X) \,\right] \neq 0$. With histograms, it means that we do not consider bases with more and more vectors where $s = 0$. Note that (Ad) holds with histograms if $s \geq \rho > 0$ on $[0,1]$.

(Pol) simply comes from the definition of the polynomial complexity. At most, the cardinality of the set of models with dimension D is polynomial in D. Such an assumption is satisfied with nested models for instance (Section 2.1.2).

Theorem 6.2.1. *Let s denote a bonded density on [0,1] and $X_1, \ldots, X_n$ be n i.i.d. random variables drawn from s. Set $\{\varphi_\lambda\}_{\lambda \in \Lambda_n}$ a finite family of bounded functions on $[0,1]$ such that for any $m \in \mathcal{M}_n$, S_m denotes the vector space of dimension D_m, spanned by the orthonormal family $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$. Let us assume that (Reg) , $(Reg2)$, (Ad) and (Pol) hold.*
For $n \geq 29$, set $0 < \epsilon < 1$ such that

$$\frac{4\zeta(\epsilon)}{1 + 3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1) - 2} < 1\,, \tag{6.1}$$

where $\zeta(\epsilon) = \left[1 - (1+\epsilon)^{-8}\right]$. Furthermore, let us choose $1 \leq p \leq n-1$ satisfying

$$(Ran) \qquad \frac{4\zeta(\epsilon)}{1 + 3\zeta(\epsilon)} + \frac{2}{n}\frac{1 + \zeta(\epsilon)}{1 + 3\zeta(\epsilon)} + \alpha \leq \frac{p}{n} \leq 1 - \frac{2}{\zeta(\epsilon)(n-1) - 2} - \beta$$

with $0 < \alpha, \beta < 1$,
Then, there exists a set $\Omega_n(\epsilon)$ of probability larger than $1 - o(1/n^2)$ such that

$$\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)} \left\| s - \widehat{s}_{\widehat{m}} \right\|^2\right] \leq \Gamma(\epsilon, \alpha, \beta) \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\left\| s - \widehat{s}_m \right\|^2\right] + \frac{\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta)}{n},$$

where $\Gamma(\epsilon, \alpha, \beta) \geq 1$ is a constant (with respect to n) independent from s and $\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta) \geq 0$ is another constant.

REMARKS:

- The existence of ϵ satisfying the inequality (6.1) stems from a technical lemma (Lemma 6.4.1) given in the proof of Theorem 6.2.1.
- (Ran) is a *sufficient condition* for the oracle inequality to hold. In this assumption, α and β can be chosen as small as we want, but cannot vanish.
- As it is made clear from the proof of Lemma 6.4.1, the choice of ϵ is constrained. For instance, ϵ cannot be too much close to 0. This explains why the nonintuitive bounds in (Ran) cannot be easily simplified. Furthermore, this enlightens that "small values" of p could be excluded from the range of values described in (Ran) , to which the oracle inequality applies.
- In the previous result, the $o\left(1/n^2\right)$ term depends on ϵ as explained in the proof.

Without any more assumption, we also derive the following corollary in which the expectation of the left-hand side is computed over the whole space.

Corollary 6.2.1. *With the notations of the previous theorem, we have for any $\epsilon > 0$,*

$$\mathbb{E}\left[\left\| s - \widehat{s}_{\widehat{m}} \right\|^2\right] \leq C(\epsilon, \alpha, \beta) \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\left\| s - \widehat{s}_m \right\|^2\right] + \frac{\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta)}{n} + o\left(\frac{1}{n^2}\right),$$

where $C(\epsilon, \alpha, \beta) \geq 1$ is a constant independent from s and $\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta) \geq 0$ is a constant.

Indeed at the price of a $o\left(1/n^2\right)$, we may derive the same inequality as in Theorem 6.2.1, but on the whole space rather than on an event of high probability.
REMARKS:

- This remainder term turns out to be $o\left(1/n^\gamma\right)$, for any $\gamma > 0$. Indeed, it is directly related to the probability in Lemma 6.2.4. Actually, we have an explicit (non-asymptotic) expression for it as we can check in the proof of Theorem 6.2.1.
- $o\left(1/n^2\right)$ depends on ϵ as well as some other constants of the problem.

Preliminaries The proof of Theorem 6.2.1 is rather technical and relies on several intermediate results. In order to enlighten as clearly as possible our reasoning rather than the involved technicalities, we provide these results beforehand. Some of the proofs are postponed to Section 6.4.

We give a few notations. For any $p \in \{1, \ldots, n-1\}$ the Lpo risk estimator associated with the estimator $\widehat{s}_m$ is denoted by $\widehat{R}_p(m)$. For the sake of clarity, we define

$$\forall m, \qquad L_p(m) = \mathbb{E}\widehat{R}_p(m),$$

such that $L_p(\widehat{m}) := \mathbb{E}\left[\widehat{R}_p(m)\right]_{|m=\widehat{m}}$. For each m, $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ denotes an orthonormal basis of S_m. Moreover, we set

$$\begin{aligned}
\phi_m &= \sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 && \text{and} && V_m = \mathbb{E}\left[\,\phi_m(X)\,\right],\\
s_m &= \sum_{\lambda\in\Lambda(m)} \beta_\lambda\varphi_\lambda && \text{and} && \beta_\lambda = P\varphi_\lambda,\\
\widehat{s}_m &= \sum_{\lambda\in\Lambda(m)} \widehat{\beta}_\lambda\varphi_\lambda && \text{and} && \widehat{\beta}_\lambda = P_n\varphi_\lambda,\\
\chi^2(m) &= \;\|s_m - \widehat{s}_m\|^2 && = \sum_\lambda \nu_n^2\left(\varphi_\lambda\right), &&\\
E_m &= \mathbb{E}\left[\,\chi^2(m)\,\right] && \text{and} && \theta_{n,p} = \frac{2n-p}{(n-1)(n-p)}\,.
\end{aligned}$$

Remark: $\chi^2(m)$ is not a true χ^2 statistic, but is only somewhat similar to it.

We also stress two elementary but useful properties that will be repeatedly used in the following. For any $a, b \geq 0$,

$$\begin{aligned}
&(Roo) && \sqrt{a+b} \leq \sqrt{a} + \sqrt{b},\\
&(Squ) && 2ab \leq \eta a^2 + \eta^{-1}b^2, \quad \forall \eta > 0.
\end{aligned}$$

The first intermediate result deals with the relationship between $\widehat{R}_p$ and its expectation for each model.

Lemma 6.2.3. *For any $m \in \mathcal{M}_n$,*

$$\begin{aligned}
L_p(m) - L_p(\widehat{m}) &= \frac{n}{n-p}\left[\,E_m - E_{\widehat{m}}\,\right] - \left(\|s - s_{\widehat{m}}\|^2 - \|s - s_m\|^2\right),\\
\widehat{R}_p(m) - L_p(m) &= \theta_{n,p}\,\nu_n(\phi_m) - (1+\theta_{n,p})\left[\,\chi^2(m) - E_m\,\right] - 2(1+\theta_{n,p})\nu_n(s_m).
\end{aligned}$$

In Lemma 6.2.3, we see that $\nu_n\left(\phi_m\right)$ appears in the expressions. The following Proposition enables to upper bound the deviation of this quantity. It is a consequence of Bernstein's inequality [26].

Proposition 6.2.2. *With the above notations, let $z > 0$ and $C > 0$ be any positive constants and for each m, let us define $y_m = z + C\,nE_m$. Then, we have*

$$\forall m \in \mathcal{M}_n, \quad \mathbb{P}\left[\left|\nu_n\left(\phi_m\right)\right| \geq \sqrt{2V_m\frac{\|\phi_m\|_\infty}{n}y_m} + \frac{\|\phi_m\|_\infty}{n}y_m\right] \leq 2e^{-y_m}.$$

Moreover if (Ad) holds, we have

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \left|\nu_n\left(\phi_m\right)\right| \geq \sqrt{2V_m\frac{\|\phi_m\|_\infty}{n}y_m} + \frac{\|\phi_m\|_\infty}{n}y_m\right] \leq \Sigma_1 e^{-z},$$

where Σ_1 is a positive constant independent from n.

We now recall that $\chi^2(m) = \sum_\lambda \nu_n^2(\varphi_\lambda)$. A handy way to study this χ^2-like statistic is to introduce an event of large probability on which we are able to get some control. That is the reason why we introduce the event $\Omega_n(\epsilon)$ for any $\epsilon > 0$.

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\},$$

where $\kappa(t) = 2(t^{-1} + 1/3)$.
Another use of Bernstein's inequality provides the following Lemma.

Lemma 6.2.4. *Set $\epsilon > 0$ and assume that (Reg) , (Reg2) and (Pol) hold. Then,*

$$\forall \alpha > 0, \qquad \mathbb{P}\left[\Omega_n^c(\epsilon)\right] \leq 2n^{2+\delta} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2} = o\left(\frac{1}{n^\alpha}\right),$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

This Lemma turns out to be useful in order to assess the concentration of $\chi^2(m)$ around its expectation. This result may be found in Massart [26] and is a consequence of Talagrand's inequality.

Proposition 6.2.3. *With the above notations, set $\epsilon > 0$ and for any $C', z > 0$, $x_m = z + C' nE_m$. Assume that (Reg) , (Reg2) and (Pol) are fulfilled. Then,*

$$\forall m \in \mathcal{M}_n, \qquad \mathbb{P}\left[\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty x_m}\right)\right] \leq e^{-x_m}.$$

Furthermore if (Ad) holds,

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty x_m}\right)\right] \leq \Sigma_2 e^{-z},$$

where $\Sigma_2 > 0$ denotes a positive constant independent from n.

Finally, in Lemma 6.2.3, it remains $\nu_n(s_m)$ for which nothing has already been made. The control of this quantity comes from an upper bound, which results from the following lemma.

Lemma 6.2.5. *Set $m, m' \in \mathcal{M}_n$. Then for any $\rho > 0$,*

$$\sup_{t \in S_m + S_{m'}} \nu_n^2\left(\frac{t}{\|t\|}\right) \leq (1+\rho)\chi^2(m) + (1+\rho^{-1})\chi^2(m').$$

Proofs

Proof. (Theorem 6.2.1)
We are now in position to give the main inequality from which we derive Theorem 6.2.1.
From the definition of $\widehat{m}$ as $\widehat{m} = \text{Argmin}_{m \in \mathcal{M}_n} \widehat{R}_p(m)$, we deduce that

$$\forall m \in \mathcal{M}_n, \qquad \widehat{R}_p(\widehat{m}) \leq \widehat{R}_p(m),$$

which implies

$$\left[\widehat{R}_p(\widehat{m}) - L_p(\widehat{m})\right] \leq \left[\widehat{R}_p(m) - L_p(m)\right] + \left[L_p(m) - L_p(\widehat{m})\right]. \tag{6.2}$$

Then, we apply Lemma 6.2.3 to (6.2) and get

$$\begin{aligned} \|s - s_{\widehat{m}}\|^2 + n\theta_{n,p}E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \quad \leq \quad & \|s - s_m\|^2 + n\theta_{n,p}E_m - (1+\theta_{n,p})\chi^2(m) + \\ & \theta_{n,p}\nu_n(\phi_m - \phi_{\widehat{m}}) + 2(1+\theta_{n,p})\nu_n(s_{\widehat{m}} - s_m). \end{aligned} \tag{6.3}$$

Remarks:

- An upper bound for $\nu_n(\phi_m - \phi_{\widehat{m}})$ may be obtained through Bernstein's inequality, so that we may relate $\nu_n(\phi_{\widehat{m}})$ to $E_{\widehat{m}}$. This is reached thanks to Proposition 6.2.2.
- Ideally in the oracle inequality we have in mind, the left-hand side of the final inequality is something like $\mathbb{E}\left[\, \|s - \widehat{s}_{\widehat{m}}\|^2 \right]$, which is equal to $\mathbb{E}\left[\, \|s - s_{\widehat{m}}\|^2 \right] + \mathbb{E}\left[\, \chi^2(\widehat{m}) \,\right]$ with the present notations. However in (6.3), we observe that the left-hand side is $\mathbb{E}\left[\, \|s - s_{\widehat{m}}\|^2 \right] + \mathbb{E}\left[\, E_{\widehat{m}} \,\right]$. In order to relate $\mathbb{E}\left[\, E_{\widehat{m}} \,\right]$ to $\mathbb{E}\left[\, \chi^2(\widehat{m}) \,\right]$, we will uniformly control the discrepancy $E_m - \chi^2(m)$ over $\mathcal{M}_n$ thanks to both Lemma 6.2.4 and Proposition 6.2.3.
- Finally, $\nu_n(s_{\widehat{m}} - s_m)$ may be upper bounded thanks to Lemma 6.2.5, independently from $E_{\widehat{m}}$ and will therefore be dealt with later.

According to the preceding remarks, we first apply Proposition 6.2.2 to $\nu_n(\phi_m - \phi_{\widehat{m}})$. The successive use of (Reg) , (Squ) with any $\eta > 0$, and (Roo) provides

$$\begin{aligned}\sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} &\leq \sqrt{2V_m \Phi y_m},\\ &\leq \eta \Phi V_m + \eta^{-1} y_m.\end{aligned}$$

Moreover, note that

$$V_m = \sum_\lambda \mathbb{E}\left[\, \varphi_\lambda^2(X) \,\right] = nE_m + \|s_m\|^2 \leq nE_m + \|s\|^2 .$$

Hence with $y_m = z + C\, nE_m$,

$$\sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} \leq \left[\, \eta\Phi + \eta^{-1} C \,\right] nE_m + \eta\Phi \|s\|^2 + \eta^{-1} z.$$

Similarly, (Reg) entails that

$$\frac{\|\phi_m\|_\infty}{3n} y_m \leq \frac{\Phi C}{3} nE_m + \frac{\Phi}{3} z,$$

which leads us to

$$\begin{aligned}|\nu_n(\phi_m - \phi_{\widehat{m}})| \;\leq\; & nE_m \left[\eta\Phi + C\eta^{-1} + \Phi\frac{C}{3}\right] + nE_{\widehat{m}} \left[\eta\Phi + C\eta^{-1} + \Phi\frac{C}{3}\right] + 2z\left[\frac{\Phi}{3} + \eta^{-1}\right] + \\ & 2\eta\Phi \|s\|^2 ,\end{aligned}$$

except on an event of probability less than $\Sigma_1 e^{-z}$.
Set $\epsilon'' > 0$ and let us choose $\eta = \epsilon''/(3\Phi)$ and $C = 2\epsilon''/\left(\eta^{-1} + \Phi/3\right)$. Then it comes that

$$|\nu_n(\phi_m - \phi_{\widehat{m}})| \leq nE_m\epsilon'' + nE_{\widehat{m}}\epsilon'' + 2z\Phi\left[\frac{1}{3} + \frac{3}{\epsilon''}\right] + 2\frac{\epsilon''}{3}\|s\|^2 ,$$

Plugging this into (6.3) provides

$$\begin{aligned}\|s - s_{\widehat{m}}\|^2 + n\theta_{n,p}(1-\epsilon'')E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \leq \|s - s_m\|^2 + n\theta_{n,p}(1+\epsilon'')E_m - (1+\theta_{n,p})\chi^2(m) + \\ 2(1+\theta_{n,p})\nu_n(s_{\widehat{m}} - s_m) + \theta_{n,p}\left(2z\Phi\left[\frac{1}{3} + \frac{3}{\epsilon''}\right] + 2\frac{\epsilon''}{3}\|s\|^2\right), \qquad (6.4)\end{aligned}$$

except on an event of probability less than $\Sigma_1 e^{-z}$.

On the other hand, Proposition 6.2.3 implies that for a given $\epsilon > 0$, except on a set of probability less than $\Sigma_2 e^{-z}$, we have

$$\forall m \in \mathcal{M}_n, \quad \sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty x_m}\right).$$

Using $x_m = z + C'nE_m$ and (Roo) , we get

$$\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)\left(\sqrt{nE_m}\left[1+\sqrt{2\left\|s\right\|_\infty C'}\right]+\sqrt{2\left\|s\right\|_\infty z}\right),$$

which in turn, combined with (Squ) , implies for any $x > 0$

$$\chi^2(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)^2\left((1+x)E_m\left[1+\sqrt{2\left\|s\right\|_\infty C'}\right]^2+(1+x^{-1})\frac{2\left\|s\right\|_\infty}{n}z\right).$$

It holds for the particular choices $x = \epsilon$ and $C' = \left(1-\sqrt{1+\epsilon}\right)^2 / \left(2\left\|s\right\|_\infty\right)$, which results in

$$\frac{1-\epsilon''}{(1+\epsilon)^4}\chi^2(\widehat{m})\mathbb{1}_{\Omega_n(\epsilon)} \leq (1-\epsilon'')E_{\widehat{m}} + \frac{1-\epsilon''}{\epsilon(1+\epsilon)}\frac{2\left\|s\right\|_\infty}{n}z.$$

with probability larger than $1 - \Sigma_2 e^{-z}$.
From the above result and (6.4), it comes that on $\Omega_n(\epsilon)$, with probability larger than $1 - (\Sigma_1 + \Sigma_2)\, e^{-z}$, we have

$$\begin{aligned}\left\|s-s_{\widehat{m}}\right\|^2+\left(n\theta_{n,p}\frac{1-\epsilon''}{(1+\epsilon)^4}-(1+\theta_{n,p})\right)\chi^2(\widehat{m}) \quad\leq\quad & \left\|s-s_m\right\|^2+n\theta_{n,p}(1+\epsilon'')E_m-(1+\theta_{n,p})\chi^2(m)+\\ & 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}}-s_m\right)+\\ & \theta_{n,p}z\left(\frac{1-\epsilon''}{\epsilon(1+\epsilon)}2\left\|s\right\|_\infty+2\Phi\left[\frac{1}{3}+\frac{3}{\epsilon''}\right]\right)+\\ & 2\theta_{n,p}\frac{\epsilon''}{3}\left\|s\right\|^2.\end{aligned}$$

Now for any $\epsilon > 0$, we define $\epsilon' > 0$ such that $\sqrt{1-\epsilon'} = (1+\epsilon)^{-4}$ and let us take ϵ'' satisfying $1 - \epsilon'' = \sqrt{1-\epsilon'}$. Then, the above inequality becomes

$$\begin{aligned}\left\|s-s_{\widehat{m}}\right\|^2+\left[n\theta_{n,p}\left(1-\epsilon'\right)-(1+\theta_{n,p})\right]\chi^2(\widehat{m}) \quad\leq\quad & \left\|s-s_m\right\|^2+n\theta_{n,p}\left[2-\sqrt{1-\epsilon'}\right]E_m-(1+\theta_{n,p})\chi^2(m)+\\ & 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}}-s_m\right)+\\ & \theta_{n,p}z\left(\frac{\sqrt{1-\epsilon'}}{\epsilon(1+\epsilon)}2\left\|s\right\|_\infty+2\Phi\left[\frac{1}{3}+\frac{3}{1-\sqrt{1-\epsilon'}}\right]\right)+\\ & 2\theta_{n,p}\frac{1-\sqrt{1-\epsilon'}}{3}\left\|s\right\|^2. \end{aligned} \tag{6.5}$$

The following point consists in deriving an upper bound for $\nu_n\left(s_{\widehat{m}} - s_m\right)$. It results from the following inequalities and Lemma 6.2.5. Indeed, we have

$$\begin{aligned}2\nu_n\left(s_{\widehat{m}}-s_m\right) \quad&\leq\quad 2\nu_n\left(\frac{s_{\widehat{m}}-s_m}{\left\|s_{\widehat{m}}-s_m\right\|}\right)\left\|s_{\widehat{m}}-s_m\right\|,\\ &\leq\quad 2\sup_{t\in S_{\widehat{m}}+S_m}\nu_n\left(\frac{t}{\|t\|}\right)\left\|s_{\widehat{m}}-s_m\right\|.\end{aligned}$$

Moreover, $\left\|s_{\widehat{m}} - s_m\right\| \leq \left\|s_{\widehat{m}} - s\right\| + \left\|s - s_m\right\|$ and a double use of (Squ) give for any $x > 0$:

$$2\nu_n\left(s_{\widehat{m}}-s_m\right) \quad\leq\quad (1+x)\sup_{t\in S_{\widehat{m}}+S_m}\nu_n^2\left(\frac{t}{\|t\|}\right)+\frac{2}{2+x}\left\|s_{\widehat{m}}-s\right\|^2+\frac{2}{x}\left\|s_m-s\right\|^2.$$

Finally, Lemma 6.2.5 yields that for any $\rho > 0$, we have

$$2\nu_n\left(s_{\widehat{m}}-s_m\right) \quad\leq\quad (1+x)\left[(1+\rho)\chi^2(\widehat{m})+(1+\rho^{-1})\chi^2(m)\right]+\frac{2}{2+x}\left\|s_{\widehat{m}}-s\right\|^2+\frac{2}{x}\left\|s_m-s\right\|^2.$$

With $x = \epsilon'$ and $\rho = \epsilon'(1+\epsilon')^{-1}$, we get

$$2\nu_n\left(s_{\widehat{m}}-s_m\right) \quad\leq\quad \left(1+2\epsilon'\right)\chi^2(\widehat{m})+(1+\epsilon')\frac{1+2\epsilon'}{\epsilon'}\chi^2(m)+\frac{2}{2+\epsilon'}\left\|s_{\widehat{m}}-s\right\|^2+\frac{2}{\epsilon'}\left\|s_m-s\right\|^2.$$

Plugging this in (6.5) yields:
On the event $\Omega_n(\epsilon)$, with probability larger than $1 - (\Sigma_1 + \Sigma_2)\, e^{-z}$, we have for any $m \in \mathcal{M}_n$

$$\begin{aligned}\left[\frac{\epsilon' - 2\theta_{n,p}}{2+\epsilon'}\right] \|s - s_{\widehat{m}}\|^2 + \left[\, n\theta_{n,p}\,(1-\epsilon') - 2(1+\theta_{n,p})(1+\epsilon')\,\right] \chi^2(\widehat{m}) \leq \left[1 + \frac{2}{\epsilon'}(1+\theta_{n,p})\right] \|s - s_m\|^2 + \\ n\theta_{n,p}\left[2 - \sqrt{1-\epsilon'}\right] E_m + \\ \left[\frac{1+2\epsilon'+2\epsilon'^2}{\epsilon'}\right](1+\theta_{n,p})\chi^2(m) + \\ \theta_{n,p}(Az+B),\end{aligned} \tag{6.6}$$

where $A = \left(\frac{\sqrt{1-\epsilon'}}{\epsilon(1+\epsilon)} 2\,\|s\|_\infty + 2\Phi\left[\frac{1}{3} + \frac{3}{1-\sqrt{1-\epsilon'}}\right]\right)$ and $B = 2\frac{1-\sqrt{1-\epsilon'}}{3}\,\|s\|^2$.
Then, Lemma 6.2.2 allows us to take the expectation and get the following result.

$$(\psi_1 \wedge \psi_2)\, \mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)}\, \|s - \widehat{s}_{\widehat{m}}\|^2\right] \leq (\psi_3 \vee \psi_4)\, \mathbb{E}\left[\|s - \widehat{s}_m\|^2\right] + \theta_{n,p}\left[\, A\,(\Sigma_1 + \Sigma_2) + B\,\right], \tag{6.7}$$

where

$$\begin{aligned}\psi_1 &= \frac{\epsilon' - 2\theta_{n,p}}{2+\epsilon'}\\ \psi_2 &= n\theta_{n,p}\,(1-\epsilon') - 2(1+\theta_{n,p})(1+\epsilon')\\ \psi_3 &= 1 + \frac{2}{\epsilon'}(1+\theta_{n,p})\\ \psi_4 &= n\theta_{n,p}\left[2 - \sqrt{1-\epsilon'}\right] + (1+\theta_{n,p})\left[\frac{1+2\epsilon'+2\epsilon'^2}{\epsilon'}\right].\end{aligned}$$

In order to obtain a meaningful inequality, a necessary requirement is $\psi_1,\ \psi_2,\ \psi_3,\ \psi_4 \geq 0$. This is already satisfied for ψ_3 and ψ_4. We have only to check it for both ψ_1 and ψ_2.
It turns out that if $\epsilon' > 2/(n-1)$, then p must satisfy

$$\frac{4\epsilon'}{1+3\epsilon'} + \frac{2}{n}\frac{1+\epsilon'}{1+3\epsilon'} \leq \frac{p}{n} \leq 1 - \frac{2}{\epsilon'(n-1)-2}, \tag{6.8}$$

provided

$$\frac{4\epsilon'}{1+3\epsilon'} + \frac{2}{n}\frac{1+\epsilon'}{1+3\epsilon'} \leq 1 - \frac{2}{\epsilon'(n-1)-2},$$

which is established by Lemma 6.4.1 for $n \geq 29$.
Remark: In (6.8) since $0 < \epsilon' \leq 1$ by definition, we have $\frac{4\epsilon'}{1+3\epsilon'} \leq 1$.

Finally to assert the existence if the constant Γ in Theorem 6.2.1, we need to make sure that the ratio $(\psi_3 \vee \psi_4)\,/\,(\psi_1 \wedge \psi_2)$ is bounded.
It may be easily checked that all ψ_ks may be reshaped as

$$\psi_k = \frac{F(p,n)}{1 - p/n},$$

where F is a bounded quantity. Moreover by construction, the bounds in (6.8) lead to $\psi_1 = 0$ and $\psi_2 = 0$, which should be prohibited since we would like to consider the ratio $(\psi_3 \vee \psi_4)\,/\,(\psi_1 \vee \psi_2)$. That is the reason why p/n must be slightly larger (resp. lower) than each one of the above bounds, hence (Ran) . Furthermore since no bound depend on s, (Ran) gives the required constant Γ. A similar reasoning shows that it exists a constant $\kappa > 0$ depending on s and the constants of the problem but independent from n, such that

$$\frac{\theta_{n,p}}{\psi_1 \wedge \psi_2} \leq \frac{\kappa}{n},$$

which enables to conclude the proof. □

Proof. (Corollary 6.2.1) From theorem 6.2.1 for any $\epsilon > 0$, we have

$$\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)} \left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right] \leq \Gamma \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\left\|s - \widehat{s}_m\right\|^2\right] + \frac{\kappa}{n}.$$

Then, we simply add the missing term $\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)^c} \left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right]$ to both sides of the inequality. It only remains to show that this term is of the right order.

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right] &\leq \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \left\|s - s_{\widehat{m}}\right\|^2\right] + \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \left\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\right\|^2\right],\\ &\leq \|s\|^2\, \mathbb{P}\left[\Omega_n^c(\epsilon)\right] + \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \sum_{\lambda\in\widehat{m}} \left[\nu_n(\varphi_\lambda)^2\right]\right],\end{aligned}$$

From Lemma 6.2.4, the first term in the right-hand side inequality is $o\left(1/n^2\right)$.
For the second one, Jensen's inequality gives

$$\mathbb{E}\left[\sum_{\lambda\in\widehat{m}} \nu_n^2\left(\varphi_\lambda\right) \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \leq \mathbb{E}\left[\sum_{\lambda\in\widehat{m}} \left(\varphi_\lambda(X) - P\varphi_\lambda\right)^2 \mathbb{1}_{\Omega_n^c(\epsilon)}\right].$$

Moreover, (Squ) with any $\eta > 0$ provides

$$\left(\varphi_\lambda(X) - P\varphi_\lambda\right)^2 \leq (1+\eta)\varphi_\lambda^2(X) + (1+\eta^{-1})P\varphi_\lambda^2.$$

Finally, $\sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 = \phi_m$ and $P\phi_{\widehat{m}} \leq \left\|\phi_{\widehat{m}}\right\|_\infty$ lead to

$$\mathbb{E}\left[\sum_{\lambda\in\widehat{m}} \nu_n^2\left(\varphi_\lambda\right) \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \leq \left(2+\eta+\eta^{-1}\right) \mathbb{E}\left[\left\|\phi_{\widehat{m}}\right\|_\infty \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \leq \left(2+\eta+\eta^{-1}\right) \frac{\Phi n}{(\log n)^2} \mathbb{P}\left[\Omega_n^c(\epsilon)\right]$$

thanks to (Reg) , and Lemma 6.2.4 enables to conclude. □

6.2.3 Adaptivity result

In this section, we apply the results of Section 6.2.2 to derive an adaptivity result for densities in some Hölder space.

Beforehand, we introduce some notations and specify the models we work with. In the sequel, s denotes a density on $[0,1]$. For a given partition of $[0,1]$ in D regular intervals $(I_\lambda)_{\lambda\in\Lambda(m)}$ of length $1/D$, let us define the model

$$S_m = \left\{ t \mid t = \sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda,\ (\alpha_\lambda)_\lambda \in \mathbb{R} \right\},$$

where $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$ and $|I_\lambda|$ denotes the length of I_λ. S_m is the vector space of dimension $D_m = D$ spanned by $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$. It is made of all piecewise constant functions defined on the partition $I = (I_1, \ldots, I_{D_m})$. The histogram estimator built from model S_m is defined by

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda\, \varphi_\lambda = \sum_{\lambda\in\Lambda(m)} \frac{n_\lambda}{n} \frac{\mathbb{1}_{I_\lambda}}{\sqrt{|I_\lambda|}},$$

where $n_\lambda = \mathrm{Card}\left(\{i \mid X_i \in I_\lambda\}\right)$. We consider the collection $S = \cup_{m\in\mathcal{M}_n} S_m$ such that $m \mapsto D_m$ is a one-to-one application from $\mathcal{M}_n$ towards $\mathcal{D} = \{D_m \mid m \in \mathcal{M}_n\}$ and

$$\max(\mathcal{D}) = N_n \leq \Phi\, n/\left(\log n\right)^2.$$

REMARKS:

- This assumptions is the same as (Reg) in the previous section. Since $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$,

$$\begin{aligned}\|\phi_m\|_\infty &= \sum_{t\in[0,1]}\left(\sum_{\lambda\in\Lambda(m)}\varphi_\lambda^2(t)\right),\\ &= \max_{\lambda\in\Lambda(m)}\frac{1}{|I_\lambda|} = D_m.\end{aligned}$$

  Then $\sup_{m\in\mathcal{M}_n}\|\phi_m\|_\infty \le \Phi\, n/\left(\log n\right)^2$ amounts to require that

$$\max_m D_m = N_n \le \Phi\, n/\left(\log n\right)^2.$$

  It means that on average, there are at least about $\left(\log n\right)^2/n$ points in each interval of any considered partition.

- Note that by construction, (Pol) is satisfied since our collection is made of only one model for each dimension.

In the sequel, we further assume that (Ad) and (Ran) hold so that Theorem 6.2.1 applies.

Our purpose is to show that the Lpo-based approach leads to an adaptivity property when s belongs to an unknown Hölder space $\mathcal{H}(L^*,\alpha^*)$ for $L^*>0$ and $\alpha^*\in(0,1]$. A function $f:\ [0,1]\to\mathbb{R}$ belongs to $\mathcal{H}(L^*,\alpha^*)$ if

$$\forall x,y\in[0,1],\quad |f(x)-f(y)|\le L^*\,|x-y|^{\alpha^*}.$$

We refer to [18] for an extensive study of a wide range of functional spaces.
An estimator is said to be *adaptive for* (L^*,α^*) if, without knowing (L^*,α^*), it "works as well as" any estimator which would exploit this knowledge. For us, the effectiveness measure is the L^2-risk and we say that an estimator enjoys such an adaptivity property if its risk is nearly the same (up to some constants) as the minimax risk with respect to $\mathcal{H}\left(L^*,\alpha^*\right)$:

$$\inf_{\widehat{s}}\sup_{s\in\mathcal{H}(L^*,\alpha^*)}\mathbb{E}\left[\|s-\widehat{s}\|^2\right]\le\sup_{s\in\mathcal{H}(L^*,\alpha^*)}\mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right]\le C\inf_{\widehat{s}}\sup_{s\in\mathcal{H}(L^*,\alpha^*)}\mathbb{E}\left[\|s-\widehat{s}\|^2\right],$$

where the infimum is taken over all possible estimators. Note that very often, $C\ge 1$ depends on the unknown parameters (L^*,α^*), but neither from s nor from n.
If this property holds for every parameters $(L,\alpha)=\theta$ in a set Θ, then $\widehat{s}_{\widehat{m}}$ is said to be *adaptive in the minimax sense with respect to the family* $\{\mathcal{H}(\theta)\}_{\theta\in\Theta}$. For a unified presentation about various notions of adaptivity, see Barron *et al.* [5].

As for the problem of density estimation on $[0,1]$ when s belongs to some Hölder space, it is known since the early 80s [21] that the minimax rate with respect to $\mathcal{H}\left(L,\alpha\right)$ for the quadratic risk is of order $L^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}$, with any $L>0$ and $\alpha>0$.
Remark: However when the problem is the estimation over $\mathbb{R}$, things turn out to be very different. For instance, the minimax rate now depends on the value of regularity parameter α with respect to the parameter p of the L^p-norm used for the assessment [22].

The following results settles that the Lpo-based procedures yields an adaptive in the minimax sense estimator of the density on $[0,1]$.

Theorem 6.2.2. *With the above notations, assume that* (Reg) , (Ad) *and* (Ran) *hold and that the collection of models is that one previously described in this section. Furthermore, assume that the target density* $s\in\mathcal{H}\left(L,\alpha\right)$ *for* $L>0$ *and* $\alpha\in(0,1]$*. Then,*

$$\sup_s\mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right]\le K_\alpha\,L^{\frac{2}{1+2\alpha}}\,n^{-\frac{2\alpha}{1+2\alpha}}+O\left(\frac{1}{n}\right),\tag{6.9}$$

for a given constant K_α *independent from* n *and* s.

Since the minimax risk is of order $L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}$, we deduce from this result that $\widehat{s}_{\widehat{m}}$ is adaptive in the minimax sense with respect to $\{\mathcal{H}(L,\alpha)\}_{L>0,\alpha\in(0,1]}$.

Remark: As we will see in the proof, this result remains true with any polynomial collection of models satisfying the requirements of Theorem 6.2.1, and including models with dimension of the order of $L^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}}$.

Proof. The idea is simply to use Theorem 6.2.1 and to derive the upper bound from

$$\mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] = \|s-s_m\|^2 + \mathbb{E}\left[\|s_m-\widehat{s}_m\|^2\right].$$

For the bias term, we have

$$\begin{aligned}
\|s-s_m\|^2 &= \sum_{\lambda\in\Lambda(m)} \frac{1}{|I_\lambda|^2}\int_{I_\lambda}\left(\int_{I_\lambda}[\,s(t)-s(x)\,]\,dx\right)^2 dt,\\
&\leq \sum_{\lambda\in\Lambda(m)} L^2 D_m^2 \int_{I_\lambda}\left(\int_{I_\lambda}|t-x|^\alpha\,dx\right)^2 dt \qquad (s\in\mathcal{H}(L,\alpha)),\\
&\leq C_\alpha\, L^2\, D_m^{-2\alpha} \quad \text{(after integration)},
\end{aligned}$$

where $C_\alpha = 4(\alpha+2)\left[(1+\alpha)^2(2\alpha+3)\right]^{-1}$.
On the other hand,

$$\begin{aligned}
\mathbb{E}\left[\|s_m-\widehat{s}_m\|^2\right] &= \frac{V_m-\|s_m\|^2}{n},\\
&\leq \frac{V_m}{n},\\
&\leq \frac{\|\phi_m\|_\infty}{n},\\
&= \frac{\sup_{x\in[0,1]}\sum_{\lambda\in\Lambda(m)}\varphi_\lambda^2(x)}{n} = \frac{D_m}{n}.
\end{aligned}$$

Hence under the same assumptions as Theorem 6.2.1, we get that it exists $C\geq 1$ and $\kappa>0$ such that

$$\forall m\in\mathcal{M}_n,\qquad \mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right] \leq C\left(C_\alpha\, L^2\, D_m^{-2\alpha}+\frac{D_m}{n}\right)+\frac{\kappa}{n}+o\left(\frac{1}{n^2}\right).$$

Now, let us define the sequence $\{D_{m_n}\}_n$ such that for each n,

$$\frac{1}{2}L^{\frac{1}{1+2\alpha}}\, n^{\frac{1}{1+2\alpha}} \leq D_{m_n} \leq 2L^{\frac{1}{1+2\alpha}}\, n^{\frac{1}{1+2\alpha}}.$$

Then, we derive that it exists $K'_\alpha>0$ such that

$$\inf_{m\in\mathcal{M}_n}\mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] \leq C_\alpha\, L^2\, D_m^{-2\alpha}+\frac{D_m}{n} \leq K'_\alpha\, L^{\frac{2}{1+2\alpha}}\, n^{-\frac{2\alpha}{1+2\alpha}},$$

hence the expected result. □

6.3 Exponential complexity

6.3.1 Limitations of the CV approach

From Section 6.2.1, we know that Lpo overpenalizes. According to Section 2.2.1 and as we will see in Section 6.5, it enables to somewhat balance the overfitting phenomenon that may result from a too high collection complexity, even in the polynomial complexity framework. Our present purpose is to

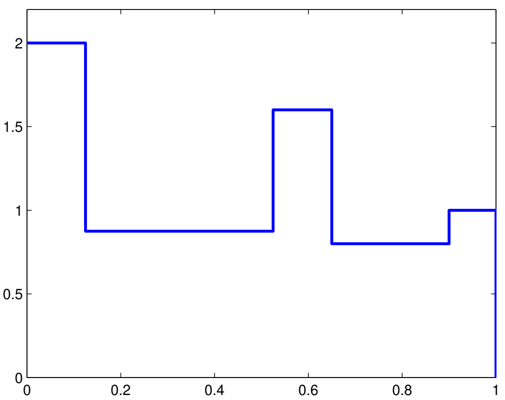

Figure 6.1: Plot of the density s.

illustrate that this amount of overpenalization is not sufficient to cope with an exponential complexity in a satisfactory way, at least for small values of p.
Indeed, this idea is already suggested through the comparison of the Lpo with other penalized criteria, at least in the linear regression framework. The Loo is shown to be asymptotically equivalent to Mallows' C_p by Li [24], while the general Lpo pursues the same (asymptotic) consistency purpose as BIC, provided $p/n \to 1$ [31]. Zhang [41] concludes that Lpo and FPE_α with $\alpha > 2$ are asymptotically equivalent if $\alpha = (2-\lambda)/(1-\lambda)$, where $p/n = \lambda + o(1)$. Finally, Shao [32] carries out a wider comparison between several penalized criteria, leading to the asymptotic equivalence of the general GIC_λ criterion and the Lpo when $\lambda = (2-p/n)/(1-p/n)$. It states that the Lpo-based procedure follows several behaviors, depending on the choice of p/n. For p/n close to 0, it is very similar to C_p-like criteria and should therefore be well suited in the polynomial framework, while the choice $p/n = (\log n - 2)/\log n - 1 < 1$ provides a penalty almost the same as BIC, which means that the Lpo tries to recover (the best approximation of) the true model.

However in the exponential complexity framework, all the C_p-like penalties turn out to be useless. This point is stressed by Breiman [14] or Birgé and Massart [10] to name but a few. In such frameworks, the latter authors have developed some penalties like for instance

$$\text{pen}(m) \quad = c_1 \frac{D_m}{n} + c_2 \frac{D_m}{n} \log\left(\frac{n}{D_m}\right)$$

in the change-points detection, where $c_1, c_2 > 0$ are universal constants. These penalties have demonstrated very good performance in a wide range of applications [23]. Moreover, very similar penalties have been independently advocated by others researchers (see [1] for some of them). Subsequently even if a $\log n$ term may be reached by the Lpo procedure for an adequate choice of p/n, it may only appear as a rough approximation of the above penalties since the dimension cannot be taken into account in the choice of p for instance. Furthermore, the intuition says that trying to estimate a density with several sharp thin picks by removing nearly all the data from the training set may prevent us from recovering an estimate with the right number of picks for instance. Nevertheless, more insight in the relationship between the choice of p and the complexity is required before drawing any definitive conclusion.

In what follows, we provide an illustration of this phenomenon where the problem is the estimation of the density s, displayed in Figure 6.1. In this example, we consider a subdivision of $[0,1]$: $0 = t_0 < t_1 < \ldots, t_K = 1$ where $K = 100$ and $t_i = i/K$. All the partitions of $[0,1]$ derived from this subdivision with $D_m \leq 50$ intervals are used. In this setting, a model is uniquely associated with such a partition, which defines our collection of models. The size of the sample is $n = 1000$.
The procedure consists in the following steps:

Algorithm 6.3.1.

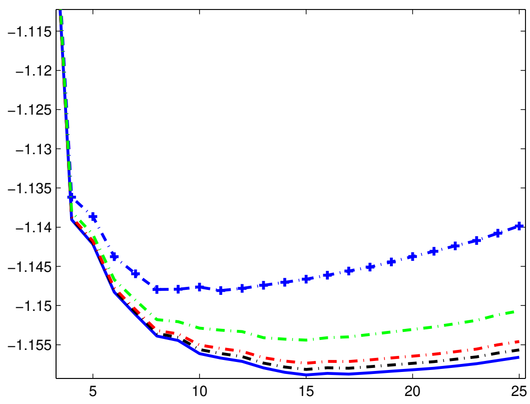

Figure 6.2: Plot of the Lpo risk estimators for $p = 1$ (bottom plain line), $p = 50$, $p = 100$, $p = 250$ (dashed lines) and $p = 500$ (top '-+-' line) in the exponential setting.

1. *The Lpo risk estimate is computed for each* $m \in \mathcal{M}_n$*.*
2. *Models are gathered according to their dimension.*
3. *For each dimension* $1 \leq D \leq 50$*, we choose the model with the lowest estimated risk, which provides us with a sequence of models* $\{\widehat{m}_D\}_{1 \leq D \leq 50}$*.*

Since this procedure is computationally intractable, we use the dynamic programming algorithm introduced by Bellman and Dreyfus [7], which yields the optimal model for each dimension with a quadratic instead of exponential algorithmic complexity. At the end of the process for each p, we get the sequence $\left\{\widehat{R}_p(D)\right\}_{1 \leq D \leq 50}$ of the Lpo estimates for each $\widehat{m}_D$.

Let us focus on Figure 6.2, which depicts the curve $D \mapsto \widehat{R}_p(D)$ for each value of $p = 1, 50, 100, 250,$ and 500. From a general point of view, we observe that after the bias has disappeared ($D \leq 5$), all the curves still decrease for a while before starting their growth more or less late ($D \approx 15$). Note that the dimension of the oracle model is 5, which is also that of the true model according to Figure 6.1. Therefore, the conclusion is that the Loo as well as Lpo for $p \leq n/2$ suffers some possibly high overfitting and would lead to poor estimators.
We also observe that for a given dimension (larger than that of the oracle), the Lpo estimate is an increasing function of p, which confirms the ability of Lpo to somewhat balance this deficiency of the Loo.

