Model selection in density estimation via cross-validation

Alain Celisse

Density estimation *via* cross-validation: Model selection point of view

Alain Celisse

April 14, 2009

Abstract

The problem of model selection by cross-validation is addressed in the density estimation framework. Extensively used in practice, cross-validation (CV) remains poorly understood, especially in the non-asymptotic setting which is the main concern of this work.
A recurrent problem with CV is the computation time it involves. This drawback is overcome here thanks to closed-form expressions for the CV estimator of the risk for a broad class of widespread estimators: projection estimators.
In order to shed new lights on CV procedures with respect to the cardinality p of the test set, the CV estimator is interpreted as a penalized criterion with a random penalty. For instance, the amount of penalization is shown to increase with p.
A theoretical assessment of the CV performance is carried out thanks to two oracle inequalities applying to respectively bounded or square-integrable densities. For several collections of models, adaptivity results with respect to Hölder and Besov spaces are derived as well.

Keywords: Density estimation, cross-validation, model selection, leave-p-out, random penalty, oracle inequality, projection estimators, adaptivity in the minimax sense, Hölder, Besov.

1 Introduction

The main concern of this paper is the analysis of cross-validation procedures when employed to perform model selection in the density estimation context. This analysis results in a new understanding of CV behaviour as well as in several optimality results. Before entering into details, let us briefly describe related works in the model selection area.

1.1 Model selection

Model selection via penalization has been introduced by the seminal papers of Mallows (1973) on C_p, and Akaike (1973) about AIC, and also by Schwarz (1978) who proposed the BIC criterion. AIC and BIC have an asymptotic flavour, which makes their performance

depend on the model collection in hand as well as on the sample size (see Baraud et al., 2009).

More recently, Birgé and Massart (1997, 2001, 2006) have developed a non-asymptotic approach, inspired from the pioneering work of Barron and Cover (1991). It aims at choosing a model among a countable family $\{S_m\}_{m\in\mathcal{M}_n}$ where $\mathcal{M}_n$ is allowed to depend on the sample size n. From this point of view, an estimator $\widehat{s}_m$ is associated with each model S_m, and a penalized criterion is designed and then minimized to provide a final estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$. The goal of this approach is *efficiency*, that is the risk of $\widetilde{s}$ is as small as the smallest achievable risk by any of the estimators in the collection. Actually, this cannot be reached in the non-asymptotic setting and the quality assessment of the procedure is made through an oracle inequality. Such an inequality instead asserts that the risk of $\widetilde{s}$ is *almost* the same as that of the smallest achievable one up to a multiplicative constant $C_n \geq 1$ and a remainder term. When C_n converges to 1 as n tends to infinity, the model selection procedure is said *asymptotically efficient*.

In the density estimation framework, Barron et al. (1999) developed a general approach based on deterministic penalties, leading to an oracle inequality involving Kullback-Leibler divergence and Hellinger distance. This result has been adapted to the particular case of histograms by Castellan (1999, 2003) and further studied in Birgé and Rozenholc (2006). With the quadratic risk, the penalties proposed by Birgé and Massart (1997) and Barron et al. (1999) also enjoy some optimality properties when applied to projection estimators. The resulting estimators exhibit some adaptivity in the minimax sense with respect to Besov spaces for several appropriate functional bases (see Birgé and Massart, 1997).

1.2 Cross-validation

Unlike the aforementioned approaches relying on some deterministic penalties, the main concern of the present work is the use of cross-validation (CV) as a model selection procedure in the density estimation context. "Cross-validation" refers to a family of resampling-based procedures, resulting from a heuristic argument. The cross-validation procedures have been first studied in a regression context by Stone (1974, 1977) for the leave-one-out (Loo) and Geisser (1974, 1975) for the V-fold cross-validation (VFCV), and by Rudemo (1982) and Stone (1984) in the density estimation framework.
Since these algorithms can be computationally demanding or even intractable, Rudemo (1982) and Bowman (1984) 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 (2008b) to the leave-p-out cross-validation (Lpo).

Most of theoretical results about the performance of CV procedures are asymptotic and mainly concern the regression framework. For a fixed model, Burman (1989, 1990) expands several CV estimators of the risk of $\widehat{s}_m$ and concludes that Loo is the best one in terms of bias and variance. Besides several comparisons are pursued between CV and various penalized criteria: Li (1987) and Zhang (1993) in view of *asymptotic efficiency*, and Shao

(1993) and Yang (2007) on *model consistency*, that is recover the "true model". Interested readers are referred to Shao (1997) for an extensive review about asymptotic optimality properties in terms of efficiency and model consistency of some penalized criteria as well as CV procedures.

As for non-asymptotic results in the density setting, Birgé and Massart (1997) have settled an oracle inequality that relies on a conjecture and may be applied to the Loo procedure. However to the best of our knowledge, no result of this type has already been proved for the Lpo procedure in the density estimation setup. Recently in the regression setting, Arlot (2007b) established oracle inequalities for V-fold penalties, while Arlot and Celisse (2009) have carried out an extensive simulation study in the change-point detection problem with heteroscedastic data.

1.3 Main contributions

The present paper is devoted to study CV procedures as a means to perform model selection in the density estimation framework.

A constant drawback of CV—and resampling strategies in general—is the computation time such procedures involve. Indeed pursuing Loo with a large data set can be computationally prohibitive. Closed-form expressions are provided for the Lpo estimator of the L^2-risk of the broad class of projection estimators, demonstrating the wide applicability of these results. More insight is given into the behaviour of CV risk estimators thanks to these expressions, which drastically reduce the computation time.

CV estimator is then embedded into the penalized criterion framework. It emphasizes the tight relationship between the choice of p and the amount of penalization resulting from this choice. In the model selection setting, the interest of choosing $p > 1$ raises as a way to balance overfitting phenomenon.

Several non-asymptotic optimality results are also derived in terms of oracle inequality as well as adaptivity results in the minimax sense. To the best of our knowledge, these are the first theoretical non-asymptotic results of this type applying to Lpo in the density estimation setting.

The paper is organized as follows. The next section describes the statistical framework and notation. CV is presented as a special case of resampling procedures and some examples of famous CV procedures are provided. Closed-form expressions are then derived in Section 3 with several examples. Some bias and variance calculations are also yielded for various CV risk estimators.

The main concern of Section 4 is model selection. The Lpo estimator of the risk is interpreted as a penalized criterion with a random penalty. The amount of penalization is quantified with respect to p, which stresses the interest of choosing $p > 1$ as a means to overcome overfitting. Two oracle inequalities are then derived that warranty the good non-asymptotic performance of Lpo as model selection procedure with polynomial collections of models.

Section 5 is devoted to adaptivity results in the minimax sense with respect to Hölder as well as Besov spaces. Different collections of models are considered such as piecewise and trigonometric polynomials. A discussion with some possible prospects then follows in Section 6. Finally, proofs are collected in Section 7.

2 Leave-p-out cross-validation

Resampling-based strategies such as CV are usually time-consuming and can even be computationally intractable. The interest of the forthcoming approach is to derive closed-form expressions for the CV-based estimator of the risk of projection estimators, which are widespread in the density estimation community (Rudemo (1982); Donoho et al. (1996); Birgé and Massart (1997); Barron et al. (1999)).

First, the statistical framework is described. A definition of projection estimators is yielded and illustrated by several examples. Second, the CV heuristics is detailed with an emphasis on the relationship between CV and resampling procedures. Several famous CV procedures are also recalled.

2.1 Statistical framework

Let us start with introducing the framework and some notation which are repeatedly used throughout the paper.

2.1.1 Notation

In the sequel, $X_1, \ldots, X_n \in [0,1]$ are independent and identically distributed random variables drawn from a probability distribution P of density $s \in L^2([0,1])$ with respect to Lebesgue's measure on $[0,1]$.

Let $\mathcal{S}^*$ denote the set of mesurable functions on $[0,1]$. The distance between s and any $u \in \mathcal{S}^*$ is measured thanks to the quadratic *loss* denoted by $\ell(\cdot,\cdot)$ satisfying

$$\ell : \ (s,u) \mapsto \ell(s,u) := \|s-u\|^2 .$$

Since this quantity depends on s that is unknown, let us introduce the associated *contrast* function

$$\gamma : \ (u,x) \mapsto \gamma(u,x) := \|u\|^2 - 2u(x).$$

This contrast is related to the loss function by $\ell(s,u) = P\gamma(u) - P\gamma(s)$, where $P\gamma(u) = \mathbb{E}[\gamma(u,X)]$ and $X \sim P$ for any $u \in \mathcal{S}^*$. The *empirical risk* at point $u \in \mathcal{S}^*$, which estimates $\ell(s,u)$ up to a constant term, is defined by

$$\gamma_n(u) := P_n\gamma(u) = \frac{1}{n}\sum_{i=1}^{n} \gamma(u,X_i),$$

where $P_n = 1/n \sum_{i=1}^{n} \delta_{X_i}$ denotes the empirical measure. The quality assessment of an estimator $\widehat{s} = \widehat{s}(X_1, \ldots, X_n)$ of s is made through the corresponding *quadratic risk*

$$R_n(\widehat{s}) := \mathbb{E}\left[\ell(s, \widehat{s})\right] = \mathbb{E}\left[\|s - \widehat{s}\|^2\right].$$

Let $\mathcal{M}_n$ denote a countable set of indices. For every $m \in \mathcal{M}_n$, S_m is a set of candidate functions to estimate s, called a *model* in the following. Since every model S_m is uniquely determined by its index, m is also called a model.

In every model S_m, $\widehat{s}_m$ denotes an estimator of s defined as the empirical risk minimizer over S_m

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

The resulting collection of estimators $\{\widehat{s}_m\}_{m \in \mathcal{M}_n}$ corresponds to the collection of models $\{S_m\}_{m \in \mathcal{M}_n}$.

2.1.2 Projection estimators

Let Λ_n be a set of countable indices and $\{\varphi_\lambda\}_{\lambda \in \Lambda_n}$ a family of vectors in $L^2([0,1])$ such that for every $m \in \mathcal{M}_n$, there exists $\Lambda(m) \subset \Lambda_n$ and $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$ is an orthonormal family of $L^2([0,1])$. Then, let S_m denote the linear space spanned by $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$ and $D_m = \dim(S_m)$ for every m.

The orthogonal projection of s onto S_m is denoted by s_m

$$s_m := \text{Argmin}_{u \in S_m} P\gamma(u) = \sum_{\lambda \in \Lambda(m)} P\varphi_\lambda \, \varphi_\lambda, \qquad \text{with} \qquad P\varphi_\lambda = \mathbb{E}\left[\varphi_\lambda(X)\right].$$

Definition 2.1. *An estimator $\widehat{s} \in L^2([0,1])$ is a projection estimator if there exists a family $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ of orthonormal vectors of $L^2([0,1])$ such that*

$$\widehat{s} = \sum_{\lambda \in \Lambda} \alpha_\lambda \, \varphi_\lambda, \qquad \text{with} \qquad \alpha_\lambda = \frac{1}{n} \sum_{\lambda \in \Lambda} H_\lambda(X_i),$$

where $\{H_\lambda(\cdot)\}_{\lambda \in \Lambda}$ depends on the family $\{\varphi_\lambda\}_{\lambda \in \Lambda}$.

Therefore, it turns out that for every $m \in \mathcal{M}_n$, the empirical risk minimizer over S_m is a projection estimator since

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

Examples of projection estimators (see DeVore and Lorentz, 1993)

- Histograms:
  For every $m \in \mathcal{M}_n$, let $\{I_\lambda\}_{\lambda\in\Lambda(m)}$ be a partition of $[0,1]$ in $\mathrm{Card}(\Lambda(m)) = D_m$ intervals. Set $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$ for every $\lambda \in \Lambda(m)$, with $|I_\lambda|$ denotes the Lebesgue measure of I_λ. Then, the empirical risk minimizer, which is a histogram, is a projection estimator

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\mathbb{1}_{I_\lambda}\,\frac{\mathbb{1}_{I_\lambda}}{|I_\lambda|}\,.$$

- Trigonometric polynomials:
  Let $\{\varphi_\lambda\}_{\lambda\in\mathbb{Z}}$ be the orthonormal basis of $L^2([0,1])$ such that $t \mapsto \varphi_\lambda(t) = e^{2\pi i\lambda t}$. For any finite $\Lambda(m) \subset \mathbb{Z}$, the trigonometric polynomial

$$t \mapsto \widehat{s}_m(t) = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda\; e^{2\pi i\lambda t}$$

  is a projection estimator.

