Journal of Machine Learning Research 17 (2016) 1-50 Submitted 7/14; Revised 7/15; Published 12/16

Choice of V for V-Fold Cross-Validation in Least-Squares Density Estimation

Sylvain Arlot SYLVAIN.ARLOT@MATH.U-PSUD.FR
*Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay
91405 Orsay, France*

Matthieu Lerasle MLERASLE@UNICE.FR
*CNRS
Univ. Nice Sophia Antipolis LJAD CNRS UMR 7351
06100 Nice France*

Editor: Xiaotong Shen

Abstract

This paper studies V-fold cross-validation for model selection in least-squares density estimation. The goal is to provide theoretical grounds for choosing V in order to minimize the least-squares loss of the selected estimator. We first prove a non-asymptotic oracle inequality for V-fold cross-validation and its bias-corrected version (V-fold penalization). In particular, this result implies that V-fold penalization is asymptotically optimal in the nonparametric case. Then, we compute the variance of V-fold cross-validation and related criteria, as well as the variance of key quantities for model selection performance. We show that these variances depend on V like $1 + 4/(V - 1)$, at least in some particular cases, suggesting that the performance increases much from $V = 2$ to $V = 5$ or 10, and then is almost constant. Overall, this can explain the common advice to take $V = 5$—at least in our setting and when the computational power is limited—, as supported by some simulation experiments. An oracle inequality and exact formulas for the variance are also proved for Monte-Carlo cross-validation, also known as repeated cross-validation, where the parameter V is replaced by the number B of random splits of the data.

Keywords: V-fold cross-validation, Monte-Carlo cross-validation, leave-one-out, leave-p-out, resampling penalties, density estimation, model selection, penalization

1. Introduction

Cross-validation methods are widely used in machine learning and statistics, for estimating the risk of a given statistical estimator (Stone, 1974; Allen, 1974; Geisser, 1975) and for selecting among a family of estimators. For instance, cross-validation can be used for model selection, where a collection of linear spaces is given (the models) and the problem is to choose the best least-squares estimator over one of these models. Cross-validation is also often used for choosing hyperparameters of a given learning algorithm. We refer to Arlot and Celisse (2010) for more references about cross-validation for model selection.

Model selection can target two different goals: (i) *estimation*, that is, minimizing the risk of the final estimator, which is the goal of AIC and related methods, or (ii) *identification*, that is, identifying the smallest true model in the family considered, assuming it exists and it is unique, which is the goal of BIC for instance; see the survey by Arlot and Celisse (2010) for more details about this distinction. These two goals cannot be attained simultaneously in general (Yang, 2005).

We assume throughout the paper that the goal of model selection is estimation. We refer to Yang (2006, 2007) and Celisse (2014) for some results and references on cross-validation methods with an identification goal.

Then, a natural question arises: which cross-validation method should be used for minimizing the risk of the final estimator? For instance, a popular family of cross-validation methods is V-fold cross-validation (Geisser, 1975, often called k-fold cross-validation), which depends on an integer parameter V, and enjoys a smaller computational cost than other classical cross-validation methods. The question becomes (1) which V is optimal, and (2) can we do almost as well as the optimal V with a small computational cost, that is, a small V? Answering the second question is particularly useful for practical applications where the computational power is limited.

Surprisingly, few theoretical results exist for answering these two questions, especially with a non-asymptotic point of view (Arlot and Celisse, 2010). In short, it is proved in least-squares regression that at first order, V-fold cross-validation is suboptimal for model selection (with an estimation goal) if V stays bounded, because V-fold cross-validation is biased (Arlot, 2008). When correcting for the bias (Burman, 1989; Arlot, 2008), we recover asymptotic optimality whatever V, but without any theoretical result distinguishing among values of V in second order terms in the risk bounds (Arlot, 2008).

Intuitively, if there is no bias, increasing V should reduce the variance of the V-fold cross-validation estimator of the risk, hence reduce the risk of the final estimator, as supported by some simulation experiments (Arlot, 2008, for instance). But variance computations for unbiased V-fold methods have only been made in the asymptotic framework for a fixed estimator, and they focus on risk estimation instead of model selection (Burman, 1989).

This paper aims at providing theoretical grounds for the choice of V by two means: a non-asymptotic oracle inequality valid for any V (Section 3) and exact variance computations shedding light on the influence of V on the variance (Section 5). In particular, we would like to understand why the common advice in the literature is to take $V = 5$ or 10, based on simulation experiments (Breiman and Spector, 1992; Hastie et al., 2009, for instance).