Let us consider Figure 6.3. The same curves are displayed, but moreover with several larger values of $p = 750, 900$ and 950 (top lines). We observe that larger values of p lead to "corrected" curves. The Lpo algorithm with $p = 750$ still suffers some overfitting, while $p = 950$ does not. Whereas $p = 900$ seems to completely balance overfitting in the left panel, the right one contradicts this conclusion. Besides, when the correction is large enough to overcome overfitting ($p \gtrsim 900$), the dimension of the selected model is 4. Indeed according to Figure 6.1, all the successive values of the true density but one are well separated. Thus, removing too much points from the training set prevents us from detecting this harder to detect difference, and results in a partition with only 4 intervals.

Open questions:

- Subsequently, p very close to n acts as a $\log n$ term and balances overfitting even if any recipe about this choice cannot yet be given here. On the other hand, such a high p may lead to selection of a model with a too small dimension with respect to the oracle. That is the reason why the use of Lpo in the exponential complexity framework deserves some further work.

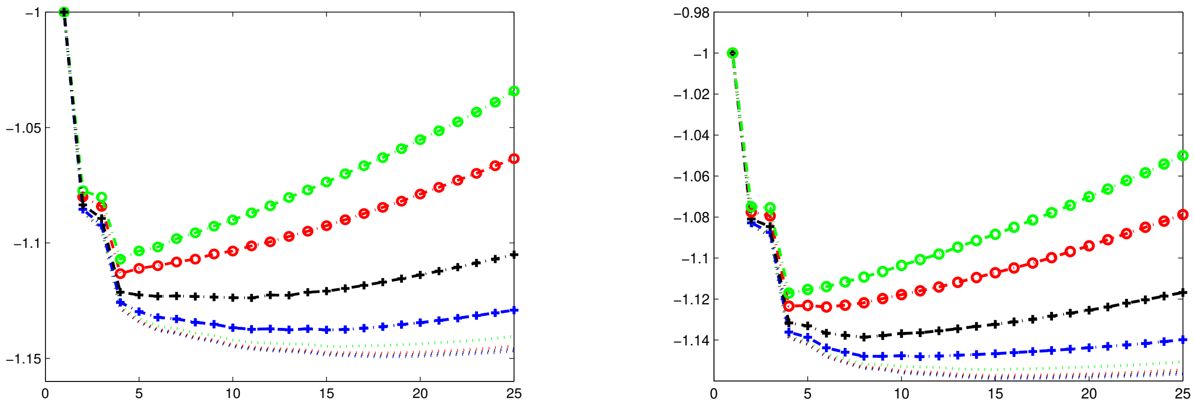

Figure 6.3: Plot of the Lpo risk estimators for $p = 1$ (bottom plain line), $p = 50$, $p = 100$, $p = 250$ (bottom middle dashed lines), $p = 500$, $p = 750$ (top middle '-+-' lines), and $p = 900$, $p = 950$ (top '-o-' lines) in the exponential setting. Left and right panels depict results with two different samples.

- However, it still appears as a rough criterion in this setting, in comparison to more refined log-based penalties. Indeed whereas these penalties independently penalize a given model for its own complexity ($c_1 D_m/n$) and also for the richness of the whole collection ($c_2 D_m/n \log(n/D_m)$), Proposition 6.2.1 demonstrates that there is not such distinction in the Lpo penalty.

6.3.2 Oracle inequality

Section 6.3.1 has enlighten an issue encountered by the Lpo in the exponential complexity framework. Indeed, the Lpo criterion turns out to be not well suited to take into account a too large collection complexity. Following the last remark of the previous section, we propose to add a new term to the Lpo risk estimator, designed to cope with exponential collection complexity.
In order to design this "complexity term", we restart the proof of Theorem 6.2.1 from the beginning, but in the exponential setting. This strategy is applied to the specific case of histograms essentially for the sake of simplicity, but it could be generalized to some kinds of localized bases such as piecewise polynomials for instance (with appropriate assumptions).

Main result In the sequel, we define a regular partition of $[0, 1]$ in N_n intervals of length $1/N_n$. Let us consider the set of all the partitions of $[0, 1]$ with $1 \leq D \leq N_n$ intervals $(I_1, \ldots, I_D)$ we may build from the thinnest one in N_n intervals. To each partition corresponds a model S_m of piecewise constant functions. S_m is the finite vector space of dimension D_m spanned by $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ where $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$. For each $m \in \mathcal{M}_n$, $\{\varphi_\lambda\}_\lambda$ is then an orthonormal basis of S_m. We recall that

$$\theta_{n,p} = \frac{2n-p}{(n-1)(n-p)},$$
$$\phi_m = \sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 \qquad \text{and} \qquad V_m = \mathbb{E}\phi_m(X),$$

where $X \sim s$.
The description of our procedure starts with the definition of $\widehat{m}$ for each $1 \leq p \leq n-1$:

$$\widehat{m} = \operatorname{Argmin}_{m\in\mathcal{M}_n} \left\{ \widehat{R}_p(m) + \operatorname{pen}(m) \right\},$$

where $\operatorname{pen}(\cdot)\ :\ \mathcal{M}_n \to \mathbb{R}_+$ denotes the penalty term which is to be designed. The following result warranties the performance of this procedure through an oracle inequality with a properly chosen penalty.

Theorem 6.3.1. *With the same notations as above, let s denote a bounded density on $[0,1]$ and $X_1,\ldots,X_n$ n i.i.d. random variables drawn from s. Let us assume that there exists $\Phi>0$ independent from n satisfying*

$$(Reg) \qquad \sup_{m\in\mathcal{M}_n} \|\phi_m\|_\infty \le \Phi \frac{n}{(\log n)^2},$$

and that there exists $\Sigma>0$ and a set $\{L_m\}_{m\in\mathcal{M}_n}$ of nonnegative weights such that

$$\sum_{m\in\mathcal{M}_n} e^{-L_m} \le \Sigma < +\infty \tag{6.10}$$

where Σ does not depend on n.
Let us assume $n\ge 29$ and set $0<\epsilon<1$ such that

$$\frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} < 1,$$

where $\zeta(\epsilon) = \left[1-(1+\epsilon)^{-8}\right]$. Furthermore, let us choose $1\le p\le n-1$ satisfying

$$(Ran) \qquad \frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n}\frac{1+\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \alpha \le \frac{p}{n} \le 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} - \beta$$

with $0<\alpha,\beta<1$.
As a penalty, set

$$\forall m\in\mathcal{M}_n, \qquad \mathrm{pen}(m) = \theta_{n,p}\, L_m \frac{\Phi}{3}\left(\frac{4\,(1+\epsilon)^4-1}{(1+\epsilon)^4-1}\right)\left[1+\frac{6}{\Phi}\frac{\|s\|_\infty}{(1+\epsilon)^4\left[4\,(1+\epsilon)^4-1\right]}\right].$$

Then, there exists a set $\Omega_n(\epsilon)$ of probability larger than $1-o(1/n^2)$ such that

$$\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)}\|s-\widehat{s}_{\widehat{m}}\|^2\right] \le \Gamma\left(\mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] + \frac{L_m}{n}\left[1+\frac{\|s\|_\infty}{\Phi\,(1+\epsilon)^4}\right]\right) + \frac{\kappa}{n}, \tag{6.11}$$

where $\Gamma=\Gamma(\epsilon,\alpha,\beta,\Phi)\ge 1$ is a constant independent from s and $\kappa=\kappa(s,\epsilon,\alpha,\beta,\Phi,\Sigma)\ge 0$ is a constant.

Remarks:

- The penalty depends on Φ and $\|s\|_\infty$. Whereas Φ is known from a given basis of functions, $\|s\|_\infty$ is *a priori* unknown and has to be estimated beforehand for the procedure to be applicable. Note that the same issue arises in [26].
- Assumption (Ad) from Theorem 6.3.1 is related to (6.10), since provided the collection is polynomial and the weights $L_m = n\mathbb{E}\chi^2(m)$, (6.10) is implied by (Ad) .
- Note that by construction, the cardinality of the models with the same dimension D is $\binom{N-1}{D-1}$, which corresponds to the exponential complexity framework.

For the same reasons as Corollary 6.2.1, we deduce from Theorem 6.3.1 the following result in which the expectation of the left-hand side is computed over the whole space.

Corollary 6.3.1. *With the notation of the previous theorem and under the same assumptions, we have for any $\epsilon>0$,*

$$\mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right] \le \Gamma \inf_{m\in\mathcal{M}_n}\left(\mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] + \frac{\|s\|_\infty}{n}L_m\right) + \frac{\kappa}{n} + o\left(\frac{1}{n^2}\right),$$

where $\Gamma=\Gamma(\epsilon,\alpha,\beta,\Phi)\ge 1$ is a constant independent from s and $\kappa=\kappa(s,\epsilon,\alpha,\beta,\Phi,\Sigma)\ge 0$ is a constant.

Proof. The proof is exactly the same as that of Corollary 6.2.1. ∎

Comments on the proof of Theorem 6.3.1 The proof of Theorem 6.3.1 is very similar in the spirit to that of Theorem 6.2.1. For the sake of completeness, it is nevertheless provided in Section 6.4.2.
In the sequel, we detail the main differences in the strategy.

- Unlike Theorem 6.2.1, the goal of this proof is to design a penalty term, which is able to take into account the complexity of rich collections of models.
- The proof relies on the same concentration inequalities as that of Theorem 6.2.1 (Proposition 6.4.1 and Proposition 6.4.2). In both these propositions without any further specification of the weights L_ms, we choose $y_m = z + L_m$ and respectively $x_m = z + L_m$.
- Due to the exponential complexity setting and the specific choice of a localized basis, we define a set $\Omega_n(\epsilon)$ such that
$$\mathbb{P}\left[\,\Omega_n^c(\epsilon)\,\right] \leq 2n\, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}$$
instead of
$$\mathbb{P}\left[\,\Omega_n^c(\epsilon)\,\right] \leq 2n^{2+\delta}\, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}.$$

Choice of weights L_m We now study which are the minimal weights we may choose so that (6.10) is fulfilled. If we make the classical assumption that $L_m = L(D_m)$ for any $m \in \mathcal{M}_n$ [8, 10, 26], we get

$$\begin{aligned}
\sum_{m\in\mathcal{M}_n} e^{-L_m} &= \sum_{m\in\mathcal{M}_n} e^{-L(D_m)},\\
&= \sum_{D=1}^{N_n} \sum_{m/\ D_m=D} e^{-L(D)},\\
&= \sum_{D=1}^{N_n} e^{-L(D)+\log|\mathcal{M}(D)|},
\end{aligned}$$

where $\mathcal{M}(D) = \{m \in \mathcal{M}_n \mid D_m = D\}$.
In our setup,

$$|\mathcal{M}(D)| = \binom{N-1}{D-1},$$

which implies that (6.10) becomes

$$\sum_{m\in\mathcal{M}_n} e^{-L_m} = \sum_{D=1}^{N_n} e^{-L(D)+\log\binom{N-1}{D-1}} \leq \Sigma < +\infty.$$

Let us notice that

$$\sum_{D=1}^{N_n} \frac{1}{\binom{N_n-1}{D-1}^{\alpha}}$$

is a convergent numerical series for $\alpha > 0$. Indeed for any $\alpha > 0$, set

$$\forall n \in \mathbb{N}^*, \qquad S_n(\alpha) = \sum_{D=1}^{n} \binom{n-1}{D-1}^{-\alpha}.$$

Then, we see that $S_{n+1}(\alpha) \leq 1 + S_n(\alpha)$ and since $S_n(\alpha) > 0$, we conclude that for any $\alpha > 0$, $(S_n(\alpha))_{n\in\mathbb{N}^*}$ is a convergent sequence, hence the convergence of any subsequence of it. Thus, $(S_{N_n}(\alpha))_n$ converges for any increasing $(N_n)_n$, in particular such that $N_n \leq \Phi n/(\log n)^2$.
Therefore, the "minimal" choice of weight is

$$\forall m \in \mathcal{M}_n, \qquad L_m = C \log\binom{N_n-1}{D_m-1},$$

with any constant $C > 1$.
REMARK: Note that "minimal" refers to the fact that smaller weights do not provide the result, while large wrights give a rougher inequality (6.11).

OPEN QUESTIONS:

- We have pointed out earlier that the penalty definition has the drawback of depending on $\|s\|_\infty$, which has to be estimated beforehand. We do not know whether it is possible to remove this type of dependency, probably with another scheme of proof.

- Moreover, note that the choice of weight mentioned above induces another (universal) constant, that has to be determined. A similar problem occurs with generalizations of Mallows' C_p penalties [8, 15, 16, 26]. When the empirical contrast is used, instead of the Lpo risk estimator for us, an data-driven approach has been developed by Birgé and Massart [10] who establish its theoretical properties in a gaussian setting. This strategy is called the slope heuristics and has been successfully applied by Lebarbier [23] for instance. In a general heteroscedastic framework Arlot and Massart [3] have proved this strategy to be effective with not too rich collections. We may hope that a similar approach could be used in our setting.

6.4 Proofs

6.4.1 Polynomial complexity

Proof of Useful Lemma

Proof. Set $Z = X - Y - K_2$. We have

$$\mathbb{P}\left(Z \geq K_1 z\right) \leq \Sigma e^{-z}.$$

Then,

$$\begin{aligned}
\mathbb{E}Z &\leq \mathbb{E}\left[\int_0^{+\infty} \mathbb{1}_{(t\leq Z)}\, dt\right], \\
&= \int_0^{+\infty} \mathbb{E}\left[\, \mathbb{1}_{(t\leq Z)}\,\right]\, dt, \\
&= \int_0^{+\infty} \mathbb{P}\left[\, t \leq Z\,\right]\, dt, \\
&\leq K_1 \int_0^{+\infty} \Sigma e^{-z}\, dz = K_1\Sigma.
\end{aligned}$$

□

Proof of Proposition 6.2.2

Proof. Bernstein's inequality [26] states

$$\forall x > 0, \qquad \mathbb{P}\left[\left|\nu_n\left(\phi_m\right)\right| \geq \frac{1}{n}\sqrt{2v\, x} + \frac{b}{3n}x\right] \leq e^{-x},$$

with $b \geq |\phi_m(X_i) - \mathbb{E}\phi_m(X_i)|$ and $v = \sum_{i=1}^n \mathrm{Var}\left[\, \phi_m(X_i)\,\right]$.
Since X_i are *i.i.d.* and $\phi_m \geq 0$, we have

$$b = \|\phi_m\|_\infty \qquad \text{and} \qquad v \leq n\, V_m \left\|\phi_m\right\|_\infty,$$

hence the first part of the proposition.
For the second part of the result, a union bound provides

$$\begin{aligned}
&\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid |\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m\right] \\
&\leq \sum_{m \in \mathcal{M}_n} \mathbb{P}\left[|\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m\right], \\
&\leq \sum_{m \in \mathcal{M}_n} e^{-y_m}, \\
&\leq e^{-z} \sum_{m \in \mathcal{M}_n} e^{-C\,nE_m} \qquad (y_m = z + C\,nE_m), \\
&\leq e^{-z} \sum_{m \in \mathcal{M}_n} e^{-C\xi D_m}, \qquad (Ad) \\
&\leq e^{-z} \sum_{D \geq 1} e^{-C\xi D + \delta \log(D)}, \qquad (Pol), \\
&\leq \Sigma_1 e^{-z}.
\end{aligned}$$

□

Proof of Lemma 6.2.4

Proof. We recall that

$$\Omega_n(\epsilon) = \left\{\forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}}\right\}.$$

Then, we deduce that

$$\begin{aligned}
\mathbb{P}\left[\Omega_n^c(\epsilon)\right] &= \mathbb{P}\left[\left\{\exists m \in \mathcal{M}_n,\ \exists \lambda \in \Lambda(m) \mid |\nu_n(\varphi_\lambda)| \geq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}}\right\}\right], \\
&\leq \sum_{m \in \mathcal{M}_n} \sum_{\lambda \in \Lambda(m)} \mathbb{P}\left[\left\{|\nu_n(\varphi_\lambda)| \geq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}}\right\}\right], \\
&\leq \sum_{m \in \mathcal{M}_n} D_m\, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}, \quad \text{(Bernstein and } (\mathit{Reg2})) \\
&\leq \sum_{D \geq 1} D^{\delta+1} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}, \quad (Pol) \\
&\leq n^{\delta+2} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}, \quad (D \leq n)
\end{aligned}$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

□

Proof of Proposition 6.2.3

Proof. First, we notice that $\chi(m) = \sqrt{\chi^2(m)}$ may be also expressed as

$$\chi(m) = \sup_{a / \sum_{\lambda \in \Lambda(m)} \alpha_\lambda^2 = 1} \left|\nu_n\left(\sum_{\lambda \in \Lambda(m)} a_\lambda \varphi_\lambda\right)\right| \geq \sup_{a \in A} \left|\nu_n\left(\sum_{\lambda \in \Lambda(m)} a_\lambda \varphi_\lambda\right)\right|,$$

where A is dense subset of

$$\left\{a = (a_1, \ldots, a_{D_m}) \in \mathbb{R}^{D_m} \mid \sum_{\lambda \in \Lambda(m)} \alpha_\lambda^2 = 1 \text{ and } \sum_{\lambda \in \Lambda(m)} |\alpha_\lambda| \leq \frac{t}{z}\right\}.$$

Moreover, if we define the event

$$\Omega = \left\{ \sup_{\lambda \in \Lambda(m)} \nu_n \left(\varphi_\lambda \right) \leq t \right\}$$

for $t > 0$, then we deduce that

$$\chi(m) \leq \sup_{a \in A} \left| \nu_n \left(\sum_{\lambda \in \Lambda(m)} a_\lambda \varphi_\lambda \right) \right| \tag{6.12}$$

on $\Omega \cap \{\chi(m) \geq z\}$.
Then, Talagrand's inequality to $\sup_{a \in A} \left| \nu_n \left(\sum_{\lambda \in \Lambda(m)} a_\lambda \varphi_\lambda \right) \right|$ gives for $\epsilon > 0$

$$\forall x > 0, \qquad \mathbb{P} \left[\mathbb{1}_\Omega \sup_{a \in A} \left| \nu_n \left(\sum_{\lambda \in \Lambda(m)} a_\lambda \varphi_\lambda \right) \right| \geq (1+\epsilon) \left(\sqrt{\chi^2(m)} + \sqrt{\frac{2 \left\| s \right\|_\infty}{n} x} \right) \right] \leq e^{-x},$$

with $z = \sqrt{2 \left\| s \right\|_\infty / n}$ and $t = 2\epsilon \left\| s \right\|_\infty \left[\kappa(\epsilon) \Phi n / (\log n)^2 \right]^{-1}$.
Finally, the first result comes from both (6.12) and $\Omega_n(\epsilon) = \Omega$.

As for the second inequality,

$$\begin{aligned}
&\mathbb{P} \left[\exists m \in \mathcal{M}_n \mid \sqrt{n} \chi(m) \mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon) \left(\sqrt{n E_m} + \sqrt{2 \left\| s \right\|_\infty x_m} \right) \right] \\
&\leq \sum_{m \in \mathcal{M}_n} e^{-x_m}, \\
&\leq e^{-z} \sum_{m \in \mathcal{M}_n} e^{-C' \, n E_m}, \quad (x_m = C' \xi \, D_m + z) \\
&\leq e^{-z} \sum_{D \geq 1} e^{-C' \, \xi D + \delta \log D}, \quad (Ad) \text{ and } (Pol) \\
&\leq \Sigma_2 \, e^{-z}.
\end{aligned}$$

□

Proof Lemma 6.4.1

Lemma 6.4.1. *For $n \geq 29$, there exists $0 < \epsilon < 1$ such that*

$$\zeta(\epsilon) > \frac{2}{n-1} \qquad \text{and} \qquad \frac{4\zeta(\epsilon)}{1 + 3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1) - 2},$$

where $\zeta(\epsilon) = \left[1 - (1+\epsilon)^{-8} \right]$.

Proof. The first part is obvious since for a given n, we can choose $0 < \epsilon < 1$ such that $\zeta(\epsilon) > 2/(n-1)$. Then with $\delta = \zeta(\epsilon) - 2/(n-1)$, we have

$$\delta(n-1) = \zeta(\epsilon)(n-1) - 2.$$

After some calculations, it is easy to see that

$$\begin{aligned}
&\frac{4\zeta(\epsilon)}{1 + 3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1) - 2}, \\
\Leftrightarrow \quad &\delta^2 \, \frac{n+6}{n} - \delta \, \frac{n-10}{n} + \frac{2n+10}{(n-1)^2} < 0,
\end{aligned}$$

which is a polynomial of degree 2 in δ.
For $n \geq 29$, the discriminant is positive and any δ between the two distinct zeros yields a value for $\zeta(\epsilon)$ such that

$$\frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2},$$

which enables to conclude. $\square$

6.4.2 Exponential complexity

Proof of Theorem 6.3.1

Preliminaries

Proposition 6.4.1. *With the same notations as Theorem 6.3.1, let $z > 0$ be any positive constant and for each m, let us define $y_m = z + L_m$. Then, we have*

$$\forall m \in \mathcal{M}_n, \quad \mathbb{P}\left[|\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m \right] \leq 2e^{-y_m}.$$

Moreover if (Ad) holds, we have

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid |\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m \right] \leq 2\Sigma e^{-z},$$

where $\Sigma > 0$ is the positive constant independent from n given in Theorem 6.3.1.

We now provide a similar result to Lemma 6.4.2 to introduce the event $\Omega_n(\epsilon)$ for any $\epsilon > 0$.

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n, \ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon \log n \, \|s\|_\infty}{\kappa(\epsilon)\sqrt{\Phi n}} \right\},$$

where $\kappa(t) = 2(t^{-1} + 1/3)$.
Bernstein's inequality then provides

Lemma 6.4.2. *Set $\epsilon > 0$ and assume that (Reg) hold. Then,*

$$\forall \alpha > 0, \qquad \mathbb{P}\left[\Omega_n^c(\epsilon) \right] \leq 2n \, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi} (\log n)^2} = o\left(\frac{1}{n^\alpha} \right),$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

This lemma is useful in the following concentration result.

Proposition 6.4.2. *With the above notations, set $\epsilon > 0$ and for any C', $z > 0$, $x_m = z + L_m$. Assume that (Reg) is fulfilled. Then,*

$$\forall m \in \mathcal{M}_n, \qquad \mathbb{P}\left[\sqrt{n}\chi(m) \mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty x_m} \right) \right] \leq e^{-x_m}.$$

Furthermore if (6.10) holds,

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \ \sqrt{n}\chi(m) \mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty x_m} \right) \right] \leq \Sigma \, e^{-z},$$

where $\Sigma > 0$ is the same positive constant as in Theorem 6.3.1.

Proof We are now in position to give the proof of the main result.

Proof. (Theorem 6.3.1) From Lemma 6.2.3, we still derive

$$\begin{aligned}\|s - s_{\widehat{m}}\|^2 + n\theta_{n,p}E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \quad &\leq \quad \|s - s_m\|^2 + n\theta_{n,p}E_m - (1+\theta_{n,p})\chi^2(m) + \\ &\theta_{n,p}\nu_n\left(\phi_m - \phi_{\widehat{m}}\right) + 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right) + \Delta(m, \widehat{m}),\end{aligned}$$

where $\Delta(m, m') = \mathrm{pen}(m) - \mathrm{pen}(m')$, which is nearly the same as (6.3).
We first apply Proposition 6.4.1. To do so, let us notice that we have

$$\begin{aligned}\sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} \quad &\leq \quad \eta\Phi\, nE_m + \eta^{-1}L_m + \eta^{-1}z + \eta\Phi\ \|s\|_\infty^2\,, \\ \frac{\|\phi_m\|_\infty}{3n} y_m \quad &\leq \quad \frac{\Phi}{3}\left(z + L_m\right),\end{aligned}$$

which entails that with probability at least $1 - 2\Sigma e^{-z}$, for any $\eta > 0$,

$$\begin{aligned}\forall m \in \mathcal{M}_n, \qquad \left|\nu_n\left(\phi_m - \phi_{\widehat{m}}\right)\right| \quad &\leq \quad \eta\Phi\, nE_{\widehat{m}} + \eta\Phi\, nE_m + \left(\eta^{-1} + \frac{\Phi}{3}\right)\left(L_m + L_{\widehat{m}}\right) + \\ &2z\left(\eta^{-1} + \frac{\Phi}{3}\right) + 2\eta\Phi\ \|s\|_\infty^2\,.\end{aligned}$$

Set $\epsilon'' > 0$ and let us choose $\eta = \epsilon''\Phi^{-1}$. It provides

$$\begin{aligned}\|s - s_{\widehat{m}}\|^2 + n\theta_{n,p}\left(1-\epsilon''\right)E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \quad &\leq \quad \|s - s_m\|^2 + n\theta_{n,p}\left(1+\epsilon''\right)E_m - (1+\theta_{n,p})\chi^2(m) + \\ &\theta_{n,p}\,\Phi\left(\frac{1}{\epsilon''} + \frac{1}{3}\right)\left(L_m + L_{\widehat{m}}\right) + 2z\theta_{n,p}\,\Phi\left(\frac{1}{\epsilon''} + \frac{1}{3}\right) + \\ &2\theta_{n,p}\,\epsilon''\,\|s\|_\infty^2 + 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right) + \Delta(m, \widehat{m}).\end{aligned}$$

Then, Proposition 6.4.2 gives that with probability at least $1 - \Sigma\, e^{-z}$, we have

$$n\theta_{n,p}\left(1-\epsilon''\right)E_{\widehat{m}} \quad \geq \quad n\theta_{n,p}\chi^2(\widehat{m})\frac{1-\epsilon''}{(1+\epsilon)^4}\mathbb{1}_{\Omega_n(\epsilon)} - 2\|s\|_\infty\theta_{n,p}\frac{1-\epsilon''}{(1+\epsilon)^2 - 1}\left(L_{\widehat{m}} + z\right).$$

Now, if we set $\epsilon'' > 0$ such that $1 - \epsilon'' = (1+\epsilon)^{-4}$ and define $0 < \epsilon' < 1$ such that $\sqrt{1-\epsilon'} = (1+\epsilon)^{-4}$, we obtain that on the event $\Omega_n(\epsilon)$,

$$\begin{aligned}\|s - s_{\widehat{m}}\|^2 + n\theta_{n,p}\left(1-\epsilon'\right)\chi^2(\widehat{m}) - (1+\theta_{n,p})\chi^2(\widehat{m}) \quad &\leq \quad \|s - s_m\|^2 + n\theta_{n,p}\left(1+\epsilon''\right)E_m - (1+\theta_{n,p})\chi^2(m) + \\ &\theta_{n,p}\,\Phi\left(\frac{1}{\epsilon''} + \frac{1}{3}\right)\left(L_m + L_{\widehat{m}}\right) + 2z\theta_{n,p}\,\Phi\left(\frac{1}{\epsilon''} + \frac{1}{3}\right) + \\ &2\|s\|_\infty\theta_{n,p}\frac{1-\epsilon'}{1-\sqrt{1-\epsilon'}}\left(L_{\widehat{m}} + z\right) + \\ &2\theta_{n,p}\,\epsilon''\,\|s\|_\infty^2 + 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right) + \Delta(m, \widehat{m}).\end{aligned}$$

In the same way as in the proof of Theorem 6.3.1 and thanks to Lemma 6.2.5, we have that

$$2\nu_n\left(s_{\widehat{m}} - s_m\right) \quad \leq \quad \left(1+2\epsilon'\right)\chi^2(\widehat{m}) + \left(1+\epsilon'\right)\frac{1+2\epsilon'}{\epsilon'}\chi^2(m) + \frac{2}{2+\epsilon'}\|s_{\widehat{m}} - s\|^2 + \frac{2}{\epsilon'}\|s_m - s\|^2\,.$$

From this, we derive that on $\Omega_n(\epsilon)$, with probability at least $1-3\Sigma\, e^{-z}$,

$$\begin{aligned}\frac{\epsilon'-2\theta_{n,p}}{2+\epsilon'}\left\|s-s_{\widehat{m}}\right\|^2+\left[n\theta_{n,p}\left(1-\epsilon'\right)-2\left(1+\theta_{n,p}\right)\left(1+\epsilon'\right)\right]\chi^2(\widehat{m}) \quad\leq\quad & \frac{\epsilon'+2\left(1+\theta_{n,p}\right)}{\epsilon'}\left\|s-s_m\right\|^2+\\ & n\theta_{n,p}\left(1+\epsilon''\right)E_m+\\ & \left(1+\theta_{n,p}\right)\left[2\left(1+\epsilon'\right)+\frac{1}{\epsilon'}\right]\chi^2(m)+\\ & \theta_{n,p}\,L_{\widehat{m}}\left[\Phi\left(\frac{1}{\epsilon''}+\frac{1}{3}\right)+2\left\|s\right\|_\infty\frac{1-\epsilon'}{1-\sqrt{1-\epsilon'}}\right]+\\ & \theta_{n,p}\,L_m\left[\Phi\left(\frac{1}{\epsilon''}+\frac{1}{3}\right)\right]+\\ & \theta_{n,p}\,z\left[\Phi\left(\frac{1}{\epsilon''}+\frac{1}{3}\right)+2\left\|s\right\|_\infty\frac{1-\epsilon'}{1-\sqrt{1-\epsilon'}}\right]+\\ & 2\theta_{n,p}\epsilon''\left\|s\right\|_\infty^2+\Delta(m,\widehat{m}).\end{aligned}$$

If we choose

$$\mathrm{pen}(m)=\theta_{n,p}\left[\Phi\left(\frac{1}{\epsilon''}+\frac{1}{3}\right)+2\left\|s\right\|_\infty\frac{1-\epsilon'}{1-\sqrt{1-\epsilon'}}\right]L_m,$$

we observe that the term in $L_{\widehat{m}}$ disappears and we obtain on $\Omega_n(\epsilon)$ with high probability

$$\begin{aligned}\frac{\epsilon'-2\theta_{n,p}}{2+\epsilon'}\left\|s-s_{\widehat{m}}\right\|^2+\left[n\theta_{n,p}\left(1-\epsilon'\right)-2\left(1+\theta_{n,p}\right)\left(1+\epsilon'\right)\right]\chi^2(\widehat{m}) \quad\leq\quad & \frac{\epsilon'+2\left(1+\theta_{n,p}\right)}{\epsilon'}\left\|s-s_m\right\|^2+\\ & n\theta_{n,p}\left(1+\epsilon''\right)E_m+\\ & \left(1+\theta_{n,p}\right)\left[2\left(1+\epsilon'\right)+\frac{1}{\epsilon'}\right]\chi^2(m)+\\ & 2\theta_{n,p}\,L_m\left[\Phi\left(\frac{1}{\epsilon''}+\frac{1}{3}\right)+\left\|s\right\|_\infty\frac{1-\epsilon'}{1-\sqrt{1-\epsilon'}}\right]+\\ & \theta_{n,p}\,z\left[\Phi\left(\frac{1}{\epsilon''}+\frac{1}{3}\right)+2\left\|s\right\|_\infty\frac{1-\epsilon'}{1-\sqrt{1-\epsilon'}}\right]+\\ & 2\theta_{n,p}\epsilon''\left\|s\right\|_\infty^2.\end{aligned}$$

We now apply Lemma 6.2.2 so that we get

$$\begin{aligned}(\psi_1\wedge\psi_2)\,\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)}\left\|s-\widehat{s}_{\widehat{m}}\right\|^2\right]\leq(\psi_3\vee\psi_4)\,\mathbb{E}\left[\left\|s-\widehat{s}_m\right\|^2\right]+\\ 2\theta_{n,p}\,L_m\frac{\Phi}{3}\left(\frac{4\left(1+\epsilon\right)^4-1}{\left(1+\epsilon\right)^4-1}\right)\left[1+\frac{3}{\Phi}\frac{\left\|s\right\|_\infty}{\left(1+\epsilon\right)^4\left[4\left(1+\epsilon\right)^4-1\right]}\right]+\\ 3\theta_{n,p}\,\Sigma\frac{\Phi}{3}\left(\frac{4\left(1+\epsilon\right)^4-1}{\left(1+\epsilon\right)^4-1}\right)\left[1+\frac{6}{\Phi}\frac{\left\|s\right\|_\infty}{\left(1+\epsilon\right)^4\left[4\left(1+\epsilon\right)^4-1\right]}\right]+\\ 2\theta_{n,p}\epsilon''\left\|s\right\|_\infty^2,\end{aligned}$$

with ψ_1,ψ_2,ψ_3 and ψ_4 the same as in the proof of Theorem 6.3.1. Subsequently provided (Ran) holds, there exists $\Gamma\geq 1$ independent from n and $\kappa>0$ depending on s such that

$$\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)}\left\|s-\widehat{s}_{\widehat{m}}\right\|^2\right]\leq\Gamma\left(\mathbb{E}\left[\left\|s-\widehat{s}_m\right\|^2\right]+\frac{L_m}{n}\left[1+\frac{\left\|s\right\|_\infty}{\Phi\left(1+\epsilon\right)^4}\right]\right)+\frac{\kappa}{n}\cdot$$

□

Proof of Proposition 6.4.1

Proof. The proof is the same as that of Proposition 6.2.2, except that we do not require (Ad) and (Pol) to hold, but only set $y_m = z + L_m$ with the assumption that $\sum_{m \in \mathcal{M}_n} e^{-L_m} \leq \Sigma$ where Σ is independent from n.

□

Proof of Lemma 6.4.2

Proof. We recall that

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\}.$$

Thanks to the localization of the basis, we have that

$$\left\{ \exists m \in \mathcal{M}_n,\ \exists \lambda \in \Lambda(m) \mid |\nu_n(\varphi_\lambda)| \geq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\} = \left\{ \exists \lambda \in \Lambda(m_N) \mid |\nu_n(\varphi_\lambda)| \geq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\},$$

where m_N is the index corresponding to the thinnest partition.

Thus, it comes straightforwardly that

$$\begin{aligned}
\mathbb{P}\left[\Omega_n^c(\epsilon)\right] &\leq \sum_{\lambda \in \Lambda(m_N)} \mathbb{P}\left[\left\{ |\nu_n(\varphi_\lambda)| \geq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\} \right], \\
&\leq \sum_{\lambda \in \Lambda(m_N)} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi} (\log n)^2}, \quad \text{(Bernstein)}, \\
&\leq N_n e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi} (\log n)^2}, \\
&\leq \Phi \frac{n}{(\log n)^2} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi} (\log n)^2}, \quad (Reg)
\end{aligned}$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$, hence the result.

□

Proof of Proposition 6.4.2

Proof. The proof is the same as that of Proposition 6.2.3, except that $x_m = z + L_m$ and the L_ms are required to satisfy $\sum_{m \in \mathcal{M}_n} e^{-L_m} \leq \Sigma$, with Σ independent from n.

□

6.5 p as a regularization parameter

6.5.1 Overfitting in the polynomial framework

Two main situations are often distinguished in model selection, depending on the complexity of the collection of models [8, 15, 5, 26]. Indeed, Birgé and Massart [10] prove that polynomial complexity problems can be efficiently dealt with, thanks to penalties like $c_1 D_m/n$, whereas exponential complexity problems definitely require penalties of the type $c_1 D/n + c_2 D/n \log(n/D)$, since the former ones are shown to be misleading in this setting.
As already explained, these penalties result from upper bounds and depends on some constants (c_1, c_2), which must be determined afterwards, for instance by an intensive simulation study [23].

As an alternative to this approach, CV methods rely on the estimation (rather than majorization) of the risk. Moreover, such methods are known to provide reliable results in a wide range of frameworks (density estimation [33, 8], regression [37, 38], classification [12, 39]) and it is probably not the least contribution to their success.

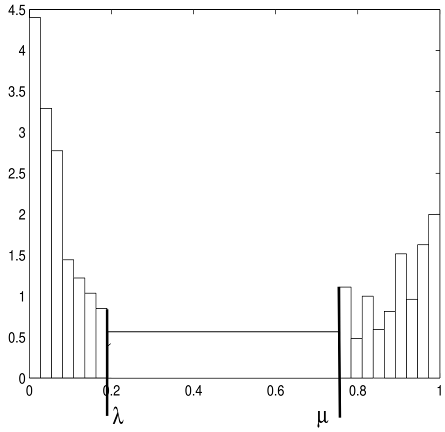

Figure 6.4: Example of histogram built from a given irregular partition, which belongs to $\mathcal{I}_{16,40}$. $\lambda = k/40$ and $\mu = \ell/40$.

However, one drawback of these methods is their sensitivity to the collection complexity, as in the exponential setting for instance (see Section 6.3.1), where it results in some misleading overfitting. The main point here is that the overfitting issue already arises in the polynomial framework. To our knowledge, this aspect is usually not mentioned in the literature.

A case example Let us start with an example to illustrate this phenomenon. Beforehand, we describe the collection of models we use.
From a regular partition of $[0,1]$ into N intervals of length $1/N$, let us consider all the partitions we can build from the regular one, made of k regular intervals of length $1/N$ from 0 to k/N, a wide central interval, and $N-\ell$ regular intervals of length $1/N$ from ℓ/N to 1, with $0 \leq k < \ell \leq N$. The set of all partitions of $[0,1]$ into D intervals built from the regular one is denoted by $\mathcal{I}_{D,N}$.
Furthermore, several values of N are used so that we define the set $\mathcal{I}$ of all the partitions of $[0,1]$ we consider by

$$\mathcal{I} = \cup_{1 \leq D \leq N_{\max}} \mathcal{I}(D) \quad \text{with} \quad \mathcal{I}(D) = \cup_{1 \leq N \leq N_{\max}} \mathcal{I}_{D,N},$$

where N_{max} denotes the maximal number of intervals of the largest regular partition of $[0,1]$ we consider. An example of the histograms we consider is given on Figure 6.4. This collection of partitions has already been used in the multiple testing context (Section 5).
Now, with each partition $I(m) \in \mathcal{I}$ is associated a model S_m of dimension D_m, spanned by the corresponding family $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$, where $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$ with $I_\lambda \in I(m)$.
Remark: For a given dimension D and $N > D$, $\text{Card}\,[\,\mathcal{I}_{D,N}\,] = D-2$, so that for any $N_{\max} > D$, $\text{Card}\,[\,\mathcal{I}(D)\,] = (D-2)(N_{\max}-D) = \text{Card}\,[\,\mathcal{M}(D)\,]$. Subsequently, the polynomial complexity assumption holds.

Our example consists in the estimation of three piecewise constant densities with respectively 2, 4 and 6 pieces as we can see on Figure 6.5. Note that these densities all belong to the collection of models we use. We have compared the loss of the histogram chosen according to the Loo to that of the oracle from Nb = 200 samples of size $n = 250$. These comparisons have been carried out for three values of $N_{\max}$: 20,40 and 60. Note that as underlined by the preceding remark, $N_{\max}$ determines the complexity of the collection of models we consider: the higher $N_{\max}$, the more complex the collection. The same simulations have been performed with $n = 1000$ as well (data not shown here), but results lead to the same conclusion.

Results are given in Table 6.1 where D^* denotes the dimension of the oracle. We observe that for a given experimental condition, the increase of $N_{\max}$ induces the growth of the average loss of the Loo-based histogram. This trend arises with each one of the three densities, which confirms our intuition about the

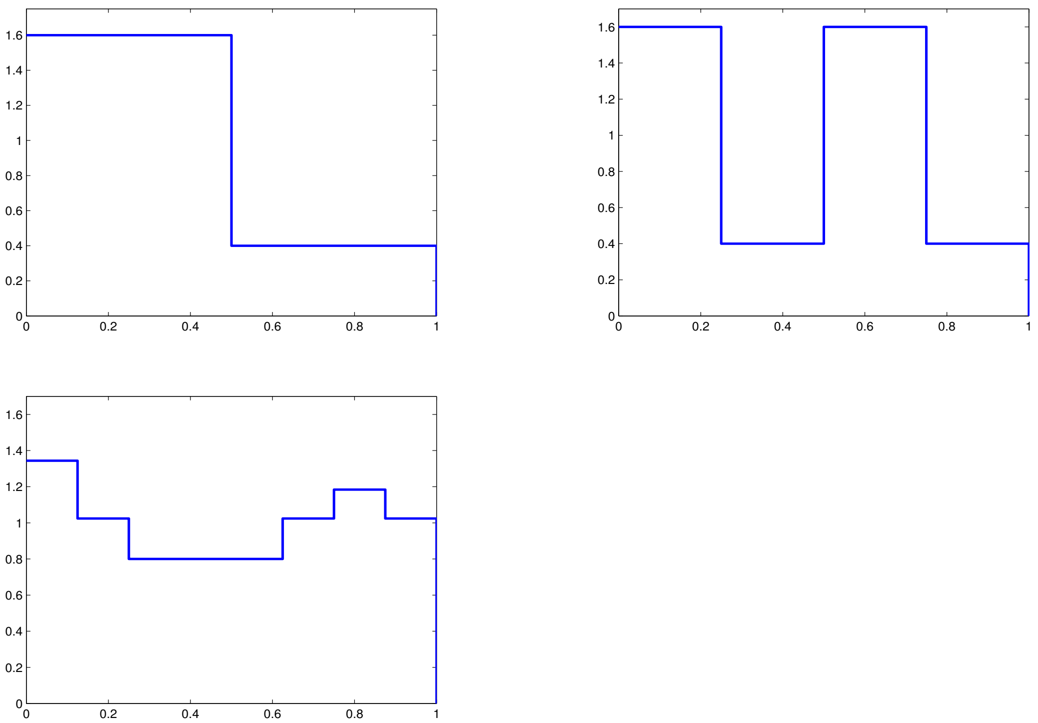

Figure 6.5: Graphs of three densities s built from a partition of $[0,1]$ in $D^* = 2, 4$ and 6 intervals.

overfitting in a polynomial setting. We think that this illustration is representative of how misleading Loo may be when no attention is paid for collection complexity.

Heuristics If we come back to Theorem 6.2.1, we see that we have required that it exists $\zeta > 0$ independent from n such that for any $m \in \mathcal{M}_n$, $nE_m \geq \zeta\, D_m$. Indeed since we use concentration inequalities like that of Proposition 6.2.2

$$\forall m \in \mathcal{M}_n, \quad \mathbb{P}\left[\left|\nu_n\left(\phi_m\right)\right| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m \right] \leq 2e^{-y_m},$$

D^*	$N_{\max}$	$\ell\left(s,\, \widehat{s}_{\,Loo}\right)$
2	20	0.0368 ± 0.042
	40	0.0699 ± 0.061
	60	0.0702 ± 0.073
4	20	0.0307 ± 0.048
	40	0.0435 ± 0.063
	60	0.0458 ± 0.074
6	20	0.0402 ± 0.023
	40	0.0589 ± 0.038
	60	0.0703 ± 0.047

Table 6.1: Average loss of the Loo-based histogram estimator $\ell\left(s,\, \widehat{s}_{\,Loo}\right)$ for 200 trials. Results are given with $\pm$ one standard deviation. D^* denotes the dimension of the oracle. The sample size is $n = 250$.

it is necessary to make sure that

$$\sum_{m\in\mathcal{M}_n} e^{-y_m} \leq \Sigma < +\infty \tag{6.13}$$

for a constant $\Sigma > 0$ independent from n.