- Wavelet basis:
  Set $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$ an orthonormal basis of $L^2([0,1])$ made of compact supported wavelets, where $\Lambda_n = \big\{(j,k) \mid j \in \mathbb{N}^* \text{ and } 1 \le k \le 2^j\big\}$. For every subset $\Lambda(m)$ of Λ_n, the empirical risk minimizer associated with $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ is

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

2.2 Cross-validation

First, CV is presented as a particular instance of *subsampling*, which enables to yield a unified description of CV procedures. Then, several CV procedures are detailed with an emphasis on *leave-p-out cross-validation* (Lpo) that will be further studied in the following of the paper.

2.2.1 Resampling

A *resampling* procedure consists in generating new sets of observations—the *resamples*—from the original sample according to a given scheme. Resampling corresponds to *subsampling* when the resample cardinality is less than that of the original sample.

Among first resampling procedures, a primitive version of CV has been performed by Larson (1931) at the early 30s, while jackknife was introduced by Quenouille (1949) and also studied by Tukey (1958). However, resampling procedures have only emerged as a worthwhile matter of study following the work by Efron (1979, 1982) on bootstrap.

Interested readers are referred to (Arlot, 2007a, Introduction) and (Celisse, 2008, Introduction) for a general point of view about resampling and also Giné (1997) for some more references.

Following the heuristics described by Efron (1979), resampling approximates the unknown distribution of a statistics by that of "resampled statistics", that is statistics computed from the resamples, given original data (see Mason and Newton, 1992, for examples of theoretical results).

Let $X_{1,n} = \{X_1, \ldots, X_n\}$ denote original observations and $X^*_{1,N} = \{X^*_1, \ldots, X^*_N\}$ for $N \leq n$, some resampled data. Then the empirical distribution of $X^*_{1,N}$ is equal to

$$\frac{1}{N}\sum_{i=1}^{N} \delta_{X^*_i} = \frac{1}{n}\sum_{i=1}^{n} W_{n,i}\delta_{X_i} =: P_n^W$$

where $W_n = (W_{n,1}, \ldots, W_{n,n})$ denotes a weight vector, which is specific of the resampling scheme. Weights $W_{n,i}$ are random variables drawn independently from X_is according to a known distribution.

2.2.2 Cross-validation rationale

Unlike bootstrap, CV does not aim at recovering the distribution of a given statistics. More precisely, CV is devoted to estimate the risk of an estimator $\widehat{s}(P_n)$ of s. Notation $\widehat{s}(P_n)$ stresses the dependence of $\widehat{s}$ on original observations through empirical measure. The risk of $\widehat{s}$ can be expressed as

$$r_n(\widehat{s}) = \mathbb{E}_{X_{1,n}}\left[\mathbb{E}_X\left[\gamma\left(\widehat{s}\left(P_n\right), X\right)\right]\right], \tag{1}$$

where X denotes a new observation, independent from $X_{1,n}$ and identically distributed. $\mathbb{E}_X$ and $\mathbb{E}_{X_{1,n}}$ are expectations with respect to X and respectively $X_{1,n}$.

The crux in the CV heuristics is the independence between X and $X_{1,n}$, which arises from (1). This point is at the core of any CV procedure and justifies the splitting of $X_{1,n}$ into a *training set*—used to compute the estimator— and a *test set*—used to assess the quality of the latter estimator.

Since the training set plays the same role as the initial sample but with cardinality less than n, it acts as a subsample of $X_{1,n}$. Therefore, this subsampling scheme can be defined by the choice of some random weights $W_n = (W_{n,1}, \ldots, W_{n,n})$. Details about CV procedures and corresponding weights are provided in Section 2.2.3.

Let P_n^W and $P_n^{\overline{W}}$ respectively denote empirical measures of data in the training set, resp. in the test set. For a given split of the data, the CV estimate satisfies

$$P_n^{\overline{W}}\gamma\left(\widehat{s}\left(P_n^W\right)\right) \approx \mathbb{E}_X\left[\gamma\left(\widehat{s}(P_n), X\right)\right].$$

The left-hand side quantity depends on the realization of the random weights, which can be removed by integrating with respect to them:

$$\widehat{R}_{CV,\mathrm{W}} := \mathbb{E}_W\left[P_n^{\overline{W}}\gamma\left(\widehat{s}\left(P_n^W\right)\right)\right] \approx \mathbb{E}_{X_{1,n}}\left[\mathbb{E}_X\left[\gamma\left(\widehat{s}\left(P_n\right), X\right)\right]\right] = r_n(\widehat{s}),$$

where $\mathbb{E}_W$ means integration is carried out with respect to the weights.

$\widehat{R}_{CV,\mathrm{W}}$ denotes the CV estimate of the risk of $\widehat{s}$, up to a constant. The notation points out that it depends on the choice of the weight distribution (see Section 2.2.3).

In the sequel, $r_n(\widehat{s})$ is repeatedly used and referred to as the risk of $\widehat{s}$.

2.2.3 Cross-validation botany

Several CV procedures are described with their associated weights. A distinction is made between time-consuming and computationally efficient ones. VFCV and Lpo are then compared one another in several respects.

In the following, for any $1 \leq p \leq n-1$, $\mathcal{E}_p$ denotes the set of all possible subsets of $\{1, \ldots, n\}$ with cardinality p.

Hold-out From a historical point of view, *simple validation* also called *Hold-out* (Ho) has been introduced at the early 30s. For instance, it is employed by Larson (1931) in his empirical analysis. Hold-out simply consists in randomly splitting observations into a training set of cardinality $n-p$ and a test set of cardinality p, with $1 \leq p \leq n-1$. Data splitting is only made once, which results in additional variability.

Since it is easy to analyse, hold-out has been often studied: see for instance Bartlett et al. (2002); Blanchard and Massart (2006) in classification, and Lugosi and Nobel (1999); Wegkamp (2003) in regression.

For any random choice of $e \in \mathcal{E}_p$, the hold-out estimator of $r_n(\widehat{s})$ is

$$\widehat{R}_{\mathrm{Ho},p}(\widehat{s}) := P_n^e \gamma\left(\widehat{s}(P_n^{\overline{e}})\right) = \frac{1}{p}\sum_{i \in e} \gamma\left(\widehat{s}(X_{1,n}^{\overline{e}}), X_i\right),$$

where P_n^e (resp. $P_n^{\overline{e}}$) denotes the empirical distribution of data in the test set (resp. in the training set). Hold-out corresponds to the random choice of W_n such that for every i, $W_{n,i} \in \{0, n/p\}$, $\sum_{i=1}^n W_{n,i} = n$, and W_n is drawn from a Dirac measure over the $\binom{n}{p}$ such vectors.

Leave-one-out *Leave-one-out* (Loo) was the first CV procedure, since strictly speaking CV starts when simple validation is carried out for several splits of the data. It was first formalized by Mosteller and Tukey (1968), and then studied in the model selection framework by Stone (1974). It consists in successively removing each observation from the original data, using the $n-1$ remaining ones to compute the estimator. The performance of the latter estimator is then assessed thanks to the removed point. The Loo risk estimator is defined as the average performance assessment over the n possible splits:

$$\widehat{R}_1(\widehat{s}) = \frac{1}{n}\sum_{i=1}^{n} \gamma(\widehat{s}(X_{1,n}^{(i)}), X_i),$$

where $X_{1,n}^{(i)}$ represents $X_{1,n}$ from which X_i has been removed.

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 $\sum_{j=1}^{n} W_{n,j} = n$.

Leave-p-out *Leave-p-out* (Lpo) generalizes Loo to the case where $1 \leq p \leq n-1$ observations are removed from original data at each split.

It is studied in linear regression setup by Shao (1993) and Zhang (1993), and in the change-point detection setting by Arlot and Celisse (2009). In density estimation, Celisse and Robin (2008b) derive closed-form expressions for the Lpo estimator with histograms and kernels.

The Lpo estimator of $r_n(\widehat{s})$ consists in the same procedure as Loo except that at each one of the $\binom{n}{p}$ possible splits, p observations are removed:

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

The corresponding weights satisfy $W_{n,i} \in \{0, n/p\}$ for any i, $\sum_{i=1}^{n} W_{n,i} = n$ and the probability of any such vector is $\binom{n}{p}^{-1}$.

Remark 1. *A naïve implementation of Lpo has computational complexity of order $\binom{n}{p}$ times that of the $\widehat{s}$ computation, which is intractable as soon as n is large and $p > 1$. Even Loo, that is Lpo with $p = 1$, can be time-consuming.*

V-fold cross-validation Due to the high computational burden of the previous procedures, Geisser (1974, 1975) has introduced an alternative procedure named *V-fold cross-validation* (VFCV). For instance, it has been studied in Burman (1989, 1990) who suggests a correction to remove some bias.

VFCV relies on a preliminary (random or not) choice of a partition of the data into 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 performance assessment. The V-fold risk estimator is the average over 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 $\sum_{i=1}^{n} W_{n,i} = n$.

Let $e_1, \ldots, e_V$ denote the partition of $\{1, \ldots, n\}$ into V blocks. Then, the VFCV estimator of $r_n(\widehat{s})$ is

$$\widehat{R}_{\mathrm{VFCV},V}(\widehat{s}) = \frac{1}{V} \sum_{v=1}^{V} \left[\frac{V}{n} \sum_{i \in e_v} \gamma\left(\widehat{s}\left(P_n^{\overline{e_v}}\right), X_i\right) \right].$$

2.2.4 Lpo versus VFCV

Nowadays, it is usual to deal with a large amount of data. For instance, biology as well as computer vision are perfect illustrations of this statement.

As explained in Section 2.2.3, the Loo computational complexity is n times that of $\widehat{s}$, which can be highly time-consuming. With this respect, provided $V \ll n$, VFCV (Geisser, 1974, 1975) is by far less computationally demanding than Loo.

However, VFCV relies on a *random* partitioning of the data into V subsets. This additional randomness induces some more variability with respect to Loo and Lpo, which both carry out exhaustive splitting of the observations. In the density estimation framework, Celisse and Robin (2008b) has theoretically quantified the amount of randomness induced in applying VFCV instead of Lpo.

As it does not introduce any additional variability, Lpo can be seen as a "gold standard" among CV procedures. VFCV turns out as an approximation of the "ideal Lpo", which is unachievable due to prohibitive computation-time. Indeed, the Lpo computation requires to explore $\binom{n}{p}$ resamples, which is intractable even for not too large n when $p > 2$. Therefore, with full generality, Lpo cannot be performed and one has to use approximations.

Other approximations to Lpo exist like *repeated learning-testing cross-validation* (RLT), introduced by Breiman et al. (1984) and then studied in Burman (1989) and Zhang (1993).

The purpose of the next section is to describe a broad range of settings in which closed-form expressions of the Lpo estimator can be derived. On the one hand, such formulas drastically reduce the Lpo computational complexity from exponential—for a naïve implementation of Lpo—to linear. Furthermore, such formulas make Lpo preferable to VFCV since the latter is more variable and expensive to perform.

On the other hand, these closed-form formulas yield more insight in the general behaviour of CV as an estimator of the risk. The study of CV as a model selection procedure is the main concern of Section 4.

3 Closed-form expressions

Closed-form expressions of the Lpo risk estimator are provided for the broad family of projection estimators. First, such formulas enable the efficient computation of Lpo. Second, they provide some information about the quality of the CV estimator as an estimator of the risk. Indeed, closed-form expressions for bias and variance of the CV estimator are also derived.

3.1 Leave-p-out risk estimator

Here is an elementary and essential lemma that leads to these closed-form expressions. This result is obtained thanks to combinatorial calculations.

Lemma 3.1. *Let $\widehat{s}_m(X^{\overline{e}}_{1,n})$ denote a generic projection estimator based on model S_m and computed from the training set $X^{\overline{e}}_{1,n}$. Then,*

$$\sum_{e\in\mathcal{E}_p}\left\|\widehat{s}_m(X^{\overline{e}}_{1,n})\right\|_2^2 = \frac{1}{(n-p)^2}\left[\binom{n-1}{p}\sum_{k=1}^{n}\sum_{\lambda\in\Lambda(m)}\varphi_\lambda^2(X_k) + \binom{n-2}{p}\sum_{k\neq\ell}\sum_{\lambda\in\Lambda(m)}\varphi_\lambda(X_k)\varphi_\lambda(X_\ell)\right], \tag{2}$$

$$\sum_{e\in\mathcal{E}_p}\sum_{i\in e}\widehat{s}\,(X^{\overline{e}}_{1,n})(X_i) = \frac{1}{n-p}\binom{n-2}{p-1}\sum_{i\neq j}\sum_{\lambda\in\Lambda(m)}\varphi_\lambda(X_i)\varphi_\lambda(X_j). \tag{3}$$

The proof of Lemma 3.1 is deferred to Section 7.