The results of the paper are proved in the least-squares density estimation framework, because we can then benefit from explicit closed-form formulas and simplifications for the V-fold criteria. In particular, we show that V-fold cross-validation and all leave-p-out methods are particular cases of V-fold penalties in least-squares density estimation (Lemma 1).

The first main contribution of the paper (Theorem 5) is an oracle inequality with leading constant $1 + \varepsilon_n$, with $\varepsilon_n \to 0$ as $n \to \infty$ for unbiased V-fold methods, which holds for any value of V. To the best of our knowledge, Theorem 5 is the first non-asymptotic oracle inequality for V-fold methods enjoying such properties: the leading constant $1 + \varepsilon_n$ is new in density estimation, and the fact that it holds whatever the value of V had never been obtained in any framework. Theorem 5 relies on a new concentration inequality for the V-fold penalty (Proposition 4). Note that Theorem 5 implicitly assumes that the oracle loss is of order $n^{-\alpha}$ for some $\alpha \in (0, 1)$, that is, the setting is nonparametric; otherwise, Theorem 5 may not imply the asymptotic optimality of V-fold penalization. Let us also emphasize that the leading constant is $1+\varepsilon_n$ whatever V for unbiased V-fold methods, with ε_n independent from V in Theorem 5. So, second-order terms must be taken into account for understanding how the model selection performance depends on V. Section 4 proposes a heuristic for comparing these second order terms thanks to variance comparisons. This motivates our next result.

The second main contribution of the paper (Theorem 6) is the first non-asymptotic variance computation for V-fold criteria that allows to understand precisely how the *model selection performance* of V-fold cross-validation or penalization depends on V. Previous results only focused on the variance of the V-fold criterion (Burman, 1989; Bengio and Grandvalet, 2005; Celisse, 2008, 2014; Celisse and Robin, 2008), which is not sufficient for our purpose, as explained in Section 4. In our setting, we can explain, partly from theoretical results, partly from a heuristic argument, why taking, say, $V > 10$ is not necessary for getting a performance close to the optimum, as supported by experiments on synthetic data in Section 6.

An oracle inequality and exact formulas for the variance are also proved for other cross-validation methods: Monte-Carlo cross-validation, also known as repeated cross-validation, where the parameter V is replaced by the number B of random splits of the data (Section 8.1), and hold-out penalization (Section 8.2).

Notation. For any integer $k \geqslant 1$, $\llbracket k \rrbracket$ denotes $\{1, \ldots, k\}$.

For any vector $\xi_{\llbracket n \rrbracket} := (\xi_1, \ldots, \xi_n)$ and any $B \subset \llbracket n \rrbracket$, ξ_B denotes $(\xi_i)_{i\in B}$, $|B|$ denotes the cardinality of B and $B^c = \llbracket n \rrbracket \setminus B$.

For any real numbers t, u, we define $t \vee u := \max\{t, u\}$, $u_+ := u \vee 0$ and $u_- := (-u) \vee 0$.

All asymptotic results and notation $\mathrm{o}(\cdot)$ or $\mathcal{O}(\cdot)$ are for the regime when the number n of observations tends to infinity.

2. Least-Squares Density Estimation and Definition of V-Fold Procedures

This section introduces the framework of the paper, the main procedures studied, and some useful notation.

2.1 General Statistical Framework

Let $\xi, \xi_1, ..., \xi_n$ be independent random variables taking value in a Polish space $\mathcal{X}$, with common distribution P and density s with respect to some known measure μ. Suppose that $s \in L^\infty(\mu)$, which implies that $s \in L^2(\mu)$. The goal is to estimate s from $\xi_{\llbracket n \rrbracket} = (\xi_1, \ldots, \xi_n)$, that is, to build an estimator $\widehat{s} = \widehat{s}(\xi_{\llbracket n \rrbracket}) \in L^2(\mu)$ such that its loss $\|\widehat{s} - s\|^2$ is as small as possible, where for any $t \in L^2(\mu)$, $\|t\|^2 := \int_{\mathcal{X}} t^2 \,\mathrm{d}\mu$.