However, the choice $y_m = nE_m$ (if we drop z) may somehow seem arbitrary and we may search for the minimal weight which fulfills the requirement in (6.13). Let us assume that for any dimension $\text{Card}\,[\,\mathcal{M}(D)\,] = D^\delta$, where δ comes from the polynomial complexity definition and $\mathcal{M}(D)$ denotes the set of all the models of dimension D. Then assuming that $y_m = y(D_m)$ only depends on the model m through its dimension, we have

$$\sum_{m\in\mathcal{M}_n} e^{-y_m} \quad = \quad \sum_{D=1}^{N_n} e^{-y(D)+\log(\text{Card}[\,\mathcal{M}(D)\,])},$$

where $N_n = \max\{D_m \mid m \in \mathcal{M}_n\}$. It suggests a choice of $y_m = y(D_m)$ of the shape $y(D) = \alpha \log(\text{Card}\,[\,\mathcal{M}(D)\,])$, which gives $y_m = \alpha\delta\log D$ due to the polynomial framework. Moreover assuming that $\delta \geq 1$, we deduce that $\alpha > 2$ provides the desired control

$$\sum_{m\in\mathcal{M}_n} e^{-y_m} \quad \leq \quad \sum_{D\geq 1} e^{-(\alpha-1)\delta\log D} \leq \Sigma < +\infty.$$

Thus under all these requirements, we could choose $y_m = \alpha \log \text{Card}\,[\,\mathcal{M}(D_m)\,]$ with $\alpha > 2$. If we now rewrite the summation so as we make nE_m appear, it comes that

$$\sum_{m\in\mathcal{M}_n} e^{-y_m} \quad = \quad \sum_{m\in\mathcal{M}_n} e^{-nE_m\,[\,\alpha\,\delta\,\log D_m/(nE_m)\,]}.$$

Then, the quantity in square brackets is increasing with respect to δ and could be related to the constant C in Proposition 6.2.2. Remembering that in the proof of Theorem 6.2.1, ϵ may be written as an increasing function of C, we could interpret ϵ as an increasing function of the complexity instead of a user specified constant. Thus, we notice that the bounds in condition (*Ran*) in Theorem 6.2.1 are increasing functions of the complexity as well. We see that an increase (at least in the polynomial range) of the collection complexity requires the choice of large values of p for the oracle inequality to apply.

Remark: Note that in no way we state the above reasoning as a mathematical established result. We only provide this heuristics because of the interesting relationship between the complexity and the choice of p it enables.

6.5.2 Simulations

In order to go further in the study of the overfitting phenomenon as well as to provide a potential solution to this issue, we show the results of the simulation experiment we have carried out. The setting is exactly the same as that described in the small illustration of Section 6.5.1. Actually, the difference lies in the wider range of Lpo algorithms we explored in order to understand to what extent an increase in p may overcome the overfitting issue.

Results are displayed in Table 6.2 and have been obtained for a sample size $n = 1000$. Very similar conclusions may be drawn from the same simulations with $n = 250$.

We may distinguish two different behaviours with respect to the amount of complexity. For the small values of p ($p \leq n/2$), we always observe an increase in the average loss as the complexity level grows. On the contrary, this is not always true for large values of p ($p > n/2$) where we may observe a decrease ($p = 750$).

Besides in the easiest problems ($D^* = 2$ and $D^* = 4$), overpenalizing with very large p enables to be very close to the oracle performance. Thus for a given experimental condition and given $N_{\max}$, we notice a constant decrease in the average loss as we rise p. This is no longer true when the problem is harder, that

D^*	$N_{\max}$	$\ell(s, \widehat{s}_{Loo})$	$\ell(s, \widehat{s}_{0.1})$	$\ell(s, \widehat{s}_{0.2})$	$\ell(s, \widehat{s}_{0.5})$	$\ell(s, \widehat{s}_{0.75})$	$\ell(s, \widehat{s}_{0.9})$
2	20	9.39 ± 0.64	8.52 ± 0.63	7.27 ± 0.6	3.49 ± 0.47	0.631 ± 0.2	$\mathbf{0} \pm 0$
	40	13.3 ± 1.1	12 ± 1.1	9.39 ± 0.98	4.19 ± 0.68	1.12 ± 0.39	$\mathbf{0.027} \pm 0.027$
	60	15.5 ± 1.2	13.8 ± 1.2	11.5 ± 1.1	4.53 ± 0.6	1.45 ± 0.35	$\mathbf{0.025} \pm 0.025$
4	20	6.11 ± 0.62	5.49 ± 0.6	4.88 ± 0.58	2.97 ± 0.47	0.665 ± 0.22	$\mathbf{0} \pm 0$
	40	8.98 ± 1	7.44 ± 0.93	5.8 ± 0.84	1.9 ± 0.37	0.0743 ± 0.074	$\mathbf{0} \pm 0$
	60	11.9 ± 1.3	10.5 ± 1.2	8.43 ± 1.1	2.2 ± 0.45	0.218 ± 0.17	$\mathbf{0.156} \pm 0.16$
6	20	9.89 ± 0.59	9.89 ± 0.59	9.79 ± 0.59	$\mathbf{9.75} \pm 0.58$	18.2 ± 0.37	20.8 ± 0.23
	40	12.1 ± 0.77	11 ± 0.71	10.3 ± 0.69	$\mathbf{9.95} \pm 0.51$	17.1 ± 0.38	21.1 ± 0.2
	60	14.2 ± 1	13.2 ± 0.97	11.8 ± 0.88	$\mathbf{10.6} \pm 0.61$	17.3 ± 0.46	20.7 ± 0.24

Table 6.2: Average loss of various Lpo-based histogram estimators: Loo ($\ell(s, \widehat{s}_{Loo})$), and Lpo for $p = 100$ ($\ell(s, \widehat{s}_{0.1})$), $p = 200$ ($\ell(s, \widehat{s}_{0.2})$), $p = 500$ ($\ell(s, \widehat{s}_{0.5})$), $p = 750$ ($\ell(s, \widehat{s}_{0.75})$) and $p = 900$ ($\ell(s, \widehat{s}_{0.9})$). Nb $= 200$ trials have been used for the averages and the sample size $n = 1000$. Results are given with $\pm$ one standard deviation and all the values have been multiplied by 1000. D^* denotes the dimension of the oracle. **Bold text** stresses the minimum value of the line, while underlined text points out the significantly worse results of the line.

is when several thin picks have to be detected without large difference between the level of one another ($D^* = 6$). Thus, removing too much data prevents us from recovering a good estimate, which explains we observe the optimal performance for $p = n/2$ and then an increase of the average loss as p grows.
Anyway in the explored examples, small values of p definitely lead to very poor estimates.

OPEN QUESTION:

- There is not yet any guideline to appropriately choose the parameter p in any real situation.
- However, it turns out to be a promising direction to further explore, especially because there are still meaningful things to understand in the way collection complexity interferes in the CV algorithm.

6.6 Two-step algorithm

6.6.1 Description

In Sections 6.3 and 6.5, we have pointed out some issues, which may be encountered by cross-validation procedures such as the popular Loo for instance. Some of them may be (at least partially) overcome by the Lpo, thanks to an appropriate choice of p, in the polynomial setting for instance. In the exponential framework, the addition of a complexity term has been proposed and results in a procedure whose performance is warranted by an oracle inequality.
However, two main criticisms can be formulated about these two proposals. In the polynomial setting, we do not yet give any recipe to properly choose p in any realistic situation. As for the exponential complexity, the proposed penalty could be applied, but at the price of a preliminary estimation step, since this penalty depends on the unknown $\|s\|_\infty$.
In the sequel, we tackle these questions and remarks from the practical point of view by describing a fully data-driven procedure, which adapts itself to the problem complexity and provides reliable results according to our simulation experiments.

As we described in Section 6.3.1, the classical way to cope with exponential complexity of a collection of models is to gather models with the same dimension, so that we get a family $\{\widehat{m}(D)\}_D$. The complexity of the resulting models is then evaluated through, for instance, the addition of an appropriate penalty term. The addition of this penalty should be understood as an attempt to estimate the risk of each of the corresponding estimators $\widehat{s}_{\widehat{m}(D)}$:

$$\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(D)}\right\|^2\right].$$

Following the cross-validation heuristics (Section 3.2.1), we see that an alternative way to estimate this risk is to use CV. Due to the lack of closed-form expression for the Lpo risk estimator of any of the $\widehat{s}_{\widehat{m}(D)}$s, we decide to use the VFCV algorithm. Indeed for a given $1 \leq V \leq n$, we obtain

$$\mathbb{E}\left[\left\| s - \widehat{s}_{\widehat{m}(D)} \right\|^2 \right] \approx \frac{1}{V} \sum_{v=1}^{V} P_n^{e_v} \gamma \left(\widehat{f}_D^{e_v^c} \right) =: \widehat{R}_{\mathrm{VFCV},V}(D) \qquad \text{with} \qquad \widehat{f}_D = \widehat{s}_{\widehat{m}(D)},$$

where $e_1, \ldots, e_V$ denotes the V elements of the partition of $\{1, \ldots, n\}$ in subsets of cardinality n/V (assuming that V divides n). $\widehat{f}_D^{e_v}$ denotes the estimator $\widehat{f}_D$ computed from the data in the training set $\{X_i \mid i \in e_v^c\}$, and $P_n^{e_v} = V/n \sum_{i \in e_v} \delta_{X_i}$. Thus, the final dimension is chosen by minimization of the VFCV criterion as a function of D.

The above algorithm, denoted by $\mathcal{A}$, can therefore be summarized in the following way:

Algorithm 6.6.1.

1. *For each dimension D,*

$$\widehat{m}(D) = \mathrm{Argmin}_{m \in \mathcal{M}(D)} P_n \gamma(\widehat{s}_m),$$

2. *Then,*

$$\widehat{D} = \mathrm{Argmin}_D \widehat{R}_{\mathrm{VFCV},V}(D),$$

where $\widehat{R}_{\mathrm{VFCV},V}(D)$ denotes the V-fold estimator of the risk of $\widehat{s}_{\widehat{m}(D)}$.

3. *Finally,*

$$\widetilde{s} = \widehat{s}_{\widehat{m}(\widehat{D})}.$$

6.6.2 Simulations

Our present purpose is to illustrate through some simulations that this data-dependent algorithm outperforms Lpo for small values of p in both the polynomial and the exponential complexity frameworks.

Polynomial setting

The experimental design is the same as that of Section 6.5.2, except we used two more densities with respective dimension 3 and 4, displayed on Figure 6.6.
The comparison between Loo and algorithm $\mathcal{A}$ is therefore carried out on 6 densities (see Figure 6.5 and Figure 6.6), from samples of size $n = 1000$. The collection of models we use is the same as that of Section 6.5.2.
The results of this comparison are displayed in Table 6.3 where we see the average loss of the estimator selected by each algorithm in the different experimental conditions.

We observe that algorithm $\mathcal{A}$ nearly always outperforms the Loo. The discrepancy is even significant in half of the simulation conditions. We also notice that the loss of algorithm $\mathcal{A}$ seems more stable as the complexity grows than that of Loo, which suggests that $\mathcal{A}$ is not as sensitive to an increase in the polynomial complexity as Loo.
However, we also see that the gain in the use of $\mathcal{A}$ is no longer that high in the last experimental condition where both criterion are very similar. We conclude that there may be a great interest in the use of algorithms such as $\mathcal{A}$, even in the polynomial setting. Nevertheless, some further (theoretical and experimental) study is required to get insight in the relationship between resampling in general and complexity.

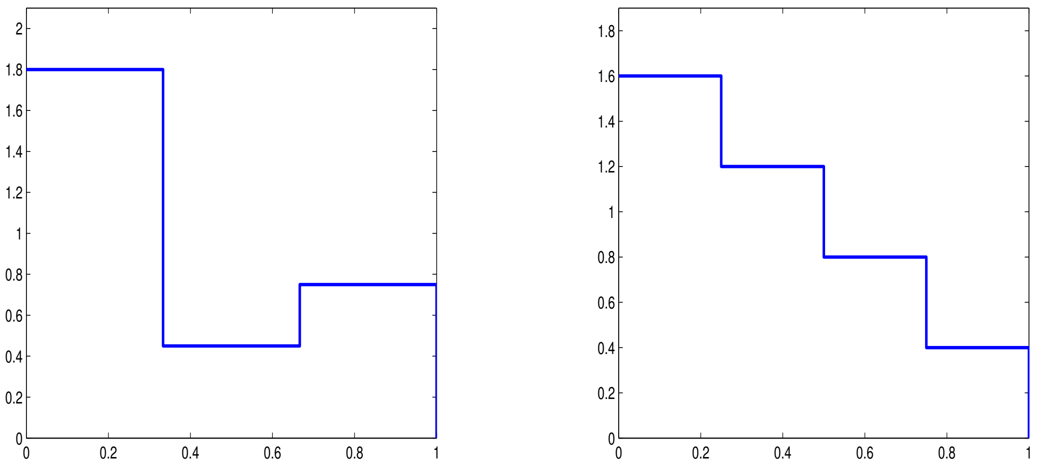

Figure 6.6: Examples of piecewise constant densities built from a partition in 3 and respectively 4 intervals

D^*	$N_{\max}$	$\ell(s, \widehat{s}_{Loo})$	$\ell(s, \widehat{s}_{\mathcal{A}})$
2	20	5.21 ± 0.56	**4.68** ± 0.55
	40	$\underline{8.02}$ ± 1	**5.92** ± 0.88
	60	6.95 ± 0.84	**6.45** ± 0.81
3	20	4.05 ± 0.56	**3.71** ± 0.58
	40	$\underline{6.51}$ ± 0.83	**3.76** ± 0.66
	60	$\underline{7.35}$ ± 0.94	**3.71** ± 0.64
4	20	$\underline{5.89}$ ± 0.62	**3.42** ± 0.52
	40	$\underline{6.87}$ ± 0.85	**4.34** ± 0.74
	60	$\underline{8.09}$ ± 1	**5.74** ± 0.85
4 bis	20	5.22 ± 0.7	**4.21** ± 0.63
	40	6.61 ± 0.99	**5.03** ± 0.83
	60	$\underline{6.02}$ ± 0.77	**3.92** ± 0.66
6	20	**10.4** ± 0.59	10.5 ± 0.61
	40	12.2 ± 0.77	**11.4** ± 0.75
	60	14.2 ± 1	**14.1** ± 0.94

Table 6.3: Average loss of the histogram estimators chosen by Loo ($\ell(s, \widehat{s}_{Loo})$) and algorithm $\mathcal{A}$ ($\ell(s, \widehat{s}_{\mathcal{A}})$). $\widehat{s}_{\mathcal{A}}$ denotes the estimator resulting from algorithm $\mathcal{A}$. Nb $= 200$ trials have been used for the averages and the sample size $n = 1000$. Results are given with ± one standard deviation and all the values have been multiplied by 1000. D^* denotes the dimension of the oracle. **Bold text** stresses the minimum value of the line, while underlined text points out the significantly worse results of the line.

Exponential setup

In Section 6.3.1, we have pointed out some deficiencies of the Lpo with small values of p in the exponential setting. Indeed Figure 6.2 has highlighted that Loo suffers from some strong overfitting with such "rich" collections of models.
We perform a few simulations that clearly illustrate this deficiency of the Lpo when $p \leq n/2$, and enlightens the weaker sensitivity of algorithm $\mathcal{A}$ to collection complexity.

We generate samples of size $n = 250$ and $n = 1000$ from two of the densities displayed on Figure 6.5 with respective dimensions $D = 4$ and 6. We used the same subdivision $t_0, \ldots, t_K$ of $[0, 1]$ with $t_i = i/K$ and $K = 100$ as in the illustration provided in Section 6.3.1. All the partitions of $[0, 1]$ defined from this subdivision are used. For each experimental condition, we draw Nb = 200 trials and we only consider models of dimension less or equal to 50. Results are provided in Table 6.4.

D^*	n	$\mathcal{A}$	$p = 1$	$p/n = 0.1$	$p/n = 0.2$	$p/n = 0.5$
4	250	**11.5** $\pm$ 0.88	25 $\pm$ 0.5	23 $\pm$ 0.48	21.5 $\pm$ 0.48	15 $\pm$ 0.37
	1000	**9.45** $\pm$ 0.65	25.1 $\pm$ 0.54	23.2 $\pm$ 0.53	21.5 $\pm$ 0.53	14.9 $\pm$ 0.41
6	250	**9.03** $\pm$ 0.46	19.5 $\pm$ 0.32	18.2 $\pm$ 0.3	17 $\pm$ 0.29	11.9 $\pm$ 0.22
	1000	**5.87** $\pm$ 0.33	20.7 $\pm$ 0.33	19.4 $\pm$ 0.3	18.3 $\pm$ 0.3	12.6 $\pm$ 0.25

Table 6.4: Average loss values for the different tested algorithms $\mathcal{A}$, and Lpo with $p = 1$, $p = 0.1\,n$, $p = 0.2\,n$ and $p = 0.5\,n$. In each experimental condition, the sample sizes are $n = 250$ and $n = 1000$.

Let us consider Table 6.4 where we give the average loss value of the final estimator selected by each algorithm.
At first glance, we notice that the values of the best algorithm are quite high. It must be related to the unavoidable $\log n$ term, which is the price to pay for such a high complexity ([9, 10]). Especially, it explains the increase in the loss as the sample size grows.
We observe that all the considered Lpo-based algorithms have significantly worse performances than algorithm $\mathcal{A}$, and strongly suffer from overfitting.
In order to illustrate the behaviour of $\mathcal{A}$, we also provide Figure 6.7, which depicts the estimated risk of each estimator $\widehat{s}_{\widehat{m}(D,alg)}$ with respect to the dimension for one trial, where alg denotes one of the competing algorithms. This graph is obtained from the estimation of the density in Figure 6.5 with dimension $D^* = 6$.
We notice that unlike the Lpo curves for $p \leq n/2$, which still decrease for a while after the bias vanishes, that of $\mathcal{A}$ reaches its minimum nearly the dimension of the true model ($D^* = 6$).

6.7 Discussion

One of the main interest of this chapter is the interaction between the Lpo algorithm and the complexity of the collection of models we consider. Throughout this study, several aspects have been explored.

In the case of a polynomial complexity, the Lpo algorithm enjoys some optimality properties in terms of an oracle inequality, which holds for any projection estimator. In the example of the histogram estimator, an adaptivity result in the minimax sense has been obtained.
If we consider the exponential setup, we observe that the widespread leave-one-out algorithm fails dramatically as well as some Lpo algorithms when p is small with respect to n. In order to take into account the richness of the collection of models, we propose to add a penalty term to the Lpo estimator of the risk. An oracle inequality is also derived for this procedure.
In this study, we also consider some practical aspects through some simulation experiments, which highlight the overfitting suffered by the Loo algorithm, even in the polynomial framework. The Lpo with larger values of $p > 1$ is shown to balance this overfitting phenomenon. We also propose a data-driven procedure, which exhibits an automatic adaptation property to the complexity of the model collection.

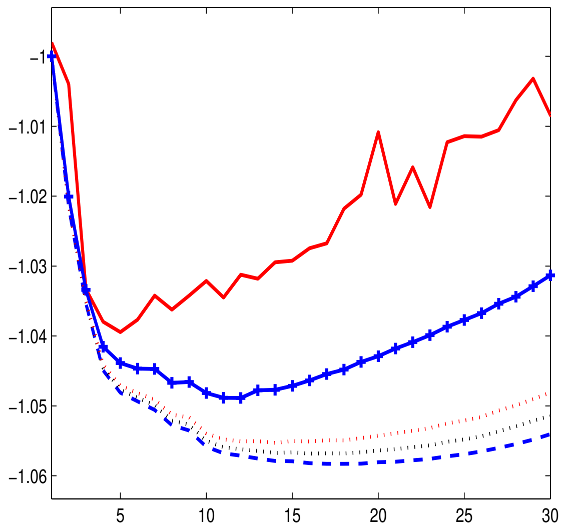

Figure 6.7: Plot of the Lpo risk estimators versus the dimension: Loo ('–' bottom dashed line), Lpo with $p = 100, 200$ (':' middle dotted lines), Lpo with $p = 500$ ('-+-' line). The estimated value provided by algorithm $\mathcal{A}$ corresponds to the '-' top plain line. The sample size is $n = 1000$.

Bibliography

[1] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to Unknown Sparsity by controlling the False Discovery Rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[2] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning*, page submitted, 2008.

[4] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with unknown variance. *The Annals of Statistics*, 00:00, 2008.

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

[6] A. Barron and T. M. Cover. Minimum Complexity Density Estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

[7] R. E. Bellman and S. E. Dreyfus. *Applied Dynamic Programming*. Princeton, 1962.

[8] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgensen, and G. Yang, editors, *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

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

[10] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 2006.

[11] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram? *ESAIM Probab. Statist.*, 10:24–45, 2006.

[12] G. Blanchard and P. Massart. Discussion: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2664–2671, 2006.

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

[14] L. Breiman. The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. *J. Amer. Statist. Assoc.*, 87(419):738–754, 1992.

[15] G. Castellan. Modified Akaike's criterion for histogram density estimation. Technical Report 99.61, Université de Paris-Sud, 1999.

[16] G. Castellan. Density estimation via exponential model selection. *IEEE transactions on information theory*, 49(8):2052–2060, 2003.

[17] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

[18] R.A. DeVore and G.G. Lorentz. *Constructive Approximation.* Springer, 1993.

[19] S. Geisser. A predictive approach to the random effect model. *Biometrika*, 61(1):101–107, 1974.

[20] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[21] I. Ibragimov and R. Khas'minskij. *Statistical Estimation. Asymptotic Theory.* Springer-Verlag, Berlin, 1981.

[22] A. Juditsky and S. Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. *Bernoulli*, 10(2):187–220, 2004.

[23] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

[24] K.-C. Li. Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975, 1987.

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

[26] P. Massart. *Concentration Inequalities and Model Selection.* Lecture Notes in Mathematics. Springer, 2007.

[27] J. Rissanen. Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431, 1983.

[28] M. Rudemo. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.

[29] Marie Sauvé. Histogram selection in non gaussian regression. Technical Report 5911, INRIA, may 2006.

[30] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[31] J. Shao. Model Selection by Cross-Validation. *Journal of the American Statis. Association*, 88(422):486–494, 1993.

[32] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[33] C.J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.

[34] 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.

[35] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *JRSS B*, 39(1):44–47, 1977.

[36] A. B. Tsybakov. *Introduction à l'estimation non-paramétrique.* Mathématiques et Applications. Springer-Verlag, 2003.

[37] M. Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003.

[38] Y. Yang. Regression with multiple candidate model: selection or mixing? *Statist. Sinica*, 13:783–809, 2003.

[39] Y. Yang. Comparing Learning Methods for Classification. *Statistica Sinica*, 16:635–657, 2006.

[40] Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

[41] P. Zhang. Model selection via multifold cross-validation. *The Annals of Statistics*, 21(1):299–313, 1993.

Chapter 7

Change-points detection *via* resampling

This chapter is a joint work with Sylvain Arlot initiated since January 2008, and is devoted to the change-points detection problem. Our approach is based on two remarks. A wide range of existing methods rely on a homoscedasticity assumption and the collection we consider is rich (exponential complexity). Moreover, we know that resampling strategies provide reliable results with both homoscedastic and heteroscedastic data, but only with not too rich collections. Therefore, our idea is simply to use appropriate resampling-based algorithms as a means to cope with heteroscedasticity in the unusual framework of exponential complexity.
We first describe the statistical framework to point out that we are interested by breakpoints in the mean with possibly heteroscedastic data. Through a parallel with the Birgé and Massart penalty [16] derived in the homoscedastic setup, we interpret the penalty as a complexity measure of the collection in hand, which is supported by the simulations. The proposed algorithm turns out perform almost as well as Birgé and Massart's penalty under homoscedasticity in choosing the number of breakpoints, whereas it outperforms the latter penalty under heteroscedasticity.
Similarly, it is also possible to enhance the procedure by using the resampling to choose the best segmentation for each dimension, instead of the empirical risk minimization.
Finally, we apply our algorithms to a real data set and the results are compared with those yielded by other well known methods, indicating promising performances.

7.1 Introduction

The change-points detection (also called one-dimensional segmentation) deals with a stochastic process the distribution of which abruptly changes at some unknown instants. The purpose is then to recover the location of these changes. Due to a wide range of applications, from voice recognition and time-series analysis in the financial area [29], to biology and CGH (Comparative Genomic Hybridization) data

analysis [37], the literature about change-points detection is very abundant. We refer to Basseville and Nikiforov [13] and to Brodsky and Darkhovsky [18] for a wide range of applications.
Two different approaches have been developed on the matter. The first one aims at detecting any potential change from the *on-line* signal, that is as the observations are sequentially given. On the contrary, the other is said *off-line* since the data are assumed to be entirely observed at the same time and the problem is to recover any change *a posteriori*. The present work belongs to the *off-line* setting.
The first papers on the subject were devoted to the search for the location of a unique change-point (also named break-point) [36]. Looking for multiple change-points is a harder task and has been studied later. For instance, Yao [48] performs the multiple change-points detection from a gaussian signal by use of the BIC criterion, while Miao et Zhao [35] propose an approach relying on rank statistics.

Two different viewpoints have arisen in the multiple change-point detection problem: one is frequentist [24, 27] while the other is bayesian [26]. It is interesting to notice that these different conceptions do not provide the same kind of information about the signal in hand. Indeed, the first one essentially reduces to find the "best" segmentation, that is the list of the breakpoints which is the solution of an optimization problem, whereas the bayesian approach rather intends to estimate the distribution of the change-point at a given location as well as the corresponding confidence interval. Since both of them present some interesting aspects, Lavielle [25] combines the two strategies by putting an *a priori* distribution on segmentations, which is based on the penalized criterion preliminary used to define the best segmentation. Throughout this chapter, we adopt the frequentist viewpoint and the problem is to recover the potential multiple change-points locations from the observed signal.

There are numerous asymptotic results [35, 28]. Provided a "true segmentation" exists and belongs to the set of candidate segmentations, such approaches attempt to estimate consistently the true segmentation. On the contrary, our approach is fully non-asymptotic, in the same line as that of Lebarbier [30] and Baraud *et al.* [10], which are among the few works on this aspect.
In (*off-line*) change-points detection, we are given n points $(t_1, Y_1), \ldots, (t_n, Y_n) \in [0,1] \times \mathbb{R}$, which are n successive observations of a signal Y_i at point t_i, and satisfying

$$Y_i = s(t_i) + \sigma_i \epsilon_i, \quad (\epsilon_i)_i \ \textit{i.i.d.}\,, \ \mathbb{E}\left[\, \epsilon_i \,\right] = 0, \ \mathrm{Var}(\epsilon_i) = 1 \quad \text{and} \quad (\sigma_i)_i \in (\mathbb{R}_+)^n\,, \tag{7.1}$$

where s denotes the unknown regression function. Usually in the change-points detection framework, s is assumed to be piecewise constant. That is why piecewise constant estimators are used in this setup. Besides as pointed out by Lavielle [25], several change-points detection problems may be distinguished, depending on whether we are interested in changes in the mean under constant variance, in the variance with constant mean or in both the mean and the variance.
In this work, the question we address is different from the previous ones in two respects. Indeed, we focus on the heteroscedastic case in which we intend to detect changes in the mean of the signal, and not in its variance. To this end, we use piecewise constant functions as approximation of s. To a certain extent in this setting, variance could be considered as a nuisance parameter. Moreover, the regression functions we consider are more general since s may be non piecewise constant. With the latter situation, we have in mind very smooth (but non necessarily constant) signals with abrupt changes in the mean.

Our approach relies on model selection, which has been introduced in the early 70s by Akaike [2, 3] for FPE and AIC, Mallows [33] for the C_p criterion and Schwartz [41] for BIC, to name but a few. From a finite (or at most countable) family of models $(S_m)_{m \in \mathcal{M}_n}$ and given a set of associated estimators $(\widehat{s}_m)_{m \in \mathcal{M}_n}$, the purpose of model selection is to provide $\widehat{m}$ such that $\widehat{s}_{\widehat{m}}$ reaches the best performance according to our criterion among all the estimators $\widehat{s}_m$.
In model selection, two main objectives can be pursued. We may intend to design a model selection procedure which finds the estimator with the smallest quadratic risk. This strategy is that of AIC and Mallows' C_p for instance. The quality of such a procedure is then measured through an oracle inequality like

$$\mathbb{E}\left[\, \|s - \widehat{s}_{\widehat{m}}\|^2 \,\right] \leq C \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\, \|s - \widehat{s}_m\|^2 \,\right] + R_n \tag{7.2}$$

where $C \geq 1$ is a constant and R_n denotes a remainder term. Of course, the closer C to 1 and R_n to 0, the better the procedure. On the other hand if we assume that s belongs to one of the S_ms, then we may

intend to recover it with a probability as high as possible. It is the purpose of BIC for instance (see [47] for a distinction of these two problems).
Actually, the present work belongs to the former strategy. Indeed, there is not necessarily any true model in our setting since smooth regression functions with abrupt changes are allowed, while in the same time we use piecewise constant functions. Besides, we do not try to detect a change-point, which is hidden by the noise.

In the change-point detection problem requires, the collection of models $\mathcal{M}_n$ can depend on n. Indeed, with n points in the design, there are $\binom{n}{D}$ different partitions of $[0,1]$ into D pieces. Moreover, if s is not piecewise constant, it is natural to use more parameters to estimate it. Consequently, an oracle inequality (7.2) has to be satisfied from the non-asymptotic viewpoint: we cannot assume that the collection of model is fixed and the sample size n tends to infinity.

Two broad situations may be distinguished in the model selection area, depending on the "small" or "large" number of models with the same dimension in the collection. Typically, the "small" case is that of polynomial collection complexity. It is the situation where AIC, C_p and their variants lead to good performance [9, 16, 11, 17]. However, these criteria heavily rely on a homoscedastic assumption which may be violated in various situations with true data. In contrast, there is only very few results in the heteroscedastic polynomial setting, where the previous penalties have been recently shown to be suboptimal [7]. Gendre [22] has proposed a model selection procedure that seems to perform quite well, but he requires that the residuals are gaussian variables. In a recent work in the heteroscedastic framework, Arlot [5] shows that resampling penalties give promising results without any assumption on the type of distribution of residuals. Unfortunately, the results he provides only apply in the polynomial setting.
As for the exponential complexity framework, Birgé and Massart [16, 17] advocate penalties like

$$\mathrm{pen}(m) = c_1 \frac{D_m}{n} + c_2 \frac{D_m}{n} \log\left(\frac{n}{D_m}\right), \tag{7.3}$$

where c_1 and c_2 are positive universal unknown constants, which may be determined through a simulation step as successfully carried out by Lebarbier [30]. Similar penalties have been found independently by several other authors [1] and are shown to provide satisfactory results, except their dependence on the homoscedastic assumption. Indeed having in mind the deficiency of linear penalties in the heteroscedastic framework under polynomial complexity [7], we may conjecture that the same issue will arise in the exponential setting, which we observe in what follows.

The purpose of the present work is therefore to analyze with some theoretical arguments, but mainly through an extensive simulation study several resampling strategies such as CV in the heteroscedastic change-points detection setting. We propose a fully data-dependent algorithm, which automatically takes into account the complexity of the collection of models in hand. In a homoscedastic setting, these new algorithms provide similar results as (7.3), whereas they turn out to outperform (7.3) in the heteroscedastic setting. Thus since we do not know in advance whether the noise is constant or heteroscedastic with real data, these new algorithms appear as reliable and more robust alternatives to (7.3). Besides, a common strategy in model selection [16] consists in gathering models according to their dimensions and then, choosing one model for each dimension thanks to the empirical contrast minimization. The final estimator is selected among the resulting family of estimators $\left(\widehat{s}_{\widehat{m}(D)}\right)_{D\in\mathcal{D}}$, thanks to penalization. Following the same approach, several simulations experiments are carried out with different regression functions and heteroscedastic noises in order to understand the behaviour of the considered algorithms at each step of the described strategy. Thus, we observe that the V-fold cross-validation (VF) enables to automatically fit the collection complexity, while in many circumstances, resampling at the first step of the above strategy outperforms upon empirical risk minimization.

The chapter is organized as follows. In Section 7.2, we first describe the change-points detection from the model selection viewpoint, and then provide non-asymptotic results about the Lpo risk estimator. It is shown that increasing p is a means to overpenalize, which is a means to balance overfitting. in this section, we also describe our strategy to study the problem as well as the simulation experiments we pursue. Section 7.3 is devoted to the description and analysis of resampling as measures of the collection complexity in order to choose the dimension of the final model. In particular, some comparisons are made with penalties like (7.3), which turn out to be in favour of resampling. Besides, we also compare the empirical risk and resampling algorithms as means to choose the best model for each dimension in Section 7.4. The proposed algorithms are also compared to widespread methods on CGH microarray data in Section 7.6. Some comments on alternative complexity measures and model selection strategies are provided in Section 7.7 as well as some starting points for further works.

7.2 Overview of the problem

7.2.1 Model selection view on change-points detection

From the efficiency viewpoint, a classical strategy in model selection consists in estimating the risk of each model and then to choose that one with the smallest estimated error. It is the rationale of penalized criteria such as AIC (Akaike Information Criterion, Akaike [3]), Mallows' C_p [33], but also that of CV algorithms (Allen [4], Stone [44], Geisser [21]) and more generally of resampling penalties (see the introduction of [6] for an overview).

A large number of asymptotic results have been established for such criteria when the number of competing models in not too high. The asymptotic efficiency of C_p has been proved by Shibata [43], while the asymptotic equivalence between the C_p and Loo is shown in Li [31]. We refer the interested reader to Shao [42] and the following discussion papers for an extensive review of asymptotic results derived for linear regression models.
Among the first non-asymptotic results in a regression setting with penalized criteria, we find Barron *et al.* [12] and Birgé and Massart [16] in a gaussian homoscedastic framework. Some of these results apply with "rich" collections of models such as those encountered in all subset selection problem in variable selection for instance. In a situation where the collection is not too rich, Baraud [9, 11] extends the preceding results (oracle inequality with $C > 1$) with a homoscedastic but unknown noise level. Moreover, he replaces the gaussian assumption by some requirements on the moments of the residuals. A noticeable enhancement of these results (constant $C_n \to 1$) in a gaussian homoscedastic setting has been recently obtained by Birgé and Massart [17] with various family of models.

As for CV, Stone [45] has proved the asymptotic equivalence of Loo and AIC, while more recently, Dudoit and van der Laan [20] have shown the asymptotic optimality of a model selection procedure based on various CV algorithms except Loo (see Györfi *et al.* [23] for more asymptotic results). Some non-asymptotic results arise in Lugosi et Nobel [32] in a prediction framework and for regression in Wegkamp [46] who makes some assumptions on the residuals moments. In a wide comparison of VF with the named resampling penalties, Arlot [6, 8] enabled a better understanding of CV techniques with respect to the cardinality of the test set. Celisse and Robin [19] recently proposed a data dependent method to choose this parameter.

In the change-point detection problem, an additional difficulty lies in the richness of the collection in hand. Indeed, we would like to be able to detect a breakpoint at any of the design points, which leads us to consider up to 2^n models, where n denotes the sample size. Moreover for each dimension D, we still have $\binom{n-1}{D-1}$ models of dimension D, which is too large for a criterion similar to the C_p, if we allow the dimension to increase with n. Indeed, Birgé and Massart [17] proved that the quadratic risk of an estimator resulting for such a criterion in this framework can go to infinity. This is essentially the consequence of the fact that Mallows' C_p unbiasedly estimates the risk of an estimator, but has a non null probability to choose a wrong model. When these models are too numerous, this probability becomes sufficiently high to induce the frequent choice of very bad models.

A classical way to solve this issue is overpenalization, that is the overestimation of the risk of the largest models in order to prevent us from any mistaken choice of one of them. Depending on the collection complexity, several penalties have been proposed for instance by Birgé and Massart [16, 17] and Sauvé [40]. Since such penalties depend on some unknown constants, a calibration methodology have been developed in [17] and applied to the change-points detection problem by Lebarbier [30].

Since these penalties may be written as a function $F_n(D_m)$ of the dimension D_m for each m, we may also describe these procedures in the following way. On the one hand, for each dimension $D \in \{1, \ldots, n\}$, we choose the estimator $\widehat{s}_{\widehat{m}(D)}$ minimizing the empirical risk over

$$\widetilde{S}_D := \bigcup_{D_m = D} S_m \ .$$

Then, we choose a dimension $\widehat{D}$ which minimizes the sum of the empirical risk and of the penalty term $F_n(D)$. the final model corresponds to $\widehat{m} = \widehat{m}(\widehat{D})$. Thus, everything happens as if we had chosen a model among $\left(\widetilde{S}_D\right)_{1 \leq D \leq n}$, which is not a rich collection of models since there is only one model $\widetilde{S}_D$ by dimension. The reason why we cannot keep a penalty like Mallows' C_p is that for each dimension D the *aggregated model* $\widetilde{S}_D$ is more complex that any of its components S_m. The above penalties are essentially devoted to measuring this "complexity".
Remark: Note that if the S_ms are linear spaces of dimension D, then the corresponding $\widetilde{S}_D$ is no longer a vector space, but only the union of the S_ms. Although these two sets have a different nature, they are both called models. Furthermore, if the complexity structure of a given vector space S_m is easily described by its dimension, the effective complexity of $\widetilde{S}_D$ is by far more difficult to characterize, even if some upper bounds on metric entropy of such sets can be obtained, which lead to the same penalty term as that of [16].

The major drawback of penalized criteria such as that of Birgé and Massart [16] and Sauvé [40] is that their justification heavily relies on the homoscedasticity of the data. For instance, Mallows' penalty $2\sigma^2 D_m n^{-1}$ applies as long as the noise level remains constant equal to σ. As soon as this assumption is violated, the corresponding criterion does no longer unbiasedly estimate the risk, except very peculiar situations. Furthermore, Arlot [7] recently proved that linear penalties in the dimension of the model only lead to suboptimal procedures, even if the knowledge of the true distribution is used. On simulated data, it results in very bad performances for Mallows' C_p.
Subsequently, it seems dangerous to use penalties derived from theoretical results which strongly depend on the homoscedastic assumption. For not too rich collection of models, that is polynomial in the sample size, a natural idea is the use of CV and more generally of resampling techniques, known to be more robust to such hard tasks. For instance, Arlot [5, 8] has established that resampling penalties enable to perform an optimal model selection in a heteroscedastic framework, without any assumption on the shape of the noise level, or any distributional assumption. These results suggest that CV algorithm should enjoy similar properties, provided the size of the test set is large enough with respect to the sample size.

However, these results do not apply anymore in the change-points detection framework, due to collection complexity. In the sequel, we aim at studying how such resampling methods could be adapted to this specific problem, in particular when the data are of heteroscedastic nature.

Beforehand, several questions already arise:

- Do the penalties proposed in [16] still work from a practical viewpoint, even in the heteroscedastic case?

- Does resampling work better than the previous penalties?

- Provided resampling strategy enables some enhancement (for instance in the heteroscedastic setting), several resampling algorithms may be applied. Is there an "optimal" choice of one of them?

We will attempt to answer some of these questions by theoretical considerations and mainly on the basis of an extensive simulation study.

7.2.2 Purpose and strategy

Notations

Let $(t_1, Y_1), \ldots, (t_n, Y_n) \in [0,1] \times \mathbb{R}$ denote n random variables $Z_i = (t_i, Y_i)$ which are n successive observations of a signal Y at points $(t_i)_{1 \leq i \leq n}$. Moreover, let us assume that the following model holds

$$Y_i = s(t_i) + \sigma(t_i)\epsilon_i, \quad (\epsilon_i)_i \ \textit{i.i.d.}\,, \ \mathbb{E}\left[\,\epsilon_i\,\right] = 0,\ \mathrm{Var}(\epsilon_i) = 1, \tag{7.4}$$

where the t_is are deterministic points of $[0,1]$ and $\sigma: \ [0,1] \to \mathbb{R}_+$. Set $P_n = n^{-1}\sum_{i=1}^n \delta_{(t_i, Y_i)}$, the empirical measure of $Z_1, \ldots, Z_n$ and P, the distribution of a "new observation" $Z = (t, Y)$. Then,

$$P = P_n^t \otimes P_{Y|t}$$

where $P_n^t = n^{-1}\sum_{i=1}^n \delta_{t_i}$ and $P_{Y|t}$ denotes the conditional distribution of Y given t.
Set γ, the quadratic contrast such that for any *predictor* $f: \ [0,1] \to \mathbb{R}$, the prediction error is denoted by $P\gamma(f) = \mathbb{E}_{Z\sim P}\left[\,\gamma(f; Z)\,\right]$, where $\gamma(f; z) = (f(x) - y)^2$ with $z = (x, y)$. As we know, the prediction error reaches its minimum for $f = s$. Thus, the excess risk of f (also named the loss of f) is defined by

$$\ell\,(s, f\,) = P\gamma(f) - P\gamma(s) \ \geq 0 \ .$$

Remark: With the quadratic contrast, we have

$$\begin{aligned}
\ell\,(s, f\,) &= \mathbb{E}\left[\,(f(t) - Y)^2 - (s(t) - Y)^2\,\right] = \mathbb{E}\left[\,(f(t) - s(t))\,(f(t) + s(t) - 2Y)\,\right], \\
&= \mathbb{E}\left\{(f(t) - s(t))\,\mathbb{E}\left[\,(f(t) + s(t) - 2Y) \mid t\,\right]\right\}, \\
&= \mathbb{E}\left[\,(f(t) - s(t))^2\,\right].
\end{aligned}$$

Thus, prediction and estimation lead to the same estimator.
Given a model S_m and a set of observations (represented by the empirical measure P_n), the mean squares estimator is denoted by

$$\widehat{s}_m = \widehat{s}_m(P_n) = \mathrm{ERM}(S_m, P_n) := \arg\min_{t \in S_m}\left\{P_n\gamma(t)\right\} \ .$$

The notation ERM stresses that we use an algorithm (the minimization of the empirical risk) which takes in input a model and some data and outputs an estimator $\widehat{s}_m$.

Let $\mathcal{I} = \{I(m) \mid m \in \mathcal{M}_n\}$ denote the set of all the partitions $I(m)$ of $[0,1]$, built from the subdivision $t_0 = 0 < t_1 < \ldots < t_n = 1$. For each $I(m) \in \mathcal{I}$, we define the model S_m of piecewise constant functions built from $I(m) = (I_\lambda)_{\lambda \in \Lambda(m)}$, where $\Lambda(m)$ denotes the set of indices associated with m. Set D_m the dimension of the model S_m, $\mathcal{D} = \{D_m \mid m \in \mathcal{M}_n\}$ and $\mathcal{M}(D) = \{m \mid D_m = D\}$.

As for now, we are looking for an algorithm A: $P_n \mapsto A(P_n) = \widehat{m}(P_n)$ such that $\widetilde{s}(P_n) = \mathrm{ERM}\left(S_{\widehat{m}(P_n)}, P_n\right)$ has a prediction error satisfying an oracle inequality

$$\ell\,(s, \widetilde{s}(P_n)\,) \leq C \inf_{m \in \mathcal{M}_n}\left\{\ell\,(s, \widehat{s}_m(P_n)\,) + R(n,m)\right\} \tag{7.5}$$

with high probability and C close to 1 if possible, $R(n,m)$ being a remainder term.