From the previous lemma, the closed-form expression for the Lpo estimator of the risk is derived. This expression holds with the quadratic loss and projection estimators.

Proposition 3.1. *For any $m \in \mathcal{M}_n$, let $\widehat{s}_m$ denote the projection estimator onto the model S_m, spanned by the orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$. Then for any $p \in \{1,\dots,n-1\}$,*

$$\widehat{R}_p(m) = \frac{1}{n(n-p)}\sum_{\lambda\in\Lambda(m)}\left[\sum_j\varphi_\lambda^2(X_j) - \frac{n-p+1}{n-1}\sum_{j\neq k}\varphi_\lambda(X_j)\varphi_\lambda(X_k)\right]. \tag{4}$$

The computation cost of (4) is of order $\mathcal{O}(n)$.

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

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

Besides, the general projection estimator is

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

The simple application of (2) and (3) provides the expected conclusion.

□

Examples We are now in position to specify the expression of the Lpo risk estimator in Proposition 3.1 for several projection estimators.

1. Histograms:

Corollary 3.1. *Let us 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 respective length $|I_\lambda|$. Then for $p \in \{1, \ldots, n-1\}$,*

$$\widehat{R}_p(m) = \frac{1}{(n-1)(n-p)} \sum_{\lambda=1}^{D_m} \frac{1}{|I_\lambda|} \left[(2n-p)\, \frac{n_\lambda}{n} - n(n-p+1) \left(\frac{n_\lambda}{n} \right)^2 \right], \quad (5)$$

where $n_\lambda = \sharp \{i \,|\, X_i \in I_\lambda\}$.

Proof. (5) comes simply from the application of (4) with $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$. □

2. Trigonometric polynomials:

Corollary 3.2. *Let φ_λ denote either $t \mapsto \cos(2\pi k t)$, if $\lambda \in 2\mathbb{N}$ or $t \mapsto \sin(2\pi k t)$, if $\lambda \in 2\mathbb{N}+1$.*
Let us further assume that $\Lambda(m) = \{0, \ldots, 2K\}$ for an integer $K > 0$. Then,

$$\widehat{R}_p(m) = \frac{(p-2)(K+1)}{(n-1)(n-p)} - \frac{n-p+1}{n(n-1)(n-p)} \sum_{k=0}^{K} \left[\left\{ \sum_{j=1}^{n} \cos(2\pi k X_j) \right\}^2 + \left\{ \sum_{j=1}^{n} \sin(2\pi k X_j) \right\}^2 \right].$$

3. Haar basis:

Corollary 3.3. *Set $\varphi : t \mapsto \mathbb{1}_{[0,1]}$ and $\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j \cdot - k)$, where $j \in \mathbb{N}$ and $0 \le k \le 2^j - 1$.*
For any $m \in \mathcal{M}_n$, let us define $\Lambda(m) \subset \left\{(j,k) \mid j \in \mathbb{N},\ 0 \le k \le 2^j - 1\right\}$. Then,

$$\widehat{R}_p(m) = \frac{1}{(n-1)(n-p)} \sum_{(j,k) \in \Lambda(m)} 2^j \left[(2n-p) \frac{n_{j,k}}{n} - n(n-p+1) \left(\frac{n_{j,k}}{n} \right)^2 \right],$$

where $n_{j,k} = \mathrm{Card}\left(\left\{ i \mid X_i \in [k/2^j, (k+1)/2^j] \right\}\right)$.

3.2 Moment calculations

As a consequence of the closed-form expressions settled in the previous section, similar expressions are also available for expectation and variance. A precise assessment of the performance of the Lpo estimator as an estimator of the risk is thus available thanks to these closed-form expressions.

Proposition 3.2. *With the same notations as in Proposition 3.1, we have for any* $1 \leq p \leq n-1$,

$$\mathbb{E}\widehat{R}_p(m) = \frac{1}{n-p}\sum_{\lambda\in\Lambda(m)}\left[\mathbb{E}\varphi_\lambda^2(X) - (\mathbb{E}\varphi_\lambda(X))^2\right] - \sum_{\lambda\in\Lambda(m)}(\mathbb{E}\varphi_\lambda(X))^2,$$

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

where $P\varphi_\lambda = \mathbb{E}\varphi_\lambda(X)$, $\alpha = n-1$, $\beta = n-p+1$, $t_1 = n(n-1)$, *and* $t_2 = t_1(n-2)$.

The technical proof is given in Section 7. Note that these formulas may be derived provided $P|\varphi_\lambda|^3 < +\infty$ for any $\lambda \in \Lambda(m)$, which is satisfied if s is assumed to be bounded and $\int|\varphi_\lambda|^3 < +\infty$ (φ_λ continuous and compact supported for instance).

The bias of the Lpo risk estimator may be a more interesting quantity to work with. Its expression straightforwardly results from Proposition 3.2.

Corollary 3.4. *For any projection estimator, the bias of the Lpo estimator is equal to*

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

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

Illustration By application of Proposition 3.2 to histogram estimators, the following expressions are derived for expectation and variance of the Lpo risk estimator (see Celisse and Robin, 2008b):

Corollary 3.5. *For every* $\lambda \in \Lambda(m)$, *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

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

4 Model selection

Although CV is extensively used in practice, very few is known about its non-asymptotic behaviour as a model selection procedure. In particular, there is no theoretical and non-asymptotic guideline about the optimal choice of p with respect to the model selection goal one pursue.

The purpose of the present section is first to analyze CV as a penalized criterion, which enables a new interpretation of the choice of p. Second, the performance of CV in terms of model selection procedure is quantified by non-asymptotic optimality results. To the best of our knowledge, these oracle inequalities are the first results of this type.

4.1 Random penalty

This section sheds new lights on the behaviour of CV as model selection procedure with respect to the choice of p. On the one hand, CV is embedded in the framework of model selection via penalized criteria. It is shown that the choice of p determines the amount of penalization.

On the other hand, several conclusions are drawn about the appropriate—non-asymptotic—use of CV, depending on the value of p. For instance, since it behaves like Mallows'C_p, Loo must not be employed as a model selection procedure with exponential collections of models.

4.1.1 Ideal and Lpo penalties

Given the *a priori* knowledge of the target s, a countable collection $\{S_m\}_{m\in\mathcal{M}_n}$ is chosen so that the S_ms are assumed to be close to s. The purpose of model selection is to design a procedure providing a candidate model $\widehat{m}$ such that the final estimator $\widehat{s}_{\widehat{m}}$ is as close as possible to the target s.

For instance, the choice of $\widehat{m}$ is made by minimizing a penalized criterion $\mathrm{crit}(\cdot)$ (Barron et al., 1999) defined by

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m), \tag{6}$$

where $P_n\gamma\left(\widehat{s}_m\right)$ is the empirical risk of $\widehat{s}_m$. $\mathrm{pen}(\cdot) : \mathcal{M}_n \to \mathbb{R}_+$ denotes the penalty term, which takes into account the complexity of model S_m.

On the one hand, the optimal criterion to minimize over $\mathcal{M}_n$ is the ideal random quantity

$$\mathrm{crit}_{id}(m) = P\gamma(\widehat{s}_m) := \mathbb{E}\gamma\left(\widehat{s}_m, X\right) \tag{7}$$

where the expectation is taken with respect to $X \sim P$, which is independent from the original data. The minimization of the *ideal criterion* crit_{id} over $\mathcal{M}_n$ would systematically yield the best estimator one can achieve among $\{\widehat{s}_m\}_{m\in\mathcal{M}_n}$, that is the *oracle*. The link between (6) and (7) can be clarified by rewriting

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

so that the *ideal penalty* is defined by

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

The ideal penalty is what must be added to the empirical risk to recover the ideal criterion.

On the other hand 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 $\widehat{m}$ satisfies

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

The existence of a strong relationship between penalized criteria and CV procedures is strongly supported by the large amount of literature about (asymptotic) comparisons of these two model selection procedures (see for instance Stone, 1977; Li, 1987; Zhang, 1993). Therefore, the CV strategy can be embedded into penalized criterion minimization procedures:

$$\widehat{m} = \mathrm{Argmin}_{m\in\mathcal{M}_n} \left\{ P_n\gamma(\widehat{s}_m) \; + \; \mathrm{pen}_p(m) \right\},$$

where $\mathrm{pen}_p(m)$ is called the *Lpo penalty* of model m and satisfies for every m $\mathrm{pen}_p(m) := \widehat{R}_p(m) - P_n\gamma(\widehat{s}_m)$. A somewhat related approach applied to Loo can be found in Birgé and Massart (1997).

4.1.2 Lpo overpenalization

This embedding of CV into penalized criteria provides more insight in the behaviour of CV procedures with respect to parameter p. In particular, some features in the behaviour of

pen_p as function of p arise from the comparison between pen_{id} and pen_p. This comparison is carried out through expectations of these penalties. The next results hold with general projection estimators.

Let us start with a preliminary lemma:

Lemma 4.1. *With any projection estimator $\widehat{s}_m$ onto S_m, we obtain*

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

This enables to precisely evaluate, the discrepancy between Lpo and ideal penalties:

Proposition 4.1. *For every $m\in\mathcal{M}_n$, 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, for every $m\in\mathcal{M}_n$ and $1\le p\le n-1$,*

$$\mathbb{E}\left[\,\text{pen}_p(m)-\text{pen}_{id}(m)\,\right]=\frac{p}{n(n-p)}\sum_{\lambda\in\Lambda(m)}\text{Var}\left(\varphi_\lambda(X)\right)\ge 0. \tag{8}$$

Whatever $1\le p\le n-1$, the Lpo penalty remains larger than the ideal one by an amount that increases with p. Furthermore, this amount of penalization can vary within a wide range of values. Indeed, (8) yields

$$\mathbb{E}\left[\,\text{pen}_p(m)\,\right]=C_{\text{over}}(p)\,\mathbb{E}\left[\,\text{pen}_{id}(m)\,\right],$$

where $C_{\text{over}}(p)=(2n-p)/(2n-2p)$. Therefore with $p=1$, $C_{\text{over}}(1)=1+1/(2n-2)$ leads to a nearly unbiased estimator of the ideal penalty, while $C_{\text{over}}(n/2)=3/2$ indicates that the Lpo penalty overpenalizes by an amount of the same order as the ideal penalty. A $\log n$ factor can even be achieved by $C_{\text{over}}(p)$ provided $p\approx(1-1/(2\log n-1))\,n$.

At this stage, an important distinction must be made between risk estimation and model selection. On the one hand, if the purpose is the estimation of the risk of a given estimator, for instance, an unbiased (or nearly unbiased) estimator of this risk can be desirable. Therefore, the choice $p=1$ seems the most appropriate one, provided the variance of the resulting estimator remains at a reasonable level. As for "optimal" risk estimation, Celisse and Robin (2008b) has developed a strategy aiming at providing the Lpo risk estimator with the smallest mean square error. In Celisse and Robin (2008a), it is shown that the proposed estimator asymptotically amounts to Loo as n tends to infinity. This is also consistent with the results of Burman (1989) who shows in the regression setting that Loo is asymptotically the best risk estimator among CV ones in terms of bias and variance.

On the other hand, model selection requires to choose the closest model to the target, even at the price of a worse estimation of the risk for some models in the collection.

For instance, minimizing a penalized criterion over $\mathcal{M}_n$ leads to misleading models with a probability increasing with $\mathcal{M}_n$, provided the oracle remains the same. This results in random downwards deviations of the minimized criterion for some "bad models". A classical way to balance these unwanted deviations is to overpenalize by an amount that depends on the structure of the considered collection of models. Subsequently, choosing $p = 1$ is not necessary desirable in model selection as already noticed in Breiman and Spector (1992) and recently in (Celisse, 2008, Chap. 6) where Loo has been empirically shown to suffer from overfitting with polynomial collections of models.