Projection estimators are among the most classical estimators in this framework (see, for example, DeVore and Lorentz, 1993 and Massart, 2007). Given a separable linear subspace S_m of $L^2(\mu)$ (called a model), the projection estimator of s onto S_m is defined by

$$\widehat{s}_m := \operatorname*{argmin}_{t \in S_m} \left\{ \|t\|^2 - 2P_n(t) \right\} \ , \tag{1}$$

where P_n is the empirical measure; for any $t \in L^2(\mu)$, $P_n(t) = \int t dP_n = \frac{1}{n} \sum_{i=1}^n t(\xi_i)$. The quantity minimized in the definition of $\widehat{s}_m$ is often called the empirical risk, and can be denoted by

$$P_n\gamma(t) = \|t\|^2 - 2P_n(t) \qquad \text{where} \quad \forall x \in \mathcal{X}, \forall t \in L^2(\mu), \quad \gamma(t;x) = \|t\|^2 - 2t(x) \enspace .$$

The function γ is called the least-squares contrast. Note that $S_m \subset L^1(P)$ since $s \in L^2(\mu)$.

2.2 Model Selection

When a finite collection of models $(S_m)_{m\in\mathcal{M}_n}$ is given, following Massart (2007), we want to choose from data one among the corresponding projection estimators $(\widehat{s}_m)_{m\in\mathcal{M}_n}$. The goal is to design a model selection procedure $\widehat{m} : \mathcal{X}^n \mapsto \mathcal{M}_n$ so that the final estimator $\widetilde{s} := \widehat{s}_{\widehat{m}}$ has a quadratic loss as small as possible, that is, comparable to the oracle loss $\inf_{m\in\mathcal{M}_n} \|\widehat{s}_m - s\|^2$. This goal is what is called the estimation goal in the Introduction. More precisely, we aim at proving that an oracle inequality of the form

$$\|\widehat{s}_{\widehat{m}} - s\|^2 \leqslant C_n \inf_{m\in\mathcal{M}_n} \left\{ \|\widehat{s}_m - s\|^2 \right\} + R_n$$

holds with a large probability. The procedure $\widehat{m}$ is called asymptotically optimal when R_n is much smaller than the oracle loss and $C_n \to 1$, as $n \to +\infty$. In order to avoid trivial cases, we will always assume that $|\mathcal{M}_n| \geqslant 2$.

In this paper, we focus on model selection procedures of the form

$$\widehat{m} := \operatorname*{argmin}_{m\in\mathcal{M}_n} \left\{ \mathrm{crit}(m) \right\} \enspace ,$$

where crit : $\mathcal{M}_n \mapsto \mathbb{R}$ is some data-driven criterion. Since our goal is to satisfy an oracle inequality, an ideal criterion is

$$\mathrm{crit}_{\mathrm{id}}(m) = \|\widehat{s}_m - s\|^2 - \|s\|^2 = -2P(\widehat{s}_m) + \|\widehat{s}_m\|^2 = P\gamma(\widehat{s}_m) \enspace .$$

Penalization is a popular way of designing a model selection criterion (Barron et al., 1999; Massart, 2007)

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

for some penalty function pen : $\mathcal{M}_n \to \mathbb{R}$, possibly data-driven. From the ideal criterion $\mathrm{crit}_{\mathrm{id}}$, we get the ideal penalty

$$\begin{aligned} \mathrm{pen}_{\mathrm{id}}(m) &:= \mathrm{crit}_{\mathrm{id}}(m) - P_n\gamma(\widehat{s}_m) = (P - P_n)\gamma(\widehat{s}_m) = 2(P_n - P)(\widehat{s}_m) \qquad (2) \\ &= 2(P_n - P)(\widehat{s}_m - s_m) + 2(P_n - P)(s_m) = 2\|\widehat{s}_m - s_m\|^2 + 2(P_n - P)(s_m) \enspace , \end{aligned}$$

$$\text{where} \quad s_m := \operatorname*{argmin}_{t\in S_m} \left\{ P\gamma(t) \right\} = \operatorname*{argmin}_{t\in S_m} \left\{ \|t - s\|^2 \right\}$$

is the orthogonal projection of s onto S_m in $L^2(\mu)$. Let us finally recall some useful and classical reformulations of the main term in the ideal penalty (2), that proves in particular the last equality in Eq. (2): If $\mathbb{B}_m = \{t \in S_m \text{ s.t. } \|t\| \leqslant 1\}$ and $(\psi_\lambda)_{\lambda\in\Lambda_m}$ denotes an orthonormal basis of S_m in $L^2(\mu)$, then

$$\begin{aligned} (P_n - P)(\widehat{s}_m - s_m) &= \sum_{\lambda\in\Lambda_m} \left[(P_n - P)(\psi_\lambda) \right]^2 \\ &= \|\widehat{s}_m - s_m\|^2 = \sup_{t\in\mathbb{B}_m} \left[(P_n - P)(t) \right]^2 \enspace , \end{aligned} \qquad (3)$$

where the last equality follows from Eq. (30) in Appendix A.