Remark: We could have considered a model different from (7.4), called "random-design":

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i, \qquad (\epsilon_i)_i \ \textit{i.i.d.}\ , \mathbb{E}\left[\,\epsilon_i \mid X_i\,\right] = 0,\ \mathrm{Var}\left[\,\epsilon_i \mid X_i\,\right] = 1, \tag{7.6}$$

where $\sigma: \ [0,1] \to \mathbb{R}_+$and $\{(X_i, Y_i)\}_i$ *i.i.d.* .
Several remarks could be made about this model.

- The Lpo estimator of Section 7.2.3 remains the same with both models.
- The random design model is no longer that one used in change-points detection where the design is essentially fixed.

- However, both (7.4) and (7.6) are meaningful.
  The random design model corresponds to a prediction purpose: given some *i.i.d.* random observations, what is the value of the signal for a new one drawn from the same distribution?
  The fixed design is rather the expression of a technical reality: the signal is observed at points, which are more or less the consequence of some experimental features (CGH for instance). Thus, the observation instants are likely not random.

Algorithm description

In Section 7.2.1, we have described the classical procedure with rich collections of models. In a first step, models with the same dimension are "aggregated" to get $\left(\widetilde{S}_D\right)_{D\in\mathcal{D}}$. Then for each dimension $D \in \mathcal{D}$, the empirical risk minimization provides $\widehat{m}_D$. The second step relies on another algorithm consisting in the minimization of the penalized empirical contrast $\mathrm{crit}\left(\widehat{m}_D\right)$ over $\mathcal{D}$, which provides us with $\widehat{m}$. Actually, we have an algorithm $\mathcal{A}(\cdot,\cdot)$ with in input the collection $(S_m)_{m\in\mathcal{M}_n}$ and the observations P_n and outputs $\widehat{m}$. It may be formalized as follows.

Algorithm 7.2.1. Algorithm $\mathcal{A}$*:*
Input: $(S_m)_{m\in\mathcal{M}_n}$, P_n
1st Step:

$$\forall D \in \mathcal{D}, \qquad \widehat{m}_D := \mathrm{Argmin}_{m\in\mathcal{M}(D)} \mathrm{crit}_1(S_m, P_n),$$

2nd Step:

$$\widehat{D} := \mathrm{Argmin}_{D\in\mathcal{D}} \mathrm{crit}_2(D, P_n),$$

Output: $\mathcal{A}\left((S_m)_{m\in\mathcal{M}_n}, P_n\right) = \widehat{m}_{\widehat{D}}$,
where crit_1 *and* crit_2 *both denote two given algorithms.*

This procedure may be generalized by replacing the ERM and penalized empirical contrast minimizations by other algorithms involving resampling for instance. This is the purpose of the following description.

First step: Model choice algorithm for each dimension
Let crit_1 be a model choice criterion, that is an algorithm with in input some data (P_n) and a predictor referred to by S_m in the algorithm, which is the least squares estimator associated with model S_m. In output, crit_1 returns a real number $\mathrm{crit}_1(S_m, P_n)$. Ideally, crit_1 measures the prediction error of $\widehat{s}_m(P_n)$. Let us assume that this minimizer exists and is uniquely defined. Then for each dimension $D \in \mathcal{D}$, we set

$$\widehat{m}(D, P_n) := \mathrm{Argmin}_{m\in\mathcal{M}(D)} \left\{\mathrm{crit}_1(S_m, P_n)\right\} \quad . \tag{7.7}$$

There are several possible choices for this criterion:

Id : the ideal criterion (for comparison):

$$\mathrm{crit}_{1,\mathrm{Id}}(S_m, P_n) := P\gamma\left(\mathrm{ERM}(S_m, P_n)\right)$$

It equals the prediction error of the least squares estimator built from model S_m.

Emp: Empirical risk ("resubstitution" error):

$$\mathrm{crit}_{1,\mathrm{Emp}}(S_m, P_n) := P_n\gamma\left(\mathrm{ERM}\left(S_m, P_n\right)\right)$$

Lpo_p: Leave-p-out with parameter p:

$$\mathrm{crit}_{1,\mathrm{Lpo}}(S_m, P_n, p) := \mathbb{E}_e\left[P_n^{(e)}\gamma\left(\mathrm{ERM}\left(S_m, P_n^{(e^c)}\right)\right)\right]$$

where e follows the uniform distribution the subsets of $\{1,\ldots,n\}$ of size p, $P_n^{(e)} = \mathrm{Card}(e)^{-1}\sum_{i\in e}\delta_{(X_i,Y_i)}$ et $P_n^{(e^c)} = (n-\mathrm{Card}(e))^{-1}\sum_{i\notin e}\delta_{(X_i,Y_i)}$.

penRad$_C$: Rademacher resampling penalty with overpenalization coefficient $C \geq 1$. This resampling scheme is defined in Arlot [5]. The weights $W_{n,1}, \ldots, W_{n,n}$ are independent and $pW_{n,i} \sim \mathcal{B}(p)$, $p \in (0,1)$. A classical choice is $p = 1/2$:

$$\mathrm{crit}_{1,\mathrm{penRad}}(S_m, P_n) := P_n\gamma\left(\mathrm{ERM}\left(S_m, P_n\right)\right) + C\mathbb{E}_W\left[\left(P_n - P_n^W\right)\gamma\left(\mathrm{ERM}\left(S_m, P_n^W\right)\right)\right]$$

where $\mathbb{E}_W\left[\,\cdot\,\right] = \mathbb{E}\left[\,\cdot \mid (X_i, Y_i)_{1\leq i\leq n}\,\right]$.

Remark: Note that for each dimension $D \in \mathcal{D}$, the algorithm consisting in the minimization of $\mathrm{crit}_1(S_m, P_n)$ over $\mathcal{M}(D)$ and which returns the associated estimator may be denoted by $\mathcal{A}_D(P_n)$. Thus,

$$\forall D \in \{1, \ldots, n\}, \qquad \mathcal{A}_D(P_n) = \widehat{s}_{\widehat{m}(D,P_n)}(P_n).$$

Second step: Choice of the dimension

Let crit_2 denote a criterion dedicated to the "choice of an algorithm", that is an algorithm taking in input some data (P_n), an algorithm $\mathcal{A}_D$, and outputs a real number $\mathrm{crit}_2(\mathcal{A}_D, P_n)$. Ideally, crit_2 measures the prediction error of the estimator associated with $\mathcal{A}_D(P_n)$.

Set $\mathcal{A}_D := \widehat{s}_{\widehat{m}(D,\cdot)}(\cdot)$ for any $D \in \mathcal{D}$ (*i.e.* the algorithm chosen at the first step), we then define

$$\widehat{D}\left((\mathcal{A}_D)_{D\in\mathcal{D}}, P_n\right) := \mathrm{Argmin}_{D\in\mathcal{D}}\left\{\mathrm{crit}_2(\mathcal{A}_D, P_n)\right\} \quad . \tag{7.8}$$

Several choices for this algorithm may be considered as well:

Id : For comparison, the ideal criterion

$$\mathrm{crit}_{2,\mathrm{Id}}(\mathcal{A}_D, P_n) := P\gamma\left(\mathcal{A}_D(P_n)\right)$$

Same : We keep the value of $\mathrm{crit}_{1,\cdot}$ at the first step, without paying attention to the complexity of $\mathcal{M}_n$.

$$\mathrm{crit}_{2,\mathrm{Same}}(P_n) := \mathrm{crit}_{1,\cdot}(\widehat{m}_D) = \min_{\mathcal{M}(D)} \mathrm{crit}_{1,\cdot}(m)$$

VF$_V$: "V-fold" CV risk estimator with V subsets:

$$\mathrm{crit}_{2,\mathrm{VF}}(\mathcal{A}_D, P_n, V) := \frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\gamma\left(\mathcal{A}_D\left(P_n^{(-j)}\right)\right)$$

where B_j denotes the $j-$th element of the partition of the data in V subsets of approximately equal size n/V, $P_n^{(j)} = \mathrm{Card}(B_j)^{-1}\sum_{i\in B_j}\delta_{(X_i,Y_i)}$ et $P_n^{(-j)} = (n - \mathrm{Card}(B_j))^{-1}\sum_{i\notin B_j}\delta_{(X_i,Y_i)}$.

BM: Penalty derived from a theoretical framework relying on the homoscedasticity assumption, and justified by Birgé and Massart [16, 17, 30]:

$$\mathrm{crit}_{2,\mathrm{BM}}(\mathcal{A}_D, P_n) := P_n\gamma\left(\mathcal{A}_D\left(P_n\right)\right) + \mathrm{pen}_{\mathrm{BM}}(\mathcal{A}_D)$$

where $\mathrm{pen}_{\mathrm{BM}}(\mathcal{A}_D) := \widehat{C}(D/n(5 + 2\log(n/D))$ and $\widehat{C}$ denotes a positive constant determined by slope heuristics.

C_p : Mallows' penalized criterion, derived from a homoscedastic and asymptotic framework by Mallows [33] :

$$\mathrm{crit}_{2,C_p}(\mathcal{A}_D, P_n) := P_n\gamma\left(\mathcal{A}_D\left(P_n\right)\right) + 2\widehat{\sigma}^2\frac{D}{n},$$

where $\widehat{\sigma}^2$ is an estimator of the variance.

This second step is therefore devoted to the estimation of the prediction error of $\mathrm{ERM}(S_{\widehat{m}(D,P_n)}, P_n) = \widehat{s}_{\widehat{m}(D,P_n)}(P_n)$ for every D.

Remark: $\mathrm{crit}_{2,\mathrm{Id}}$ is completely different from $\mathrm{crit}_{1,\mathrm{Id}}$ in that it applies to a family of estimators predetermined by the $(\mathcal{A}_D)_{D\in\mathcal{D}}$ algorithms, whereas $\mathrm{crit}_{1,\mathrm{Id}}$ is minimized over the whole collection. In particular, $\mathrm{crit}_{2,\mathrm{Id}}$ can only lead to the choice of the best possible model among $(\widehat{m}(D))_{D\in\mathcal{D}}$, which is potentially very different from the best over $\mathcal{M}_n$.

Global algorithm

Finally, the whole algorithm is a function $\mathcal{A}(\cdot,\cdot)$, which takes in input a family of models $(S_m)_{m\in\mathcal{M}_n}$ and a set of observations (P_n), and outputs

$$\widetilde{s} = \mathcal{A}\left((S_m)_{m\in\mathcal{M}_n}, P_n\right) := \mathrm{ERM}(S_{\widehat{m}_{\widehat{D}}(\widehat{D}((\mathcal{A}_D)_{D\in\mathcal{D}},P_n),P_n)}, P_n).$$

Then, $\mathcal{A}$ may be seen as the superimposition of two algorithms (that of the first step and that of the second one), chosen among the above propositions. For this reason and for the sake of simplicity in the sequel, we denote the $\mathcal{A}$ algorithm as "$1x2y$" where x refers to the first step algorithm and y, to that of the second one. For instance, the classical procedure of Birgé and Massart [16] with rich collections will be denoted by 1Emp2BM.

7.2.3 Non-asymptotic results

In this section, we aim at deriving a closed-form formula for the Lpo risk estimator, which is a necessary requirement to apply our strategy. Moreover in our setting, the richness of the collection of models prevents us from computing the risk estimator associated with each model. A widespread effective algorithm to circumvent this problem is the dynamic programming [14]. Thus, deriving a closed-form expression enables us to use dynamic programming.
Analogous results for resampling penalties such as Rademacher penalties may be found in Arlot [5].

Closed-form formulas

In the present work, we use piecewise constant functions on $[0,1]$ in order to approximate the regression function s.

Let $I(m) = (I_1,\ldots,I_{D_m})$ be a partition of $[0,1]$ in D_m intervals indexed by m and let S_m denote the vector space of all piecewise constant functions on $I(m)$. Then, the regressogram associated with S_m is

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda}, \quad \text{where} \quad \widehat{\beta}_\lambda = \frac{\sum_{j=1}^n Y_j \mathbb{1}_{I_\lambda}(X_j)}{\sum_{k=1}^n \mathbb{1}_{I_\lambda}(X_k)} \cdot$$

Note that this estimator is uniquely defined if and only if there is at least one observation in each interval I_λ of the partition. In the sequel, we only consider models for which this requirement holds.

By inverting the sums, we can rewrite $\widehat{s}_m$ so that

$$\forall 1\leq i\leq n, \quad \widehat{s}_m(X_i) = \frac{1}{n}\sum_{j=1}^n \left(\sum_{\lambda\in\Lambda(m)} \frac{H_m^\lambda(Z_j,Z_i)}{G_m^\lambda(Z_{1,n},Z_i)} \right),$$

where $H_m^\lambda(Z_j,Z_i) = Y_j\mathbb{1}_{I_\lambda}(X_j)\mathbb{1}_{I_\lambda}(X_i)$ and $G_m^\lambda(Z_{1,n,Z_i}) = 1/n\sum_{j=1}^n \mathbb{1}_{I_\lambda}(Z_j)$.
Remark: An important issue is to make sure that these estimators are uniquely defined. Indeed, the denominator could vanish if there is no observation in at least one interval of the partition for the regressogram. In the sequel, we assume that there is at least one in each interval of the considered partitions.

Once these estimators are well defined, another issue occurs when applying the Lpo. Since this resampling scheme consists in removing p observations from the n original ones, if p is larger than the number of points in a given interval and since the resampling is made exhaustively, some intervals are emptied. In the sequel, we describe a possible solution to this problem.
First notice that when working with regressograms,

$$\forall 1\leq i\leq n, \quad \widehat{s}_m(X_i) = \frac{1}{n}\sum_{j=1}^n \frac{H_m^{\lambda_i}(Z_j,Z_i)}{G_m^{\lambda_i}(Z_{1,n},Z_i)},$$

where λ_i denotes the index λ such that $X_i \in I_\lambda$ ($I = (I_1, \ldots, I_n)$ is a partition of $[0,1]$). Since the Lpo estimator may be written as

$$\widehat{R}_p(m) = \frac{1}{p} \left[\mathrm{Card}\,(\mathcal{E}_p) \right]^{-1} \sum_{i=1}^{n} \left[\sum_{e \in \mathcal{E}_p} \mathbb{1}_{(i \in e)} \left(Y_i - \widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i) \right)^2 \right],$$

the question amounts to warranty that $\widehat{s}_m(Z_{1,n}^{\overline{e}})(X_i)$ is well defined for each point X_i. Our idea is to replace $\mathcal{E}_p$ by a subset denoted by $\mathcal{E}'_p(i)$ *depending on* i, such that $G_m^{\lambda_i}(Z_{1,n}^{\overline{e}}, Z_i)$ is non null for every $e \in \mathcal{E}'_p(i)$. In other words for each i, we are approximating the global expectation with respect to the resampling

$$\mathbb{E}_W \left[\gamma(\widehat{s}_m(P_n^{\overline{W}}), Z_i) \right]$$

by a conditional version of it

$$\mathbb{E}_W \left[\gamma(\widehat{s}_m(P_n^{\overline{W}}), Z_i) \mid G_m^{\lambda_i}(P_n^{\overline{W}}, Z_i) \neq 0 \right].$$

Thus, all the forthcoming closed-form expressions follow this convention.
We now provide the result for the regressogram. For the sake of simplicity, we denote the current estimator by $\widehat{s}_m(X_i) = 1/n \sum_{j=1}^{n} H_m(Z_j, Z_i)/G_m(Z_{1,n}, Z_i)$.

Proposition 7.2.1. *With the above notations, let $\widehat{s}_m$ denote the regressogram in S_m built from the partition $\{I_\lambda\}$. Then, we have*

$$H_m(Z_j, Z_i) = Y_j \mathbb{1}_{I_{\lambda_i}}(X_j) \mathbb{1}_{I_{\lambda_i}}(X_i) \quad \text{and} \quad G_m(Z_{1,n}, Z_i) = 1/n \sum_{k=1}^{n} \mathbb{1}_{I_{\lambda_i}}(X_k) \mathbb{1}_{I_{\lambda_i}}(X_i).$$

Moreover for any $1 \leq p \leq n-1$, the Lpo estimator is

$$\widehat{R}_p(m) = \frac{1}{p} \sum_{i=1}^{n} \frac{1}{N_i} \left[\mathbb{1}_{(n_i=1)} \{+\infty\} + \mathbb{1}_{(n_i \geq 2)} \left\{ Y_i^2 A_i - 2 \sum_{j \neq i} Y_i H_m(Z_j, Z_i) B_i + \sum_{j \neq i} \left(H_m(Z_j, Z_i)\right)^2 C_i \right\} + \right.$$
$$\left. \mathbb{1}_{(n_i \geq 3)} \sum_{j \neq i} \sum_{k \neq j, k \neq i} H_m(Z_j, Z_i) H_m(Z_k, Z_i) D_i \right],$$

where $n_i = n\, G_m(Z_{1,n}, Z_i)$ and

$$\begin{aligned}
N_i &= 1 - \mathbb{1}_{(p \geq n_i)} \binom{n - n_i}{p - n_i} \Big/ \binom{n}{p}, \\
A_i &= V_i(0) - \frac{V_i(1)}{n_i} \quad \text{and} \quad B_i = \frac{A_i}{n_i - 1}, \\
C_i &= \frac{V_i(-1)}{n_i - 1} - \frac{V_i(0)}{n_i(n_i - 1)}, \\
D_i &= \frac{(n_i + 1)V_i(0) - V_i(1) - n_i V_i(-1)}{n_i(n_i - 1)(n_i - 2)}.
\end{aligned}$$

For any $k \in \{-1, 0, 1\}$,

$$V_i(k) = \sum_{r = 1 \vee (p - n_i)}^{n_i \wedge (n-p)} r^k \frac{\binom{n-p}{r}\binom{p}{n_i - r}}{\binom{n}{n_i}}.$$

The proof has already been given in Section 3.5.
Remark: With regressograms, N_i, A_i, B_i, C_i, D_i and V_i only depends on i through the interval X_i belongs to. These quantities may therefore be calculated beforehand, whereas it is no longer the case with the kernel estimator.

These expressions may be further developed, which yields the following corollary:

Corollary 7.2.1. *With the same notations as in Proposition 7.2.1, we get*

$$\widehat{R}_p(m) = \sum_{\lambda\in\Lambda(m)} \frac{1}{pN_\lambda}\left[\mathbb{1}_{(n_\lambda=1)}\left\{+\infty\right\} + \mathbb{1}_{(n_\lambda\geq 2)}\left\{S_{\lambda,2}\left(A_\lambda + C_\lambda(n_\lambda-1)\right) - 2B_\lambda\left(S_{\lambda,1}^2 - S_{\lambda,2}\right)\right\} + \right.$$
$$\left.\mathbb{1}_{(n_\lambda\geq 3)}\left\{D_\lambda(n_\lambda-2)\left[\,S_{\lambda,1}^2 - S_{\lambda,2}\,\right]\right\}\right],$$

for regressograms, with $N_\lambda, A_\lambda, \ldots, D_\lambda$ being the same as in Proposition 7.2.1 when $X_i \in I_\lambda$. $S_{\lambda,1} = \sum_{j=1}^n Y_j \mathbb{1}_{I_\lambda}(X_j)$ and $S_{\lambda,2} = \sum_{j=1}^n Y_j^2 \mathbb{1}_{I_\lambda}(X_j)$.

We now provide expectations of the Lpo risk estimator for the regressograms. We point out that these expectations are actually conditional expectations given the design points. Indeed, we are not able to derive closed-form expressions in the random design setting. The best we could do in such a case is an approximation of these expectations as in the recent work of Arlot [6]. The proof of this proposition is given in Section 3.5.

Proposition 7.2.2. *In the fixed-design setting, let assume we observe n random variables $Y_i = s(X_i)+\sigma_i\epsilon_i$, where $(\epsilon_i)_i$ are i.i.d. centered random variables such that $\mathbb{E}\epsilon_i^2 = 1$. With the notations of Corollary 7.2.1, the Lpo risk expectation for regressograms equals*

$$\begin{aligned}\mathbb{E}\left[\,\widehat{R}_p(m)\,\right] \quad = \quad & \sum_{\lambda\in\Lambda(m)} \frac{1}{pN_\lambda}\left[\mathbb{1}_{(n_\lambda=1)}\left\{+\infty\right\} + \mathbb{1}_{(n_\lambda\geq 2)}\left\{\left[\,F_{\lambda,2} + n_\lambda\left(\sigma_\lambda^r\right)^2\right]\left(A_\lambda + C_\lambda(n_\lambda-1)\right) - \right.\right. \\ & \left.\left. 2B_\lambda\left(F_{\lambda,1}^2 - F_{\lambda,2}\right)\right\} + \mathbb{1}_{(n_\lambda\geq 3)}\left\{D_\lambda(n_\lambda-2)\left[\,F_{\lambda,1}^2 - F_{\lambda,2}\,\right]\right\}\right],\end{aligned}$$

*where $F_{\lambda,2} = \sum_{i=1}^n s^2(X_i)\mathbb{1}_{I_\lambda}(X_i)$, $F_{\lambda,1} = \sum_{i=1}^n s(X_i)\mathbb{1}_{I_\lambda}(X_i)$ and $\left(\sigma_\lambda^r\right)^2 = \sum_{i=1}^n \sigma_i^2 \mathbb{1}_{I_\lambda}(X_i)/n_\lambda$.
In the same way for kernels, we get*

$$\begin{aligned}\mathbb{E}\left[\,\widehat{R}_p(m)\,\right] \quad = \quad & \sum_{i=1}^n \frac{\mathbb{1}_{(n_i\geq 2)}}{pN_i}\left\{\left(s^2(X_i)+\sigma_i^2\right)A_i - 2s(X_i)\sum_{j\neq i} s(X_j)K_m(X_j-X_i)B_i + \right. \\ & \sum_{j\neq i}\left(s^2(X_j)+\sigma_j^2\right)K_m^2(X_j-X_i)C_i + \\ & \left.\mathbb{1}_{(n_i\geq 3)}\left\{\sum_{j\neq i}\sum_{k\neq i, k\neq j} s(X_j)s(X_k)K_m(X_j-X_i)K_m(X_k-X_i)\right\}\right\} + \sum_{i=1}^n \mathbb{1}_{(n_i=1)}\left\{+\infty\right\}.\end{aligned}$$

Overpenalization

In model selection, our goal is to find the "best" model among $(S_m)_m$, in terms of a given criterion from a family of estimators $(\widehat{s}_m)_m$. A very common way to reach this goal is the minimization of a penalized criterion $\mathrm{crit}(\cdot)$ defined as

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m),$$

where γ is the contrast function that measures the quality of an estimator and $\mathrm{pen}(\cdot):\ \mathcal{M}_n \to \mathbb{R}_+$ denotes the penalty term, which takes into account the complexity of the model m.
Ideally, the optimal criterion we would like to minimize over $\mathcal{M}_n$ is the random quantity

$$\mathrm{crit}_{id}(m) = P\gamma(\widehat{s}_m) := \mathbb{E}\gamma\left(\widehat{s}_m, Z\right),$$

with the expectation taken with respect to $Z \sim P$. The link between the two criteria can be made through rewriting the latter as follows

$$\mathrm{crit}_{id}(m) = P_n\gamma(\widehat{s}_m) \;+\; \left[\,P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m)\,\right].$$

The quantity in square brackets is named the ideal penalty

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{pen}_{id}(m) := P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m).$$

Following the CV strategy, we perform model selection by minimizing the Lpo risk estimator over $\mathcal{M}_n$, provided $\mathcal{M}_n$ is not too large. Thus for a given $1 \leq p \leq n-1$, the candidate $\widehat{m}$ is

$$\widehat{m} = \operatorname{Argmin}_{m \in \mathcal{M}_n} \widehat{R}_p(m).$$

The idea that there is a strong relationship between penalized criteria and CV is strongly supported by the large amount of literature about the comparison of these two aspects [45, 31, 49]. Therefore, we may try to include the CV strategy into the wider scope of penalized criteria minimization. Thus,

$$\widehat{m} = \operatorname{Argmin}_{m \in \mathcal{M}_n} P_n\gamma(\widehat{s}_m) \, + \, \left[\widehat{R}_p(m) - P_n\gamma(\widehat{s}_m) \right].$$

In the above expression, the quantity in square brackets is called the Lpo penalty:

$$\forall m \in \mathcal{M}_n, \qquad \operatorname{pen}_p(m) := \widehat{R}_p(m) - P_n\gamma(\widehat{s}_m).$$

This Lpo penalty may be subsequently understood as a random penalty. Note that a similar approach applied to the Loo can be found in Birgé and Massart [15].
Thanks to this parallel between CV and penalized criteria, we attempt to get more insight in the behaviour of CV techniques, for instance with respect to the parameter p.

In this section, we aim at making comparison between pen_{id} and pen_p, so that we would like to characterize some features in the behaviour of pen_p wit respect to p. This comparison is carried out through the expectations of these criteria, which are both random variables.

Calculations involving the regressogram are by far more intricate and come from Proposition 7.2.2 and the following lemmas.

Lemma 7.2.1. *With the same notations as Proposition 7.2.2, we have*

$$\mathbb{E}\left[\operatorname{pen}_{id}(m)\right] = \frac{2}{n}\sum_{\lambda\in\Lambda(m)} \left(\sigma_\lambda^r\right)^2 \qquad \text{and} \qquad \mathbb{E}\left[P_n\gamma\left(\widehat{s}_m\right)\right] = \sum_\lambda \frac{n_\lambda - 1}{n}\left[\left(\sigma_\lambda^r\right)^2 + \frac{n_\lambda}{n_\lambda - 1}\left(\sigma_\lambda^b\right)^2\right],$$

$$\mathbb{E}S_{\lambda,2} = n_\lambda\left[\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2 + \beta_\lambda^2\right] \qquad \text{and} \qquad \mathbb{E}S_{\lambda,1}^2 = n_\lambda\left(\sigma_\lambda^r\right)^2 + n_\lambda^2\beta_\lambda^2\,,$$

where $\left(\sigma_\lambda^r\right)^2 := 1/n_\lambda \sum_{i=1}^n \sigma_i^2 \mathbb{1}_{I_\lambda}(X_i)$ *and* $\left(\sigma_\lambda^b\right)^2 := 1/n_\lambda \sum_{i=1}^n \left[s(X_i) - \beta_\lambda\right]^2 \mathbb{1}_{I_\lambda}(X_i)$.

We also need some further details about coefficients $V_\lambda(k)$, which is given by the following result.

Lemma 7.2.2. *For any $\lambda \in \Lambda(m)$, let X denote a random variable following a hypergeometric distribution $X \sim \mathcal{H}(n_\lambda, n-p, n)$. Then,*

$$\begin{aligned}
V_\lambda(0) &= \mathbb{P}\left[X \in \{1 \vee p - n_\lambda, \ldots, n_\lambda \wedge n - p\}\right] = 1 - \mathbb{1}_{(p \geq n_\lambda)}\binom{p}{n_\lambda}\Big/\binom{n}{n_\lambda}, \\
V_\lambda(1) &= \mathbb{E}\left[X\right] = \frac{n_\lambda(n-p)}{n} \qquad \text{and} \qquad V_\lambda(-1) = \mathbb{E}\left[X^{-1}\mathbb{1}_{(X>0)}\right], \\
V_\lambda(1)V_\lambda(-1) &\geq 1.
\end{aligned}$$

with $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

Due to the high complexity of expressions in hand, the following result is not given in full generality, but we make a few assumptions in order to enlighten the underlying phenomenon as clearly as possible.

Proposition 7.2.3. *With the notations of Proposition 7.2.2, let assume that for any $\lambda \in \Lambda(m)$, $n_\lambda \geq 3$. Then if $p < n_\lambda$, we have*

$$\begin{aligned}
\mathbb{E}\left[\operatorname{pen}_p(m)\right] &= \sum_{\lambda\in\Lambda(m)}\left\{\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2\frac{n_\lambda}{n_\lambda - 1}\right\}\left(\frac{p-n}{np} + \frac{n}{p(n-p)}V_\lambda(1)V_\lambda(-1)\right), \\
&\geq \frac{n - p/2}{n-p}\,\mathbb{E}\left[\operatorname{pen}_{id}(m)\right].
\end{aligned}$$

Moreover, assume that $p \geq \left(n_\lambda \vee \left[\sqrt{a_\lambda} - 1\right] n/\sqrt{a_\lambda}\right)$ *for any* $\lambda \in \Lambda(m)$*, where* $a_\lambda = n_\lambda / V_\lambda(1) V_\lambda(-1)$. *Then,*

$$\begin{aligned}
\mathbb{E}\left[\,\mathrm{pen}_p(m)\,\right] &= \sum_{\lambda \in \Lambda(m)} \left\{ \left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2 \frac{n_\lambda}{n_\lambda - 1} \right\} \left[(n_\lambda - 1) \frac{n-p}{np} + \frac{1}{N_\lambda} \left(\frac{n V_\lambda(1) V_\lambda(-1)}{p(n-p)} - \frac{n_\lambda (n-p)}{np} \right) \right], \\
&\geq \frac{n - p/2}{n-p}\, \mathbb{E}\left[\,\mathrm{pen}_{id}(m)\,\right].
\end{aligned}$$

The proof has been given in Section 3.5.
Thanks to the above proposition, we have characterized the behaviour of the Lpo penalty as an overpenalizing criterion both for small and large values of p. But as previously described (Section 2.2.1), this overpenalization does not necessary turn out to be a drawback of the approach. Indeed whereas a biased criterion may be misleading when trying to accurately estimate the risk of an estimator, this bias may prevent some well known troubles in the model selection framework.

7.2.4 Overview of simulations

The following sections are devoted to some simulation studies intending to assess the performance of a large number of candidate algorithms in different respects. Here, we provide an overview of the purposes of these studies.

From the description of the classical model selection procedure in Section 7.2.1 and from the two-steps algorithm of Section 7.2.2, we see that the candidate algorithms at step 2 aim at taking into account the complexity of models $\widetilde{S}_D$ for each dimension D. The main goal of the first simulation study is the assessment of these algorithms as complexity measures.
To this end, we use 1Emp (*i.e.* the empirical risk minimization at the first step) to provide the family of models $\{\widehat{m}(D)\}_D$ to which the candidate algorithms will be applied.
Finally, we compare 1Emp2VF (*i.e.* VF at step 2) with $V = 5$, 1Emp2BM and 1Emp2Id. 1Emp2C_p has also been included in the comparison in order to see to what extent a bad measure of the complexity can be misleading.
Each of these algorithms provides us with an estimation of $\ell\,(s, \widehat{s}_{\widehat{m}}(D)\,)$ for each dimension D and a candidate model for which the the true loss may be computed. Thus, we have two criteria to measure the performance: the loss of the final model provided by each algorithm and the curve of the estimated risk versus the dimension. Note that an ideal complexity measure would provide a curve which could be superimposed on the curve $D \mapsto \ell\,(s, \widehat{s}_{\widehat{m}}(D)\,)$.

In the second simulation study, the goal is the assessment of the quality of candidate algorithms at the first step. We point out that these algorithms yield a family $\{\widehat{m}(D)\}_{D \in \mathcal{D}}$. In the change-points detection setup, it amounts to a collection of partitions of $[0,1]$ in $D \in \{1, \ldots, n\}$ intervals. We then use the ideal algorithm 2Id at the second step in order to compute the loss of each associated estimator. Therefore, we have two criteria at our disposal to assess the performance of each algorithm. The loss of the best model among $\{\widehat{m}(D)\}_D$ for each algorithm should enable to distinguish between algorithms, as well as the curves $D \mapsto \ell\,(s, \widehat{s}_{\widehat{m}}(D)\,)$. For the latter criterion, an ideal algorithm would provide the uniformly lowest curve.
The comparison is carried out between 1Emp2Id, 1Lpo$_p$2Id with different values of p, and 1penRad$_C$2Id with several values of the overpenalization constant C.

From the two first simulation experiments, we should have an idea about the behaviour of the different algorithms. However in the second study, 2Id has been used in order to focus on the performance of the first step algorithms. The third simulation experiment is performed with 2VF$_5$ instead of 2Id to check whether there is any interaction between the algorithm at the first step and that of the second one. Subsequently, the experimental design is exactly the same as that of the previous study.

In each simulation study, the comparison is made in both homoscedastic and heteroscedastic settings, with several types of noises. This answers to the questions of Section 7.2.1. Thus, the first experiment shows

that 1Emp2BM does not work anymore in the heteroscedastic setting. Furthermore, it highlights that 1Emp2VF works almost as well as 1Emp2BM in the homoscedastic case, by far better in the heteroscedastic case.
The two other simulation studies are devoted to compare 1Emp and resampling strategies 1Lpo and 1penRad, in order to check whether it is worth using resampling at the first step. For instance, does resampling yield more relevant breakpoints for a given dimension?

7.3 Resampling to take into account collection complexity

7.3.1 Theoretical considerations

Complexity measure Applying a penalized criterion like that of Birgé and Massart [16, 17] (*i.e.* 1Emp2BM) to a rich collection of models amounts to use 1Emp with models $\widetilde{S}_D$ and then, to penalize with 2BM, in order to take into account the complexity of these new models.
As suggested in Section 7.2.2, we may still aggregate models by dimension, but use other algorithms (1Emp2*) to take into account the complexity, where * refers to one of the candidate algorithms: VF_V, Lpo_p, penVF_V or penRad_C. For instance, $1Emp2\mathrm{VF}_V$ means that we estimate the prediction error the least squares estimator over each of the models $\widetilde{S}_D$ using V-fold. $\widehat{D}$ is then defined as the minimizer of this criterion over $\mathcal{D}$.
In what follows, we assess the behaviour of algorithms at step 2 as complexity measures.

General remarks On the one hand, 2BM has been justified in a homoscedastic framework and the associated penalty has required an intensive simulation step [30] in order to calibrate the constants. Its behavior with heteroscedastic data has not been studied yet.
On the other hand, 2VF relies upon the general cross-validation heuristics. It is known to be naturally robust to heteroscedasticity in some regression framework [8].

Two behaviours are therefore expected. In the homoscedastic framework, we hope that resampling algorithms will provide similar results to that obtained with 2BM, which has been optimized for this framework. On the contrary, 2BM should be a quite poor criterion in the heteroscedastic setup. Such a phenomenon has already been observed by Arlot [6] in the case of penalties linear in the dimension, when the collection of models is not too rich.

We also consider 1Emp2C_p so as to highlight the extent to which underpenalization may be misleading. Indeed, applying 2C_p amounts to choose

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n\gamma(\widehat{s}_m) + \frac{2\hat{\sigma}^2 D_m}{n} \right\} ,$$

that is to consider that $\widetilde{S}_D = \cup_{m\in\mathcal{M}(D)} S_m$ has the same complexity as any S_m with $D_m = D$.

7.3.2 Simulations

Simulation design

In this simulations experiment, the comparison is carried out between 1Emp2Id, 1Emp2VF_5, 1Emp2BM and 1Emp2C_p. We use 5 different regression functions which have been plotted in Figure 7.1. All of them but s_5 are piecewise constant: s_1 and s_3 both belong to five dimensional vector spaces, whereas s_2 and s_4 are of dimension 10. s_5 is the Heavisine function, very popular in signal processing literature [34].
The deterministic design points are taken equal to $t_i = i/n$ for every $i \in \{1, \ldots, n\}$. Of course, the fewer points we have in the vicinity of a change-point the less ability we have to detect it.
In these simulations, we only consider gaussian noise.
Both the homoscedastic and the heteroscedastic settings are explored. In the latter case, σ is either piecewise constant or sinusoidal.

- With $s_1, \ldots, s_4$:

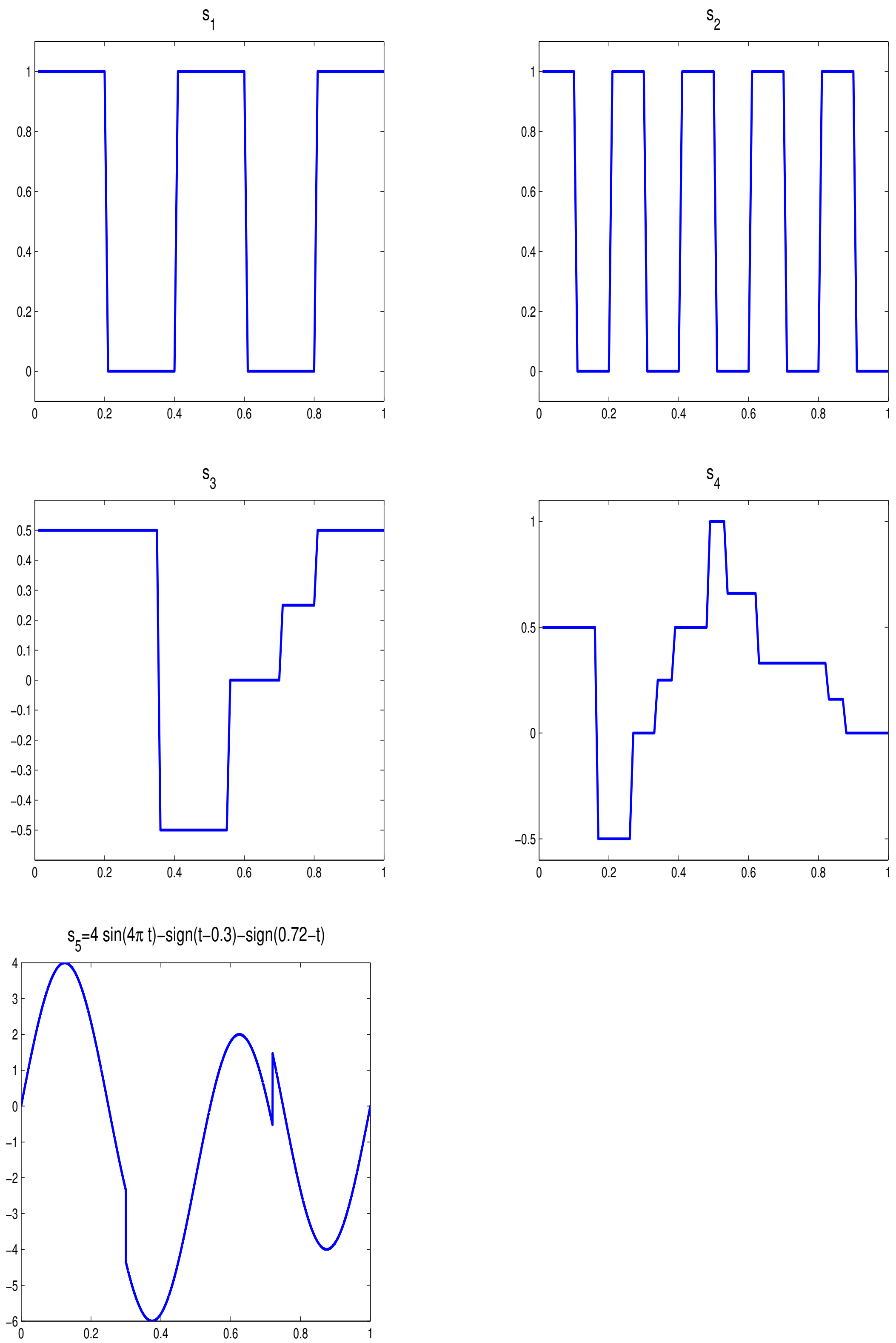

Figure 7.1: Plots of regression functions s_1, s_2, s_3, s_4, and s_5.

- Homoscedastic (**c** for constant noise level): $\sigma_{c,1} = 0.1\mathbb{1}_{[0,1]}$, $\sigma_{c,2} = 0.25\,\mathbb{1}_{[0,1]}$ and $\sigma_{c,3} = 0.5\,\mathbb{1}_{[0,1]}$.
- Heteroscedastic:
  1. (**pc** for piecewise constant noise level) $\sigma_{pc,4} = 0.5\,\mathbb{1}_{[0,1/2]}$, $\sigma_{pc,5} = 0.1\left(4\,\mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}\right)$,
  2. (**s** for sinusoidal noise level) $\sigma_{s,6} : \ t \mapsto 0.3\sin(t\pi/4)$, $\sigma_{s,7} : \ t \mapsto 0.5\sin(t\pi/4)$ and $\sigma_{s,8} : t \mapsto 0.8\sin(t\pi/4)$.

- With s_5:
  - Homoscedastic: $\sigma_{c,1} = 3\mathbb{1}_{[0,1]}$, $\sigma_{c,2} = 5\,\mathbb{1}_{[0,1]}$ and $\sigma_{c,3} = 7\,\mathbb{1}_{[0,1]}$.
  - Heteroscedastic:
    1. $\sigma_{pc,4} = 6\,\mathbb{1}_{[0,1/2]}$, $\sigma_{pc,5} = \left(6\,\mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}\right)$,
    2. $\sigma_{s,6} : \ t \mapsto 6\sin(t\pi/4)$, $\sigma_{s,7} : \ t \mapsto 8\sin(t\pi/4)$ and $\sigma_{s,8} : \ t \mapsto 10\sin(t\pi/4)$.

In each condition, we generate $n = 100$ observations and $N = 300$ trials are drawn.
All the models we consider have a dimension $D \in \{1, \ldots, D_{\max}\}$, where $D_{\max} = \lfloor n/2 \rfloor$.

Remark: We stress that a necessary and sufficient condition for the regressogram estimator to be uniquely defined is that at least one design point lies in each interval. Therefore, we only consider models based on partitions with at least one design point in each of their intervals.

Results

All the results of the above simulation design are given in Section 7.8. In this section, we only give a few results, intended to be representative of the whole simulation study.

Let us focus on Figure 7.2, where the graphs of $D \mapsto \mathrm{crit}_{2,A}$ are plotted, with A denoting one of the candidate algorithms for the second step; the first step is always 1Emp in this section.
The 1Emp2Id curve represents the loss of models $\{\widehat{m}(D)\}_D$, yielded by 1Emp. Its minimum location corresponds to the "oracle" among models $\left(S_{\widehat{m}(D)}\right)_D$ (named *relative oracle*). We notice that 1Emp provides only one or at most a few reliable models, in the neighbourhood of the relative oracle.

The curve of 1Emp2Id may be understood as the ideal complexity measure of models $\mathcal{S}_{\widehat{m}(D)}$ and subsequently taken as our gold-standard. Thus, we may call "good complexity measure" any algorithm, the curve of which may be almost superimposed on that of 1Emp2Id. Note that it includes the case where the curve associated with a given algorithm is the vertical translation of that of 1Emp2Id.
We observe that in the homoscedastic setting, the three curves are almost alike, except some small fluctuations of 2VF from one sample to another (left panel of Figures 7.2 and 7.3).
On the contrary in the heteroscedastic framework, 2BM leads to some strong overfitting, whereas 2VF provides nearly the same curve as 2Id on average, up to a vertical translation (right panel of Figures 7.2 and 7.3).
However, this "deficiency" of 2BM in the heteroscedastic case may be less strong as in Figure 7.4 but anyway, 1Emp2BM and 1Emp2Id curves cannot be superimposed anymore.