Several conclusions can be drawn from Proposition 4.1 about the CV performance as a model selection procedure. First, Loo is a nearly unbiased risk estimator, which results in similar behaviour to that of Mallows'C_p. This is consistent with asymptotic results established by Li (1987) and Zhang (1993). As a consequence, Loo only aims at yielding a reliable estimation of the target in order to perform (asymptotically) efficient model selection (see Section 1 and Li (1987)). In particular, Loo cannot be employed—with an identification purpose—to recover the "true model" with probability converging to 1 as n tends to infinity, which is the goal of BIC.

Second, with Loo—and Lpo for small values of p—only model selection over polynomial collections of models can be carried out. For instance, using Loo with exponential collections of models would systematically lead to overly large models.

Third, identification can however be pursued by CV, provided p has been chosen of the appropriate order. Indeed, $p \approx (1 - 1/(2 \log n - 1))\, n$ yields a $\log n$ term like the one in BIC penalty. This also confirms the previous asymptotic result settled by Shao (1993) in the regression setting.

4.2 Oracle inequalities

In the following, the quality of the Lpo-based model selection procedure is assessed through the statement of oracle inequalities. These results are settled in the polynomial complexity framework and hold for any projection estimator. To our knowledge, it is the first non-asymptotic results about the performance of Lpo in this framework.

Unlike the usual approach in model selection via penalized criterion, the purpose here is not to design a penalty function. Indeed, the Lpo estimator itself can be understood as a penalized criterion (see Section 4.1).

4.2.1 Preliminaries

The main results rely on several assumptions detailed and discussed in the following.

Set $X \sim s$ and for every index m,

$$\phi_m = \sum_{\lambda \in \Lambda(m)} \varphi_\lambda^2 \qquad \text{and} \qquad V_m = \mathbb{E}\, \phi_m(X).$$

Then, let us define the following assumptions:

(Reg) $\exists \Phi > 0 / \quad \sup_{m \in \mathcal{M}_n} \|\phi_m\|_\infty \leq \Phi \, n / \left(\log n\right)^2$.

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 the considered basis to the amount of data. This assumption has already been used by Castellan (2003) for instance. Let us assume we use histogram estimators based on a partition $\{I_1, \ldots, I_{D_m}\}$ of $[0,1]$ in D_m intervals, and that $\varphi_\lambda = \mathbb{1}_{I_\lambda} / \sqrt{|I_\lambda|}$, where $|I_\lambda|$ is the length of I_λ. Then, (Reg) gives a lower bound on the minimal length of any interval I_λ of the partition with respect to the number of observations. In other words, partitions made of intervals with less than $n/(\log n)^2$ observations are prohibited.

(Reg2) $\exists \Phi > 0 \mid \forall m \in \mathcal{M}_n, \quad \sup_{|a|_\infty = 1} \|\sum_\lambda a_\lambda \varphi_\lambda\|_\infty \leq \sqrt{\Phi \, n / \left(\log n\right)^2}$.

$(Reg2)$ 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 $(Reg2)$. Besides, the constant Φ is assumed to be the same in (Reg) and $(Reg2)$, which holds up to replacing one of them by their maximum. A similar requirement to $(Reg2)$ can be found in Massart (2007).

(Ad) $\exists \xi > 0 / \; \forall m \in \mathcal{M}_n$ with $D_m \geq 2, \quad n\mathbb{E}\left[\|s_m - \widehat{s}_m\|_2^2\right] \geq \xi D_m$.

Let us first notice that

$$\mathbb{E}\,\|s_m - \widehat{s}_m\|^2 = \sum_{\lambda \in \Lambda(m)} \mathbb{E}\left[\,\nu_n^2\left(\varphi_\lambda\right)\right] \quad = \sum_{\lambda \in \Lambda(m)} \frac{1}{n} \mathrm{Var}\left[\,\varphi_\lambda(X)\,\right].$$

With 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 $\mathrm{Var}\left[\,\varphi_\lambda(X)\,\right] \neq 0$. For instance a sufficient condition for (Ad) to hold with histograms is $s \geq \rho > 0$ on $[0,1]$. This assumption can also be found in Massart (2007).

(Pol) $\exists \delta \geq 0 / \; \forall D \geq 1, \; |\{m \in \mathcal{M}_n \mid D_m = D\}| \leq D^\delta$.

A model collection is said to have a polynomial complexity if (Pol) holds, that is if the cardinality of the set of models with dimension D is polynomial in D. Such an assumption is satisfied with nested models for instance (Birgé and Massart (1997)). It straightforwardly implies that $\mathrm{Card}\left(\mathcal{M}_n\right) \leq n^{\delta+1}$.

4.2.2 Main results

In the following, two oracle inequalities are settled, which warranty the ability of the Lpo procedure to select an efficient density estimator. The first result holds with bounded densities, while the second one concerns the more general case of square integrable densities at the price of an additional assumption. Several instances of bases for which the latter assumption holds are provided at the end of this section.

Bounded density

Theorem 4.1. *Let s denote a bounded density on [0,1] and $X_1, \dots, X_n$ 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

$$\frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} < 1\,, \tag{9}$$

where $\zeta(\epsilon) = \left[1 - (1+\epsilon)^{-8}\right]$. Then for any $1 \leq p \leq n-1$ 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 $0 < \alpha, \beta < 1$, we have

$$\mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2\right] \leq \Gamma(\epsilon,\alpha,\beta) \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\|s - \widehat{s}_m\|^2\right] + \frac{\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta)}{n},$$

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.

The proof of this result is deferred to Section 7.
Remarks:

- (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.
- The existence of ϵ satisfying the inequality (9) stems from a technical lemma given in the proof of Theorem 4.1.
- As it is made clear from the proof of the aforementioned technical lemma, 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.

- The independence of $\Gamma(\epsilon, \alpha, \beta)$ from s is essential in our framework since we have in mind the use of this result to derive adaptivity in the minimax sense properties.

Square-integrable density

The second result is derived following the same idea as the previous one, thanks to an additional mild assumption on the considered bases. This requirement turns out to be non restrictive at all, since it is met by a broad class of orthonormal bases.

Theorem 4.2. *Let s denote a density in $L^2([0,1])$ and $X_1, \ldots, X_n$ be n i.i.d. random variables drawn from s. We 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, and moreover that*

$$(Reg3) \qquad \exists \Phi > 0 / \ \forall m \in \mathcal{M}_n, \qquad \|\phi_m\|_\infty \leq \Phi D_m.$$

For $n \geq 29$, set $0 < \epsilon < 1$ such that

$$\frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} < 1\,,$$

where $\zeta(\epsilon) = \left[1 - (1+\epsilon)^{-8}\right]$. Then for any $1 \leq p \leq n-1$ 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 $0 < \alpha, \beta < 1$, we have

$$\mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2\right] \leq \Gamma(\epsilon, \alpha, \beta) \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\|s - \widehat{s}_m\|^2\right] + \frac{\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta)}{n},$$

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.

For the sake of clarity, the proof is also deferred to Section 7. Since it is very similar to that of Theorem 4.1, only the main differences are detailed.

Remark 2. *Assumption $(Reg3)$ is quite different from (Reg) . Whereas the latter relates the "regularity" of any basis to the number of observations uniformly over $\mathcal{M}_n$, $(Reg3)$ rather controls $\|\phi_m\|_\infty$ for every model by means of its dimension. All models with the same dimension must be somehow alike since their associated sup-norm $\|\phi_m\|_\infty$ remains upper bounded by ΦD_m. This assumption can be found in Birgé and Massart (1997) as well.*

Examples Several examples of widespread functional bases are now detailed to illustrate the high generality level of assumption $(Reg3)$.

- It is easy to check that $(Reg3)$ applies to *regular* histograms with $\Phi = 1$ (Section 5.3).

- A typical example of basis satisfying $(Reg3)$ is the trigonometric basis. For $m \in \mathbb{N}$, let $\Lambda(m) = \{0, \dots, 2m\}$ denote a set of indices where $\varphi_0 = \mathbb{1}_{[0,1]}$, $\varphi_\lambda(t) = \sqrt{2}\sin(2k\pi t)$ if $\lambda = 2k-1$ and $\varphi_\lambda(t) = \sqrt{2}\cos(2k\pi t)$ if $\lambda = 2k$.
  Then,

$$\forall t \in [0,1], \qquad \sum_{\lambda \in \Lambda(m)} \varphi_\lambda^2(t) = 1 + 2\sum_{k=1}^{m} \left(\cos^2(2k\pi t) + \sin^2(2k\pi t)\right), \\ = 2m+1.$$

  Since $D_m = 2m+1$, it comes that $\|\phi_m\|_\infty = D_m$ and $(Reg3)$ holds with $\Phi = 1$.

- Barron et al. (1999) (Lemma 7.13) proved that with piecewise polynomials on a regular partition of $[0,1]$ with degree not larger than r on each element of this partition,

$$\|\phi_m\|_\infty \leq (r+1)(2r+1)D_m.$$

  The resulting constant $\Phi = (r+1)(2r+1)$ is subsequently independent from m.

- Haar basis: For any positive integer j, we introduce $\Lambda(j) = \left\{(j,k) \mid 0 \leq k \leq 2^j - 1\right\}$. Furthermore, set $\varphi = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1]}$ and for any $\lambda = (j,k)$, let us define $\varphi_{j,k}(t) = 2^{j/2}\varphi\left(2^j t - k + 1\right)$ on $[0,1]$. For a positive integer $m \in \mathcal{M}_n$, let us consider S_m as the linear space spanned by $\{\varphi_\lambda\}_{\lambda \in \cup_{j \leq m}\Lambda(j)}$. Then, it can be seen that

$$\|\phi_m\|_\infty = D_m$$

  since for each j, there is only one $0 \leq k \leq 2^j - 1$, which contributes to the sum in ϕ_m.
  For more general wavelet bases, an upper bound—uniform with respect to m—can be established (see Birgé and Massart, 1997, for instance).

5 Adaptivity

In this section, the idea is to apply theorems of Section 4.2 to derive several adaptivity results in the minimax sense with respect to Hölder as well as Besov functional spaces.

5.1 Adaptivity in the minimax sense

Let us assume that s belongs to a set of functions $\mathcal{T}(\theta)$, indexed by a parameter $\theta \in \Theta$, and define an estimator $\widehat{s}$ of s.

An estimator $\widehat{s}$ is said to be *adaptive for* θ if, without knowing θ, it "works as well as" any estimator which would exploit this knowledge.

Definition 5.1. *An estimator $\widehat{s}$ is said to be adaptive for θ if its risk is nearly the same as the minimax risk with respect to $\mathcal{T}(\theta)$, that is if there exists $C \geq 1$ satisfying:*

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

where the infimum is taken over all possible estimators.
Furthermore if this property holds for every parameters θ in a set Θ, then $\widehat{s}$ is said to be adaptive in the minimax sense with respect to the family $\{\mathcal{T}(\theta)\}_{\theta \in \Theta}$.

Interested readers are referred to Barron et al. (1999) for a unified presentation about various notions of adaptivity.

Remark 3. *Very often, $C \geq 1$ depends on the unknown parameters θ, but neither from s nor from n.*

5.2 Description of the collections of models

Since such optimality results depend on the approximation properties of the considered models, three different model collections are described in the following, each one being defined from a specific family of vectors $\{\varphi_\lambda\}_{\lambda \in \Lambda_n}$.

5.2.1 Piecewise constant functions (Pc)

For a given partition of $[0,1]$ in D regular intervals $(I_\lambda)_{\lambda \in \Lambda(m)}$ of length $1/D$ and $m \in \mathcal{M}_n$, 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 the orthonormal family $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$. It is made of all piecewise constant functions defined on the partition $I = (I_1, \ldots, I_{D_m})$.
Thus with each index $m \in \mathcal{M}_n$, we associate the linear space S_m of piecewise constant functions defined on a *regular partition* of $[0,1]$ in D_m intervals of length $1/D_m$. Moreover, let $N_n = \max_{m \in \mathcal{M}_n} D_m$ be the maximal dimension of a model belonging to the collection.

5.2.2 Piecewise dyadic polynomials (Pp)

Set $\mathcal{M}_n = \{0, \dots, J_n\}$ and for any $m \in \mathcal{M}_n$, S_m denotes the linear space of functions

$$t = \sum_{k=0}^{2^m-1} P_k \, \mathbb{1}_{[k2^{-m},(k+1)2^{-m})},$$

where the P_ks denote polynomials of degree less than r. The dimension of S_m is subsequently defined by

$$D_m = r\, 2^m \qquad \text{and} \qquad N_n = \max_{m \in \mathcal{M}_n} D_m = r\, 2^{J_n}.$$

With this collection of models, (Pol) is satisfied since there is at most one model for each dimension.