2.3 V-Fold Cross-Validation

A standard approach for model selection is cross-validation. We refer the reader to Arlot and Celisse (2010) for references and a complete survey on cross-validation for model selection. This section only provides the minimal definitions and notation necessary for the remainder of the paper.

For any subset $A \subset [\![n]\!]$, let

$$P_n^{(A)} := \frac{1}{|A|} \sum_{i \in A} \delta_{\xi_i} \quad \text{and} \quad \widehat{s}_m^{(A)} := \operatorname*{argmin}_{t \in S_m} \left\{ \|t\|^2 - 2P_n^{(A)}(t) \right\} \ .$$

The main idea of cross-validation is data splitting: some $T \subset [\![n]\!]$ is chosen, one first trains $\widehat{s}_m(\cdot)$ with ξ_T, then test the trained estimator on the remaining data ξ_{T^c}. The hold-out criterion is the estimator of $\mathrm{crit}_{\mathrm{id}}(m)$ obtained with this principle, that is,

$$\mathrm{crit}_{\mathrm{HO}}(m,T) := P_n^{(T^c)} \gamma\left(\widehat{s}_m^{(T)} \right) = -2P_n^{(T^c)} \left(\widehat{s}_m^{(T)} \right) + \left\| \widehat{s}_m^{(T)} \right\|^2 \ , \tag{4}$$

and all cross-validation criteria are defined as averages of hold-out criteria with various subsets T.

Let $V \in \{2, \ldots, n\}$ be a positive integer and let $\mathcal{B} = \mathcal{B}_{[\![V]\!]} = (\mathcal{B}_1, \ldots, \mathcal{B}_V)$ be some partition of $[\![n]\!]$. The V-fold cross-validation criterion is defined by

$$\mathrm{crit}_{\mathrm{VFCV}}(m, \mathcal{B}) := \frac{1}{V} \sum_{K=1}^{V} \mathrm{crit}_{\mathrm{HO}}(m, \mathcal{B}_K^c) \ .$$

Compared to the hold-out, one expects cross-validation to be less variable thanks to the averaging over V splits of the sample into $\xi_{\mathcal{B}_K}$ and $\xi_{\mathcal{B}_K^c}$.

Since $\mathrm{crit}_{\mathrm{VFCV}}(m, \mathcal{B})$ is known to be a biased estimator of $\mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m)]$, Burman (1989) proposed the bias-corrected V-fold cross-validation criterion

$$\mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}) := \mathrm{crit}_{\mathrm{VFCV}}(m, \mathcal{B}) + P_n \gamma(\widehat{s}_m) - \frac{1}{V} \sum_{K=1}^{V} P_n \gamma\left(\widehat{s}_m^{(\mathcal{B}_K^c)} \right) \ .$$

In the particular case where $V = n$, this criterion is studied by Massart (2007, Section 7.2.1, p. 204–205) under the name cross-validation estimator.

2.4 Resampling-Based and V-Fold Penalties

Another approach for building general data-driven model selection criteria is penalization with a resampling-based estimator of the expectation of the ideal penalty, as proposed by Efron (1983) with the bootstrap and later generalized to all resampling schemes (Arlot, 2009). Let $W \sim \mathcal{W}$ be some random vector of $\mathbb{R}^n$ independent from $\xi_{[\![n]\!]}$ with

$$\frac{1}{n} \sum_{i=1}^{n} W_i = 1 \ ,$$

and denote by $P_n^W = n^{-1} \sum_{i=1}^n W_i \delta_{\xi_i}$ the weighted empirical distribution of the sample. Then, the resampling-based penalty associated with $\mathcal{W}$ is defined as

$$\mathrm{pen}_{\mathcal{W}}(m) := C_{\mathcal{W}} \mathbb{E}_W \left[\left(P_n - P_n^W \right) \gamma \left(\widehat{s}_m^W \right) \right] \enspace , \tag{5}$$

where $\widehat{s}_m^W \in \mathrm{argmin}_{t \in S_m} \{ P_n^W \gamma(t) \}$, $\mathbb{E}_W[\cdot]$ denotes the expectation with respect to W only (that is, conditionally to the sample $\xi_{[\![n]\!]}$), and $C_{\mathcal{W}}$ is some positive constant. Resampling-based penalties have been studied recently in the least-squares density estimation framework (Lerasle, 2012), assuming that W is exchangeable, that is, its distribution is invariant by any permutation of its coordinates.