2BM seems to perform as well as 2VF, as a measure of complexity in the homoscedastic framework. However in the heteroscedastic setup, 2BM performance may be much worse than that of 2VF. We deduce that 2BM is actually very dependent on the homoscedastic assumption it is derived from, unlike 2VF which remains reliable.
Furthermore in the heteroscedastic case, 2BM is not always that bad (see Figure 7.4), even if it does no longer follow the behaviour of 1Emp2Id. Heteroscedasticity influences the effective complexity quantification so that $D/n\log(D/n)$ is no longer an appropriate complexity measure. However, the purpose of algorithms at step 2 is to find the dimension, that is the minimum location. Thus, 2BM is perhaps not a reliable complexity measure anymore in the heteroscedastic framework but it may sometimes succeed in finding the dimension of the relative oracle.

Let us now focus on the top panel of Table 7.1, which displays the average ratios of the loss of the model selected by a given algorithm over that of the oracle $\ell\left(s, \widehat{s}_{m^\star}\right)$. For a given condition, bold quantities refers to the minimal value among the candidate algorithms, while underlined ones indicate that this value

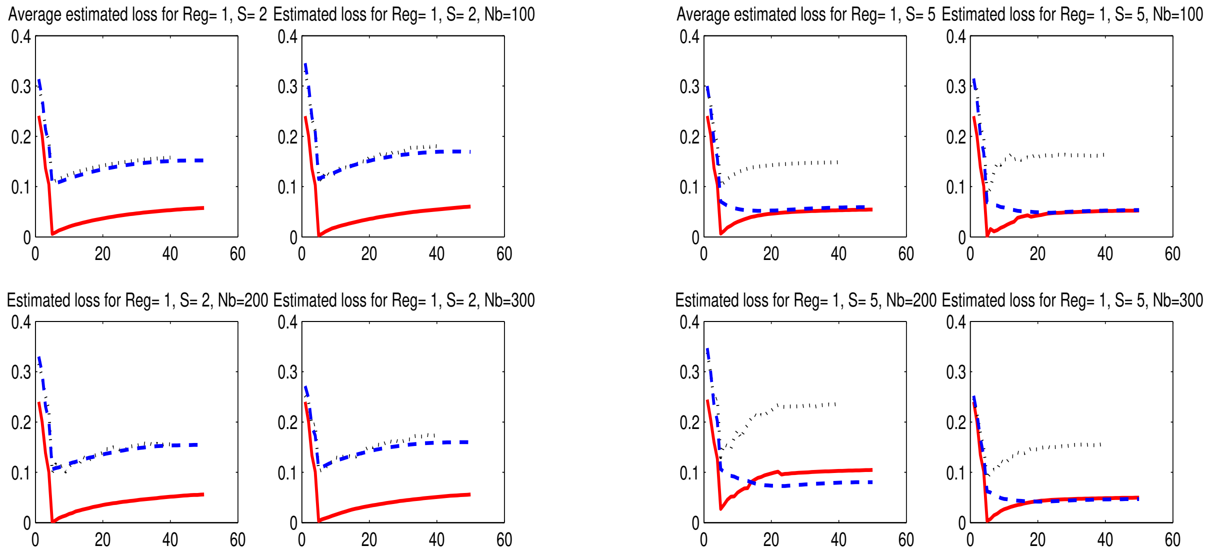

Figure 7.2: **Left panel:** Graph of $D \mapsto \mathrm{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_1 and $\sigma_{c,2}$ (homoscedastic): 1Emp2Id ('-' plain line), 1Emp2BM ('- -'dashed line) and 1Emp2VF$_5$ (':' dotted line). The top-left graph depicts the average loss over the trials, while the remaining graphs display results for three particular trials. **Right panel:** Graph of $D \mapsto \mathrm{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_1 and $\sigma_{pc,5}$ (heteroscedastic): 1Emp2Id ('-' plain line), 1Emp2BM ('- -'dashed line) and 1Emp2VF$_5$ (':' dotted line). The top-left graph depicts the average loss over the trials, while the remaining graphs display results for three particular trials.

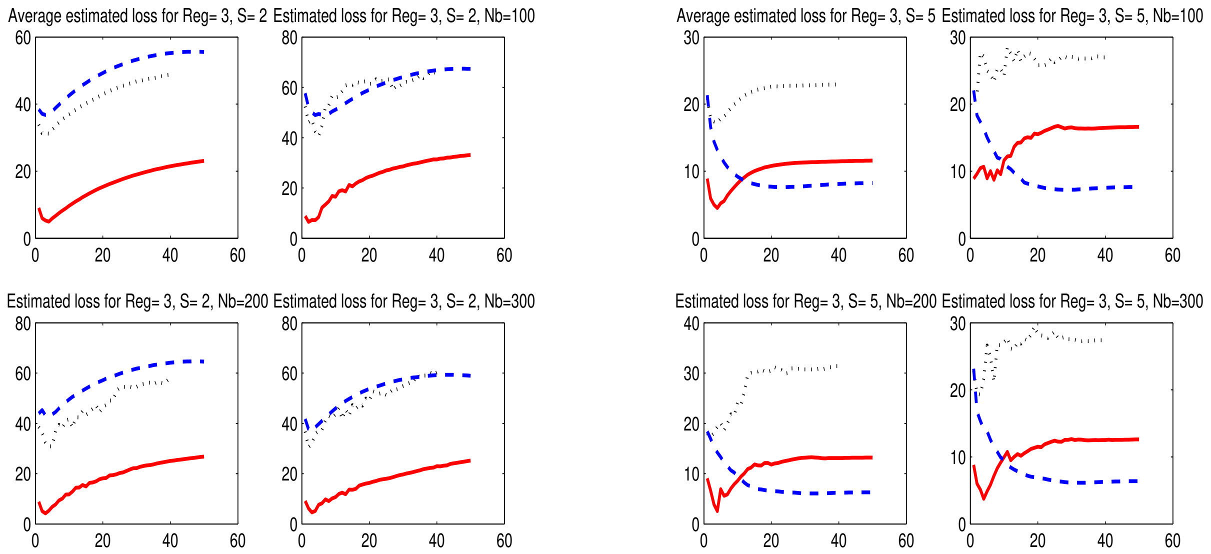

Figure 7.3: **Left panel:** Graph of $D \mapsto \mathrm{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_5 and $\sigma_{c,2}$ (homoscedastic): 1Emp2Id ('-' plain line), 1Em2BM ('- -'dashed line) and 1Emp2VF$_5$ (':' dotted line). The top-left graph depicts the average loss over the trials, while the remaining graphs display results for three particular trials. **Right panel:** Graph of $D \mapsto \mathrm{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_5 and $\sigma_{pc,5}$ (heteroscedastic): 1Emp2Id ('-' plain line), 1Em2BM ('- -'dashed line) and 1Emp2VF$_5$ (':' dotted line). The top-left graph depicts the average loss over the trials, while the remaining graphs display results for three particular trials.

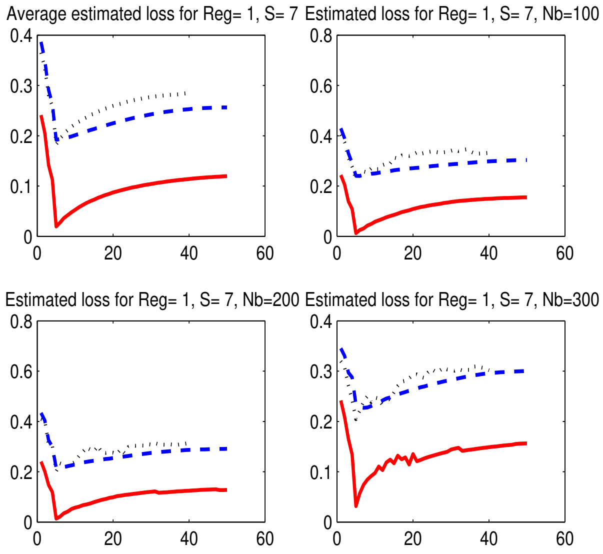

Figure 7.4: Graph of $D \mapsto \mathrm{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_5 and $\sigma_{s,7}$ (heteroscedastic): 1Emp2Id ('-' plain line), 1Em2BM ('- -'dashed line) and 1Emp2VF$_5$ (':' dotted line). The top-left graph depicts the average loss over the trials, while the remaining graphs display results for three particular trials.

is significantly worse than the minimal one. The same results for other functions s and σ are provided in Appendix (Table 7.7).

At first sight, the values of the ratios seem to be quite high, even for 1Emp2Id, which has the best possible performance for an algorithm selecting a model among $(\widehat{m}(D))_{D\in\mathcal{D}})$. This can be interpreted in two ways. Either the noise level is so high that it makes the model selection task too difficult, or 1Emp is not so suited as an algorithm to provide reliable change-points positions, which results in both cases in poor results in comparison with the absolute oracle.

Besides, we notice that in the homoscedastic framework, 2BM generally provides the best results but that very often, 2VF is not significantly worse than 2BM in this setting. On the contrary, 2VF always gets the best results under heteroscedasticity, which are always significantly better than those of 2BM.

A possible objection is that 2BM relies upon a heuristic strategy itself depending on the maximal dimension $D_{\max}$. For instance, a too large $D_{\max}$ may "artificially" induce over- or underpenalization. We also used $D_{\max} = 25$ instead of 50 (which may seem large with respect to $n = 100$). The corresponding results are displayed in the column of 1Emp2BM+. We first observe that these results are all significantly worse than the best one. Moreover if we sometimes obtain better results in the heteroscedastic case (s_2 or s_3 with $\sigma_{pc,4}$ or $\sigma_{pc,5}$, see Section 7.8), the other results especially in the homoscedastic case are suboptimal.

The bottom panel of Table 7.1 displays the average dimension of the model selected by each algorithm. The column of 1Emp2Id should be interpreted as the average dimension of the relative oracle (among models $\{\widehat{m}(D)\}_D$ provided by 1Emp).
First, we notice that $2C_p$ strongly underpenalizes since its penalty is not well suited to the high complexity. Besides as the noise level increases, the dimension of the oracle decreases in general, which means that 1Emp puts more and more meaningless change-points where the noise is strong.

Let us consider 2BM and 2VF now. Once again, two behaviours arise depending on the kind of noise. In the homoscedastic setting, 2BM turns out to slightly overpenalize since the average dimension is smaller than that of the oracle (s_3, s_5). Meanwhile, 2VF selects models with a dimension somewhat larger than that of 2BM on average. Thus whereas this slight overpenalization of 2BM with respect to 2VF turns out to be an asset for s_1 and s_3, it entails a better performance of 2VF with s_5. The model chosen according to 2BM has a too small dimension with respect to that of the oracle, while that furnished by 2VF is closer

s.	σ.	1Emp2Id	1Emp2VF$_5$	1Emp2BM	1Emp2BM+	1Emp2C_p
1	c,2	2.19 ± 0.28	3.35 ± 0.35	**2.96** ± 0.36	7.23 ± 1	62.1 ± 7.1
	pc,5	4.04 ± 0.57	**7.33** ± 0.9	38.5 ± 2.8	65.1 ± 8.8	131 ± 14
5	c,2	3.81 ± 0.077	**5.14** ± 0.13	5.62 ± 0.15	5.69 ± 0.19	11 ± 0.36
	pc,5	5.78 ± 0.19	**8.48** ± 0.29	19.8 ± 0.59	12.8 ± 0.48	16.6 ± 0.55
	s,7	4.6 ± 0.14	**6.17** ± 0.22	7.55 ± 0.29	8.49 ± 0.35	15.5 ± 0.66

s.	σ.	1Emp2Id	1Emp2VF$_5$	1Emp2BM	1Emp2C_p
1	c,2	5.11 ± 0.02	5.51 ± 0.06	5.13 ± 0.02	9.67 ± 0.53
	pc,5	5.1 ± 0.02	5.48 ± 0.08	16.1 ± 0.23	13 ± 0.74
5	c,2	3.66 ± 0.05	3.37 ± 0.09	2.65 ± 0.067	14.8 ± 0.52
	pc,5	4.04 ± 0.06	4.61 ± 0.15	22.6 ± 0.29	22.7 ± 0.78
	s,7	2.66 ± 0.05	2.93 ± 0.09	3.44 ± 0.16	16.6 ± 0.62

Table 7.1: **Top panel:** Average ratios of the loss of the model selected by each algorithm (1Emp2*) over that of the oracle over the while collection of models ± a standard deviation. Bold text denotes the minimum value for each condition, while underlined figures indicate significantly worse values. We say that a value is significantly worse than another one if it is larger than the latter and the discrepancy is larger than the sum of their respective standard deviations. **Bottom panel:** Average selected dimension for each algorithm (1Emp2*) ± one standard deviation. $N = 300$ trials are used.

to the oracle.
This overpenalization tendency of 2BM completely disappears in the heteroscedastic setting where the selected dimension is by far higher than that of the oracle, while that of 2VF remains close to it, no matter of the experimental condition.

7.4 Resampling to choose the best model for each dimension

In this section, our purpose is the study of step 1 algorithms as a means to choose the best possible model for each dimension. Essentially, this comparison study aims at knowing whether or not it is worth using resampling rather than ERM at the first step of the global algorithm. To this aim, we compare the performances of 1Emp2Id with the ones of 1*2Id.
Then, we carry out the same experiments with at the second step 2VF$_5$ instead of 2Id, in order to check whether the whole resulting algorithms actually behave as expected.
We start with some theoretical remarks.

7.4.1 Theoretical considerations

Homoscedastic setting From both Lemma 7.2.1 and Proposition 7.2.3, we deduce the following results for 1Emp and 1Lpo$_p$.
For any $m \in \mathcal{M}_n$, we have

$$\begin{aligned} \mathbb{E}\left[\operatorname{crit}_{1,Emp}(S_m, P_n)\right] &= \ell\left(s, s_m\right) - \sigma^2 \frac{D_m}{n} + \sigma^2, \\ \mathbb{E}\left[\operatorname{crit}_{1,\mathrm{Lpo}_p}(S_m, P_n)\right] &\approx \ell\left(s, s_m\right) + \sigma^2 \frac{D_m}{n-p} + \sigma^2, \end{aligned}$$

where the second approximation holds when $p < n_\lambda$ and with $V_\lambda(1)V_\lambda(-1) \approx 1$, up to the term

$$\frac{2n-p}{n(n-p)} \sum_{\lambda} \frac{n_\lambda}{n_\lambda - 1} \left(\sigma_\lambda^b\right)^2 .$$

This term remains small with respect to $\ell\left(s, s_m\right)$ as long as n_λ is large enough, that is for models with a not too large dimension.

We also recall the expression of the true risk to enable the comparison with the two above criteria:

$$\mathbb{E}\left[\|Y - \widehat{s}_m\|_n^2 \right] = \ell\left(s, s_m\right) + \sigma^2 \frac{D_m}{n} + \sigma^2. \tag{7.9}$$

Thus for a given dimension D, we essentially compare models S_m in terms of their approximation properties ($\ell\left(s, s_m\right)$). Moreover, since the above approximation holds especially for small dimensions, it is likely that no difference arises between 1Emp and Lpo_p for small models in the homoscedastic setting but still appear for larger dimensions. Moreover, (7.9) indicates that the resulting model should be not too far from the optimal.

Remark: Note that we see a major difference between 1Lpo and 1Emp since the expectation of the latter decreases as the dimension grows, which inevitably leads to overfitting, whereas 1Lpo actually performs a tradeoff between approximation and model complexity.

Heteroscedastic setup From both Lemma 7.2.1 and Proposition 7.2.3, we derive the following expressions for 1Emp and $1\mathrm{Lpo}_p$.
For any $m \in \mathcal{M}_n$, we have

$$\begin{aligned}
\mathbb{E}\left[\|Y - \widehat{s}_m\|_n^2 \right] &= \ell\left(s, s_m\right) + \frac{1}{n}\sum_{\lambda} \left(\sigma_\lambda^r\right)^2 + \frac{1}{n}\sum_{i=1}^n \sigma_i^2, \qquad (7.10)\\
\mathbb{E}\left[\mathrm{crit}_{1,Emp}(S_m, P_n) \right] &= \ell\left(s, s_m\right) - \frac{1}{n}\sum_{\lambda} \left(\sigma_\lambda^r\right)^2 + \frac{1}{n}\sum_{i=1}^n \sigma_i^2,\\
\mathbb{E}\left[\mathrm{crit}_{1,\mathrm{Lpo}_p}(S_m, P_n) \right] &\approx \ell\left(s, s_m\right) + \frac{1}{n-p}\sum_{\lambda} \left(\sigma_\lambda^r\right)^2 + \frac{1}{n}\sum_{i=1}^n \sigma_i^2,
\end{aligned}$$

under the same approximations as above.
Unlike the homoscedastic setting, we deduce that all the models with the same dimension are no longer compared with respect to their approximation properties only, but also through the way heteroscedasticity and the partitioning of $[0, 1]$ are intertwined. Thus, if we think about a noise of type $\sigma \mathbb{1}_{[0,1/3]}$ for instance, we see that when the bias term becomes small (which occurs when the main change-points have been detected, but not all of them), 1Emp should put new breakpoints in the noisy region (provided σ is large enough). On the contrary, 1Lpo is supposed to add breakpoints where the noise is the weakest. Moreover, 1Lpo is more likely to provide a model close to the oracle than 1Emp, due to the difference in their expressions.
In conclusion, the families yielded by 1Emp and 1Lpo could be quite different in the heteroscedastic case, while close to one another in the homoscedastic framework for small models.

7.4.2 Simulation setting

In the following simulations, we have the same 5 regression functions $s_1, \ldots, s_5$ as in the previous simulations.
We consider Gaussian residuals as before. Except the homoscedastic setting, we have two types of heteroscedastic noise: piecewise constant and sinusoidal. Note that this simulation design is slightly different from that of Section 7.3.2.
With regression functions $s_1, \ldots, s_4$, we have

- Homoscedastic: $\sigma_{c,1} = 0.1\,\mathbb{1}_{[0,1]}$, $\sigma_{c,2} = 0.25\,\mathbb{1}_{[0,1]}$ and $\sigma_{c,3} = 0.5\,\mathbb{1}_{[0,1]}$,
- Heteroscedastic:
  1. Piecewise constant: $\sigma_{pc,4} = 0.05\left(2\,\mathbb{1}_{[0,1/3]} + 0.5\mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,5} = 0.1\left(2\,\mathbb{1}_{[0,1/3]} + 0.5\mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,6} = 0.25\left(2\,\mathbb{1}_{[0,1/3]} + 0.5\mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,7} = 0.33\left(2\,\mathbb{1}_{[0,1/3]} + 0.5\mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,8} = 0.5\left(2\,\mathbb{1}_{[0,1/3]} + 0.5\mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,9} = 2\,\mathbb{1}_{[0,1/3]} + 0.5\mathbb{1}_{[1/3,1]}$,

2. Sinusoidal: $\sigma_{s,10} : t \mapsto 0.005 \sin(t\pi/4)$, $\sigma_{s,11} : t \mapsto 0.0075 \sin(t\pi/4)$, $\sigma_{s,12} : t \mapsto 0.01 \sin(t\pi/4)$, $\sigma_{s,13} : t \mapsto 0.025 \sin(t\pi/4)$, $\sigma_{s,14} : t \mapsto 0.033 \sin(t\pi/4)$, $\sigma_{s,15} : t \mapsto 0.05 \sin(t\pi/4)$, $\sigma_{s,16} : t \mapsto 0.1 \sin(t\pi/4)$, and $\sigma_{s,17} : t \mapsto 0.2 \sin(t\pi/4)$.

With s_5, we use

- Homoscedastic: $\sigma_{c,1} = 0.1\, \mathbb{1}_{[0,1]}$, $\sigma_{c,2} = 0.25\, \mathbb{1}_{[0,1]}$ and $\sigma_{c,3} = 0.5\, \mathbb{1}_{[0,1]}$,
- Heteroscedastic:

1. Piecewise constant: $\sigma_{pc,4} = 0.05\left(4\, \mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,5} = 0.1\left(4\, \mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,6} = 0.25\left(4\, \mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,7} = 0.33\left(4\, \mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,8} = 0.5\left(4\, \mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}\right)$, $\sigma_{pc,9} = 4\, \mathbb{1}_{[0,1/3]} + \mathbb{1}_{[1/3,1]}$,
2. Sinusoidal: $\sigma_{s,10} : t \mapsto 0.5 \sin(t\pi/4)$, $\sigma_{s,11} : t \mapsto 0.75 \sin(t\pi/4)$, $\sigma_{s,12} : t \mapsto \sin(t\pi/4)$, $\sigma_{s,13} : t \mapsto 1.25 \sin(t\pi/4)$, $\sigma_{s,14} : t \mapsto 2 \sin(t\pi/4)$, $\sigma_{s,15} : t \mapsto 3 \sin(t\pi/4)$, $\sigma_{s,16} : t \mapsto 4 \sin(t\pi/4)$, and $\sigma_{s,17} : t \mapsto 6 \sin(t\pi/4)$.

For each condition, we generate $n = 100$ observations and $N = 300$ trials are drawn. We compare 1Emp, 1Lpo$_p$ with $p = 1, 20$ and 50, and penRad$_C$ with $C = 1, 1.25$ and 1.5.
Remark: Since we use the Lpo strategy, we have to require that each model we consider has at least two points in each interval of the associated partition.

7.4.3 Study of the first step

In this subsection, we focus on the first step of the algorithm, that is the choice of some segmentation given the number of breakpoints. To this aim, we compare the performances of the procedures 1*2Id, even if they cannot be used with real data. The corresponding algorithms 1*2VF are considered in the next subsection.

Homoscedastic case

Let us consider Figure 7.5 where are plotted the average losses of estimators chosen by different algorithms 1*2Id. We stress that displayed graphs only concern s_2 and s_5, but other results, which are very alike, are provided in Appendix.

We distinguish two different behaviors. With piecewise constant functions (such as s_2), the curves are very sharp at the minimum location, which is around the dimension $D = 10$ of the smallest true model. The curves are completely different for s_5 since they are decreasing from 1 to $D_{\max}$. In the latter example, all the curves remain nearly indistinguishable, even when the noise level grows (from $\sigma_{c,2}$ to $\sigma_{c,3}$). Things seem somewhat different with s_2 since we observe some slight difference between 1Emp2Id and resampling curves for middle value dimensions ($D \simeq 30$). Furthermore, this discrepancy grows with the noise level, but only for large enough dimensions.
This difference may be explained by the fact that s_5 does not belong to any model in the family, which entails a larger bias than that of s_2. The decreasing curves therefore indicate that the noise is not large enough to balance the bias term so that the best model is always that with the largest dimension. Since the bias term dominates, there is no real interest in resampling.

We also notice that for large enough dimensions, resampling curves are lower than that of 1Emp2Id, which means that the corresponding partitions are better than those provided by 1Emp. Indeed for a given dimension, the lower the loss, the more accurate the change-points locations. It is a consequence of the overfitting of 1Emp. Indeed as 1Emp only takes into account the fit of the estimator to the data, it suffers some overfitting as the set models grows, contrary to resampling strategies, which also take into account the model complexity.
However, there does seem to be any significant difference between 1Emp and resampling at the minimum location, which means that for this dimension, the selected partitions are (nearly) the same.

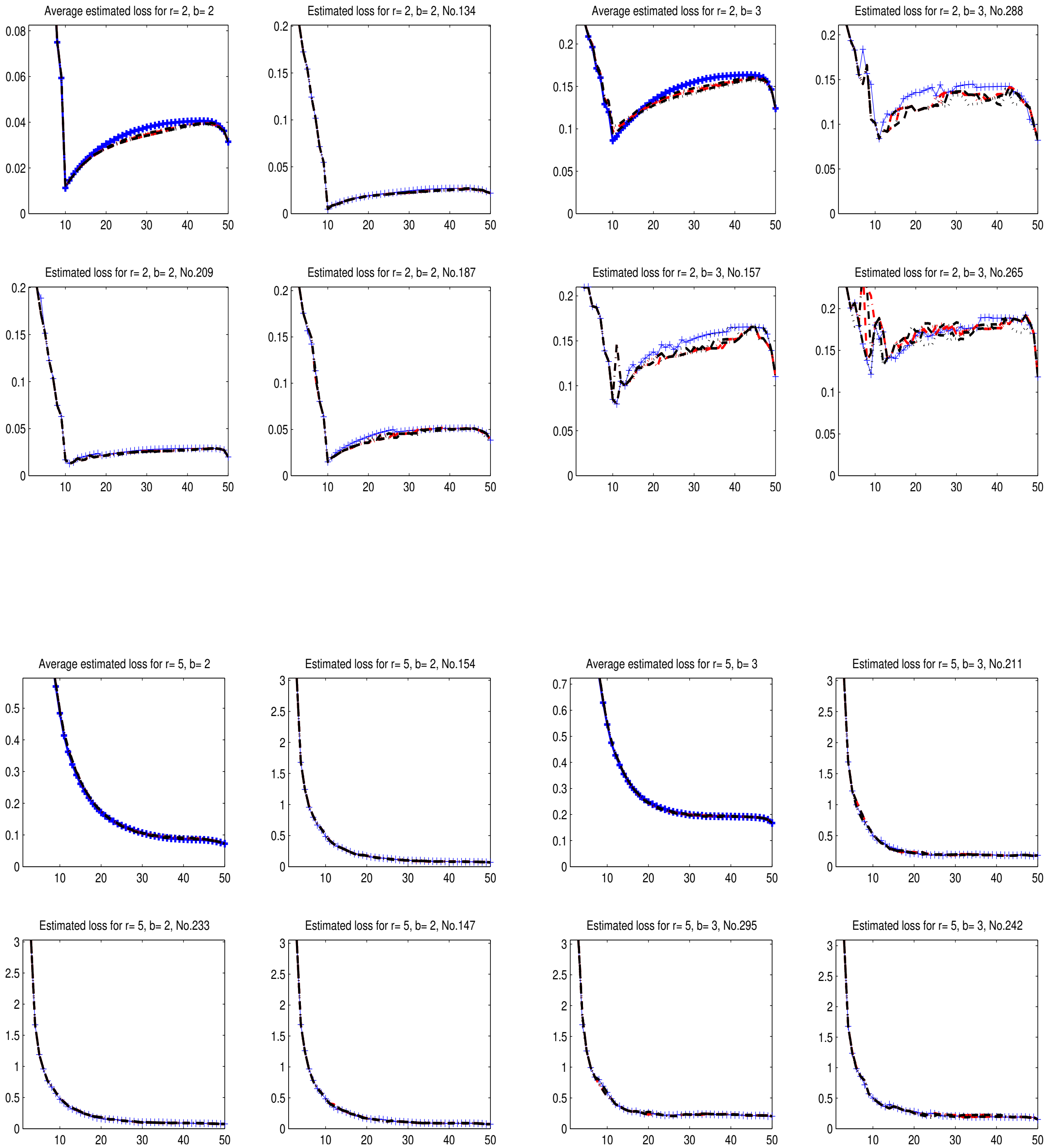

Figure 7.5: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the homoscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* ('+' blue curve), Lpo_p with $p = 1, 20, 50$ ('- -' black dashed curves) and penRad_C with $C = 1, 1.25, 1.5$ ('-.' red curves). **Top panel:** s_2 with σ_2 (left panel) and σ_3 (right panel). **Bottom panel:** s_5 with σ_2 (left panel) and σ_3 (right panel)

$s_{\cdot}$	$\sigma_{c,\cdot}$	1Emp	1Loo	1Lpo$_{20}$	1Lpo$_{50}$	1penRad$_1$	1penRad$_{1.25}$	1penRad$_{1.5}$
2	2	**1.88** $\pm$ 0.088	1.9 $\pm$ 0.091	1.89 $\pm$ 0.091	1.93 $\pm$ 0.091	1.9 $\pm$ 0.091	1.91 $\pm$ 0.09	1.92 $\pm$ 0.091
	3	**3.65** $\pm$ 0.11	3.77 $\pm$ 0.12	3.82 $\pm$ 0.12	$\underline{4.07}$ $\pm$ 0.13	3.86 $\pm$ 0.13	$\underline{3.92}$ $\pm$ 0.13	$\underline{4}$ $\pm$ 0.13
5	2	**1.75** $\pm$ 0.007	**1.75** $\pm$ 0.007	**1.75** $\pm$ 0.007	**1.75** $\pm$ 0.007	**1.75** $\pm$ 0.007	**1.75** $\pm$ 0.007	**1.75** $\pm$ 0.007
	3	**1.99** $\pm$ 0.015	**1.99** $\pm$ 0.015	**1.99** $\pm$ 0.015	**1.99** $\pm$ 0.015	**1.99** $\pm$ 0.015	**1.99** $\pm$ 0.015	**1.99** $\pm$ 0.015

Table 7.2: Average loss of algorithms 1*2Id in the homoscedastic framework for the 5 regression functions, with $\sigma_{c,1}$, $\sigma_{c,2}$ and $\sigma_{c,3}$. Bold text corresponds to minimum value in the given condition, while underlined figures indicate significantly worse results. Results for other

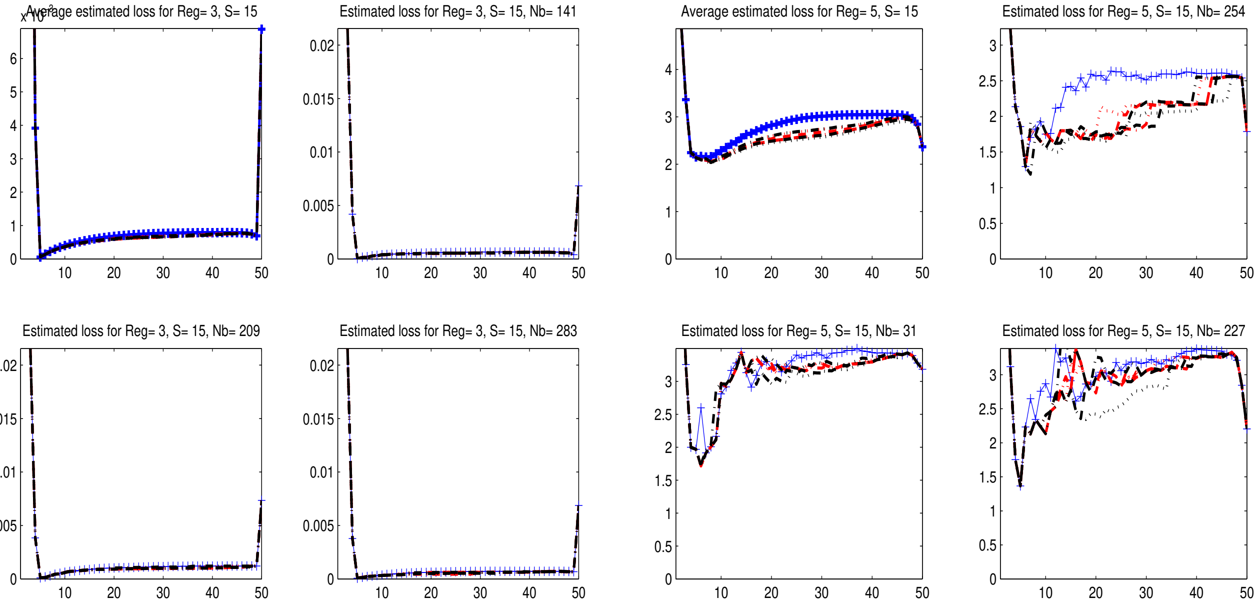

Figure 7.6: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the homoscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* ('+' blue curve), Lpo$_p$ with $p = 1, 20, 50$ ('- -' black dashed curves) and penRad$_C$ with $C = 1, 1.25, 1.5$ ('-.' red curves). **Left panel:** s_3 with $\sigma_{s,15}$. **Right panel:** s_5 with $\sigma_{s,15}$.

These remarks are confirmed by Table 7.2 where are displayed the average loss values for the model chosen by each algorithm 1*2Id. We see that even if 1Emp provides very often the minimum value among all the candidate algorithms, there is nearly no significant gap between the competing methods.

Heteroscedastic case

Sinusoidal σ Let us first focus on Figure 7.6, which displays the loss curves $D \mapsto \ell\left(s, \widehat{s}\,(A, D)\,\right)$, where A is one of the competing algorithms. This figure also shows a representative part of the results. All the curves may be found in Appendix. According to this figure, we have two different situations. With the piecewise-constant regression functions, the noise is not strong enough to enable to distinguish between the algorithms. The resulting curves are very alike and the best partition is the same for all the algorithms.
Remark: Note that we strongly zoom in on one of the left panel graphs in Figure 7.6, we get Figure 7.7. Actually, the curves are not so alike as it could have appeared. We especially see that 1Emp provides the higher curve and that all resampling methods do better. Nevertheless, the conclusion does not change since the model corresponding to the minimum location of the curves remains the same.

Besides, the case s_5 contrasts with the previous conclusion since we observe that the curves are well separated. Moreover from the average curves, we notice that for small dimensions (D is about 10), the resampling curves are below that of 1Emp. This discrepancy increases up to $D = 30$ and then decreases. If we look at the results obtained for three trials, we see that this gap between the curves may be quite high, with very different minimum locations and values. It means that in this setting, resampling provides much more reliable change-points and does not suffer the same overfitting problem as 1Emp when the number of models with the same dimension is large.

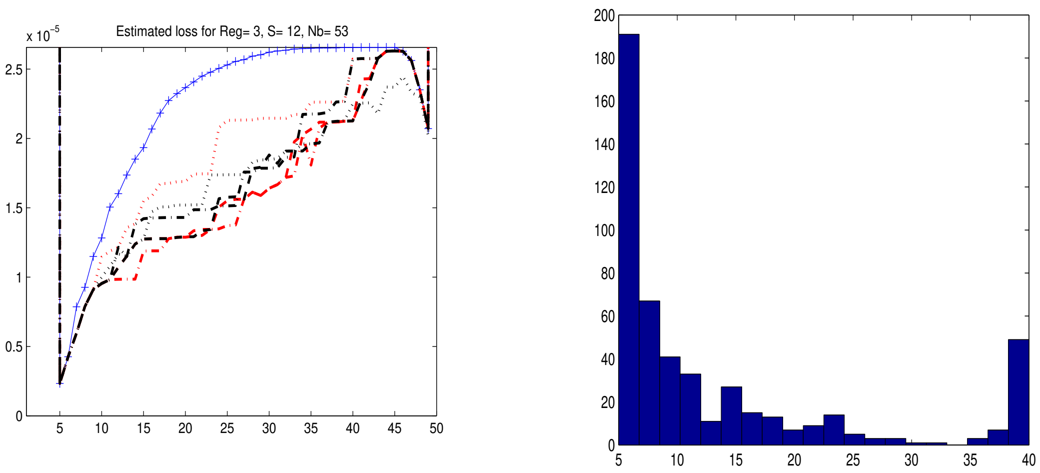

Figure 7.7: **Left panel:** Zoom in the graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ for s_3 and $\sigma_{s,12}$, where $(\widehat{m}(D))_D$ is chosen according to *Emp* ("+" blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). **Right panel:** Histogram of the average dimension of the model selected by $1\text{Emp}2\text{VF}_5$ with s_3 and $\sigma_{s,12}$. $N = 500$ trials have been made.

$s_\cdot$	$\sigma_{s,\cdot}$	1Emp	1Loo	1Lpo_{20}	1Lpo_{50}	1penRad_1	$1\text{penRad}_{1.25}$	$1\text{penRad}_{1.5}$
3	15	**1** ± 5e-017	**1** ± 5e-017	**1** ± 5e-017	**1** ± 5e-017	**1** ± 5e-017	**1** ± 5e-017	**1** ± 5e-017
	17	3.03 ± 0.2	2.8 ± 0.19	2.77 ± 0.18	**2.71** ± 0.18	2.78 ± 0.19	2.75 ± 0.18	2.72 ± 0.18
5	13	3.65 ± 0.056	3.46 ± 0.053	3.47 ± 0.053	3.48 ± 0.055	3.51 ± 0.053	3.47 ± 0.053	**3.45** ± 0.053
	15	4.65 ± 0.09	4.37 ± 0.087	4.38 ± 0.087	4.37 ± 0.087	4.42 ± 0.088	4.38 ± 0.088	**4.33** ± 0.087

Table 7.3: Ratio of the average loss obtained for model provided by 1*2Id over that of the pathwise oracle (minimizer of $\ell\left(s, \widehat{s}_m\right)$) for each regression function and each noise. $N = 300$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

If we consider Table 7.3, where is given the loss of the best model for each algorithm, we see that there is nearly no significant difference between the candidate algorithms for piecewise constant regression functions. This confirms the situation described in the left panel of Figure 7.6.
With s_5, the results are quite different: 1Emp is nearly always significantly worse than $1\text{penRad}_{1.5}$ and even than all the resampling algorithms. Thus, except when the noise level is very low or too strong (see Table 7.9 in the Appendix), resampling yields much more reliable change-points.

Piecewise-constant σ Figures 7.8 and 7.9 depicts the same curves as above, with respectively s_4 and s_5, which are representative of the main aspects. All the results may be found in Appendix. In each graph, the noise level increases from $\sigma_{pc,5}$ to $\sigma_{pc,7}$ and up to $\sigma_{pc,9}$. Thus, we may observe the behaviour of the algorithms as the noise intensity grows.
With s_3, we see that the optimal dimension decreases from $D = 10$ to 1, which indicates that the highest noise level is too strong to distinguish true change-points from the noise.

As the noise level grows, we notice the increase of the gap between the curves of 1Emp and those of resampling procedures. It is even clearer when we look at the sampled trials, where the interest of resampling may be very high from a set of observations to another. Note that resampling curves are nearly always below that of 1Emp, indicating a real interest of resampling.

If we now focus on Figure 7.9, we see although the strong bias modifies the general shape of the curve (with respect to the previous case), we observe a similar phenomenon. Indeed, the gap between 1Emp and resampling increases with the noise level. Resampling curves are below that of 1Emp and from time to time (bottom panel trial number 191), the quality of model provided by resampling is by far higher than yielded by 1Emp.

The relative performance of 1Emp and the resampling procedures differ from one experiment to another

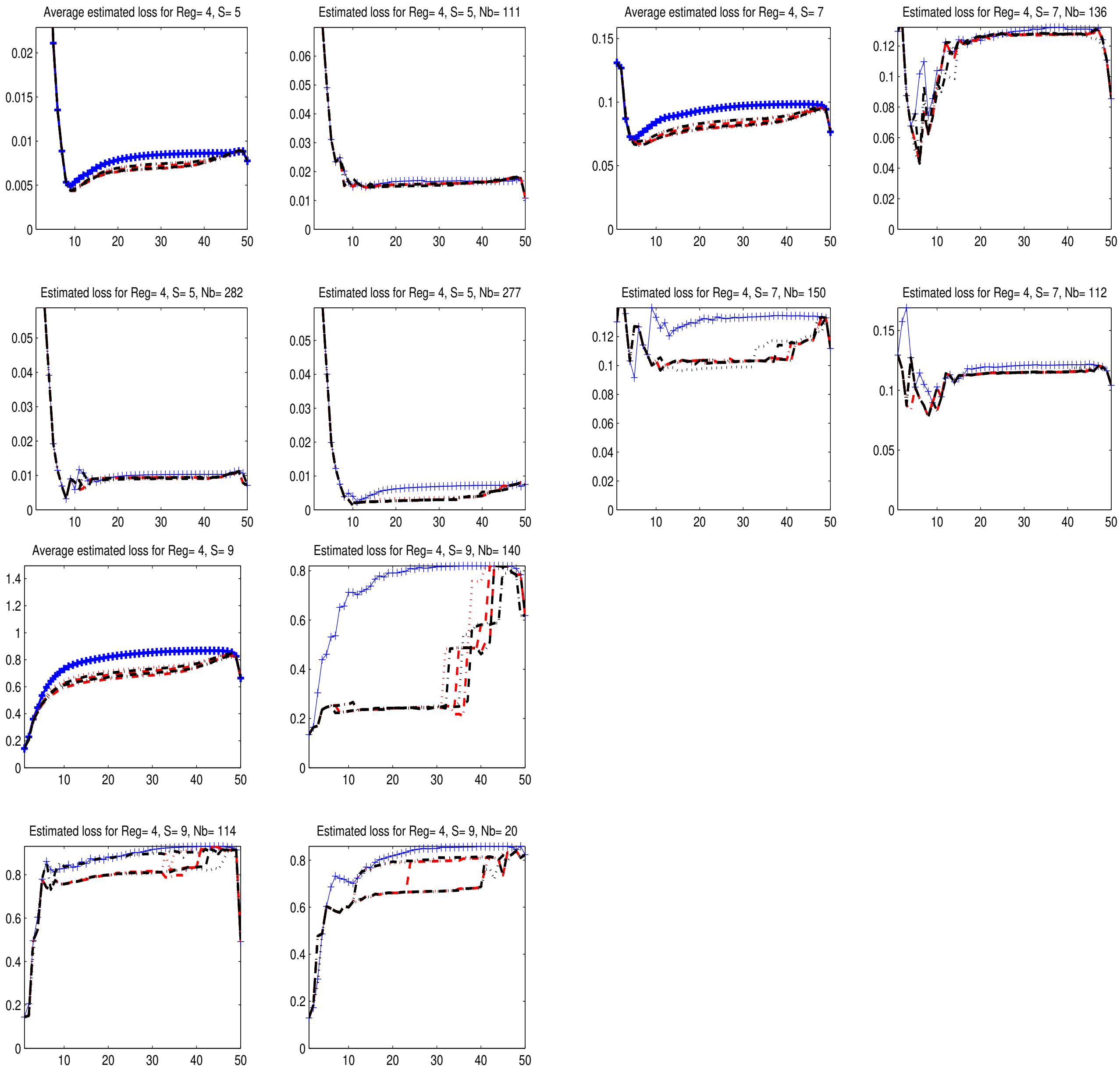

Figure 7.8: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the homoscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* ('+' blue curve), Lpo_p with $p = 1, 20, 50$ ('- -' black dashed curves) and penRad_C with $C = 1, 1.25, 1.5$ ('-.' red curves). s_3 with $\sigma_{pc,5}$, $\sigma_{pc,7}$ and $\sigma_{pc,9}$.