5.2.3 Trigonometric polynomials (Tp)

Set $\mathcal{M}_n = \{0, \dots, J_n\}$, where J_n is a positive integer. For any $m \in \mathcal{M}_n$, let $\Lambda(m) = \{0, \dots, 2m\}$ denote a set of indices such that $\varphi_0(t) = \mathbb{1}_{[0,1]}$, $\varphi_\lambda(t) = \sqrt{2}\sin(2k\pi t)$ if $\lambda = 2k-1$ and $\varphi_\lambda(t) = \sqrt{2}\cos(2k\pi t)$ if $\lambda = 2k$.
Then, S_m is the linear space spanned by $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$, of dimension $D_m = 2m+1$. Any $t \in S_m$ can be expressed as

$$\forall x \in [0,1], \qquad t(x) = a_0 + \sum_{k=1}^{m} \left[a_k \sqrt{2}\cos(2\pi k x) + b_k \sqrt{2}\sin(2\pi k x) \right],$$

the a_ks and b_ks belong to $\mathbb{R}$.
Moreover, J_n and N_n are related by the following relationship $N_n = 2J_n + 1$.

5.3 Hölder functional space

The purpose is to show that the Lpo-based approach enjoys some adaptivity when s belongs to an unknown Hölder space $\mathcal{H}(L,\alpha)$ for $L > 0$ and $\alpha \in (0,1]$. Let us recall that a function $f : [0,1] \to \mathbb{R}$ belongs to $\mathcal{H}(L,\alpha)$ with $L > 0$ and $0 < \alpha \leq 1$ if

$$\forall x, y \in [0,1], \quad |f(x) - f(y)| \leq L\,|x-y|^\alpha .$$

For an extensive study of functional spaces, (see DeVore and Lorentz, 1993).

In order to achieve this goal, s is approximated by piecewise constant functions, using the model collection (**Pc**) described in Section 5.2.1. The histogram estimator built from model S_m is defined by

$$\widehat{s}_m = \sum_{\lambda \in \Lambda(m)} P_n\varphi_\lambda \; \varphi_\lambda = \sum_{\lambda \in \Lambda(m)} \frac{n_\lambda}{n} \frac{\mathbb{1}_{I_\lambda}}{|I_\lambda|},$$

where $n_\lambda = \mathrm{Card}\left(\{i \mid X_i \in I_\lambda\}\right)$.

In the sequel, the assumptions of Theorem 4.1 are checked in order to derive the desired adaptivity property.

- With the collection (**Pc**), $m \mapsto D_m$ is a one-to-one mapping from $\mathcal{M}_n$ towards $\mathcal{D} = \{D_m \mid m \in \mathcal{M}_n\}$, which entails that (Pol) is satisfied since the collection is made of only one model for each dimension.
- Since $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$,

$$\|\phi_m\|_\infty = \sum_{t\in[0,1]} \left(\sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2(t) \right) = \max_{\lambda\in\Lambda(m)} \frac{1}{|I_\lambda|} = D_m.$$

  Thus, (Reg) amounts to require that

$$\max_m D_m = N_n \leq \Phi\, n/\left(\log n\right)^2,$$

  which means that on average, there are at least about $\left(\log n\right)^2/n$ points in each interval of any partition we consider.

- We therefore assume that (Reg) , (Ad) and (Ran) hold.

As for the problem of density estimation on $[0,1]$ when s belongs to some Hölder space, it is known since the early 80s, thanks to Ibragimov and Khas'minskij Ibragimov and Khas'minskij (1981), 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 $\alpha > 0$.

The following result settles that, applied to the collection of models (**Pc**), the Lpo-based procedure yields an adaptive in the minimax sense estimator of the density on $[0,1]$.

Theorem 5.1. *Let us assume that (Reg) , (Ad) and (Ran) hold and that the collection of models is that one denoted by (**Pc**). Furthermore, assume that the target density $s \in \mathcal{H}(L,\alpha)$ for $L > 0$ and $\alpha \in (0,1]$. Then,*

$$\sup_{s\in\mathcal{H}(L,\alpha)} \mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq K_\alpha\, L^{\frac{2}{1+2\alpha}}\, n^{-\frac{2\alpha}{1+2\alpha}} + O\left(\frac{1}{n}\right), \tag{10}$$

for a given constant K_α independent from n and s. $\widehat{m}$ derives from the Lpo risk minimization over $\mathcal{M}_n$.

Since the minimax risk is of order $L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}$, $\widehat{s}_{\widehat{m}}$ is adaptive in the minimax sense with respect to $\{\mathcal{H}(L,\alpha)\}_{L>0,\alpha\in(0,1]}$.

Remark 4. *This result still holds with any polynomial collection of models satisfying the requirements of Theorem 4.1, 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 4.1 and 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

$$\begin{aligned}
\|s-s_m\|^2 &= \sum_{\lambda\in\Lambda(m)} \frac{1}{|I_\lambda|^2}\int_{I_\lambda}\left(\int_{I_\lambda}\left[\,s(t)-s(x)\,\right]dx\right)^2 dt,\\
&\le \sum_{\lambda\in\Lambda(m)} L^2 D_m^2 \int_{I_\lambda}\left(\int_{I_\lambda}|t-x|^\alpha\,dx\right)^2 dt \qquad (s\in\mathcal{H}(L,\alpha))\,,\\
&\le C_\alpha\, L^2\, D_m^{-2\alpha} \quad \text{(after integration)}\,,
\end{aligned}$$

where $C_\alpha = 4\,(\alpha+2)\left[\,(1+\alpha)^2\,(2\alpha+3)\,\right]^{-1}$.
On the other hand,

$$\begin{aligned}
\mathbb{E}\left[\,\|s_m-\widehat{s}_m\|^2\,\right] &= \frac{V_m - \|s_m\|^2}{n} \le \frac{V_m}{n}\\
&\le \frac{\|\phi_m\|_\infty}{n} = \frac{\sup_{x\in[0,1]}\sum_{\lambda\in\Lambda(m)}\varphi_\lambda^2(x)}{n} = \frac{D_m}{n}\,.
\end{aligned}$$

Hence under the same assumptions as Theorem 4.1, we get that there exists $C\ge 1$ and $\kappa>0$ such that

$$\mathbb{E}\left[\,\|s-\widehat{s}_{\widehat{m}}\|^2\,\right] \le C\left(C_\alpha \inf_{m\in\mathcal{M}_n}\left\{L^2\,D_m^{-2\alpha}+\frac{D_m}{n}\right\}\right)+\frac{\kappa}{n}\,.$$

Now, let us define the sequence $\{D_{m_n}\}_n$ such that for each n,

$$\frac{1}{2}L^{\frac{1}{1+2\alpha}}\, n^{\frac{1}{1+2\alpha}} \le D_{m_n} \le 2L^{\frac{1}{1+2\alpha}}\, n^{\frac{1}{1+2\alpha}}.$$

Then, we derive that it exists $K'_\alpha>0$ such that

$$\inf_{m\in\mathcal{M}_n}\mathbb{E}\left[\,\|s-\widehat{s}_m\|^2\,\right] \le C_\alpha\, L^2\, D_{m_n}^{-2\alpha} + \frac{D_{m_n}}{n} \le K'_\alpha\, L^{\frac{2}{1+2\alpha}}\, n^{-\frac{2\alpha}{1+2\alpha}},$$

hence the expected result. □

5.4 Besov functional spaces

The present section aims at deriving adaptivity in the minimax sense with respect to Besov spaces. This goal is reached thanks to results of Section 4.2 as well.

5.4.1 Overview of Besov spaces

Let us start by briefly recalling in what Besov spaces and balls consist in (see DeVore and Lorentz, 1993, for an extensive presentation on this matter).

For $\alpha > 0$ and $0 < p \leq +\infty$, a function f in $L^p([0,1])$ belongs to the Besov space $\mathcal{B}^\alpha_{\infty,p} = \mathcal{B}^\alpha_\infty\left(L^p([0,1])\right)$ if $|f|_{\mathcal{B}^\alpha_{\infty,p}} < +\infty$, where

$$|f|_{\mathcal{B}^\alpha_{\infty,p}} := \sup_{t>0}\left\{t^{-\alpha}\omega_r\left(f,t\right)_p\right\}, \qquad r = [\alpha]+1,$$

with

$$\omega_r\left(f,t\right)_p := \sup_{|h|\leq t}\|\Delta^r_h\left(f,\cdot\right)\|_p, \text{ and } \quad \Delta^r_h\left(f,x\right) := \sum_{k=1}^{r}\binom{k}{r}(-1)^{r-k}f\left(x+kh\right).$$

$|\cdot|_{\mathcal{B}^\alpha_{\infty,p}}$ is a semi-norm, while the metric is provided by the following Besov norm

$$\|f\|_{\mathcal{B}^\alpha_{\infty,p}} := |f|_{\mathcal{B}^\alpha_{\infty,p}} + \|f\|_p.$$

Moreover for a given real $R > 0$, let us define the Besov ball of radius R by

$$\mathcal{B}^\alpha_{\infty,p}(R) = \left\{f \in L^p \mid \|f\|_{\mathcal{B}^\alpha_{\infty,p}} \leq R\right\}.$$

In the sequel, the particular case where $p = 2$, that is $\mathcal{B}^\alpha_{\infty,2}$ for $\alpha > 0$ is considered.

5.4.2 Piecewise and trigonometric polynomials

In the same way as in Section 5.3, the strategy consists in deriving adaptivity results from the oracle inequalities of Section 4.2. Adaptivity heavily relies on the involved model collection through its approximation properties.
The following results therefore state adaptivity in the minimax sense for both (**Pp**) and (**Tp**) collections, with respect to respectively different Besov spaces.

The next theorem settles adaptivity with respect to Besov balls $\mathcal{B}^\alpha_{\infty,2}(R)$ for $0 < \alpha < r$, where r denotes the smallest integer larger than the degree of polynomials in (**Pp**).

Theorem 5.2. *Let us consider the collection of models (**Pp**) made of piecewise polynomials of degree less than r and assume that (Reg) , (Reg3) , (Ad) , and (Ran) hold. Then for $R > 0$ and $0 < \alpha < r$,*

$$\sup_{s\in\mathcal{B}^\alpha_{\infty,2}(R)}\mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2\right] \leq C_\alpha R^{\frac{2}{1+2\alpha}}n^{-\frac{2\alpha}{1+2\alpha}} + O\left(\frac{1}{n}\right), \tag{11}$$

where C_α denotes a given constant independent from n and s.

Proof. The proof follows the same strategy as that of Theorem 5.1 in that it essentially relies on approximation properties of models in (**Pp**).

If S_m denotes a model of dyadic piecewise polynomials of degree less than r on each one of the 2^m regular dyadic intervals, the result in page 359 of DeVore and Lorentz DeVore and Lorentz (1993) states that provided $r > \alpha$,

$$\inf_{u \in S_m} \|s - u\|^2 \leq K_{\alpha,r} |s|^2_{B^\alpha_{\infty,2}} (D_m)^{2\alpha},$$

for a positive constant $K_{\alpha,r}$.
Since $s \in \mathcal{B}^\alpha_{\infty,2}(R)$, it comes

$$\|s - s_m\|^2 \leq K_{\alpha,r} R^2 (D_m)^{2\alpha}.$$

As for the variance term,

$$\mathbb{E}\left[\|s_m - \widehat{s}_m\|^2 \right] \leq \frac{\|\phi_m\|_\infty}{n} \leq \frac{\Phi D_m}{n} \qquad \text{(by } (Reg3)\text{)}.$$

Under (Reg), $(Reg3)$, (Ad) and (Ran) we apply Theorem 4.2 to derive

$$\mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq \Gamma\left(K'_{\alpha,r} \inf_{m \in \mathcal{M}_n} \left\{ R^2 D_m^{-2\alpha} + \frac{D_m}{n} \right\} \right) + \frac{\kappa}{n},$$

where $K'_{\alpha,r}$ is a positive constant.
The conclusion results from the same calculation as in the proof of Theorem 5.1 with

$$\frac{1}{2} R^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}} \leq D_{m_n} \leq 2R^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}}.$$

□

Unlike the previous result, we now turn to Besov balls $\mathcal{B}^\alpha_{\infty,2}(R)$ for any value of $\alpha > 0$, which is enabled by the use of trigonometric polynomials, that is (**Tp**).