Since computing exactly $\mathrm{pen}_{\mathcal{W}}(m)$ has a large computational cost in general for exchangeable W, some non-exchangeable resampling schemes were introduced by Arlot (2008), inspired by V-fold cross-validation: given some partition $\mathcal{B} = \mathcal{B}_{[\![V]\!]}$ of $[\![n]\!]$, the weight vector W is defined by $W_i = (1 - \mathrm{Card}(\mathcal{B}_J)/n)^{-1} \mathbb{1}_{i \notin \mathcal{B}_J}$ for some random variable J with uniform distribution over $[\![V]\!]$. Then, $P_n^W = P_n^{(\mathcal{B}_J^c)}$ so that the associated resampling penalty, called *V-fold penalty*, is defined by

$$\begin{aligned} \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, x) &:= \frac{x}{V} \sum_{K=1}^{V} \left[\left(P_n - P_n^{(\mathcal{B}_K^c)} \right) \gamma \left(\widehat{s}_m^{(\mathcal{B}_K^c)} \right) \right] \\ &= \frac{2x}{V} \sum_{K=1}^{V} \left(P_n^{(\mathcal{B}_K^c)} - P_n \right) \left(\widehat{s}_m^{(\mathcal{B}_K^c)} \right) \end{aligned} \tag{6}$$

where $x > 0$ is left free for flexibility, which is quite useful according to Lemma 1 below.

2.5 Links Between V-Fold Penalties, Resampling Penalties and (Corrected) V-Fold Cross-Validation

In this paper, we focus our study on V-fold penalties because Lemma 1 below shows that formula (6) covers all V-fold and resampling-based procedures mentioned in Sections 2.3 and 2.4.

First, when $V = n$, the only possible partition is $\mathcal{B}_{\mathrm{LOO}} = \{\{1\}, \ldots, \{n\}\}$, and the V-fold penalty is called the leave-one-out penalty $\mathrm{pen}_{\mathrm{LOO}}(m, x) := \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}_{\mathrm{LOO}}, x)$. The associated weight vector W is exchangeable, hence Eq. (6) leads to all exchangeable resampling penalties since they are all equal up to a deterministic multiplicative factor in the least-squares density estimation framework when $\sum_{i=1}^n W_i = n$, as proved by Lerasle (2012).

For V-fold methods, let us assume $\mathcal{B}$ is a regular partition of $[\![n]\!]$, that is,

$$V = |\mathcal{B}| \geqslant 2 \text{ divides } n \quad \text{and} \quad \forall K \in [\![V]\!] \, , \, |\mathcal{B}_K| = \frac{n}{V} \enspace . \tag{$\mathbf{Reg}$}$$

Then, we get the following connection between V-fold penalization and cross-validation methods.

Lemma 1 *For least-squares density estimation with projection estimators, under assumption* (**Reg**)*,*

$$\mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}) = P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1) \tag{7}$$

$$\mathrm{crit}_{\mathrm{VFCV}}(m, \mathcal{B}) = P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{VF}}\left(m, \mathcal{B}, V - \frac{1}{2}\right) \tag{8}$$

$$\begin{aligned} \mathrm{crit}_{\mathrm{LPO}}(m, p) &= P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{LPO}}\left(m, p, \frac{n}{p} - \frac{1}{2}\right) && (9) \\ &= P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{LOO}}\left(m, (n-1)\frac{n/p - 1/2}{n/p - 1}\right) && (10) \\ &= P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{VF}}\left(m, \mathcal{B}_{\mathrm{LOO}}, (n-1)\frac{n/p - 1/2}{n/p - 1}\right) \end{aligned}$$

where for any $p \in [\![n-1]\!]$, the leave-p-out cross-validation criterion is defined by

$$\mathrm{crit}_{\mathrm{LPO}}(m, p) := \frac{1}{|\mathcal{E}_p|} \sum_{A \in \mathcal{E}_p} P_n^{(A)} \gamma\left(\widehat{s}_m^{(A^c)}\right) \qquad \textit{with} \qquad \mathcal{E}_p := \left\{ A \subset [\![n]\!] \text{ s.t. } |A| = p \right\}$$

and the leave-p-out penalty is defined by

$$\forall x > 0, \quad \mathrm{pen}_{\mathrm{LPO}}(m, p, x) := \frac{x}{|\mathcal{E}_p|} \sum_{A \in \mathcal{E}_p} \left(P_n - P_n^{(A^c)} \right) \gamma\left(\widehat{s}_m^{(A^c)}\right) \ .$$

Lemma 1 is proved in Section A.1.