$s_{\cdot}$	$\sigma_{pc,\cdot}$	1Emp	1Loo	$1\mathrm{Lpo}_{20}$	$1\mathrm{Lpo}_{50}$	$1\mathrm{penRad}_1$	$1\mathrm{penRad}_{1.25}$	$1\mathrm{penRad}_{1.5}$
4	5	4.38 ± 0.23	3.68 ± 0.22	3.69 ± 0.22	3.68 ± 0.22	3.75 ± 0.22	3.67 ± 0.22	**3.61** ± 0.22
	7	6.91 ± 0.34	6.22 ± 0.31	**6.19** ± 0.3	6.44 ± 0.31	6.44 ± 0.32	6.34 ± 0.31	6.26 ± 0.3
	9	3.81 ± 0.17	**3.76** ± 0.17	3.78 ± 0.17	3.79 ± 0.17	3.8 ± 0.17	3.8 ± 0.17	3.77 ± 0.17
5	5	1.77 ± 0.012	1.77 ± 0.012	1.77 ± 0.012	**1.77** ± 0.012	1.77 ± 0.012	1.77 ± 0.012	1.77 ± 0.012
	7	2.86 ± 0.05	**2.67** ± 0.045	2.7 ± 0.045	2.74 ± 0.048	2.74 ± 0.047	2.72 ± 0.047	2.71 ± 0.046
	9	4.64 ± 0.092	4.33 ± 0.093	4.36 ± 0.09	4.37 ± 0.09	4.43 ± 0.091	4.38 ± 0.092	**4.29** ± 0.091

Table 7.4: Ratio of the average loss obtained for model provided by 1*2Id over that of the pathwise oracle (minimizer of $\ell\left(s, \widehat{s}_m\right)$) for each regression function and each noise. $N = 300$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

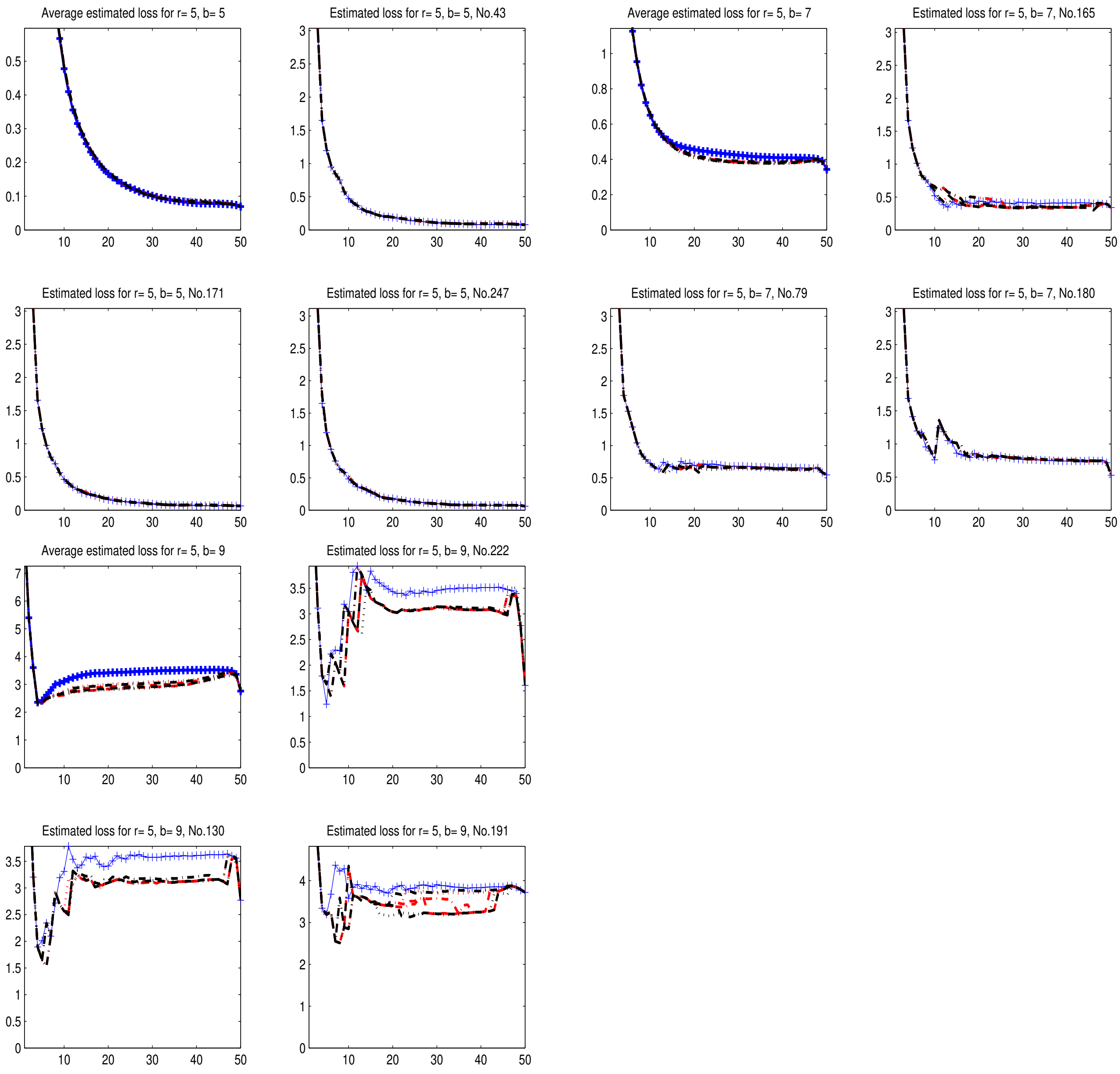

Figure 7.9: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the homoscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* ('+' blue curve), Lpo_p with $p = 1, 20, 50$ ('- -' black dashed curves) and penRad_C with $C = 1, 1.25, 1.5$ ('-.' red curves). s_5 with $\sigma_{pc,5}$, $\sigma_{pc,7}$ and $\sigma_{pc,9}$.

$s_\cdot$	$\sigma_{c,\cdot}$	1Emp	1Loo	1Lpo_{20}	1Lpo_{50}	1penRad_1	$1\text{penRad}_{1.25}$	$1\text{penRad}_{1.5}$
2	2	**2.72** $\pm$ 0.11	2.75 $\pm$ 0.11	2.77 $\pm$ 0.12	2.79 $\pm$ 0.12	2.86 $\pm$ 0.12	2.91 $\pm$ 0.12	2.87 $\pm$ 0.12
	3	**5.54** $\pm$ 0.16	$\underline{5.87}$ $\pm$ 0.17	$\underline{6.06}$ $\pm$ 0.17	$\underline{7}$ $\pm$ 0.2	$\underline{6}$ $\pm$ 0.17	$\underline{6.14}$ $\pm$ 0.17	$\underline{6.31}$ $\pm$ 0.18
5	2	**1.69** $\pm$ 0.0063	$\underline{1.86}$ $\pm$ 0.0082	$\underline{1.76}$ $\pm$ 0.0073	$\underline{1.7}$ $\pm$ 0.0067	$\underline{1.73}$ $\pm$ 0.0067	$\underline{1.74}$ $\pm$ 0.0068	$\underline{1.75}$ $\pm$ 0.0069
	3	2.21 $\pm$ 0.014	$\underline{2.23}$ $\pm$ 0.013	2.21 $\pm$ 0.014	**2.2** $\pm$ 0.014	2.2 $\pm$ 0.014	2.2 $\pm$ 0.014	2.21 $\pm$ 0.014

Table 7.5: Ratio of the average loss obtained for model provided by 1^*2VF_5 over that of the pathwise oracle (minimizer of $\ell(s, \widehat{s}_m))$ for each regression function and each noise. $N = 500$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

$s_\cdot$	$\sigma_{s,\cdot}$	1Emp	1Loo	1Lpo_{20}	1Lpo_{50}	1penRad_1	$1\text{penRad}_{1.25}$	$1\text{penRad}_{1.5}$
3	15	5.77 $\pm$ 0.56	5.89 $\pm$ 0.45	6.01 $\pm$ 0.55	**5.23** $\pm$ 0.4	5.73 $\pm$ 0.47	5.61 $\pm$ 0.46	5.69 $\pm$ 0.44
	17	5.4 $\pm$ 0.35	5.52 $\pm$ 0.35	5.37 $\pm$ 0.33	**5.09** $\pm$ 0.29	5.21 $\pm$ 0.32	5.13 $\pm$ 0.32	5.45 $\pm$ 0.37
5	13	$\underline{4.37}$ $\pm$ 0.049	**3.98** $\pm$ 0.045	4 $\pm$ 0.045	$\underline{4.1}$ $\pm$ 0.046	$\underline{4.09}$ $\pm$ 0.045	4.05 $\pm$ 0.046	4.03 $\pm$ 0.046
	15	$\underline{6.68}$ $\pm$ 0.13	**6.32** $\pm$ 0.13	6.38 $\pm$ 0.13	6.48 $\pm$ 0.13	6.39 $\pm$ 0.13	6.36 $\pm$ 0.13	6.34 $\pm$ 0.13

Table 7.6: Ratio of the average loss obtained for model provided by 1^*2VF_5 over that of the pathwise oracle (minimizer of $\ell(s, \widehat{s}_m))$ for each regression function and each noise. $N = 500$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

(see Tables 7.8 and 7.9 in the Appendix). On the one hand, when the model selection problem is easy (for instance with s_1 and s_2), there are no significant discrepancies between 1Emp and resampling. On the other hand, with s_3 and s_4, some breakpoints can be detected but this is not an easy task. Then, better performance of resampling schemes arises, as shown in Table 7.4 (and Tables 7.8 and 7.9 in the Appendix). With s_4, $1\text{penRad}_{1.5}$, $\text{Lpo}_2 0$ and Loo alternatively significantly outperform over 1Emp. As for s_5, the Loo yields the best performance, but other resampling methods are also significantly better than 1Emp.

7.4.4 Performances of 1*2VF

Homoscedastic setting

Table 7.5 displays the model selection performance of each algorithm in several experimental conditions (see also Table 7.10 in the Appendix). With 2VF_5, the losses are larger than those obtained in the same experimental conditions with 2Id (Table 7.2). This may be explained by the following ideas.
We recall that the use of 2Id in previous study means that from a family of models $\{\widehat{m}(D)\}_D$ yielded by a given algorithm, we use the true distribution to choose which one is the best. In this way, we focus on the ability of the algorithm at the first step to choose a model in $\widetilde{S}_D$ for each D, but we do not take into account the collection complexity (which stems from the comparison between dimensions). This complexity issue arises when we use 2VF_5 yet.
From the analysis of the homoscedastic framework, we know [16, 17, 30] that the price to pay for this complexity is a $\log n$ term between the risk of an estimator and that of the oracle. Moreover, this term strength grows with the amount of noise. So we may conjecture that in our case, we have also to pay something like a $\log n$ term for this complexity. Since $n = 100$, $\log(n) \simeq 4.61$ here.

As for piecewise-constant regression functions, we see that there is nearly no significant gap between the algorithms, as in Table 7.2. Nevertheless, s_5 induces a different behaviour since we now observe an almost systematic significant gap in favour of 1Emp. If we zoom in on the bottom-left panel of Figure 7.5, we get Figure 7.10, which depicts the zoom on the average loss curves of s_5 in the neighbourhood of the minimum location. As we can see close to $D = 50$, the curve of 1Emp is lower than that of any resampling algorithm. Thus when VF_5 wrongly selects a model with $D < 50$, the resulting partition is better than those provided by resampling algorithms.

Heteroscedastic setting

Let us now consider Table 7.6 (and Table 7.11 in the Appendix). We notice some differences with the previous table (see Table 7.3, and Tables 7.8 and 7.9 in the Appendix). Indeed, whereas nearly no significant gap appear between 1Emp and resampling methods in the latter tables, we observe more and more significant gaps in favour of resampling strategies as the problem becomes harder, from s_1 to s_4. If

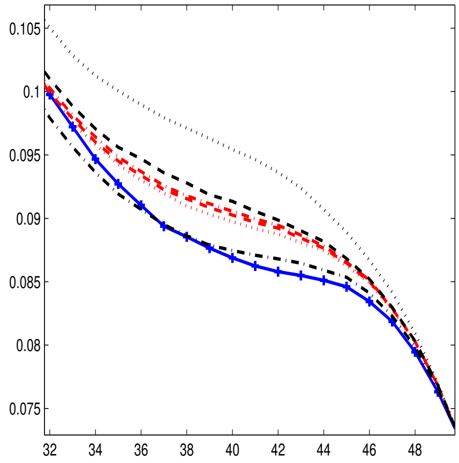

Figure 7.10: Graph of the average loss curves of s_5 with $\sigma_{c,2}$ in the neighbourhood of the minimum location.

we take another look at Figure 7.7, which displays the the loss curves for s_3 and $\sigma_{s,12}$, we see that although the minimum location models turns out to be the same, there is a large discrepancy between the partitions selected by 1Emp and those resulting from resampling in the neighbourhood of the best dimension. Thus since $2\mathrm{VF}_5$ commits some mistakes, it does not choose the best dimension, but a dimension more or less close to it as we can see on the histogram of Figure 7.7, which displays the dimension of the model selected by $1\mathrm{Emp}2\mathrm{VF}_5$. This phenomenon underlines the interest in resampling.
As for s_5, the observed results only confirms that 1Loo nearly always outperforms upon 1Emp and that $1\mathrm{Lpo}_{20}$ is close to the former.

The conclusions of Table 7.4 in favour of resampling are confirmed by the experiments with piecewise-constant noise levels (see Table 7.12 in the Appendix). A large part of the results show that 1Loo significantly outperforms upon 1Emp, but other resampling schemes like $1\mathrm{penRad}_{1.5}$ are often close to the former.

7.5 Open problems about complexity measures

We have seen that a classical strategy to cope with high collection complexity is to gather models with the same dimension. In the homoscedastic framework, this strategy leads to 2BM, which enjoys some optimality properties. However in the heteroscedastic case, the same strategy fails.
Furthermore in Section 7.4.1, we provide the expression of the true risk in both the homoscedastic and the heteroscedastic settings. Whereas the dimension arises from (7.9) as the natural quantity to work with under homoscedasticity, it turns out to be quite different under heteroscedasticity (7.10). Then, we may wonder if the dimension is the relevant quantity to take into account in the gathering procedure.
Indeed, if 2BM turns out to be an effective measure of complexity of the $\widetilde{S}_D$s in the homoscedastic setup, it is no longer the case in the general heteroscedastic setting. Since this complexity measure straightforwardly comes from the dimension as a criterion to cluster models, a more relevant quantity could lead to an effective complexity measure in the latter setting.
Arlot [7] proved the suboptimality of linear penalties in the dimension in the heteroscedastic framework for not too rich collections. Besides in the bias-variance decomposition of the risk, the variance term exhibits a linear dependance on the dimension of the model in the homoscedastic setting, which is no longer true in the heteroscedastic framework.

All these remarks lead us to search for a reliable substitute for the dimension by analogy with the homoscedastic case. In the latter setup, the complexity of the model S_m of dimension D is $\sigma^2 D/n$. A possible complexity measure for model S_m is given by the V-fold penalty for instance (see [8]). Let us denote the V-fold penalty of the model S_m by C_m. This provides an unbiased estimation of the variance term in the risk decomposition. For instance, models may be gathered according to C_m, so as we get clusters of homogeneous complexity. This would lead to the following kind of algorithm:

Algorithm 7.5.1.

1. *Compute C_m for each model.*

2. *Split $\mathcal{M}_n$ into n subsets $\mathcal{M}(1), \ldots, \mathcal{M}(n)$ with (almost) equal sizes and "homogeneous" values of C_m, that is satisfying:*

$$\forall k \in \{1, \ldots, n-1\}, \qquad \max_{m \in \mathcal{M}(k)} C_m \leq \min_{m \in \mathcal{M}(k+1)} C_m \ .$$

3. *For every $k \in \{1, \ldots, n\}$, find the model $\widehat{m}(k)$ which minimizes some criterion* $\mathrm{crit}_1(S_m, P_n)$ *(either the empirical risk or some resampling estimate of the prediction risk).*

4. *Choose k by V-fold cross-validation.*

The main concern of these data-dependent strategies is that they require the computation of the complexity measure for each model, which is computationally infeasible in the exponential setup.
A possible way to assess the relevance of these proposed strategies is to apply them in the polynomial complexity framework, where computations can be performed in a reasonable time.
Another possible idea could be to design an *ad hoc* criterion where resampling and *a priori* complexity measure would be intertwined so that it would reduce the computation burden as well as enhance the observed results of 2BM in the heteroscedastic setting.

7.6 Application to CGH microarray data

7.6.1 Biological context

CGH data Nowadays, we know that some diseases are strongly related to the presence (or absence) of some pieces of our chromosomes. The idea of CGH (Comparative Genomic Hybridization) microarray data is to enable the detection of such alterations.
Split your chromosome into as much pieces as you can. Their location and DNA sequence are perfectly known. For each of them, we find a complementary and specific sequence named *probe*. It means that theoretically, this probe will only hybridize with its complementary piece of chromosome.
Given all these probes, we would like to know whether an individual lacks any part of its chromosome (*deletion*) or on the contrary, if it has something more than expected (*amplification*). For instance, humans are supposed to have two copies of any piece of their chromosomes. If we would be able to count this number of copies, we could deduce whether something is wrong or not. Unfortunately, this count is unachievable until now and we rather use the concentration of each probe instead of their numbers.
Assume that the final concentration of a probe is a function of its cardinality. Measuring this quantity for each probe therefore provides us with some information about the number of copies of each probe. This mechanism is used to get a signal which is related to the concentration of each probe and hence, to the numbers of copies of each piece of chromosome.
Consider now the position of each "probe" on the chromosome. Provided these probes are close enough to each other, any alteration (amplification or deletion) of one of them will induce the same alteration for several of its neighbours. From CGH data, we are looking for some "homogeneous" regions on the chromosome, where the number of copies remains constant on average. Thus, an ideal *CGH profile* (which is the graph of $\{(t_i, Y_i)\}_{i=1,\ldots,n}$, where (t_i, Y_i) denotes the signal value Y_i at position t_i on the chromosome) may be seen as a piecewise constant function. Let m denote the partition on which this function is defined. Then, each interval of m may be interpreted as a homogeneous region where the copy number remains roughly the same on average.

The Bt474 cell lines Bt474 cell lines denote epithelial cells which have been obtained from human breast cancer tumors on a 60-years-old woman. A test genome of Bt474 cell lines is compared to a normal reference male genome. Several chromosomes are studied in these cell lines, but we only restrict our attention to chromosomes 1 and 9 for which some features arise [38]. The number of probes we have at our disposal for each chromosome is respectively 119 and 93.

In this study, the question is to detect some amplified or deleted regions of the chromosome, which may be related to the cancer in view of further analyses.

7.6.2 Comparison with other algorithms

Change-points in the mean or/and in the variance

CGH microarray data have recently been studied by Picard [37] and Picard *et al.* [38, 39]. Authors apply change-point detection technics in order to get an optimal segmentation of CGH data. Their methodology relies on the optimization of a penalized log-likelihood criterion in a gaussian setting. Two models were proposed, one with constant noise and the other one with a heteroscedastic noise.

Change-points in the mean (homoscedastic case) The first model they describe is homoscedastic. Let m denote a partition of $[0,1]$ in D_m intervals $I_1, \ldots, I_{D_m}$ such that the regression function s may be defined by $s = \sum_{k=1}^{D_m} \mu_k \mathbb{1}_{I_k}$. Then for any $i = 1, \ldots, n$,

$$Y_i = \mu_k + \sigma\epsilon_i,$$

if $t_i \in I_k$, where $\epsilon_i \sim \mathcal{N}(0,1)$ and $\sigma(\cdot) > 0$. In this setting, the criterion to be optimized is

$$\mathrm{crit}(m) = \sum_{k=1}^{D_m} \sum_{t \in I_k} (Y_t - \widehat{\mu}_k)^2 + \beta D_m \tag{7.11}$$

where $\widehat{\mu}_k = \sum_{t \in I_k} Y_t / n_k$ with $n_k = \sum_i \mathbb{1}_{I_k}(t_i)$, and $\beta > 0$ is a constant which is to be adaptively defined (see Lavielle [25]).

Change-points in the mean (piecewise constant noise) An alternative to the constant variance assumption is the following idea: the variability is organized along the chromosome according to the homogeneous regions of constant copy numbers. It justifies the introduction of piecewise constant variance in the model: $i = 1, \ldots, n$,

$$Y_i = \mu_k + \sigma_k \epsilon_i,$$

if $t_i \in I_k$, where $\epsilon_i \sim \mathcal{N}(0,1)$ and for any k, $\sigma_k(\cdot) > 0$. The authors minimize

$$\mathrm{crit}(m) = \sum_{k=1}^{D_m} n_k \log\left(\frac{1}{n_k} \sum_{t \in I_k} (Y_t - \widehat{\mu}_k)^2 \right) + \beta D_m \tag{7.12}$$

with the aforementioned quantities.

Comments Several remarks can be formulated beforehand about the above algorithms:

- The first criterion strongly relies on the homoscedastic setup. In the presence of heteroscedasticity, changes in the variance (which cannot be explained by the model) are likely to result in false change-points. The final number of selected breakpoints should be larger than expected.
- Unlike the previous one, this algorithm is based on a heteroscedasticity assumption since it allows piecewise constant noise. However with homoscedastic observations, some changes in the mean may be interpreted by the model as changes in the variance, resulting in less breakpoints than the oracle.
- In the heteroscedastic framework, all the models with the same dimension are penalized in the same way, no matter the amount of noise in each interval of the partition, which may lead to under- or over-penalization.

Resampling-based algorithm We would like to get a better idea of how wide the range of results provided by resampling methods may be in a real situation. To this aim, we apply the same algorithms as in our simulation study on this data set. We use 1Emp2VF$_5$, 1Loo2VF$_5$, 1Lpo$_{20}$2VF$_5$ 1penRad2VF$_5$ and 1penRad$_{1.5}$2VF$_5$.

7.6.3 Results

Let us first consider chromosome 9. We observe that all the resampling algorithms provide nearly the same change-points, like those displayed in the bottom of Figure 7.11, where we have only reported results for 1Emp2VF and 1Lpo$_{20}$2VF, which are strictly the same. If we compare resampling results and those provided by the minimization of (7.11) and (7.12), we see that the breakpoint positions are also the same. There seems to be four homogeneous regions on the chromosome 9. If we assume that the reference (normal male cell lines) owns two copies of each BAC, we may draw the following conclusions. The widest one may be interpreted as a *normal* region, that is a region with two copies at each location. On the contrary, the region corresponding to the lowest signal may be called a *deleted* region, since this low signal (with respect to the normal region) means that some copies are missing. It remains two regions with the highest signal, which could be considered as *amplified* regions of the chromosome. It means that there are more than two copies of each BAC.

As for resampling procedures, we consider that the observed agreement with existing and widespread methods in the literature dedicated to CGH microarray analysis [37, 38] is a kind of validation of the resampling strategy.

Figure 7.12 depicts results for the chromosome 1, obtained from the same methods as above. Unlike the previous example, we observe 4 different solutions. Note that the other resampling strategies result in the same change-points as either 1Emp2VF or 1Lpo$_{20}$2VF (see Appendix).

The most extreme partitions were obtained with (7.11) and (7.12) (top panel of Figure 7.12), which respectively provide the partition with the largest and the smallest dimension. Since (7.11) is based on a homoscedastic assumption, it is more sensitive than the heteroscedastic model in (7.12) to any change in the variance for instance. On the contrary the piecewise constant noise assumption in (7.12) may lead to explain a change in the mean by a change in the variance, which results in fewer change-points.
The dimensions of models selected by 1Emp2VF and 1Lpo2VF are respectively 6 and 4, which may be accounted for by the stronger penalization of the Lpo algorithm with respect to Emp as p grows. Note that both of them are less sensitive to possible outliers than the homoscedastic model (7.11) since the latter puts two breakpoints to detect a single BAC in the middle of the first widest regions, whereas neither resampling method does. On the other hand, they are more sensitive to potential change in the mean than the heteroscedastic method (7.12). In other words, we think that the proposed resampling algorithms combine assets of these two approaches, without suffering the same drawbacks.
As for the interpretation of the results, we are unable to say whether the model of dimension 4 is better or not than that of dimension 6. Thus, our conclusion is that there may be at least 4 homogeneous regions in chromosome 1, and possibly 6, with the subdivision of the last one. But some further biological experiments are required to draw any precise conclusion.

7.7 Conclusion

7.7.1 Main results

On the basis of our extensive simulation study, some conclusions can be drawn about the ability of resampling to overcome the heteroscedasticity issue with rich collections of models.

Whereas the 1Emp2BM performs very well under homoscedasticity, the same algorithm can fail seriously in a heteroscedastic setup. Cross-validation algorithms turn out to be more robust to heteroscedasticity. In comparison, 1Emp2VF remains reliable in the homoscedastic framework, but strongly outperforms 1Emp2BM under heteroscedasticity in choosing the number of breakpoints.
Furthermore, resampling can also be used for choosing a segmentation, given the number of breakpoints.

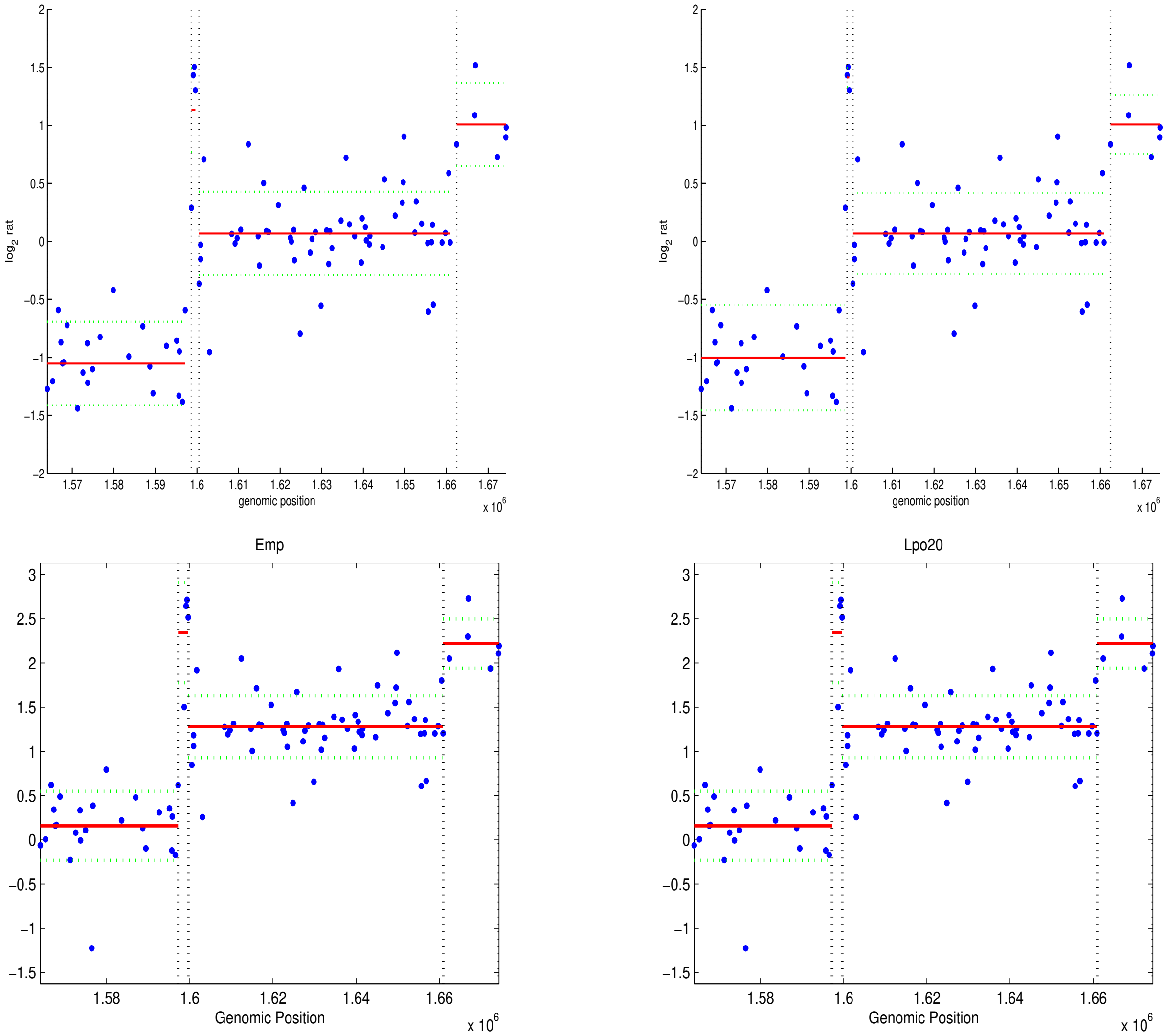

Figure 7.11: Change-points locations along Chromosome 9. The mean on each homogeneous region is indicated by plain horizontal lines. The standard deviation from this value is denoted by dotted horizontal lines. **Top panel:** Results provided by the minimization of (7.11) (left-hand side) and (7.12) (right-hand side). **Bottom panel:** Results obtained by $1\mathrm{Emp}2\mathrm{VF}_5$ on the left-hand side and $1\mathrm{Lpo}_{20}2\mathrm{VF}_5$ on the right-hand side.

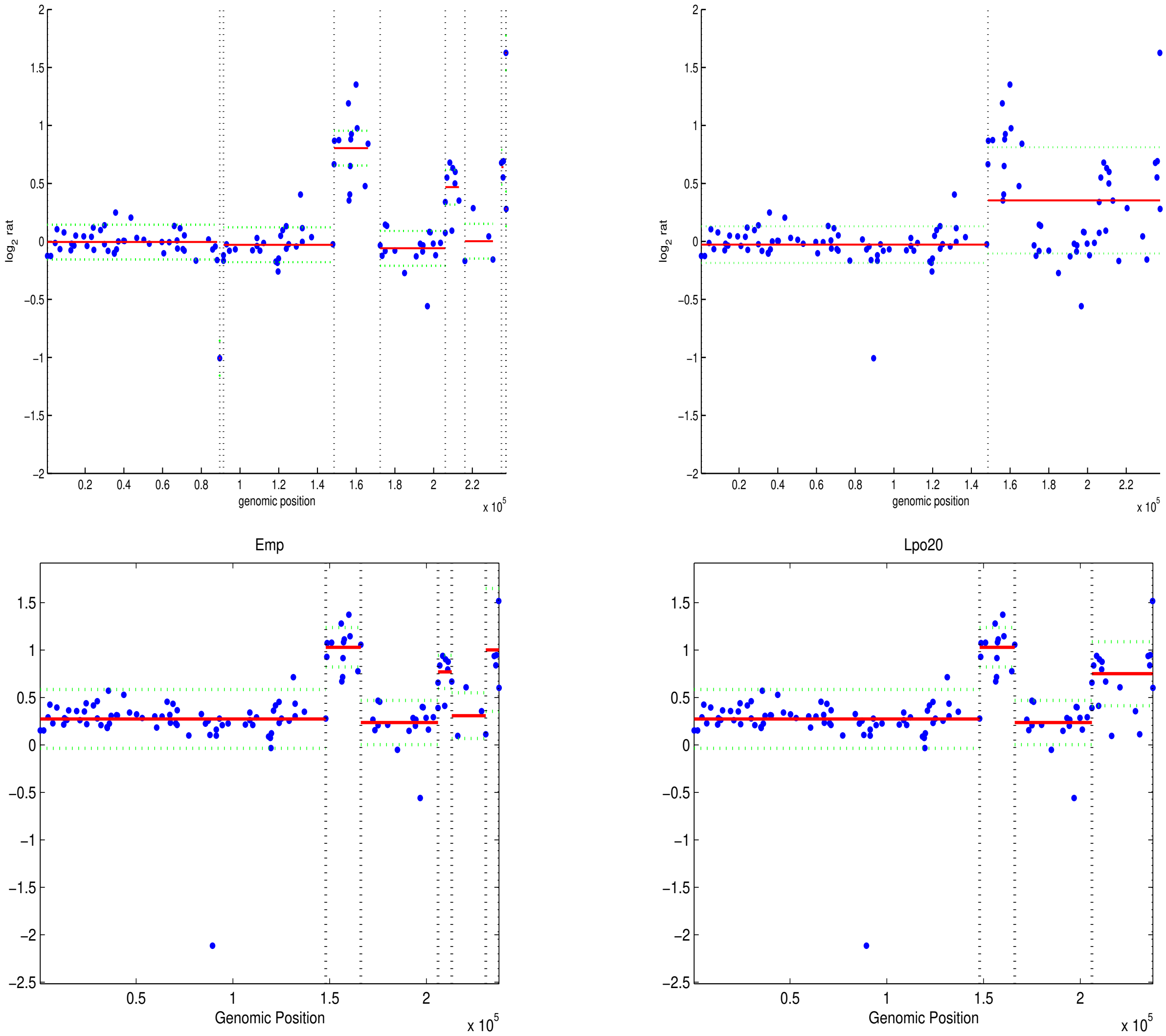

Figure 7.12: Change-points locations along Chromosome 9. The mean on each homogeneous region is indicated by plain horizontal lines. The standard deviation from this value is denoted by dotted horizontal lines. **Top panel:** Results provided by the minimization of (7.11) (left-hand side) and (7.12) (right-hand side). **Bottom panel:** Results obtained by $1\text{Emp}2\text{VF}_5$ on the left-hand side and $1\text{Lpo}_{20}2\text{VF}_5$ on the right-hand side.

Actually in presence of heteroscedasticity, 1Emp may suffer from overfitting due to the large number of competing models for each dimension. Combined with VF for choosing the number of breakpoints, this gives an algorithm with good estimation performance, even when the data are heteroscedastic.
These reliable results are confirmed by the application of this algorithm to CGH microarray data. The proposed strategy exhibits some flexibility with homoscedastic as well as heteroscedastic data, while it provides reliable results.

7.7.2 Prospects

Possible modifications of the algorithm

Unfortunately, we do not have a closed-form expression for the Lpo risk estimate of any estimator. That is the reason why in a further study, we will attempt to study to what extent the observed results of the Lpo estimator may be generalized to other cross-validation algorithms.
Besides, the V-fold penalties [5] provide an unbiased estimator of the risk, while they allow to separately control the amount of overpenalization. They therefore appear as a natural competitor with the Lpo algorithm.

Subsequently, we propose the following modifications of the first step of our algorithm:

First step

VF_V: "V-fold" CV risk estimator with V subsets:

$$\mathrm{crit}_{1,\mathrm{VF}}(S_m, P_n, V) := \frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\gamma\left(\mathrm{ERM}\left(S_m, P_n^{(-j)}\right)\right)$$

where B_j denotes the $j-$th element of the partition of the data in V subsets of approximately equal size n/V, $P_n^{(j)} = \mathrm{Card}(B_j)^{-1}\sum_{i\in B_j}\delta_{(X_i,Y_i)}$ et $P_n^{(-j)} = (n - \mathrm{Card}(B_j))^{-1}\sum_{i\notin B_j}\delta_{(X_i,Y_i)}$.

$\mathrm{penVF}_{V,C}$: VF resampling penalty with V subsets, with an overpenalization coefficient equal to $C \geq 1$ (see [8]) :

$$\mathrm{crit}_{2,\mathrm{penVF}}(S_m, P_n, V) := P_n\gamma\left(\mathcal{A}_D\left(P_n\right)\right) + \frac{C}{V}\sum_{j=1}^{V}\left[\left(P_n - P_n^{(-j)}\right)\gamma\left(\mathrm{ERM}\left(S_m, P_n^{(-j)}\right)\right)\right] \ .$$

Second step Some other resampling algorithm may be included in the second step as well as possible enhancements:

$\mathrm{penVF}_{V,C}$: VF resampling penalty with V subsets, with an overpenalization coefficient equal to $C \geq 1$ (see [8]) :

$$\mathrm{crit}_{2,\mathrm{penVF}}(\mathcal{A}_D, P_n, V) := P_n\gamma\left(\mathcal{A}_D\left(P_n\right)\right) + C\mathbb{E}_W\left[\left(P_n - P_n^{(-j)}\right)\gamma\left(\mathcal{A}_D\left(P_n^{(-j)}\right)\right)\right]$$

penRad_C : Rademacher resampling penalty with overpenalization coefficient $C \geq 1$ (see [5]):

$$\mathrm{crit}_{2,\mathrm{penRad}}(\mathcal{A}_D, P_n) := P_n\gamma\left(\mathcal{A}_D\left(P_n\right)\right) + C\mathbb{E}_W\left[\left(P_n - P_n^W\right)\gamma\left(\mathcal{A}_D\left(P_n^W\right)\right)\right]$$

where $\mathbb{E}_W\left[\,\cdot\,\right] = \mathbb{E}\left[\,\cdot \mid (X_i, Y_i)_{1\leq i\leq n}\,\right]$.

At this step of the global algorithm, we do not have any closed-form formula for the Lpo estimator as well as for penVF and penRad. Since the Lpo is intractable as soon as we do not have any closed-form expression, it is excluded from the study.
On the contrary, penVF and penRad can be computed in a very reasonable time. Including them in a future simulation study is one of our main purposes.

In principle, $2\mathrm{VF}_V$, $2\mathrm{penVF}_V$ and $2\mathrm{penRad}_C$ should behave similarly as complexity measures, except that penVF_V and penRad_C are slightly more flexible and can be more accurate [8, 5].

Other simulation settings

Non-Gaussian noise: In the simulations we carried out, we only consider a gaussian noise. There is neither theoretical nor practical reasons for this restriction. Performing a similar study with a wider panel of noise types is among our objectives.

7.8 Appendix

7.8.1 Complexity measure

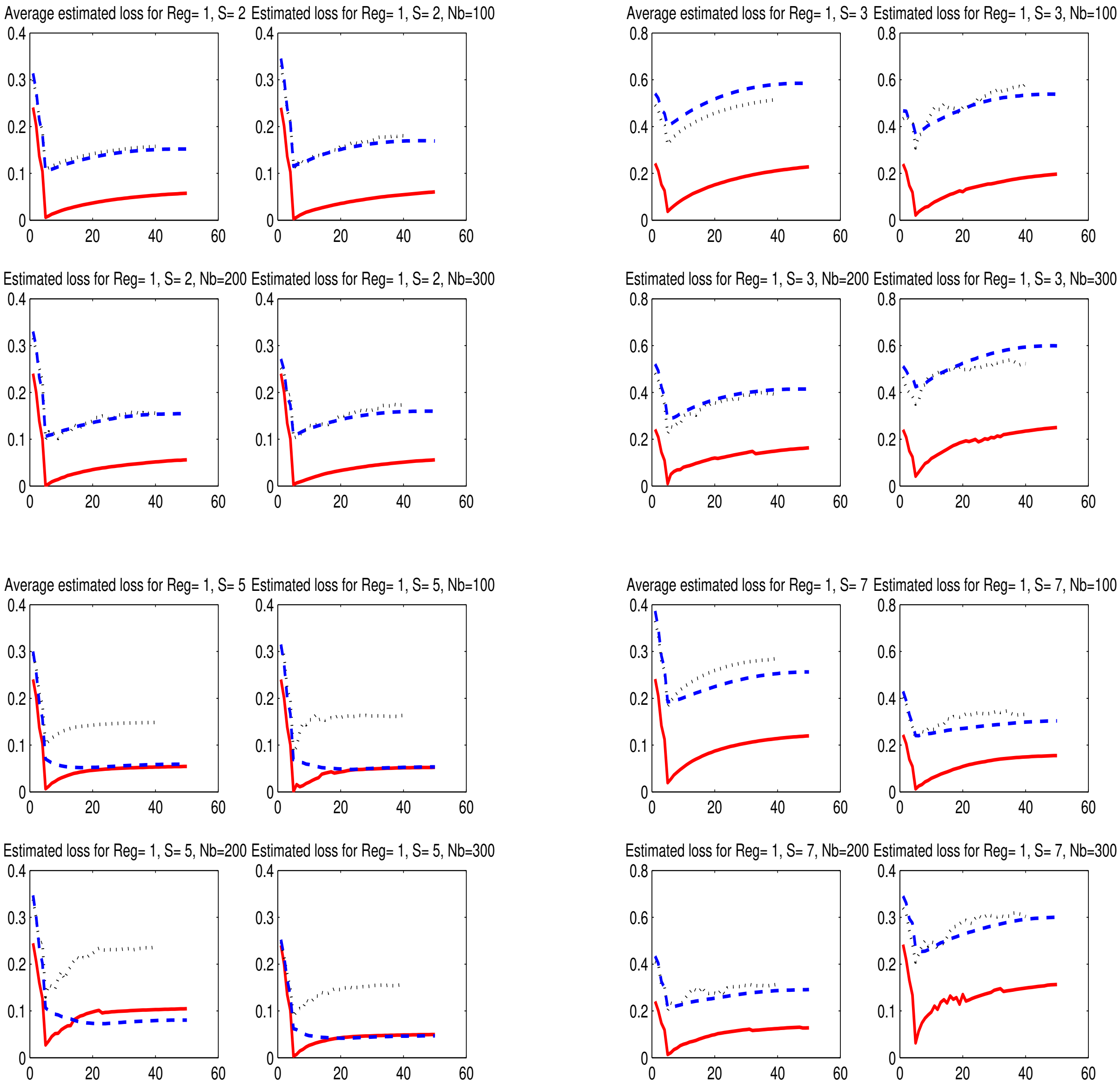

Figure 7.13: Graph of $D \mapsto \text{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_1 associated with $\sigma_{c,2}, \sigma_{c,3}, \sigma_{pc,5}$ and $\sigma_{s,7}$: 1Emp2Id ('-' plain line), 1Emp2BM ('–'dashed line) and 1Emp2VF$_5$ (':' dotted line). In each set of 4 graphs, the top-left one depicts the average loss over the trials, while the others display results for three particular trials.