Theorem 5.3. *Let us consider the collection (**Tp**) made of trigonometric polynomials and assume that (Reg), $(Reg2)$, $(Reg3)$, (Ad) and (Ran) hold.*
Then for $R > 0$ and $\alpha > 0$,

$$\sup_{s \in \mathcal{B}^\alpha_{\infty,2}(R)} \mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq C'_\alpha R^{\frac{2}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}} + O\left(\frac{1}{n}\right), \tag{12}$$

for a given constant C'_α independent from n and s.

Proof. The same scheme of proof is used, except we need for an approximation result applying to trigonometric polynomials, which is also provided in page 205 of the book by DeVore and Lorentz (1993). Indeed considering models in (**Tp**) for any $\alpha > 0$, it comes

$$\inf_{u \in S_m} \|s - u\|^2 \leq K_\alpha |s|^2_{B^\alpha_{\infty,2}} (D_m)^{2\alpha},$$

for a constant $K_\alpha > 0$. Assumption $(Reg3)$ enables to conclude as in the previous theorem.
□

6 Conclusion

6.1 Summary of main contributions

In this work, CV has been studied as a model selection procedure in the density estimation setup. First, closed-form expressions have been derived for the leave-p-out (Lpo) estimator of the risk of projection estimators. These expressions drastically reduce computation time, which is a crucial issue, and also make V-fold cross-validation (VFCV) completely useless since it is more variable and expensive to carry out than Lpo. As an estimator of the risk, closed-form expressions for bias and variance of the Lpo estimate are provided as well.

Second, the Lpo estimator is embedded in the model selection via penalized criterion framework, which enables to shed new lights on the choice of p, the cardinality of the test set, with respect to the amount of penalization. It is shown that a wide range of penalization is available from the smallest one when $p = 1$, to penalties of the same order as BIC. Loo is definitely inappropriate to recover the true model as well as to perform model selection with too rich collections of models, especially exponential ones. The conclusions drawn here are all consistent with previous empirical results such as those of Breiman and Spector (1992) for instance.

Finally, two oracle inequalities are settled in density estimation with polynomial collections of models. These optimality results hold provided the ratio $0 < p/n < 1$ is neither too small, nor too large. To the best of our knowledge, these oracle inequalities are the first non-asymptotic results applying to Lpo in the density estimation setting. Furthermore with an appropriate choice of model collections, it is shown that CV procedure leads to estimators that are adaptive in the minimax sense with respect to Hölder as well as Besov spaces.

6.2 Discussion

On the one hand, the closed-form expressions settled on the present paper address the crucial issue of resampling procedures, that is their high computational complexity. Moreover, the broad class of projection estimators for which such formulas are obtained allows extensive applications of Lpo to wavelets, piecewise polynomials, and so on...

Besides, empirical evidence has been given in Celisse (2008) of an intricate relationship between the behaviour of Lpo with respect to p as model selection procedure and the size of the polynomial collection of models. In particular, it has been shown that Loo may suffer from overfitting with a polynomial collection of models provided the latter is "large enough". The analysis of this relationship deserves further investigations in order to describe the precise settings in which such troubles can occur. For instance, specifying the minimal value from which overfitting can be avoided seems highly desirable.

7 Proofs

7.1 Closed-form Lpo estimator

7.1.1 Proof of Lemma 3.1

The first remark is that for each $e \in \mathcal{E}_p$, we have $\forall t \in [0,1]$,

$$\begin{aligned}
\widehat{s}_m(X^{\overline{e}}_{1,n})(t) &= \frac{1}{n-p}\sum_{j\in\overline{e}}\sum_{\lambda}\varphi_\lambda(X_j)\varphi_\lambda(t) = \frac{1}{n-p}\sum_{j=1}^{n}\sum_{\lambda}\varphi_\lambda(X_j)\varphi_\lambda(t)\mathbb{1}_{(j\in\overline{e})},\\
\sum_{i\in e}\widehat{s}_m(X^{\overline{e}}_{1,n})(X_i) &= \frac{1}{n-p}\sum_{i=1}^{n}\sum_{j\in\overline{e}}\sum_{\lambda}\varphi_\lambda(X_j)\varphi_\lambda(X_i)\mathbb{1}_{(i\in e)}\\
&= \frac{1}{n-p}\sum_{i\neq j}\sum_{\lambda}\varphi_\lambda(X_j)\varphi_\lambda(X_i)\mathbb{1}_{(j\in\overline{e})}\mathbb{1}_{(i\in e)}.
\end{aligned}$$

Then, the Lemma follows from the following combinatorial results

Lemma 7.1. *For any $i \neq j \neq k \in \{1,\dots,n\}$,*

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

where the sum is computed over the resamples: Indices 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 $\{1,\dots,n\}$ 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. □

7.2 Moments calculations

7.2.1 Proof of Proposition 3.2

The expectation is a straightforward consequence of (4).

The variance calculation is not difficult, but very technical: Only the main steps of this proof are yielded.

First, let us 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

$$n(n-1)(n-p)\widehat{R}_p(m) = \sum_{\lambda}\left(\alpha A_\lambda + \beta B_\lambda\right).$$

Then,

$$\left[\sum_{\lambda}\left(\alpha A_\lambda+\beta B_\lambda\right)\right]^2=\sum_{\lambda}\left(\alpha^2A_\lambda^2+\beta^2B_\lambda^2+2\alpha\beta A_\lambda\,B_\lambda\right)+\sum_{\lambda\neq\lambda'}\left(\alpha^2A_\lambda A_{\lambda'}+\beta^2B_\lambda B_{\lambda'}+2\alpha\beta A_\lambda\,B_{\lambda'}\right).$$

After some calculation, the different terms are respectively equal to

$$\begin{aligned}
\mathbb{E}\sum_{\lambda}A_\lambda^2&=\sum_{\lambda}\left[\,n\,P\varphi_\lambda^4+t_1\left(P\varphi_\lambda^2\right)^2\,\right],\\
\mathbb{E}\sum_{\lambda}B_\lambda^2&=\sum_{\lambda}\left[\,4t_2P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2+2t_1\left(P\varphi_\lambda^2\right)^2+t_3\left(P\varphi_\lambda\right)^4\,\right],\\
\mathbb{E}\sum_{\lambda}A_\lambda B_\lambda&=\sum_{\lambda}\left[\,2t_1P\varphi_\lambda^3P\varphi_\lambda+t_2P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2\,\right],\\
\mathbb{E}\sum_{\lambda\neq\lambda'}A_\lambda A_{\lambda'}&=n\left[\mathbb{E}\left(\sum_{\lambda}\varphi_\lambda^2(X)\right)^2-\sum_{\lambda}P\varphi_\lambda^4\right]+t_1\left[\left(\sum_{\lambda}P\varphi_\lambda^2\right)^2-\sum_{\lambda}\left(P\varphi_\lambda^2\right)^2\right],\\
\mathbb{E}\sum_{\lambda\neq\lambda'}B_\lambda B_{\lambda'}&=2t_1\sum_{\lambda\neq\lambda'}\left(P\varphi_\lambda\varphi_{\lambda'}\right)^2+4t_2\left[\mathbb{E}\left(\sum_{\lambda}\varphi_\lambda(X)P\varphi_\lambda\right)^2-\sum_{\lambda}P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2\right]+\\
&\qquad t_3\left[\left(\sum_{\lambda}\left(P\varphi_\lambda\right)^2\right)^2-\sum_{\lambda}\left(P\varphi_\lambda\right)^4\right],\\
\mathbb{E}\sum_{\lambda\neq\lambda'}A_\lambda B_{\lambda'}&=2t_1\sum_{\lambda\neq\lambda'}P\varphi_\lambda^2\varphi_{\lambda'}P\varphi_{\lambda'}+t_2\left[\mathbb{E}\left(\sum_{\lambda}\varphi_\lambda^2(X)\right)\sum_{\lambda'}\left(P\varphi_{\lambda'}\right)^2-\mathbb{E}\left(\sum_{\lambda}\varphi_\lambda^2(X)\left(P\varphi_\lambda\right)^2\right)\right].
\end{aligned}$$

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.

7.2.2 Proof of Corollary 3.4

For every model $m \in \mathcal{M}_n$,

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

7.3 Theorem 4.1

At the beginning of this section, several preliminary results are enumerated, which are useful in the proof of Theorem 4.1. Then, the main steps of the strategy are briefly exposed, and the complete proof of the main result is finally provided. Proofs of preliminary results are given in the sections following the proof of Theorem 4.1.

7.3.1 Preliminaries

Notation First of all, let us define some notation that will be useful in the sequel.

For every $1 \le p \le n-1$ the Lpo risk estimator associated with the estimator $\widehat{s}_m$ is denoted by $\widehat{R}_p(m)$. For every m, set

$$L_p(m) = \mathbb{E}\widehat{R}_p(m)$$

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

$$\phi_m = \sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 \qquad \text{and} \qquad V_m = \mathbb{E}\left[\,\phi_m(X)\,\right],$$
$$\chi^2(m) = \|s_m - \widehat{s}_m\|^2 = \sum_\lambda \nu_n^2\left(\varphi_\lambda\right),$$
$$E_m = \mathbb{E}\left[\,\chi^2(m)\,\right] \qquad \text{and} \qquad \theta_{n,p} = \frac{2n-p}{(n-1)(n-p)}\cdot$$

Remark 5. *$\chi^2(m)$ is not a true χ^2 statistic, but is only somewhat similar to it.*

Two elementary but useful properties are repeatedly used in the sequel: For any $a, b \ge 0$,

$$(Roo) \qquad \sqrt{a+b} \le \sqrt{a} + \sqrt{b},$$
$$(Squ) \qquad 2ab \le \eta a^2 + \eta^{-1}b^2, \quad \forall \eta > 0.$$

Preliminary results Several preliminary results are then provided. They will be repeatedly referred to within the proof of Theorem 4.1.

The first result deals with the relationship between $\widehat{R}_p$ and its expectation for each model.

Lemma 7.2. *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 7.2, we see that $\nu_n(\phi_m)$ appears in the expressions. The next Proposition enables to upper bound the deviation of this quantity. It is a consequence of Bernstein's inequality (see Massart, 2007).

Proposition 7.1. *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*

$$\forall m \in \mathcal{M}_n, \quad \mathbb{P}\left[|\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m\right] \leq 2e^{-y_m}.$$

Moreover if (Ad) holds, we have

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid |\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m\right] \leq \Sigma_1 e^{-z},$$

where Σ_1 is a positive constant independent from n.

Besides, since $\chi^2(m) = \sum_\lambda \nu_n^2(\varphi_\lambda)$, 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 of $\nu_n(\varphi_\lambda)$. The event $\Omega_n(\epsilon)$ is therefore introduced:

$$\Omega_n(\epsilon) = \left\{\forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon\,\|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}}\right\},$$

where $\kappa(t) = 2(t^{-1} + 1/3)$.
Another use of Bernstein's inequality provides the following Lemma.

Lemma 7.3. *Set $\epsilon > 0$ and assume that (Reg) , $(Reg2)$ and (Pol) hold. Then,*

$$\forall \alpha > 0, \qquad \mathbb{P}\left[\Omega_n^c(\epsilon)\right] \leq 2n^{2+\delta}\, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2},$$

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 (2007) and is a consequence of Talagrand's inequality.

Proposition 7.2. *Set $\epsilon > 0$ and for any C', $z > 0$, $x_m = z + C' nE_m$. Let us assume that (Reg) , (Reg2) and (Pol) are fulfilled. Then,*

$$\forall m \in \mathcal{M}_n, \qquad \mathbb{P}\left[\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\left\|s\right\|_\infty x_m} \right) \right] \leq e^{-x_m}.$$

Furthermore if (Ad) holds,

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \ \sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\left\|s\right\|_\infty x_m} \right) \right] \leq \Sigma_2\, e^{-z},$$

where $\Sigma_2 > 0$ denotes a positive constant independent from n.

Finally, in Lemma 7.2, it remains $\nu_n(s_m)$ for which nothing has already been made. The control of this quantity results from the following lemma.

Lemma 7.4. *Set m, $m' \in \mathcal{M}_n$. Then for any $\rho > 0$,*

$$\sup_{t \in S_m + S_{m'}} \nu_n^2\left(\frac{t}{\left\|t\right\|}\right) \leq (1+\rho)\chi^2(m) + (1+\rho^{-1})\chi^2(m').$$

7.3.2 Outline of the strategy

Let us now describe the outlines of the strategy.