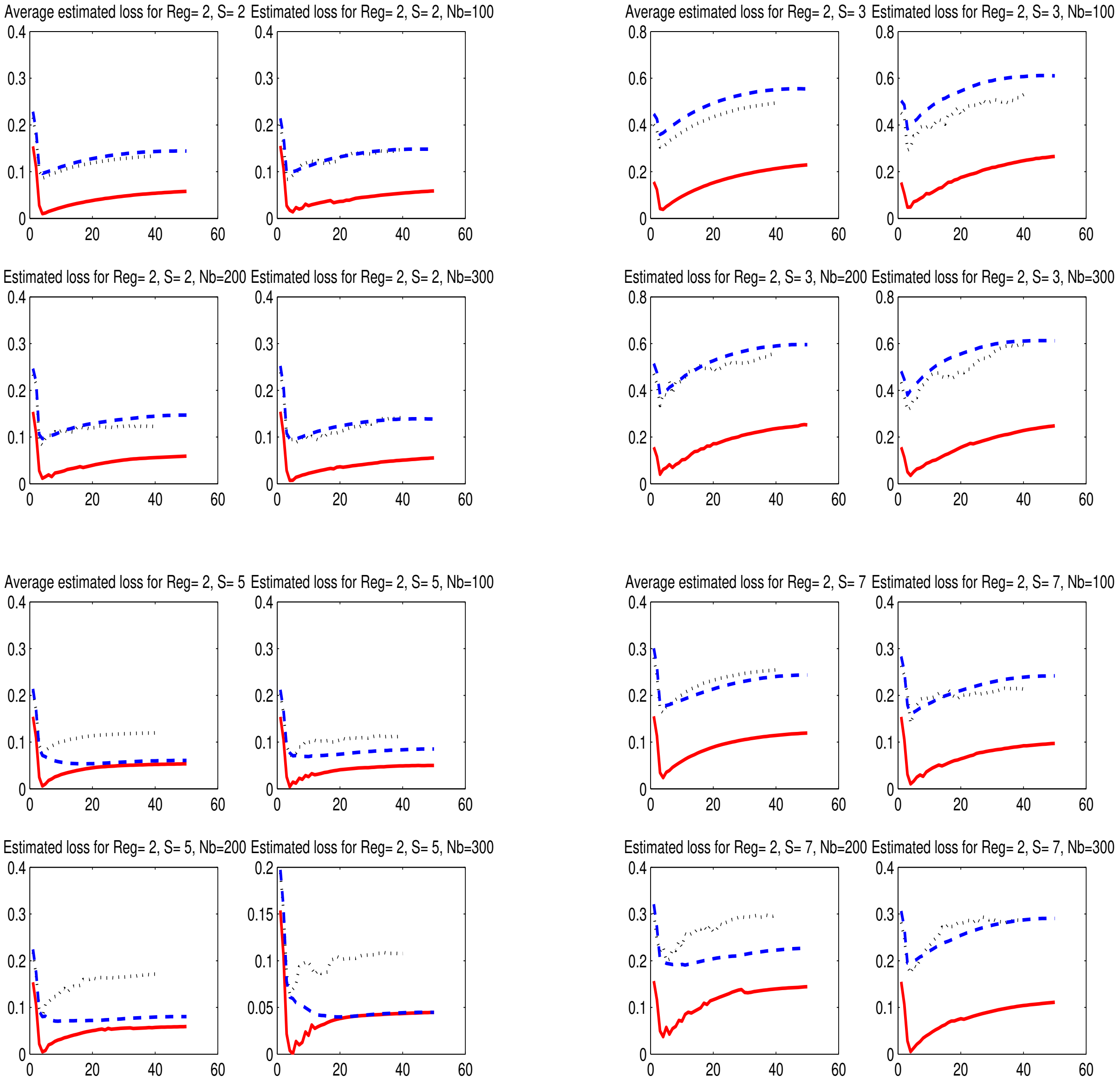

Figure 7.14: Graph of $D \mapsto \mathrm{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_2 associated with $\sigma_{c,2}, \sigma_{c,3}, \sigma_{pc,5}$ and $\sigma_{s,7}$: 1Emp2Id ('-' plain line), 1Emp2BM ('–'dashed line) and 1Emp2VF$_5$ (':' dotted line). In each set of 4 graphs, the top-left one depicts the average loss over the trials, while the others display results for three particular trials.

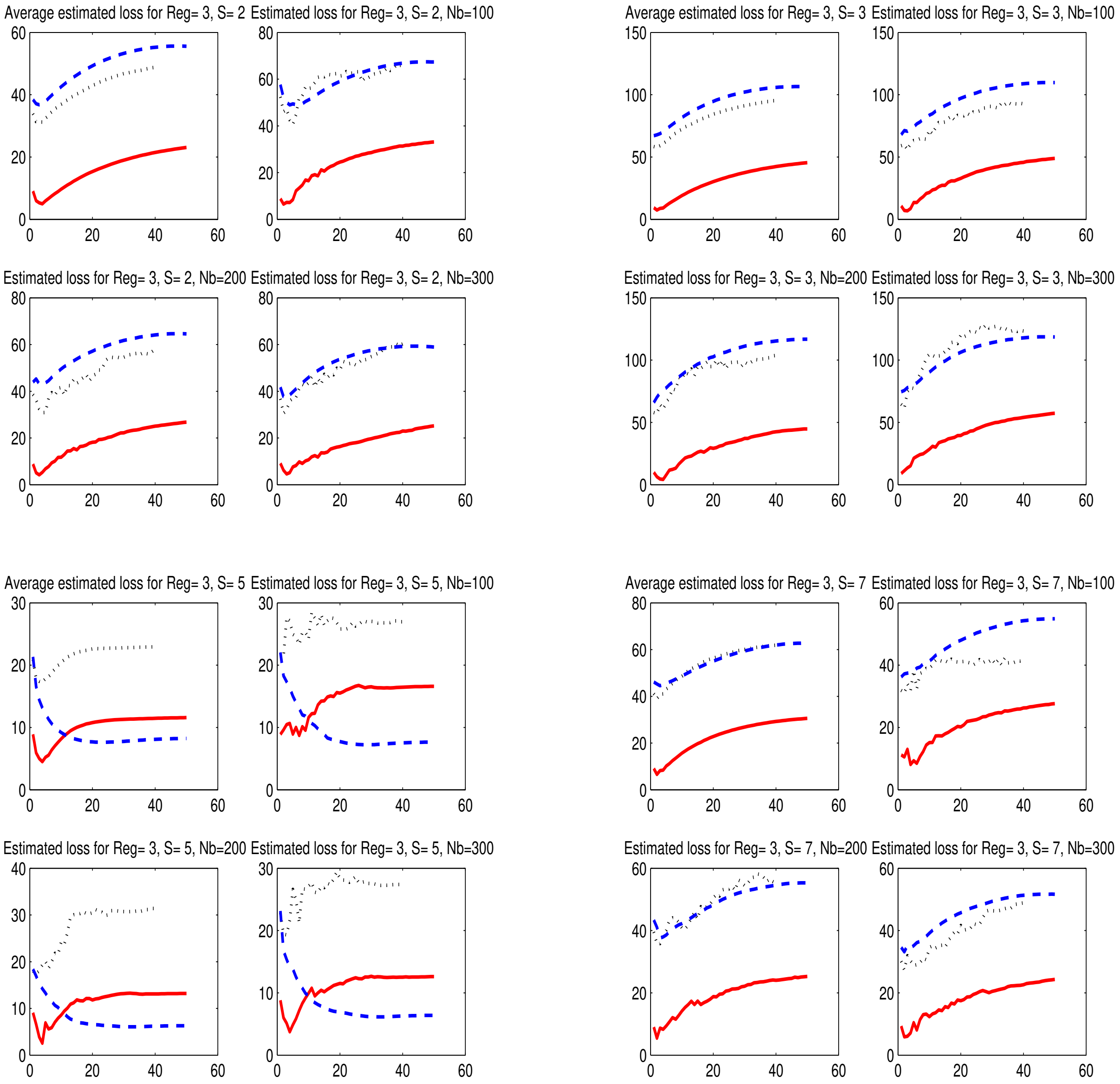

Figure 7.15: Graph of $D \mapsto \text{crit}_{2,A}$ where A denotes one the competing algorithms of the second step, with s_3 associated with $\sigma_{c,2}, \sigma_{c,3}, \sigma_{pc,5}$ and $\sigma_{s,7}$: 1Emp2Id ('-' plain line), 1Emp2BM ('–'dashed line) and 1Emp2VF$_5$ (':' dotted line). In each set of 4 graphs, the top-left one depicts the average loss over the trials, while the others display results for three particular trials.

s.	*σ.*	1Emp2Id	1Emp2VF$_5$	1Emp2BM	1Emp2BM+	1Emp2C_p
1	c,1	1 ± 6.2e-017	3.14 ± 0.34	**1.37** ± 0.088	22 ± 6.6	471 ± 80
	c,2	2.19 ± 0.28	3.35 ± 0.35	**2.96** ± 0.36	7.23 ± 1	62.1 ± 7.1
	c,3	4.09 ± 0.24	5.12 ± 0.3	**4.84** ± 0.32	6.87 ± 0.53	19.3 ± 1.2
	pc,4	7.41 ± 1.8	**9.71** ± 1.9	31.4 ± 11	47.1 ± 20	85.7 ± 21
	pc,5	4.04 ± 0.57	**7.33** ± 0.9	38.5 ± 2.8	65.1 ± 8.8	131 ± 14
	s,6	2.13 ± 0.18	**3.47** ± 0.28	5.1 ± 0.39	19 ± 3	94.7 ± 7.7
	s,7	3.83 ± 0.36	**4.92** ± 0.4	7.86 ± 1.3	9.95 ± 0.9	33 ± 3
	s,8	4.76 ± 0.29	**6.27** ± 0.36	8.2 ± 0.56	12 ± 1.2	20.3 ± 1.3
3	c,1	2.38 ± 0.18	3.82 ± 0.31	**2.84** ± 0.24	5.42 ± 0.63	15.8 ± 1.7
	c,2	4.17 ± 0.24	5.65 ± 0.33	**5.52** ± 0.37	6.28 ± 0.46	15.1 ± 0.83
	c,3	4.54 ± 0.2	**5.93** ± 0.26	6.13 ± 0.29	7.59 ± 0.54	17.1 ± 0.93
	pc,4	12 ± 1.5	**24.8** ± 4.7	54.9 ± 8.8	53.5 ± 13	72.6 ± 10
	pc,5	5.97 ± 0.49	**12.5** ± 1.2	59.4 ± 5	28.2 ± 2.4	52.4 ± 4.2
	s,6	4.95 ± 0.33	**7.34** ± 0.44	9.72 ± 0.62	10.8 ± 0.72	20.3 ± 1.2
	s,7	5.82 ± 0.41	**8.02** ± 0.51	10.3 ± 0.76	11.8 ± 1.1	21.3 ± 1.5
	s,8	4.74 ± 0.25	**6.33** ± 0.37	7.73 ± 0.5	9.85 ± 0.67	22.3 ± 1.6
5	c,1	3.67 ± 0.065	4.57 ± 0.1	**4.41** ± 0.088	4.83 ± 0.13	8.04 ± 0.21
	c,2	3.81 ± 0.077	**5.14** ± 0.13	5.62 ± 0.15	5.69 ± 0.19	11 ± 0.36
	c,3	3.76 ± 0.093	**5.3** ± 0.17	5.6 ± 0.16	6.22 ± 0.25	13.4 ± 0.55
	pc,4	6.22 ± 0.22	**8.7** ± 0.3	15.8 ± 0.76	11.6 ± 0.52	16.6 ± 0.7
	pc,5	5.78 ± 0.19	**8.48** ± 0.29	19.8 ± 0.59	12.8 ± 0.48	16.6 ± 0.55
	s,6	4.86 ± 0.15	**6.62** ± 0.24	7.95 ± 0.31	8.1 ± 0.34	13.2 ± 0.49
	s,7	4.6 ± 0.14	**6.17** ± 0.22	7.55 ± 0.29	8.49 ± 0.35	15.5 ± 0.66
	s,8	4.08 ± 0.13	**5.6** ± 0.22	7.7 ± 0.43	7.63 ± 0.34	17.1 ± 0.75

s.	*σ.*	1Emp2Id	1Emp2VF$_5$	1Emp2BM	1Emp2C_p
1	c,1	5 ± 0	6.23 ± 0.16	5.12 ± 0.024	7.06 ± 0.47
	c,2	5.11 ± 0.021	5.51 ± 0.06	5.13 ± 0.024	9.67 ± 0.53
	c,3	5.18 ± 0.03	5.46 ± 0.065	5.17 ± 0.042	13.3 ± 0.52
	pc,4	5.12 ± 0.023	5.56 ± 0.073	10.4 ± 0.45	10.5 ± 0.6
	pc,5	5.1 ± 0.019	5.48 ± 0.079	16.1 ± 0.23	13 ± 0.74
	s,6	5.12 ± 0.024	5.69 ± 0.1	6.32 ± 0.14	8.94 ± 0.57
	s,7	5.18 ± 0.032	5.53 ± 0.064	6.46 ± 0.16	12.4 ± 0.64
	s,8	5.08 ± 0.025	5.28 ± 0.077	6.19 ± 0.12	14.2 ± 0.65
3	c,1	5.06 ± 0.016	5.66 ± 0.087	5.09 ± 0.025	16.6 ± 0.54
	c,2	4.44 ± 0.038	4.67 ± 0.076	4.16 ± 0.031	15.1 ± 0.53
	c,3	3.75 ± 0.035	3.87 ± 0.071	3.24 ± 0.034	14.6 ± 0.54
	pc,4	3.95 ± 0.034	4.34 ± 0.075	9.33 ± 0.51	15.5 ± 0.65
	pc,5	4.25 ± 0.026	4.77 ± 0.14	16.3 ± 0.26	20.8 ± 0.66
	s,6	4.74 ± 0.037	5.06 ± 0.097	5.7 ± 0.15	17.9 ± 0.61
	s,7	4.11 ± 0.036	4.24 ± 0.08	4.81 ± 0.13	17.4 ± 0.61
	s,8	3.37 ± 0.028	3.44 ± 0.064	4.08 ± 0.12	16.9 ± 0.62
5	c,1	5.28 ± 0.08	5.6 ± 0.13	4.49 ± 0.06	18.2 ± 0.53
	c,2	3.66 ± 0.053	3.37 ± 0.088	2.65 ± 0.067	14.8 ± 0.52
	c,3	2.59 ± 0.058	2.51 ± 0.083	1.7 ± 0.058	13.3 ± 0.52
	pc,4	3.47 ± 0.065	3.5 ± 0.11	11 ± 0.5	15.1 ± 0.65
	pc,5	4.04 ± 0.061	4.61 ± 0.15	22.6 ± 0.29	22.7 ± 0.78
	s,6	3.49 ± 0.068	3.84 ± 0.12	4.83 ± 0.18	17.2 ± 0.62
	s,7	2.66 ± 0.052	2.93 ± 0.095	3.44 ± 0.16	16.6 ± 0.62
	s,8	2.15 ± 0.047	2.37 ± 0.1	2.96 ± 0.19	16 ± 0.65

Table 7.7: **Top panel:** Average ratios of the loss of the model selected by each algorithm (1Emp2*) over that of the oracle over the while collection of models ± a standard deviation. Bold text denotes the minimum value for each condition, while underlined figures indicate significantly worse values. We say that a value is significantly worse than another one if it is larger than the latter and the discrepancy is larger than the sum of their respective standard deviations. **Bottom panel:** Average selected dimension for each algorithm (1Emp2*) ± one standard deviation. $N = 300$ trials are used.

7.8.2 Best model choice for each dimension

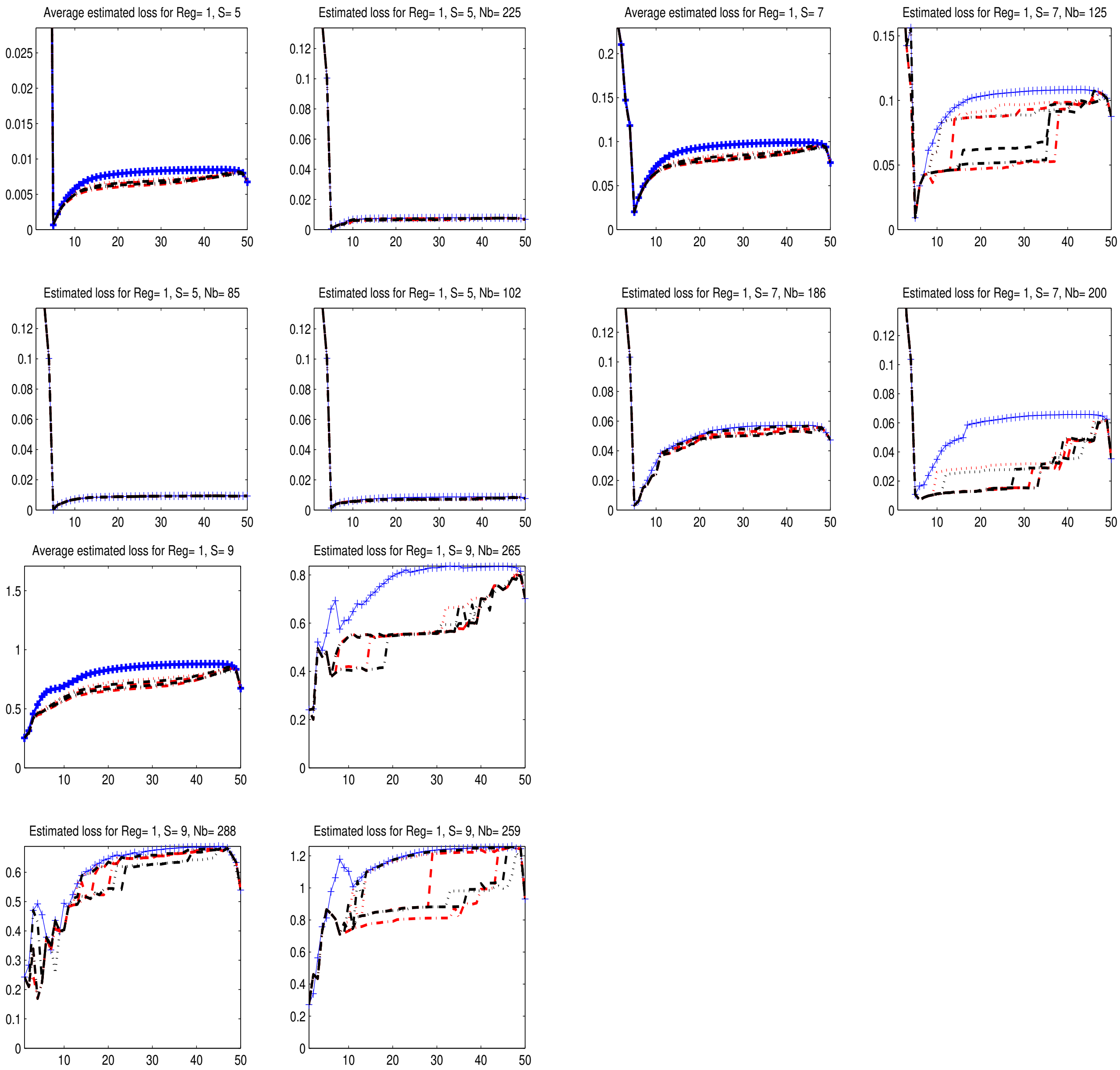

Figure 7.16: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_1, noise levels: $\sigma_{pc,5}$, $\sigma_{pc,7}$ and $\sigma_{pc,9}$.

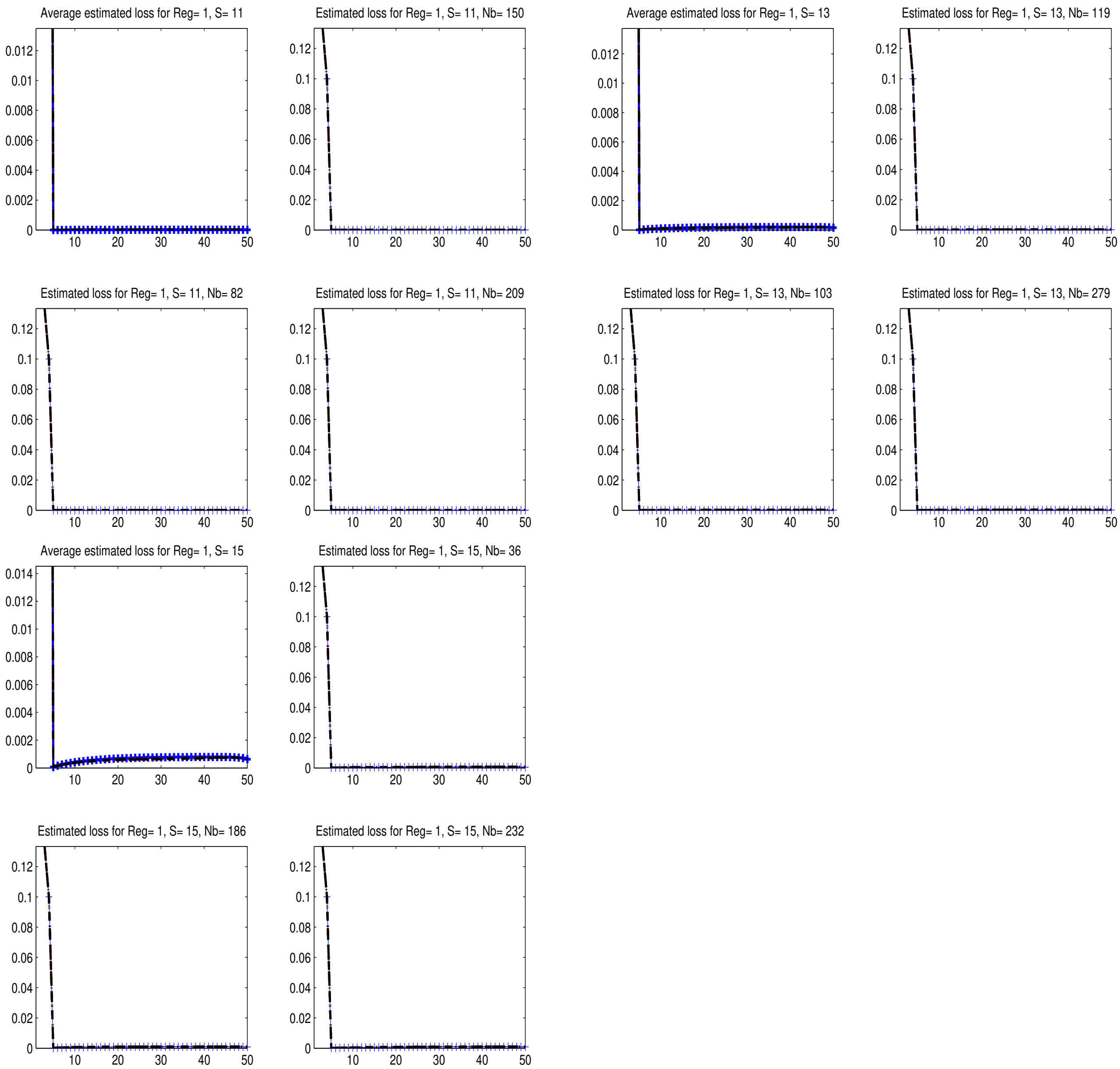

Figure 7.17: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_1, noise levels: $\sigma_{s,11}$, $\sigma_{s,13}$ et $\sigma_{s,15}$.

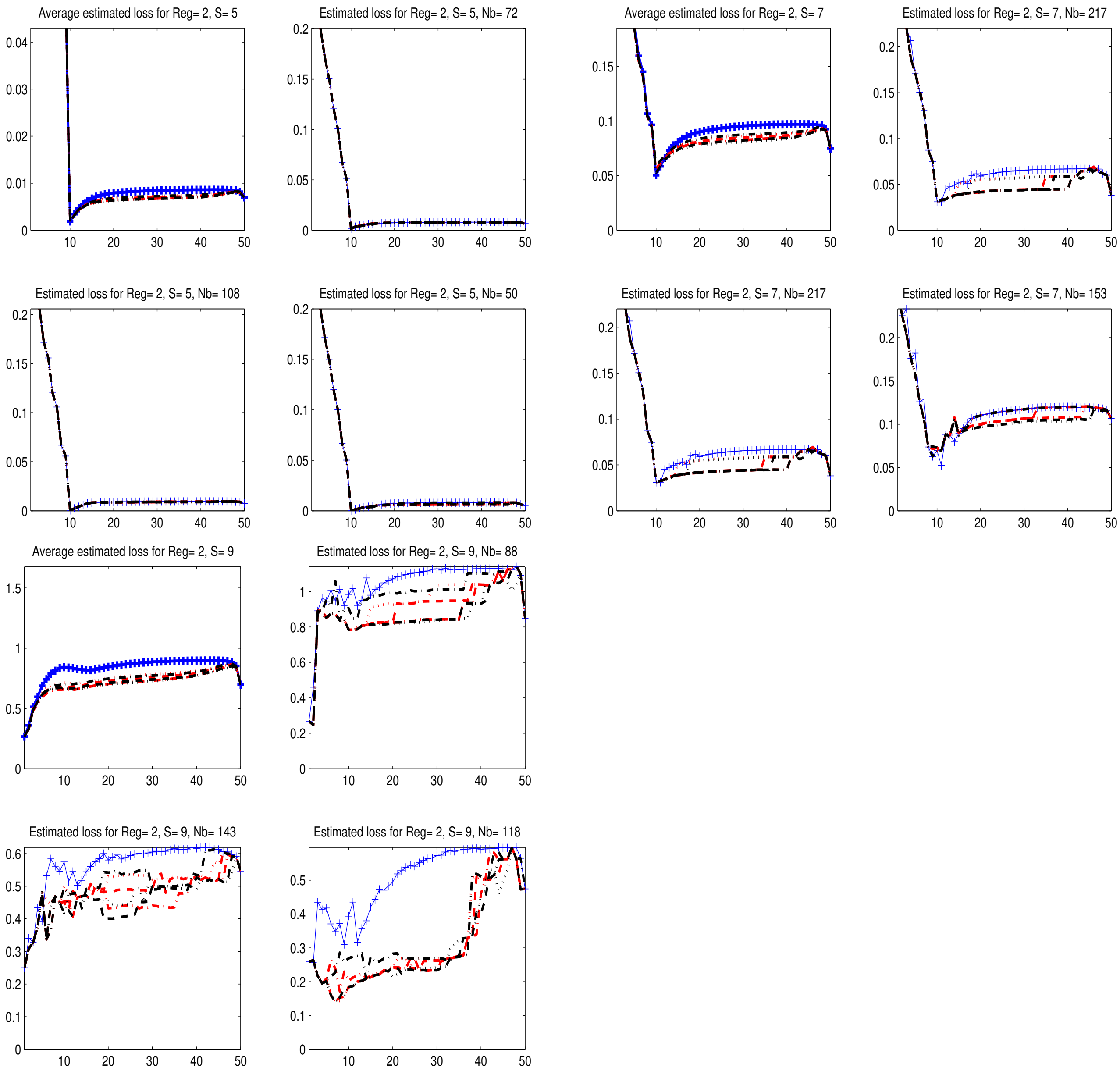

Figure 7.18: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_2, noise levels: $\sigma_{pc,5}$, $\sigma_{pc,7}$ and $\sigma_{pc,9}$.

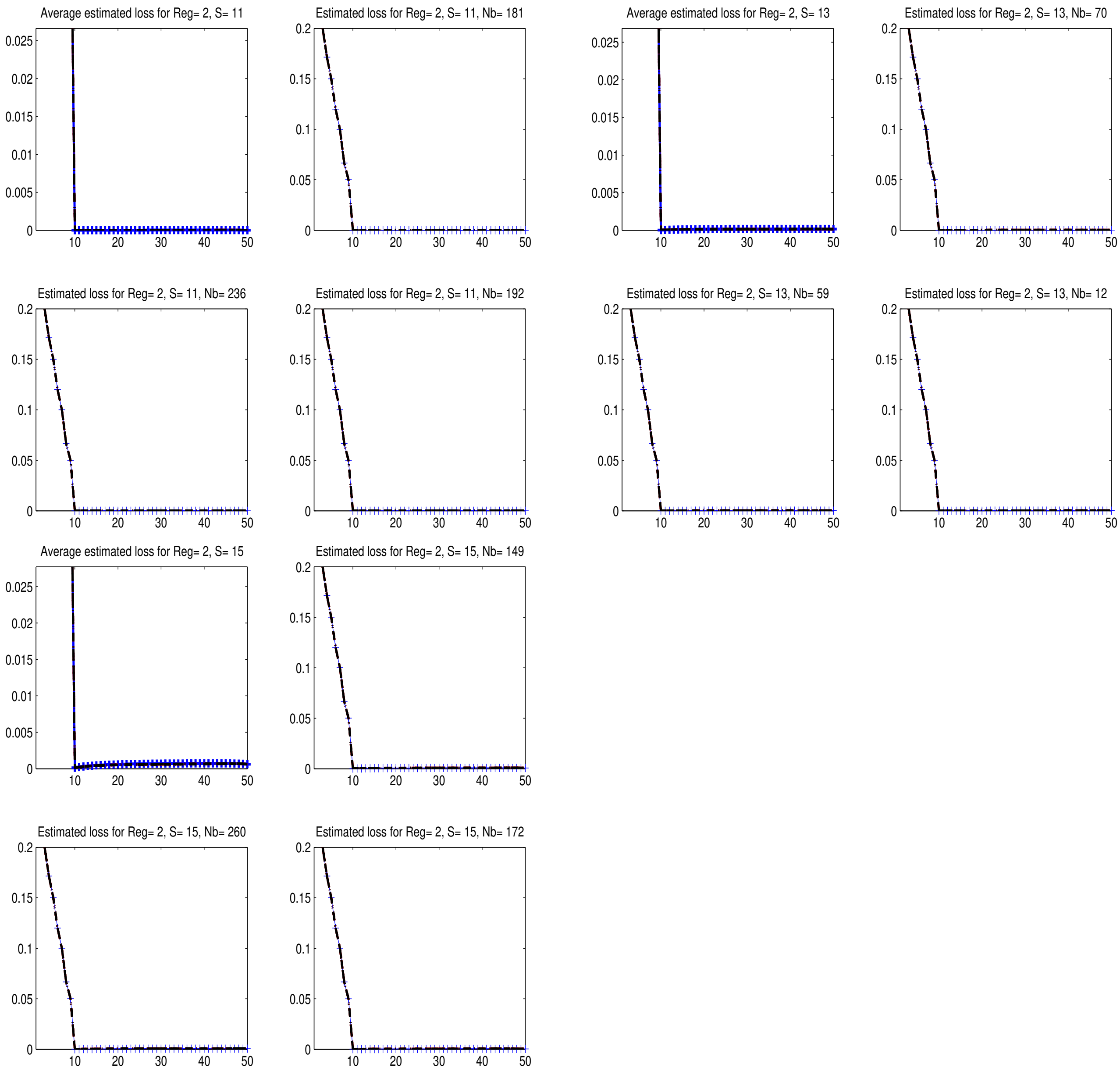

Figure 7.19: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_2, noise levels: $\sigma_{s,11}$, $\sigma_{s,13}$ et $\sigma_{s,15}$.

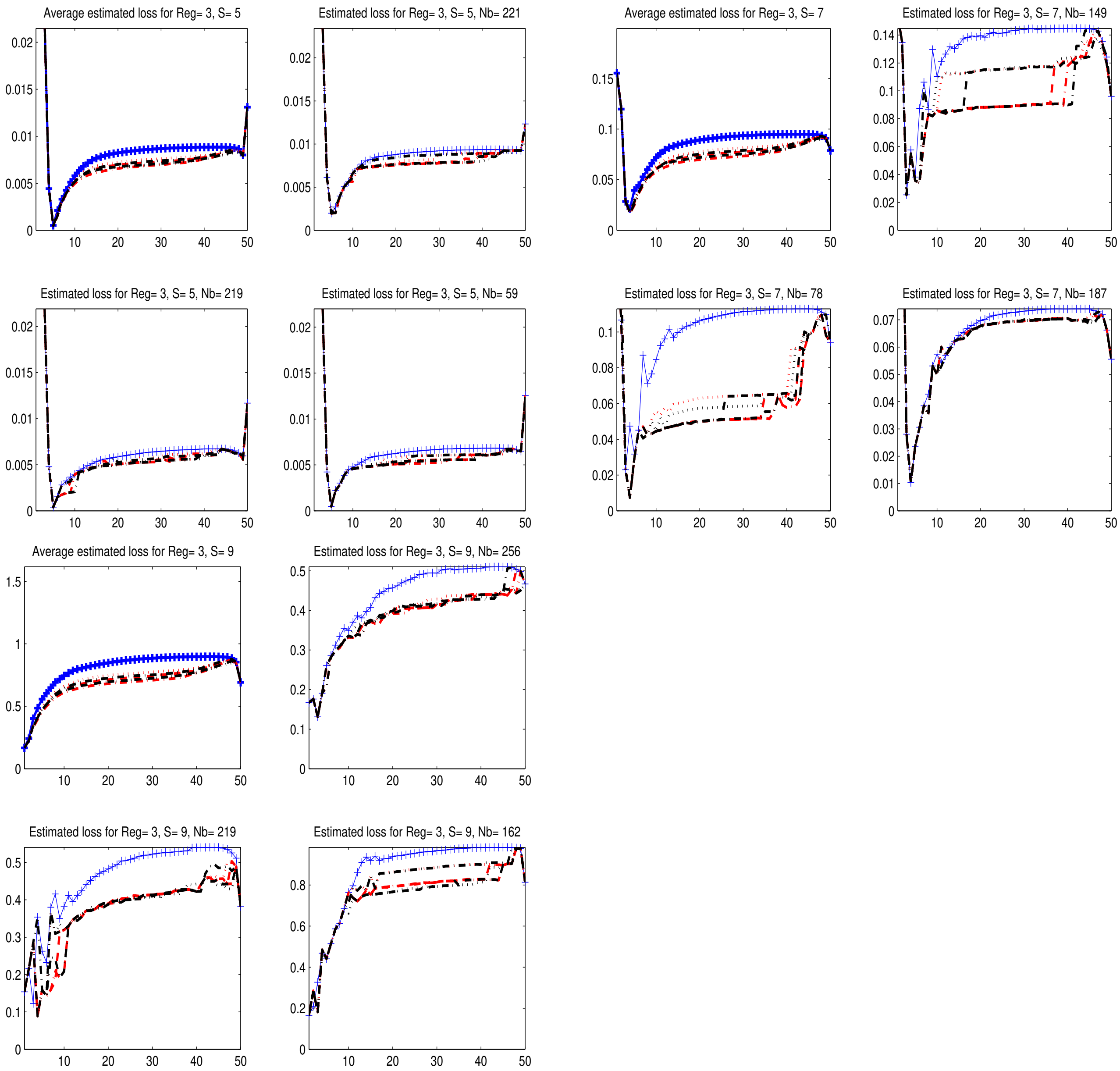

Figure 7.20: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_3, noise levels: $\sigma_{pc,5}$, $\sigma_{pc,7}$ and $\sigma_{pc,9}$.

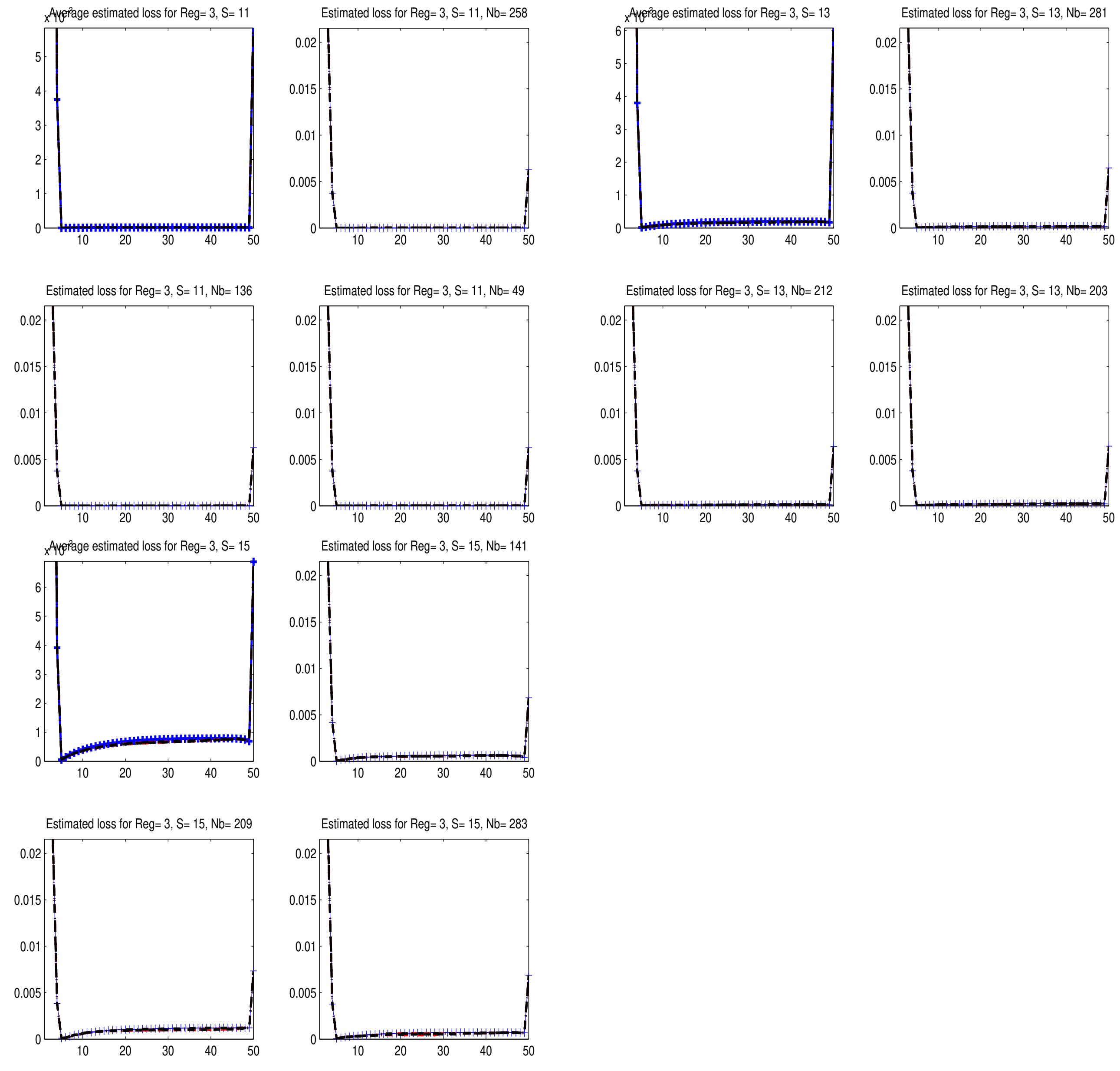

Figure 7.21: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_3, noise levels: $\sigma_{s,11}$, $\sigma_{s,13}$ et $\sigma_{s,15}$.

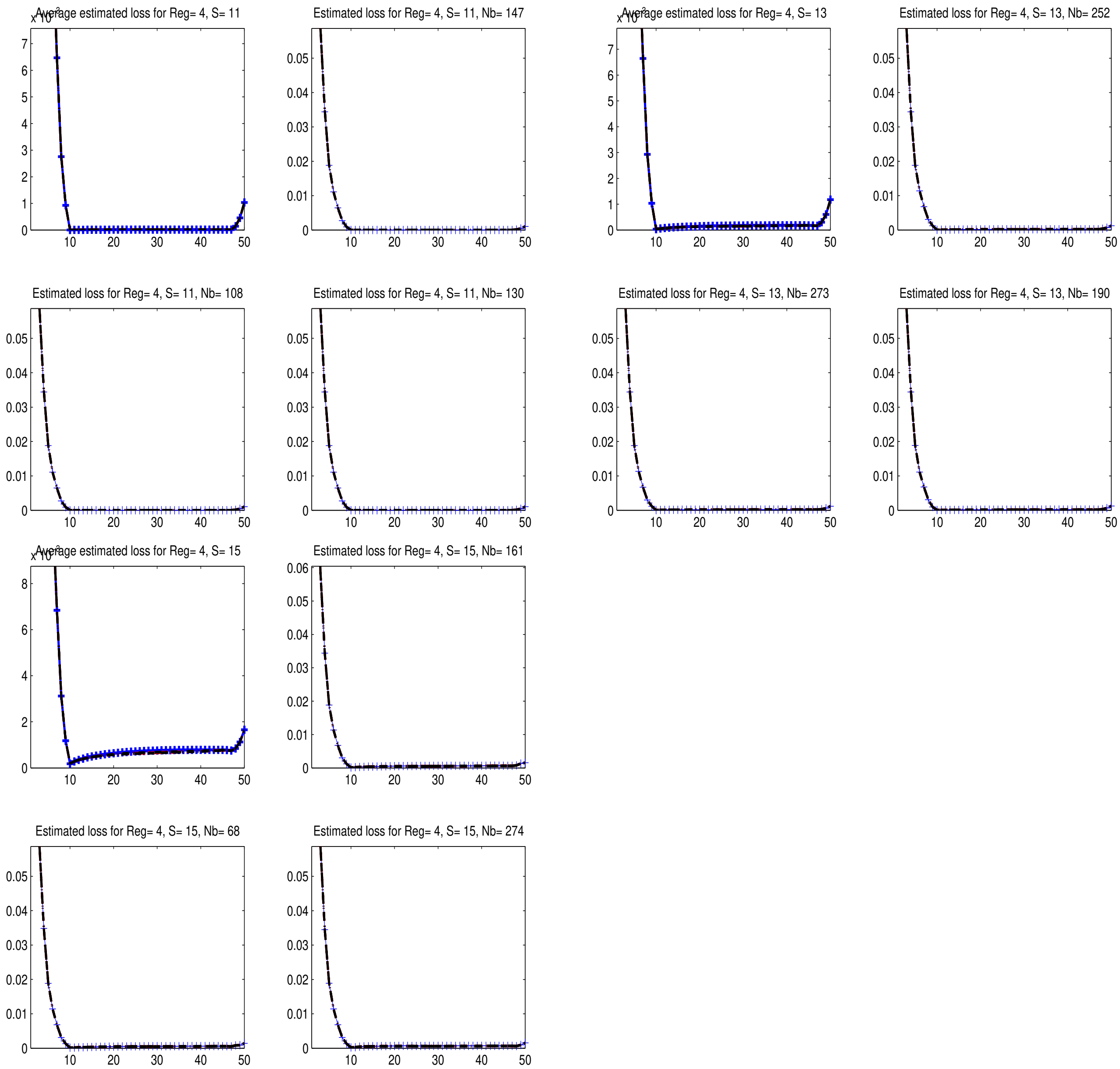

Figure 7.22: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_4, noise levels: $\sigma_{s,11}$, $\sigma_{s,13}$ et $\sigma_{s,15}$.