Since $\widehat{m} = \text{Argmin}_{m \in \mathcal{M}_n} \widehat{R}_p(m)$, it comes that for every $m \in \mathcal{M}_n$, $\widehat{R}_p(\widehat{m}) \leq \widehat{R}_p(m)$, which implies

$$\left[\widehat{R}_p(\widehat{m}) - L_p(\widehat{m}) \right] \leq \left[\widehat{R}_p(m) - L_p(m) \right] + \left[L_p(m) - L_p(\widehat{m}) \right]. \tag{13}$$

Then, Lemma 7.2 applied to (13) yields

$$\begin{aligned} \left\|s - s_{\widehat{m}}\right\|^2 + n\theta_{n,p}E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \leq \left\|s - s_m\right\|^2 + n\theta_{n,p}E_m - (1+\theta_{n,p})\chi^2(m) \\ + \theta_{n,p}\nu_n\left(\phi_m - \phi_{\widehat{m}}\right) \\ + 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right). \end{aligned} \tag{14}$$

Main steps

- In the oracle inequality one has in mind, the left-hand side of the final inequality is something like $\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}}\right\|^2 \right]$, which is equal to $\mathbb{E}\left[\left\|s - s_{\widehat{m}}\right\|^2 \right] + \mathbb{E}\left[\chi^2(\widehat{m}) \right]$ with the present notations. However in (14), the left-hand side is $\mathbb{E}\left[\left\|s - s_{\widehat{m}}\right\|^2 \right] + \mathbb{E}\left[E_{\widehat{m}} \right]$. In order to relate $\mathbb{E}\left[E_{\widehat{m}} \right]$ to $\mathbb{E}\left[\chi^2(\widehat{m}) \right]$, the discrepancy $E_m - \chi^2(m)$ will be uniformly controlled over $\mathcal{M}_n$ thanks to both Lemma 7.3 and Proposition 7.2.

- An upper bound of $\nu_n(\phi_m - \phi_{\widehat{m}})$ is obtained thanks to Proposition 7.1, so that $\nu_n(\phi_{\widehat{m}})$ is related to $E_{\widehat{m}}$.
- Finally, $\nu_n(s_{\widehat{m}} - s_m)$ may be upper bounded thanks to Lemma 7.4, independently from $E_{\widehat{m}}$ and will therefore be dealt with later.
- Combining these different steps, the desired inequality is derived except on a set of small probability (18). The conclusion results from the following lemma:

  Lemma 7.5. *Let X and Y be two random variables such that $\forall z > 0, \quad \mathbb{P}(X \geq Y + K_1 z + K_2) \leq \Sigma e^{-z}$, where $K_1,\ K_2,\ \Sigma > 0$. Then, we have*

  $$\mathbb{E}X \leq \mathbb{E}Y + K_1\,\Sigma + K_2.$$

  Proof. With $Z = X - Y - K_2$, one gets $\mathbb{P}(Z \geq K_1 z) \leq \Sigma e^{-z}$. Then,

  $$\begin{aligned} \mathbb{E}Z &\leq \mathbb{E}\left[\int_0^{+\infty} \mathbb{1}_{(t \leq Z)}\, dt\right] = \int_0^{+\infty} \mathbb{P}[\,t \leq Z\,]\, dt \\ &\leq K_1 \int_0^{+\infty} \Sigma e^{-z}\, dz = K_1 \Sigma. \end{aligned}$$

  □

7.3.3 Proof of Theorem 4.1

Proof. According to the previous remarks, Proposition 7.1 is applied to $\nu_n(\phi_m - \phi_{\widehat{m}})$. The successive use of (Reg) , (Squ) with any $\eta > 0$, and (Roo) provides

$$\sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} \leq \sqrt{2V_m \Phi y_m} \leq \eta \Phi V_m + \eta^{-1} y_m.$$

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 to

$$|\nu_n(\phi_m - \phi_{\widehat{m}})| \leq nE_m\left[\eta\Phi + C\eta^{-1} + \Phi\frac{C}{3}\right] + nE_{\widehat{m}}\left[\eta\Phi + C\eta^{-1} + \Phi\frac{C}{3}\right] + 2z\left[\frac{\Phi}{3} + \eta^{-1}\right] + 2\eta\Phi\|s\|^2,$$

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 (14) provides

$$\begin{aligned}
&\|s - s_{\widehat{m}}\|^2 + n\theta_{n,p}(1-\epsilon'')E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \\
&\leq \|s - s_m\|^2 + n\theta_{n,p}(1+\epsilon'')E_m - (1+\theta_{n,p})\chi^2(m) + 2(1+\theta_{n,p})\nu_n(s_{\widehat{m}} - s_m) \\
&\qquad + \theta_{n,p}\left(2z\Phi\left[\frac{1}{3} + \frac{3}{\epsilon''}\right] + 2\frac{\epsilon''}{3}\|s\|^2\right), \qquad (15)
\end{aligned}$$

except on an event of probability less than $\Sigma_1 e^{-z}$.

On the other hand, Proposition 7.2 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

$$\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)\left(\sqrt{nE_m}\left[1 + \sqrt{2\|s\|_\infty C'}\right] + \sqrt{2\|s\|_\infty z}\right),$$

which in turn, combined with (Squ) , implies for any $x > 0$

$$\chi^2(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)^2\left((1+x)E_m\left[1 + \sqrt{2\|s\|_\infty C'}\right]^2 + (1+x^{-1})\frac{2\|s\|_\infty}{n}z\right). \qquad (16)$$

It holds for the particular choices $x = \epsilon$ and $C' = \left(1 - \sqrt{1+\epsilon}\right)^2/\left(2\|s\|_\infty\right)$, which results in

$$\frac{1-\epsilon''}{(1+\epsilon)^4}\chi^2(\widehat{m})\mathbb{1}_{\Omega_n(\epsilon)} \leq (1-\epsilon'')E_{\widehat{m}} + \frac{1-\epsilon''}{\epsilon(1+\epsilon)}\frac{2\|s\|_\infty}{n}z,$$

with probability larger than $1-\Sigma_2 e^{-z}$.
From the above result and (15), it comes that on $\Omega_n(\epsilon)$, with probability larger than $1-(\Sigma_1+\Sigma_2)\,e^{-z}$, we have

$$\begin{aligned}
&\|s-s_{\widehat{m}}\|^2+\left(n\theta_{n,p}\frac{1-\epsilon''}{(1+\epsilon)^4}-(1+\theta_{n,p})\right)\chi^2(\widehat{m})\\
&\leq \|s-s_m\|^2+n\theta_{n,p}(1+\epsilon'')E_m-(1+\theta_{n,p})\chi^2(m)+2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}}-s_m\right)\\
&\qquad+\theta_{n,p}z\left(\frac{1-\epsilon''}{\epsilon(1+\epsilon)}2\|s\|_\infty+2\Phi\left[\frac{1}{3}+\frac{3}{\epsilon''}\right]\right)\\
&\qquad+2\theta_{n,p}\frac{\epsilon''}{3}\|s\|^2.
\end{aligned}$$

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

$$\begin{aligned}
&\|s-s_{\widehat{m}}\|^2+\left[\,n\theta_{n,p}\left(1-\epsilon'\right)-(1+\theta_{n,p})\,\right]\chi^2(\widehat{m})\\
&\leq \|s-s_m\|^2+n\theta_{n,p}\left[2-\sqrt{1-\epsilon'}\right]E_m-(1+\theta_{n,p})\chi^2(m)+2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}}-s_m\right)\\
&\qquad+\theta_{n,p}z\left(\frac{\sqrt{1-\epsilon'}}{\epsilon(1+\epsilon)}2\|s\|_\infty+2\Phi\left[\frac{1}{3}+\frac{3}{1-\sqrt{1-\epsilon'}}\right]\right)+2\theta_{n,p}\frac{1-\sqrt{1-\epsilon'}}{3}\|s\|^2.
\end{aligned}\tag{17}$$

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 7.4. Indeed, we have

$$2\nu_n\left(s_{\widehat{m}}-s_m\right)\leq 2\nu_n\left(\frac{s_{\widehat{m}}-s_m}{\|s_{\widehat{m}}-s_m\|}\right)\|s_{\widehat{m}}-s_m\|\leq 2\sup_{t\in S_{\widehat{m}}+S_m}\nu_n\left(\frac{t}{\|t\|}\right)\|s_{\widehat{m}}-s_m\|.$$

Moreover, $\|s_{\widehat{m}}-s_m\|\leq\|s_{\widehat{m}}-s\|+\|s-s_m\|$ and a double use of (Squ) give for any $x>0$:

$$2\nu_n\left(s_{\widehat{m}}-s_m\right)\quad\leq\quad(1+x)\sup_{t\in S_{\widehat{m}}+S_m}\nu_n^2\left(\frac{t}{\|t\|}\right)+\frac{2}{2+x}\|s_{\widehat{m}}-s\|^2+\frac{2}{x}\|s_m-s\|^2.$$

Finally, Lemma 7.4 yields that for any $\rho>0$, we have

$$\begin{aligned}
2\nu_n\left(s_{\widehat{m}}-s_m\right)&\leq(1+x)\left[\,(1+\rho)\chi^2(\widehat{m})+(1+\rho^{-1})\chi^2(m)\,\right]\\
&\qquad+\frac{2}{2+x}\|s_{\widehat{m}}-s\|^2+\frac{2}{x}\|s_m-s\|^2.
\end{aligned}$$

With $x=\epsilon'$ and $\rho=\epsilon'(1+\epsilon')^{-1}$, we get

$$\begin{aligned}
2\nu_n\left(s_{\widehat{m}}-s_m\right)&\leq\left(1+2\epsilon'\right)\chi^2(\widehat{m})+(1+\epsilon')\frac{1+2\epsilon'}{\epsilon'}\chi^2(m)\\
&\qquad+\frac{2}{2+\epsilon'}\|s_{\widehat{m}}-s\|^2+\frac{2}{\epsilon'}\|s_m-s\|^2.
\end{aligned}$$

Plugging this in (17) yields:
On the event $\Omega_n(\epsilon)$, with probability larger than $1-(\Sigma_1+\Sigma_2)\,e^{-z}$, we have for any $m\in\mathcal{M}_n$

$$\begin{aligned}&\left[\frac{\epsilon'-2\theta_{n,p}}{2+\epsilon'}\right]\|s-s_{\widehat{m}}\|^2+\left[\,n\theta_{n,p}\left(1-\epsilon'\right)-2(1+\theta_{n,p})(1+\epsilon')\,\right]\chi^2(\widehat{m})\\ &\leq\left[1+\frac{2}{\epsilon'}(1+\theta_{n,p})\right]\|s-s_m\|^2+n\theta_{n,p}\left[2-\sqrt{1-\epsilon'}\right]E_m\\ &\qquad+\left[\frac{1+2\epsilon'+2\epsilon'^2}{\epsilon'}\right](1+\theta_{n,p})\chi^2(m)+\theta_{n,p}(Az+B),\end{aligned}\tag{18}$$

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 7.5 allows us to take the expectation and get the following result.

$$\begin{aligned}(\psi_1\wedge\psi_2)\,\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)}\,\|s-\widehat{s}_{\widehat{m}}\|^2\right]&\leq(\psi_3\vee\psi_4)\,\mathbb{E}\left[\,\|s-\widehat{s}_m\|^2\,\right]\\ &\quad+\theta_{n,p}\left[\,A\,(\Sigma_1+\Sigma_2)+B\,\right],\end{aligned}\tag{19}$$

where $\psi_1=(\epsilon'-2\theta_{n,p})\,(2+\epsilon')^{-1}$, $\psi_2=n\theta_{n,p}\,(1-\epsilon')-2(1+\theta_{n,p})(1+\epsilon')$, $\psi_3=1+2/\epsilon'\,(1+\theta_{n,p})$, and $\psi_4=n\theta_{n,p}\left[2-\sqrt{1-\epsilon'}\right]+(1+\theta_{n,p})\left[1+2\epsilon'+2\epsilon'^2\right]/\epsilon'$.

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

$$\frac{4\epsilon'}{1+3\epsilon'}+\frac{2}{n}\frac{1+\epsilon'}{1+3\epsilon'}\leq\frac{p}{n}\leq 1-\frac{2}{\epsilon'(n-1)-2},\tag{20}$$

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 7.6 for $n\geq 29$.