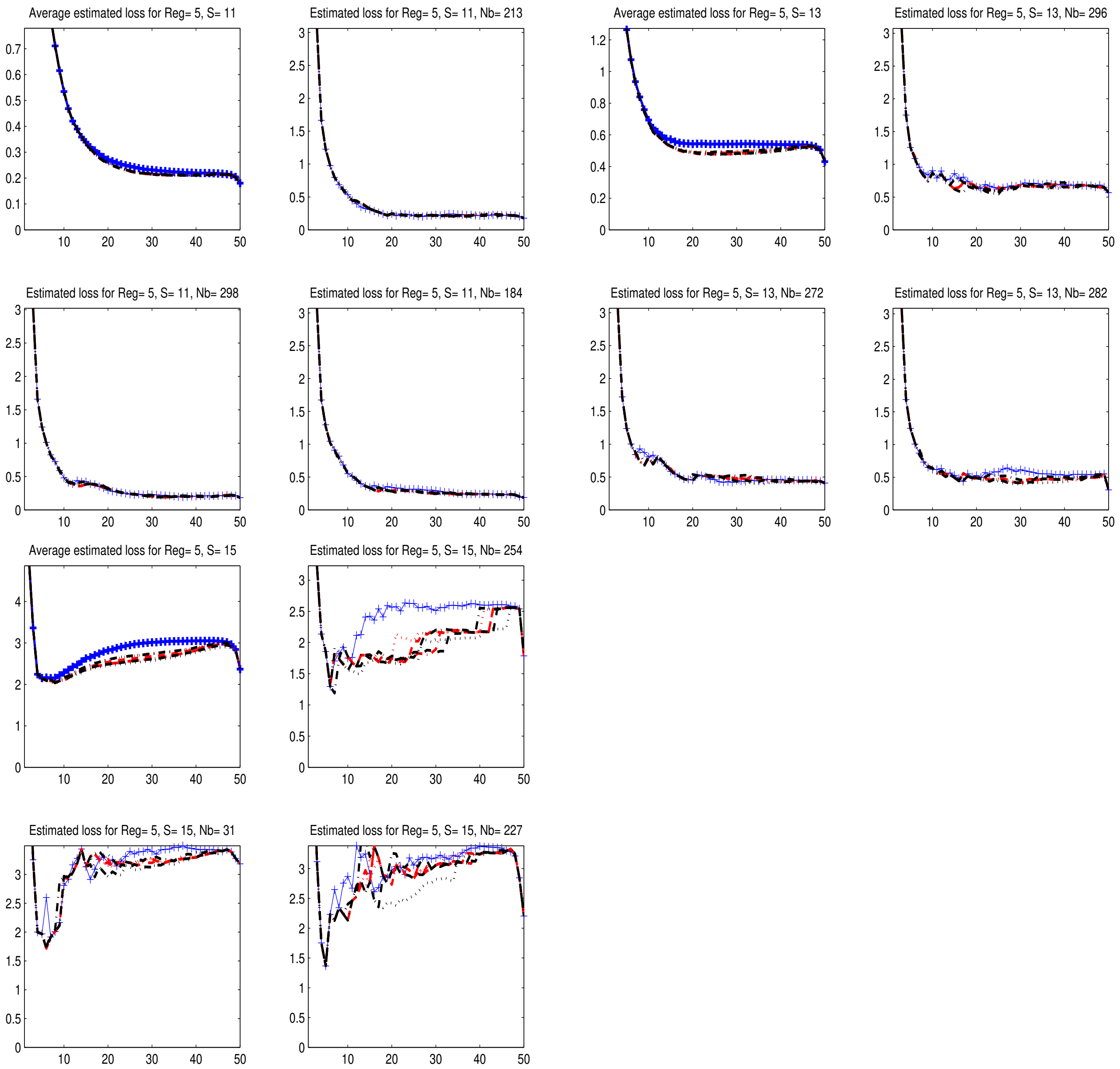

Figure 7.23: Graph of $D \mapsto \ell\left(s, \widehat{s}_{\widehat{m}(D)}\right)$ in the heteroscedastic case, where $(\widehat{m}(D))_D$ is chosen according to *Emp* (+ blue curve), Lpo_p with $p = 1, 20, 50$ (black curves) and penRad_C with $C = 1, 1.25, 1.5$ (red curves). Regression function: s_5, noise levels: $\sigma_{s,11}$, $\sigma_{s,13}$ et $\sigma_{s,15}$.

$s.$	$\sigma.$	Emp	Loo	Lpo20	Lpo50	Rad	Rad1.25	Rad1.5
1	c,1	**1** ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7e-017
	c,2	2.23 ± 0.23	2.2 ± 0.23	2.21 ± 0.23	**2.2** ± 0.23	2.21 ± 0.23	2.21 ± 0.23	2.21 ± 0.23
	c,3	4.15 ± 0.26	4.11 ± 0.26	4.12 ± 0.26	4.12 ± 0.26	4.12 ± 0.26	**4.1** ± 0.26	4.11 ± 0.26
	pc,4	**1** ± 8.8e-017	1 ± 8.8e-017	1 ± 8.8e-017	1 ± 8.8e-017	1 ± 8.8e-017	1 ± 8.8e-017	1 ± 8.8e-017
	pc,5	**1.01** ± 0.0061	1.01 ± 0.0078	1.37 ± 0.35	1.37 ± 0.35	1.37 ± 0.35	1.37 ± 0.35	1.37 ± 0.35
	pc,6	3.35 ± 0.41	3.41 ± 0.5	3.37 ± 0.49	**3.3** ± 0.47	3.31 ± 0.48	3.32 ± 0.47	3.3 ± 0.47
	pc,7	5.25 ± 0.75	**5.12** ± 0.75	6.15 ± 1.2	6.23 ± 1.2	6.15 ± 1.2	6.22 ± 1.2	6.2 ± 1.2
	pc,8	**5.83** ± 0.44	5.9 ± 0.44	5.95 ± 0.44	6.04 ± 0.43	6 ± 0.44	5.99 ± 0.44	5.95 ± 0.43
	pc,9	9.83 ± 0.82	8.59 ± 0.76	8.58 ± 0.77	8.65 ± 0.76	8.75 ± 0.77	8.56 ± 0.76	**8.45** ± 0.76
	s,10	**1** ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7,00E-17	1 ± 7e-017
	s,11	**1** ± 6.3e-017	1 ± 6.3e-017	1 ± 6.3e-017	1 ± 6.3e-017	1 ± 6.3e-017	1 ± 6.3e-017	1 ± 6.3e-017
	s,12	**1** ± 6.6e-017	1 ± 6.6e-017	1 ± 6.6e-017	1 ± 6.6e-017	1 ± 6.6e-017	1 ± 6.6e-017	1 ± 6.6e-017
	s,13	**1** ± 6.2e-017	1 ± 6.2e-017	1 ± 6.2e-017	1 ± 6.2e-017	1 ± 6.2e-017	1 ± 6.2e-017	1 ± 6.2e-017
	s,14	**1** ± 6.5e-017	1 ± 6.5e-017	1 ± 6.5e-017	1 ± 6.5e-017	1 ± 6.5e-017	1 ± 6.5e-017	1 ± 6.5e-017
	s,15	**1** ± 6.4e-017	1 ± 6.4e-017	1 ± 6.4e-017	1 ± 6.4e-017	1 ± 6.4e-017	1 ± 6.4e-017	1 ± 6.4e-017
	s,16	**1** ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017
	s,17	**1.14** ± 0.068	1.18 ± 0.076	1.18 ± 0.076	1.18 ± 0.075	1.18 ± 0.076	1.18 ± 0.076	1.18 ± 0.076
2	c,1	**1** ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017
	c,2	**1.88** ± 0.088	1.9 ± 0.091	1.89 ± 0.091	1.93 ± 0.091	1.9 ± 0.091	1.91 ± 0.09	1.92 ± 0.091
	c,3	**3.65** ± 0.11	3.77 ± 0.12	3.82 ± 0.12	4.07 ± 0.13	3.86 ± 0.13	3.92 ± 0.13	4 ± 0.13
	pc,4	**1** ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017	1 ± 6.8e-017
	pc,5	1.31 ± 0.12	1.34 ± 0.12	1.33 ± 0.12	**1.3** ± 0.12	1.31 ± 0.12	1.31 ± 0.12	1.31 ± 0.12
	pc,6	3.62 ± 0.24	3.6 ± 0.25	**3.54** ± 0.24	3.65 ± 0.24	3.59 ± 0.25	3.57 ± 0.24	3.63 ± 0.24
	pc,7	4.6 ± 0.3	**4.56** ± 0.3	4.69 ± 0.31	4.96 ± 0.31	4.75 ± 0.31	4.8 ± 0.31	4.9 ± 0.31
	pc,8	6.27 ± 0.35	**5.76** ± 0.33	5.84 ± 0.34	6.09 ± 0.36	5.97 ± 0.35	5.91 ± 0.34	5.9 ± 0.34
	pc,9	4.34 ± 0.16	4.18 ± 0.16	4.16 ± 0.15	4.13 ± 0.15	4.19 ± 0.16	4.13 ± 0.15	**4.08** ± 0.15
	s,10	**1** ± 4.6e-017	1 ± 4.6e-017	1 ± 4.6e-017	1 ± 4.6e-017	1 ± 4.6e-017	1 ± 4.6e-017	1 ± 4.6e-017
	s,11	**1** ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5e-017
	s,12	**1** ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017
	s,13	**1** ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017	1 ± 5.3e-017
	s,14	**1** ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5,00E-17	1 ± 5e-017
	s,15	**1** ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017	1 ± 4.9e-017
	s,16	**1** ± 4.7e-017	1 ± 4.7e-017	1 ± 4.7e-017	1 ± 4.7e-017	1 ± 4.7e-017	1 ± 4.7e-017	1 ± 4.7e-017
	s,17	1.09 ± 0.037	1.1 ± 0.039	1.1 ± 0.039	**1.09** ± 0.037	1.1 ± 0.039	1.1 ± 0.039	1.1 ± 0.039

Table 7.8: Ratio of the average loss obtained for model provided by 1*2Id over that of the pathwise oracle (minimizer of $\ell(s, \widehat{s}_m)$) for each regression function and each noise. $N = 300$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

$s.$	$\sigma.$	Emp	Lpo_1	Lpo_{20}	Lpo_{50}	penRad_1	$\text{penRad}_{1.25}$	$\text{penRad}_{1.5}$
3	c,1	2.33 ± 0.18	2.29 ± 0.17	2.29 ± 0.17	**2.27** ± 0.17	2.29 ± 0.17	2.29 ± 0.17	2.28 ± 0.17
	c,2	**4.87** ± 0.62	4.87 ± 0.62	4.87 ± 0.62	4.99 ± 0.62	4.9 ± 0.62	4.9 ± 0.62	4.96 ± 0.62
	c,3	**5.08** ± 0.37	5.11 ± 0.35	5.15 ± 0.35	5.34 ± 0.4	5.15 ± 0.35	5.22 ± 0.36	5.33 ± 0.4
	pc,4	**1** ± 7.4e-017	1 ± 7.4e-017	1 ± 7.4e-017	1 ± 7.4e-017	1 ± 7.4e-017	1 ± 7.4e-017	1 ± 7.4e-017
	pc,5	1.2 ± 0.081	1.2 ± 0.081	1.14 ± 0.056	**1.11** ± 0.044	1.14 ± 0.056	1.14 ± 0.056	1.11 ± 0.044
	pc,6	6.01 ± 0.53	4.14 ± 0.35	4.09 ± 0.35	3.75 ± 0.34	4.16 ± 0.35	3.82 ± 0.34	**3.66** ± 0.33
	pc,7	6.3 ± 0.61	5.27 ± 0.59	5.04 ± 0.58	4.51 ± 0.51	5.14 ± 0.59	4.65 ± 0.52	**4.43** ± 0.51
	pc,8	11.3 ± 1.1	8.86 ± 0.95	8.98 ± 0.95	9.19 ± 1.1	9.68 ± 1.1	9.12 ± 1	**8.76** ± 1
	pc,9	9.76 ± 0.73	8.54 ± 0.66	8.51 ± 0.66	8.61 ± 0.72	8.67 ± 0.66	8.53 ± 0.66	**8.37** ± 0.66
	s,10	**1** ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017
	s,11	**1** ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017
	s,12	**1** ± 6.1e-017	1 ± 6.1e-017	1 ± 6.1e-017	1 ± 6.1e-017	1 ± 6.1e-017	1 ± 6.1e-017	1 ± 6.1e-017
	s,13	**1** ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017
	s,14	**1** ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017	1 ± 5.6e-017
	s,15	**1** ± 5.1e-017	1 ± 5.1e-017	1 ± 5.1e-017	1 ± 5.1e-017	1 ± 5.1e-017	1 ± 5.1e-017	1 ± 5.1e-017
	s,16	1.44 ± 0.23	**1.4** ± 0.23	1.43 ± 0.23	1.44 ± 0.23	1.43 ± 0.23	1.43 ± 0.23	1.44 ± 0.23
	s,17	3.03 ± 0.2	2.8 ± 0.19	2.77 ± 0.18	**2.71** ± 0.18	2.78 ± 0.19	2.75 ± 0.18	2.72 ± 0.18
4	c,1	**2.66** ± 0.085	2.77 ± 0.091	2.82 ± 0.094	2.91 ± 0.095	2.82 ± 0.092	2.85 ± 0.094	2.88 ± 0.095
	c,2	**3.73** ± 0.1	3.8 ± 0.1	3.83 ± 0.1	4 ± 0.11	3.87 ± 0.1	3.92 ± 0.11	3.94 ± 0.11
	c,3	**3.92** ± 0.1	3.97 ± 0.1	4.01 ± 0.1	4.15 ± 0.1	4.04 ± 0.1	4.07 ± 0.1	4.11 ± 0.1
	pc,4	1.36 ± 0.12	1.24 ± 0.099	**1.23** ± 0.098	1.28 ± 0.1	1.31 ± 0.12	1.24 ± 0.098	1.26 ± 0.1
	pc,5	4.38 ± 0.23	3.68 ± 0.22	3.69 ± 0.22	3.68 ± 0.22	3.75 ± 0.22	3.67 ± 0.22	**3.61** ± 0.22
	pc,6	6.86 ± 0.41	**6.02** ± 0.38	6.1 ± 0.4	6.41 ± 0.44	6.4 ± 0.45	6.31 ± 0.43	6.26 ± 0.43
	pc,7	6.91 ± 0.34	6.22 ± 0.31	**6.19** ± 0.3	6.44 ± 0.31	6.44 ± 0.32	6.34 ± 0.31	6.26 ± 0.3
	pc,8	6.2 ± 0.23	5.73 ± 0.23	**5.7** ± 0.23	5.88 ± 0.23	5.83 ± 0.23	5.77 ± 0.23	5.73 ± 0.23
	pc,9	3.81 ± 0.17	**3.76** ± 0.17	3.78 ± 0.17	3.79 ± 0.17	3.8 ± 0.17	3.8 ± 0.17	3.77 ± 0.17
	s,10	**1** ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017	1 ± 5.5e-017
	s,11	**1** ± 5.9e-017	1 ± 5.9e-017	1 ± 5.9e-017	1 ± 5.9e-017	1 ± 5.9e-017	1 ± 5.9e-017	1 ± 5.9e-017
	s,12	**1** ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017
	s,13	**1** ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017	1 ± 5.7e-017
	s,14	**1.04** ± 0.03	1.05 ± 0.034	1.04 ± 0.033	1.07 ± 0.039	1.04 ± 0.03	1.06 ± 0.039	1.06 ± 0.039
	s,15	**1.36** ± 0.06	1.49 ± 0.071	1.51 ± 0.073	1.49 ± 0.07	1.5 ± 0.073	1.5 ± 0.073	1.49 ± 0.071
	s,16	**2.67** ± 0.088	2.72 ± 0.09	2.81 ± 0.093	2.91 ± 0.098	2.82 ± 0.093	2.83 ± 0.094	2.88 ± 0.096
	s,17	**3.32** ± 0.1	3.34 ± 0.1	3.39 ± 0.11	3.5 ± 0.11	3.4 ± 0.11	3.42 ± 0.11	3.45 ± 0.11
5	c,1	**1.78** ± 0.0021	1.78 ± 0.0021	1.78 ± 0.0021	1.78 ± 0.0021	1.78 ± 0.0021	1.78 ± 0.0021	1.78 ± 0.0021
	c,2	1.75 ± 0.0073	1.75 ± 0.0073	1.75 ± 0.0073	**1.75** ± 0.0072	1.75 ± 0.0073	1.75 ± 0.0073	1.75 ± 0.0073
	c,3	1.99 ± 0.015	1.99 ± 0.015	1.99 ± 0.015	**1.99** ± 0.015	1.99 ± 0.015	1.99 ± 0.015	1.99 ± 0.015
	pc,4	**1.78** ± 0.0041	1.78 ± 0.0041	1.78 ± 0.0041	1.78 ± 0.0041	1.78 ± 0.0041	1.78 ± 0.0041	1.78 ± 0.0041
	pc,5	1.77 ± 0.012	1.77 ± 0.012	1.77 ± 0.012	**1.77** ± 0.012	1.77 ± 0.012	1.77 ± 0.012	1.77 ± 0.012
	pc,6	2.32 ± 0.034	**2.22** ± 0.032	2.23 ± 0.033	2.26 ± 0.033	2.26 ± 0.033	2.24 ± 0.033	2.23 ± 0.033
	pc,7	2.86 ± 0.05	**2.67** ± 0.045	2.7 ± 0.045	2.74 ± 0.048	2.74 ± 0.047	2.72 ± 0.047	2.71 ± 0.046
	pc,8	3.62 ± 0.068	**3.32** ± 0.064	3.36 ± 0.065	3.43 ± 0.067	3.41 ± 0.066	3.38 ± 0.066	3.36 ± 0.066
	pc,9	4.64 ± 0.092	4.33 ± 0.093	4.36 ± 0.09	4.37 ± 0.09	4.43 ± 0.091	4.38 ± 0.092	**4.29** ± 0.091
	s,10	2.32 ± 0.019	2.31 ± 0.019	2.31 ± 0.019	2.31 ± 0.019	2.31 ± 0.019	2.31 ± 0.019	**2.31** ± 0.019
	s,11	2.77 ± 0.03	2.73 ± 0.029	2.72 ± 0.03	2.72 ± 0.03	2.73 ± 0.03	2.72 ± 0.03	**2.72** ± 0.03
	s,12	3.21 ± 0.044	3.07 ± 0.042	3.07 ± 0.043	3.08 ± 0.043	3.1 ± 0.043	3.08 ± 0.043	**3.06** ± 0.043
	s,13	3.65 ± 0.056	3.46 ± 0.053	3.47 ± 0.053	3.48 ± 0.055	3.51 ± 0.053	3.47 ± 0.053	**3.45** ± 0.053
	s,14	4.45 ± 0.082	4.12 ± 0.075	4.13 ± 0.076	4.16 ± 0.078	4.19 ± 0.078	4.12 ± 0.077	**4.08** ± 0.075
	s,15	4.65 ± 0.09	4.37 ± 0.087	4.38 ± 0.087	4.37 ± 0.087	4.42 ± 0.088	4.38 ± 0.088	**4.33** ± 0.087
	s,16	4.61 ± 0.11	**4.41** ± 0.1	4.42 ± 0.1	4.43 ± 0.1	4.44 ± 0.1	4.42 ± 0.1	4.42 ± 0.1
	s,17	**4.81** ± 0.14	4.84 ± 0.14	4.87 ± 0.14	4.99 ± 0.15	4.88 ± 0.14	4.9 ± 0.14	4.95 ± 0.15

Table 7.9: Ratio of the average loss obtained for model provided by 1*2Id over that of the pathwise oracle (minimizer of $\ell\left(s, \widehat{s}_m\right)$) for each regression function and each noise. $N = 300$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

$s_{\cdot}$	$\sigma_{c,\cdot}$	1Emp	1Loo	1Lpo_{20}	1Lpo_{50}	1penRad_1	$1\text{penRad}_{1.25}$	$1\text{penRad}_{1.5}$
1	1	3.17 ± 0.23	3.62 ± 0.31	3.07 ± 0.32	**3** ± 0.24	3.37 ± 0.28	3.54 ± 0.3	3.28 ± 0.25
	2	2.85 ± 0.2	**2.78** ± 0.18	3.42 ± 0.35	2.79 ± 0.18	3.08 ± 0.33	2.99 ± 0.3	3.08 ± 0.3
	3	5.82 ± 0.31	6.01 ± 0.35	**5.72** ± 0.29	5.74 ± 0.3	5.74 ± 0.3	5.79 ± 0.3	5.77 ± 0.29
2	1	2.24 ± 0.089	2.16 ± 0.086	2.18 ± 0.086	2.22 ± 0.095	2.23 ± 0.089	2.25 ± 0.088	**2.15** ± 0.086
	2	**2.72** ± 0.11	2.75 ± 0.11	2.77 ± 0.12	2.79 ± 0.12	2.86 ± 0.12	2.91 ± 0.12	2.87 ± 0.12
	3	**5.54** ± 0.16	5.87 ± 0.17	6.06 ± 0.17	7 ± 0.2	6 ± 0.17	6.14 ± 0.17	6.31 ± 0.18
3	1	4.42 ± 0.51	4.4 ± 0.51	4.21 ± 0.44	4.26 ± 0.5	4.05 ± 0.45	4.05 ± 0.44	**3.92** ± 0.42
	2	**5.85** ± 0.33	6.19 ± 0.39	6.1 ± 0.36	6.26 ± 0.38	6.03 ± 0.36	6.06 ± 0.36	6.11 ± 0.37
	3	7.11 ± 0.5	**6.9** ± 0.33	6.98 ± 0.34	7.57 ± 0.41	7.26 ± 0.53	7.34 ± 0.53	7.33 ± 0.52
4	1	4.41 ± 0.16	**4.41** ± 0.13	4.52 ± 0.14	4.55 ± 0.14	4.51 ± 0.14	4.56 ± 0.14	4.59 ± 0.14
	2	**5.05** ± 0.13	5.22 ± 0.12	5.26 ± 0.13	5.58 ± 0.12	5.33 ± 0.13	5.39 ± 0.13	5.48 ± 0.13
	3	**5.91** ± 0.15	6.09 ± 0.16	6.01 ± 0.15	6.65 ± 0.18	6.15 ± 0.16	6.25 ± 0.18	6.39 ± 0.18
5	1	**1.4** ± 0.0033	1.68 ± 0.0063	1.5 ± 0.0045	1.42 ± 0.0037	1.47 ± 0.0042	1.48 ± 0.0043	1.49 ± 0.0045
	2	**1.69** ± 0.0063	1.86 ± 0.0082	1.76 ± 0.0073	1.7 ± 0.0067	1.73 ± 0.0067	1.74 ± 0.0068	1.75 ± 0.0069
	3	2.21 ± 0.014	2.23 ± 0.013	2.21 ± 0.014	**2.2** ± 0.014	2.2 ± 0.014	2.2 ± 0.014	2.21 ± 0.014

Table 7.10: Ratio of the average loss obtained for model provided by $1{*}2\text{VF}_5$ over that of the pathwise oracle (minimizer of $\ell\left(s, \widehat{s}_m\right)$) for each regression function and each noise. $N = 500$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

$s_\cdot$	$\sigma_{s,\cdot}$	1Emp	1Loo	$1\mathrm{Lpo}_{20}$	$1\mathrm{Lpo}_{50}$	$1\mathrm{penRad}_1$	$1\mathrm{penRad}_{1.25}$	$1\mathrm{penRad}_{1.5}$
1	10	8.73 ± 0.57	**8.48** ± 0.52	8.97 ± 0.62	8.51 ± 0.61	8.93 ± 0.57	8.67 ± 0.56	8.74 ± 0.58
	11	8.94 ± 0.6	8.93 ± 0.52	8.34 ± 0.54	**8.17** ± 0.54	8.89 ± 0.57	8.75 ± 0.57	8.7 ± 0.56
	12	9.42 ± 0.65	**8.7** ± 0.59	9.61 ± 1.2	9.15 ± 0.94	9.36 ± 0.66	9.38 ± 0.69	9.19 ± 0.68
	13	7.94 ± 0.5	8.51 ± 0.75	8.76 ± 0.63	**7.3** ± 0.5	7.51 ± 0.54	7.32 ± 0.51	7.38 ± 0.51
	14	6.96 ± 0.45	6.88 ± 0.41	6.47 ± 0.39	**6.2** ± 0.38	7 ± 0.45	6.87 ± 0.44	6.8 ± 0.43
	15	7.43 ± 0.74	**5.73** ± 0.37	6.03 ± 0.41	5.75 ± 0.41	6.11 ± 0.44	6.27 ± 0.46	6.16 ± 0.44
	16	5.19 ± 0.5	5.18 ± 0.48	5.33 ± 0.48	4.9 ± 0.48	**4.61** ± 0.33	4.65 ± 0.33	4.94 ± 0.43
	17	3.05 ± 0.22	3.29 ± 0.26	3.2 ± 0.25	3.27 ± 0.27	**2.96** ± 0.24	3.11 ± 0.27	3.04 ± 0.27
2	10	3.91 ± 0.15	3.71 ± 0.14	3.9 ± 0.16	**3.67** ± 0.14	4.03 ± 0.16	4.01 ± 0.16	3.93 ± 0.15
	11	4.01 ± 0.17	3.74 ± 0.14	3.79 ± 0.14	**3.61** ± 0.13	3.66 ± 0.13	3.61 ± 0.13	3.63 ± 0.13
	12	3.98 ± 0.15	**3.86** ± 0.15	3.97 ± 0.15	3.94 ± 0.15	4.04 ± 0.16	3.94 ± 0.15	3.95 ± 0.15
	13	3.77 ± 0.15	3.49 ± 0.14	**3.4** ± 0.13	3.53 ± 0.15	3.63 ± 0.16	3.57 ± 0.15	3.57 ± 0.15
	14	3.54 ± 0.14	3.39 ± 0.16	**3.25** ± 0.14	5.2 ± 2	6.65 ± 3.2	6.55 ± 3.2	6.49 ± 3.2
	15	3.35 ± 0.14	**3.19** ± 0.12	3.29 ± 0.14	3.89 ± 0.56	3.28 ± 0.14	3.27 ± 0.14	3.26 ± 0.14
	16	2.85 ± 0.13	**2.77** ± 0.13	2.8 ± 0.13	2.96 ± 0.15	2.9 ± 0.13	2.86 ± 0.12	2.84 ± 0.13
	17	2.4 ± 0.12	2.35 ± 0.12	**2.32** ± 0.12	2.35 ± 0.12	2.4 ± 0.12	2.43 ± 0.12	2.34 ± 0.12
3	10	9.1 ± 0.55	8.93 ± 0.58	9.04 ± 0.58	8.81 ± 0.61	8.85 ± 0.54	**8.68** ± 0.56	8.73 ± 0.57
	11	9.24 ± 0.55	9.3 ± 0.54	**8.84** ± 0.6	9.12 ± 0.71	9.37 ± 0.72	9.55 ± 0.73	9.48 ± 0.72
	12	10.7 ± 0.98	9.94 ± 0.7	9.27 ± 0.59	**8.29** ± 0.5	9.37 ± 0.72	9.29 ± 0.68	9.3 ± 0.67
	13	7.6 ± 0.53	7.28 ± 0.51	**6.94** ± 0.51	7.25 ± 0.5	7.6 ± 0.55	7.48 ± 0.55	7.27 ± 0.52
	14	7.25 ± 0.73	6.49 ± 0.7	7.03 ± 0.7	**6.39** ± 0.82	6.41 ± 0.73	6.52 ± 0.73	6.44 ± 0.7
	15	5.77 ± 0.56	5.89 ± 0.45	6.01 ± 0.55	**5.23** ± 0.4	5.73 ± 0.47	5.61 ± 0.46	5.69 ± 0.44
	16	4.25 ± 0.47	3.82 ± 0.29	3.74 ± 0.29	**3.48** ± 0.25	3.8 ± 0.3	3.64 ± 0.3	3.53 ± 0.29
	17	5.4 ± 0.35	5.52 ± 0.35	5.37 ± 0.33	**5.09** ± 0.29	5.21 ± 0.32	5.13 ± 0.32	5.45 ± 0.37
4	10	6.6 ± 0.18	5.91 ± 0.16	5.9 ± 0.16	6.16 ± 0.17	6.06 ± 0.17	5.98 ± 0.17	**5.89** ± 0.17
	11	6.78 ± 0.2	**5.77** ± 0.16	5.98 ± 0.17	6.19 ± 0.18	6.08 ± 0.18	5.99 ± 0.17	5.93 ± 0.17
	12	6.38 ± 0.17	**5.64** ± 0.16	5.78 ± 0.16	6 ± 0.16	6.02 ± 0.17	5.93 ± 0.16	5.86 ± 0.16
	13	5.77 ± 0.17	5.19 ± 0.15	**5.19** ± 0.15	5.4 ± 0.16	5.56 ± 0.16	5.49 ± 0.16	5.43 ± 0.16
	14	5.64 ± 0.18	5.06 ± 0.15	**4.99** ± 0.15	5.41 ± 0.17	5.2 ± 0.16	5.1 ± 0.16	5.05 ± 0.16
	15	5.5 ± 0.26	**4.98** ± 0.21	5.12 ± 0.21	4.99 ± 0.22	5.16 ± 0.22	5.1 ± 0.21	5.05 ± 0.21
	16	5.06 ± 0.2	**5.05** ± 0.18	5.16 ± 0.18	5.31 ± 0.2	5.14 ± 0.18	5.15 ± 0.18	5.18 ± 0.18
	17	4.98 ± 0.13	5.05 ± 0.15	4.98 ± 0.14	5.08 ± 0.14	4.93 ± 0.13	**4.88** ± 0.13	4.95 ± 0.13
5	10	2.18 ± 0.013	2.22 ± 0.013	2.17 ± 0.013	**2.14** ± 0.013	2.17 ± 0.013	2.17 ± 0.013	2.17 ± 0.013
	11	2.9 ± 0.024	**2.76** ± 0.023	2.77 ± 0.023	2.81 ± 0.023	2.8 ± 0.023	2.79 ± 0.023	2.77 ± 0.023
	12	3.75 ± 0.038	**3.45** ± 0.036	3.49 ± 0.036	3.57 ± 0.037	3.55 ± 0.037	3.52 ± 0.037	3.49 ± 0.036
	13	4.37 ± 0.049	**3.98** ± 0.045	4 ± 0.045	4.1 ± 0.046	4.09 ± 0.045	4.05 ± 0.046	4.03 ± 0.046
	14	6.12 ± 0.088	**5.46** ± 0.08	5.49 ± 0.078	5.6 ± 0.081	5.68 ± 0.082	5.62 ± 0.082	5.57 ± 0.083
	15	6.68 ± 0.13	**6.32** ± 0.13	6.38 ± 0.13	6.48 ± 0.13	6.39 ± 0.13	6.36 ± 0.13	6.34 ± 0.13
	16	6.52 ± 0.15	**6.21** ± 0.14	6.39 ± 0.14	6.53 ± 0.14	6.43 ± 0.15	6.4 ± 0.15	6.39 ± 0.15
	17	**7.02** ± 0.19	7.04 ± 0.2	7.21 ± 0.2	7.41 ± 0.21	7.24 ± 0.22	7.26 ± 0.21	7.33 ± 0.22

Table 7.11: Ratio of the average loss obtained for model provided by $1{*}2\mathrm{VF}_5$ over that of the pathwise oracle (minimizer of $\ell(s, \widehat{s}_m)$) for each regression function and each noise. $N = 500$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

$s.$	$\sigma_{pc,\cdot}$	1Emp	1Loo	1Lpo_{20}	1Lpo_{50}	1penRad_{1}	$1\text{penRad}_{1.25}$	$1\text{penRad}_{1.5}$
1	4	6.3 ± 0.55	6.35 ± 0.55	6.03 ± 0.51	7.02 ± 1.1	6.18 ± 0.61	**6.02** ± 0.57	6.04 ± 0.55
	5	4.81 ± 0.47	4.56 ± 0.46	4.62 ± 0.49	4.54 ± 0.47	4.48 ± 0.43	**4.38** ± 0.42	4.44 ± 0.43
	6	6.45 ± 0.55	6.51 ± 0.55	6.43 ± 0.55	**6.41** ± 0.54	6.66 ± 0.55	6.72 ± 0.55	6.76 ± 0.55
	7	7.54 ± 0.61	**7.19** ± 0.55	7.51 ± 0.61	8.21 ± 0.84	7.26 ± 0.58	7.26 ± 0.59	7.31 ± 0.62
	8	10.9 ± 1.1	**10.1** ± 0.84	10.4 ± 0.9	11.9 ± 0.99	10.4 ± 0.86	10.7 ± 0.87	10.6 ± 0.88
	9	13.6 ± 1.2	**12.8** ± 1	13.1 ± 1.1	13.3 ± 1.1	12.9 ± 1.1	13 ± 1.1	13.1 ± 1.2
2	4	3.93 ± 0.24	3.84 ± 0.23	3.85 ± 0.23	**3.48** ± 0.2	4.06 ± 0.24	4 ± 0.24	3.8 ± 0.22
	5	3.66 ± 0.29	3.86 ± 0.31	3.6 ± 0.29	3.66 ± 0.28	**3.54** ± 0.28	3.67 ± 0.29	3.63 ± 0.28
	6	**6.19** ± 0.36	6.31 ± 0.36	6.69 ± 0.4	6.97 ± 0.39	6.53 ± 0.4	6.55 ± 0.4	6.66 ± 0.41
	7	6.96 ± 0.36	**6.93** ± 0.32	7.1 ± 0.32	8.37 ± 0.43	7.6 ± 0.4	7.71 ± 0.41	7.76 ± 0.4
	8	9.7 ± 0.38	**8.84** ± 0.35	9 ± 0.4	9.45 ± 0.37	9.14 ± 0.36	9.08 ± 0.37	8.98 ± 0.36
	9	5.78 ± 0.21	5.92 ± 0.21	**5.64** ± 0.2	5.81 ± 0.21	5.82 ± 0.21	5.79 ± 0.21	5.82 ± 0.2
3	4	6.35 ± 0.68	6.44 ± 0.66	6.23 ± 0.6	5.84 ± 0.61	5.66 ± 0.61	**5.2** ± 0.56	5.69 ± 0.79
	5	8 ± 1.3	7.58 ± 1.2	7.31 ± 1.1	7.27 ± 1.2	**6.75** ± 1.1	6.94 ± 1.1	6.85 ± 1.1
	6	13.8 ± 1.4	12.5 ± 1.5	11 ± 1.2	**10.6** ± 0.88	11.9 ± 1.4	11.8 ± 1.4	11.2 ± 1.3
	7	16.5 ± 1.9	12.1 ± 1.3	11.6 ± 1.3	**11.4** ± 1.7	13.3 ± 2.1	12.5 ± 1.9	11.8 ± 1.7
	8	23.9 ± 1.9	20.7 ± 1.9	21.8 ± 2.2	20.2 ± 1.9	21.6 ± 2.2	20.6 ± 2.2	**20.1** ± 2.1
	9	18.2 ± 1.5	18.5 ± 1.5	18.8 ± 1.6	18.2 ± 1.6	18.3 ± 1.6	**17.6** ± 1.5	17.9 ± 1.5
4	4	5.81 ± 0.32	**4.78** ± 0.24	5.09 ± 0.29	4.89 ± 0.26	5.1 ± 0.27	4.94 ± 0.27	4.96 ± 0.27
	5	7.5 ± 0.34	6.51 ± 0.31	**6.47** ± 0.3	6.86 ± 0.4	6.74 ± 0.36	6.73 ± 0.37	6.77 ± 0.37
	6	9.7 ± 0.35	**8.65** ± 0.34	8.85 ± 0.37	9.46 ± 0.37	8.94 ± 0.35	8.8 ± 0.34	8.89 ± 0.34
	7	10.6 ± 0.38	**9.7** ± 0.36	9.85 ± 0.4	10.7 ± 0.44	10.1 ± 0.41	9.99 ± 0.4	9.72 ± 0.38
	8	9.55 ± 0.32	**9.11** ± 0.32	9.23 ± 0.31	9.59 ± 0.31	9.23 ± 0.31	9.15 ± 0.31	9.19 ± 0.31
	9	5.41 ± 0.26	5.42 ± 0.29	**5.31** ± 0.26	5.4 ± 0.25	5.58 ± 0.34	5.66 ± 0.35	5.74 ± 0.37
5	4	**1.4** ± 0.0035	1.64 ± 0.0068	1.48 ± 0.0048	1.4 ± 0.0038	1.44 ± 0.0044	1.46 ± 0.0047	1.47 ± 0.0047
	5	**1.59** ± 0.0074	1.8 ± 0.0094	1.69 ± 0.009	1.6 ± 0.0076	1.65 ± 0.008	1.66 ± 0.0082	1.67 ± 0.0084
	6	2.6 ± 0.028	**2.49** ± 0.027	2.51 ± 0.028	2.54 ± 0.028	2.51 ± 0.028	2.52 ± 0.029	2.51 ± 0.027
	7	3.21 ± 0.041	**2.91** ± 0.036	2.97 ± 0.039	3.05 ± 0.039	3.02 ± 0.039	2.99 ± 0.039	2.98 ± 0.039
	8	4.6 ± 0.069	**4.02** ± 0.061	4.08 ± 0.063	4.3 ± 0.067	4.23 ± 0.067	4.17 ± 0.066	4.11 ± 0.066
	9	6.61 ± 0.15	**6.12** ± 0.14	6.16 ± 0.14	6.2 ± 0.14	6.32 ± 0.14	6.23 ± 0.14	6.19 ± 0.14

Table 7.12: Ratio of the average loss obtained for model provided by $1{*}2\text{VF}_5$ over that of the pathwise oracle (minimizer of $\ell\left(s, \widehat{s}_m\right)$) for each regression function and each noise. $N = 500$ repetitions have been made. Next to each value is indicated the corresponding empirical standard deviation.

Bibliography

[1] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to Unknown Sparsity by controlling the False Discovery Rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[2] H. Akaike. Statistical predictor identification. *Ann. Inst. Statisti. Math.*, 22:203–217, 1969.

[3] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

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

[5] S. Arlot. Model selection by resampling penalization. *Electronic journal of Statistics*, 00:00, 2008.

[6] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[7] Sylvain Arlot. Suboptimality of penalties linear in the dimension for model selection in heteroscedastic regression, June 2008. In preparation.

[8] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566.

[9] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[10] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with unknown variance. *The Annals of Statistics*, 00:00, 2008.

[11] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

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

[13] M. Basseville and N. Nikiforov. *The Detection of Abrupt Changes - Theory and Applications*. Prentice-Hall: Information and System Sciences Series, 1993.

[14] R. E. Bellman and S. E. Dreyfus. *Applied Dynamic Programming*. Princeton, 1962.

[15] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgensen, and G. Yang, editors, *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

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

[17] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 2006.

[18] B. Brodsky and B. Darkhovsky. *Methods in Change-point problems*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[19] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

[20] S. Dudoit and M. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

[21] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[22] Xavier Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression, 2008.

[23] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of non-parametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[24] M. Lavielle. Detection of multiple changes in a sequence of dependent variables. *Stocha. Proc. Appl.*, 83:79–102, 1999.

[25] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85:1501–1510, 2005.

[26] M. Lavielle and E. Lebarbier. An application of MCMC methods to the change-points problem. *Signal Processing*, 81:39–53, 2001.

[27] M. Lavielle and C. Ludena. The multiple change-points problem for spectral distribution. *Brnoulli*, 6(5):845–869, 2000.

[28] M. Lavielle and E. Moulines. Détection de ruptures multiples dans la moyenne d'un processus aléatoire. In *C.R. Acad. Sci. Paris*, volume t. 324 of *Série 1*, pages 239–243, 1997.

[29] M. Lavielle and G. Teyssière. Detection of Multiple Change-Points in Multivariate Time Series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.

[30] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

[31] K.-C. Li. Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975, 1987.

[32] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864, 1999.

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

[34] J. S. Marron, S. Adak, I. M. Johnstone, M. H. Neumann, and P. Patil. Exact Risk Analysis of Wavelet Regression. *Journal of Computational and Graphical Statistics*, 7(3):278–309, 1998.

[35] B. Q. Mia and L. C. Zhao. On detection of change points when the number is unknown. *Chinese J. Appl. Probab. Statist.*, 9(2):138–145, 1993.

[36] D. Picard. Testing and estimating change points in time series. *J. Appl. Probab.*, 17:841–867, 1985.

[37] F. Picard. *Process segmentation/clustering Application to the analysis of array CGH data*. PhD thesis, Université Paris-Sud 11, 2005.

[38] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 27(6):electronic access, 2005.

[39] Franck Picard, Stéphane Robin, Émilie Lebarbier, and Jean-Jacques Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 2007. To appear. doi:10.1111/j.1541-0420.2006.00729.x.

[40] Marie Sauvé. Histogram selection in non gaussian regression. Technical Report 5911, INRIA, may 2006.

[41] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[42] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[43] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[44] 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.

[45] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *JRSS B*, 39(1):44–47, 1977.

[46] M. Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003.

[47] Y. Yang. Regression with multiple candidate model: selection or mixing? *Statist. Sinica*, 13:783–809, 2003.

[48] Y. Yao. Estimating the number of change-points via Schwarz criterion. *Statist. Probab. Lett.*, 6:181–189, 1988.

[49] P. Zhang. Model selection via multifold cross-validation. *The Annals of Statistics*, 21(1):299–313, 1993.

L'objet de cette thèse est l'étude d'un certain type d'algorithmes de rééchantillonnage regroupés sous le nom de validation-croisée, et plus particulièrement parmi eux, du leave-p-out. Très utilisés en pratique, ces algorithmes sont encore mal compris d'un point de vue théorique, notamment sur un plan non-asymptotique. Notre analyse du leave-p-out s'effectue dans les cadres de l'estimation de densité et de la régression. Son objectif est de mieux comprendre la validation-croisée en fonction du cardinal p de l'ensemble test dont elle dépend.
D'un point de vue général, la validation-croisée est destinée à estimer le risque d'un estimateur. Dans notre cas, le leave-p-out n'est habituellement pas applicable en pratique, à cause d'une trop grande complexité algorithmique. Pourtant, nous parvenons à obtenir des formules closes (parfaitement calculables) de l'estimateur leave-p-out du risque, pour une large gamme d'estimateurs très employés.
Nous envisageons le problème de la sélection de modèles par validation-croisée sous deux aspects. L'un repose sur l'estimation optimale du risque en termes d'un compromis biais-variance, ce qui donne lieu à une procédure d'estimation de densité basée sur un choix de p entièrement fondé sur les données. Une application naturelle au problème des tests multiples est envisagée. L'autre aspect est lié à l'interprétation de l'estimateur validation-croisée comme critère pénalisé. Sur le plan théorique, la qualité de la procédure leave-p-out est garantie par des inégalités oracle ainsi qu'un résultat d'adaptativité dans le cadre de l'estimation de densité.
Le problème de la détection de ruptures est également abordé au travers d'une vaste étude de simulations, basée sur des considérations théoriques. Sur cette base, nous proposons une procédure entièrement tournée vers le rééchantillonnage, permettant de traiter le cas difficile de données hétéroscédastiques avec une complexité algorithmique raisonnable.

Mots-Clés : Rééchantillonnage, Validation-croisée, Leave-p-out, Statistique non-paramétrique, Sélection de modèles, Inégalité oracle, Adaptativité, Estimation de densité, Détection de ruptures, Tests multiples, FDR.

In this thesis, we aim at studying a family of resampling algorithms, referred to as cross-validation, and especially of one of them named leave-p-out. Extensively used in practice, these algorithms remain poorly understood, especially in the non-asymptotic framework. Our analysis of the leave-p-out algorithm is carried out both in density estimation and regression. Its main concern is to better understand cross-validation with respect to the cardinality p of the test set it relies on.
From a general point of view, cross-validation is devoted to estimate the risk of an estimator. Usually due to a prohibitive computational complexity, the leave-p-out is intractable. However, we turned it into a feasible procedure thanks to closed-form formulas for the risk estimator of a wide range of widespread estimators.
Besides, the question of model selection *via* cross-validation is considered through two approaches. The first one relies on the optimal estimation of the risk in terms of a bias-variance tradeoff, which results in a density estimation procedure based on a fully data-driven choice of p. This procedure is successfully applied to the multiple testing problem. The second approach is related to the interpretation of cross-validation in terms of penalized criterion. The quality of the leave-p-out procedure is theoretically assessed through oracle inequalities as well as an adaptivity result in the density estimation setup.
The change-points detection problem is another concern of this work. It is explored through an extensive simulation study based on theoretical considerations. From this, we propose a fully resampling-based procedure, which enables to deal with the hard problem of heteroscedasticity, while keeping a reasonable computational complexity.

Keywords: Resampling, Cross-validation, Leave-p-out, Nonparametric statistics, Model selection, Oracle inequality, Adaptivity, Density estimation, Change-points detection, Multiple testing, False Discovery Rate.