Remark 6. *In (20) since $0<\epsilon'\leq 1$ by definition, we have $\frac{4\epsilon'}{1+3\epsilon'}\leq 1$.*

Finally to assert the existence of the constant Γ in Theorem 4.1, the ratio $(\psi_3\vee\psi_4)\,/\,(\psi_1\wedge\psi_2)$ has to be bounded. One can easily check that all ψ_ks can be re-shaped as

$$\psi_k=\frac{F(p,n)}{1-p/n},$$

where F is a bounded quantity. Moreover by construction, the bounds in (20) 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 yields

$$\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)} \|s - \widehat{s}_{\widehat{m}}\|^2\right] \leq \Gamma(\epsilon,\alpha,\beta) \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\|s - \widehat{s}_m\|^2\right] + \frac{\kappa(\epsilon,s,\Phi,\alpha,\beta,\delta)}{n}.$$

We now simply add the missing term $\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)^c} \|s - \widehat{s}_{\widehat{m}}\|^2\right]$ to both sides of the above inequality. It only remains to show that this term is of the right order:

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \|s - \widehat{s}_{\widehat{m}}\|^2\right] &\leq \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \|s - s_{\widehat{m}}\|^2\right] + \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\|^2\right], \\ &\leq \|s\|^2 \mathbb{P}\left[\Omega_n^c(\epsilon)\right] + \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \sum_{\lambda\in\widehat{m}} \left[\nu_n(\varphi_\lambda)^2\right]\right].\end{aligned}$$

Lemma 7.3 then enables to deduce that the first term in the right-hand side inequality satisfies

$$\forall n, \qquad \|s\|^2 \mathbb{P}\left[\Omega_n^c(\epsilon)\right] \leq \|s\|^2 \frac{n_0}{n},$$

for an appropriate choice of $n_0 > 0$, depending on ϵ, δ and Φ.
For the second one, Jensen's inequality yields

$$\mathbb{E}\left[\sum_{\lambda\in\widehat{m}} \nu_n^2(\varphi_\lambda)\, \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \leq \mathbb{E}\left[\sum_{\lambda\in\widehat{m}} (\varphi_\lambda(X) - P\varphi_\lambda)^2\, \mathbb{1}_{\Omega_n^c(\epsilon)}\right].$$

Moreover, (Squ) with any $\eta > 0$ provides

$$(\varphi_\lambda(X) - P\varphi_\lambda)^2 \leq (1+\eta)\varphi_\lambda^2(X) + (1+\eta^{-1})P\varphi_\lambda^2.$$

Finally, $\sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 = \phi_m$ and $P\phi_{\widehat{m}} \leq \|\phi_{\widehat{m}}\|_\infty$ lead to

$$\begin{aligned}\mathbb{E}\left[\sum_{\lambda\in\widehat{m}} \nu_n^2(\varphi_\lambda)\, \mathbb{1}_{\Omega_n^c(\epsilon)}\right] &\leq \left(2+\eta+\eta^{-1}\right) \mathbb{E}\left[\|\phi_{\widehat{m}}\|_\infty \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \\ &\leq \left(2+\eta+\eta^{-1}\right) \frac{\Phi n}{(\log n)^2} \mathbb{P}\left[\Omega_n^c(\epsilon)\right]\end{aligned}$$

thanks to (Reg), and Lemma 7.3 enables to conclude.

□

7.3.4 Proof of Proposition 7.1

Proof. Bernstein's inequality Massart (2007) 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

$$b = \|\phi_m\|_\infty \qquad \text{and} \qquad v \leq n\,V_m\,\|\phi_m\|_\infty\,,$$

hence the first part of the proposition.
For the second part of the result, the union bound combined with $y_m = z + C\,nE_m$ provide

$$\begin{aligned} &\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \left|\nu_n\left(\phi_m\right)\right| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m \right] \\ \leq\;& e^{-z} \sum_{m\in\mathcal{M}_n} e^{-C\,nE_m} \leq \Sigma_1\, e^{-z} \qquad (Ad) \quad \text{and} \quad (Pol). \end{aligned}$$

□

7.3.5 Proof of Lemma 7.3

Proof. We recall that

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad \left|\nu_n\left(\varphi_\lambda\right)\right| \leq \frac{2\epsilon\,\|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\}.$$

Then, we deduce that

$$\begin{aligned} \mathbb{P}\left[\,\Omega_n^c(\epsilon)\,\right] &= \mathbb{P}\left[\left\{ \exists m \in \mathcal{M}_n,\ \exists \lambda \in \Lambda(m) \mid \left|\nu_n\left(\varphi_\lambda\right)\right| \geq \frac{2\epsilon\,\|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\} \right], \\ &\leq \sum_{m\in\mathcal{M}_n} D_m\, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2} \quad \text{(Bernstein)} \\ &\leq \sum_{D\geq 1} D^{\delta+1} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2} \quad (Pol) \\ &\leq n^{\delta+2} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2} \quad (D \leq n) \end{aligned}$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

□

7.3.6 Proof of Proposition 7.2

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

$$\left\{a = (a_1,\dots,a_{D_m}) \in \mathbb{R}^{D_m} \mid \sum_{\lambda\in\Lambda(m)} \alpha_\lambda^2 = 1 \text{ and } \sum_{\lambda\in\Lambda(m)} |\alpha_\lambda| \leq \frac{t}{z}\right\}.$$

Moreover, if we define the event

$$\Omega = \left\{\sup_{\lambda\in\Lambda(m)} \nu_n(\varphi_\lambda) \leq t\right\}$$

for $t > 0$, then we deduce that

$$\chi(m) \leq \sup_{a\in A}\left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right| \tag{21}$$

on $\Omega \cap \{\chi(m) \geq z\}$.

Then, Talagrand's inequality applied to $\sup_{a\in A}\left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right|$ gives for $\epsilon, x > 0$,

$$\mathbb{P}\left[\mathbb{1}_\Omega \sup_{a\in A}\left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right| \geq (1+\epsilon)\left(\sqrt{\chi^2(m)} + \sqrt{\frac{2\|s\|_\infty}{n}x}\right)\right] \leq e^{-x},$$

with $z = \sqrt{2\|s\|_\infty / n}$ and $t = 2\epsilon\|s\|_\infty \left[\kappa(\epsilon)\Phi n/(\log n)^2\right]^{-1}$.

Finally, the first result comes from both (21) and $\Omega_n(\epsilon) = \Omega$.

As for the second inequality, the choice $x_m = C'\xi D_m + z$ leads to

$$\begin{aligned}
&\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty x_m}\right)\right] \\
\leq& e^{-z}\sum_{m\in\mathcal{M}_n} e^{-C' nE_m} \leq e^{-z}\sum_{D\geq 1} e^{-C'\xi D + \delta\log D} \quad (Ad) \text{ and } (Pol) \\
\leq& \Sigma_2\, e^{-z}.
\end{aligned}$$

□

7.3.7 Proof of Lemma 7.6

Lemma 7.6. *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. □

7.4 Theorem 4.2

7.4.1 Intermediate results

The proof of Theorem 4.2 follows the same structure as that of Theorem 4.1. Only the main differences are reported here. These differences essentially occur in the control of the χ^2-type statistic. Since they are nearly the same, the following results are given (without or) with only short proofs.

Let us start by introducing another event of large probability on which we are able to get the desired control. For any $\epsilon > 0$,

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon \|s\|_2 \sqrt{\Phi/\xi\, n E_m}\log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\},$$

where $\kappa(t) = 2(t^{-1} + 1/3)$.

The following lemma is the counterpart of Lemma 7.3 and is devoted to control the remainder terms. It heavily relies on Bernstein's inequality.

Lemma 7.7. *Set $\epsilon > 0$ and assume that (Reg) , (Reg2) , (Reg3) (Ad) and (Pol) hold. Then,*

$$\forall \alpha > 0, \qquad \mathbb{P}\left[\,\Omega_n^c(\epsilon)\,\right] \leq 2n^{2+\delta}\, e^{-\frac{\eta(\epsilon)}{\sqrt{\Phi}}(\|s\|\vee 1)(\log n)^2},$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

Now, we are in position to give the main result providing the desired control on the χ^2-type statistic.

Proposition 7.3. *Set $\epsilon > 0$ and for any C', $z > 0$, $x_m = z + C'\sqrt{nE_m}$. Assume that (Reg) , (Reg2) , (Reg3) , (Ad) and (Pol) are fulfilled. Then, for every $m \in \mathcal{M}_n$,*

$$\mathbb{P}\left[\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi\, nE_m}\, x_m}\right)\right] \leq e^{-x_m},$$

and furthermore,

$$\mathbb{P}\left[\exists m \mid \sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi\, nE_m}\, x_m}\right)\right] \leq \Sigma_2\, e^{-z},$$

where $\Sigma_2 > 0$ denotes a positive constant independent from n.

Proof. (sketch of proof) It relies on Talagrand's inequality as well as of the following straightforward upper bound.

$$\forall m, \qquad \sup_{t\in S_m,\, \|t\|_2=1} \mathrm{Var}\left[\,t(X)\,\right] \leq \|s\|\,\|t\|_2\sqrt{\|\phi_m\|_\infty} = \|s\|\sqrt{\|\phi_m\|_\infty} \leq (\|s\|\vee 1)\sqrt{\|\phi_m\|_\infty}.$$

□

7.4.2 Outline of the proof of Theorem 4.2

The first main difference comes from the use of Proposition 16, which yields

$$\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi\, nE_m}\, x_m}\right)$$

on an event of high probability.
From several applications of (Squ) and (Roo) , with $\rho, C' > 0$, it comes

$$\begin{aligned}
&\sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi\, nE_m}\, x_m} \\
\leq &\sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi}\,\sqrt{nE_m}\, z} + \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi} C' nE_m}\,, \\
\leq &\quad \rho\sqrt{nE_m} + \rho^{-1}(\|s\|\vee 1)\sqrt{\Phi/\xi}\, z + C\sqrt{nE_m}, \\
\leq &\quad (\rho + C)\sqrt{nE_m} + \rho^{-1}(\|s\|\vee 1)\sqrt{\Phi/\xi}\, z,
\end{aligned}$$

with $C' = C\left[2(\|s\| \vee 1)\sqrt{\Phi/\xi}\right]^{-1}$.
Thus in the same way as (16), for every $x > 0$, we derive

$$\chi^2(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)^2 \left[(1+x)E_m\left[1+(\rho+C)\right]^2 + (1+x^{-1})\left(\rho^{-1}(\|s\| \vee 1)\sqrt{\Phi/\xi}\right)^2 \frac{z^2}{n}\right].$$

The following remains essentially the same and concludes the proof.

References

Akaike, H. (1973). 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.

Arlot, S. (2007a). *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

Arlot, S. (2007b). V-fold penalization: an alternative to v-fold cross-validation. In *Oberwolfach Reports*, volume 4 of *Mathematisches Forschungsinstitut*. European Mathematical Society (EMS), Zürich. Report No.50/2007. Workshop: Reassessing the Paradigms of Statistical Model Building.

Arlot, S. and Celisse, A. (2009). Segmentation in the mean of heteroscedastic data via cross-validation. Technical report, arxiv.

Baraud, Y., Giraud, C., and Huet, S. (2009). Gaussian model selection with unknown variance. *The Annals of Statistics*, 37(2):630–672.

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

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

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

Birgé, L. and Rozenholc, Y. (2006). How many bins should be put in a regular histogram? *ESAIM Probab. Statist.*, 10:24–45.

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

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

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*.

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

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

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

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

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

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

Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report 99.61, Université de Paris-Sud.

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

Celisse, A. (2008). *Model selection via cross-validation in density estimation, regression and change-points detection*. PhD thesis, University Paris-Sud 11.

Celisse, A. and Robin, S. (2008a). A leave-p-out based estimation of the proportion of null hypotheses. Technical report, arXiv: 0804.1189.

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

DeVore, R. and Lorentz, G. (1993). *Constructive Approximation*. Springer.

Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539.

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

Efron, B. (1982). 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.

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

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328.

Giné, E. (1997). *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.

Ibragimov, I. and Khas'minskij, R. (1981). *Statistical Estimation. Asymptotic Theory.* Springer-Verlag, Berlin.

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

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

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

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661–675.

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

Massart, P. (2007). *Concentration Inequalities and Model Selection.* Lecture Notes in Mathematics. Springer.

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

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

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

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

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

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

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

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147. 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.

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

Tukey, J. W. (1958). Bias and confidence in not quite large samples (Abstract). *Ann. Math. Statist.*, 29:614.

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

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

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