Remark 2 *Eq.* (7) *was first proved by Arlot (2008) in a general framework that includes least-squares density estimation, assuming only* (**Reg**). *Eq.* (10) *follows from Lerasle (2012, Lemma A.11) since* $\mathrm{pen}_{\mathrm{LPO}}$ *belongs to the family of exchangeable resampling penalties, with weights* $W_i := (1 - p/n)^{-1} \mathbb{1}_{i \notin A}$ *and A is randomly chosen uniformly over $\mathcal{E}_p$; note that* $\sum_{i=1}^n W_i = n$ *for these weights. It can also be deduced from Proposition 3.1 by Celisse (2014), see Section A.1.*

Remark 3 *It is worth mentioning here the cross-validation estimators studied by Massart (2007, Chapter 7). First, the unbiased cross-validation criterion defined by Rudemo (1982) is exactly* $\mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}_{\mathrm{LOO}})$ *(see also Massart, 2007, Section 7.2.1). Second, the penalized estimator of Massart (2007, Theorem 7.6) is the estimator selected by the penalty*

$$\mathrm{pen}_{\mathrm{LOO}}\left(m, \frac{(1+\epsilon)^6 (n-1)^2}{2\left[n - (1+\epsilon)^6\right]}\right)$$

for some $\epsilon > 0$ such that $(1+\epsilon)^6 < n$ (see Section A.1 for details).

So, in the least-squares density estimation framework and assuming only (**Reg**), Lemma 1 shows that it is sufficient to study V-fold penalization with a free multiplicative factor x in front of the penalty for studying also V-fold cross-validation ($x = V - 1/2$), corrected V-fold cross-validation ($x = V - 1$), the leave-p-out ($V = n$ and $x = (n-1)(n/p - 1/2)/(n/p - 1)$) and all exchangeable resampling penalties. For any $C > 0$ and $\mathcal{B}$ some partition of $[\![n]\!]$ into V pieces, taking $x = C(V-1)$, the V-fold penalization criterion is denoted by

$$\mathcal{C}_{(C, \mathcal{B})}(m) := P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{VF}}\big(m, \mathcal{B}, C(V-1)\big) \ . \tag{11}$$

A key quantity in our results is the bias $\mathbb{E}[\mathcal{C}_{(C,B)}(m) - \mathrm{crit}_{\mathrm{id}}(m)]$. From Lemma 13 in Section A.2, we have

$$\mathbb{E}\bigl[\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1)\bigr] = \mathbb{E}\bigl[\mathrm{pen}_{\mathrm{id}}(m)\bigr] = 2\mathbb{E}\bigl[\lVert \widehat{s}_m - s_m \rVert^2\bigr] \enspace , \tag{12}$$

so that for any $C > 0$,

$$\mathbb{E}\bigl[\mathcal{C}_{(C,\mathcal{B})}(m) - \mathrm{crit}_{\mathrm{id}}(m)\bigr] = 2(C-1)\mathbb{E}\bigl[\lVert \widehat{s}_m - s_m \rVert^2\bigr] \enspace . \tag{13}$$

In Sections 3–7, we focus our study on V-fold methods, that is, we study the performance of the V-fold penalized estimators $\widehat{s}_{\widehat{m}}$, defined by

$$\widehat{m} = \widehat{m}\bigl(\mathcal{C}_{(C,\mathcal{B})}\bigr) = \operatorname*{argmin}_{m \in \mathcal{M}_n} \bigl\{\mathcal{C}_{(C,\mathcal{B})}(m)\bigr\} \enspace , \tag{14}$$

for all values of V and $C > 1/2$. Additional results on hold-out (penalization) are given in Section 8.2 to complete the picture.

3. Oracle Inequalities

In this section, we state our first main result, that is, a non-asymptotic oracle inequality satisfied by V-fold procedures. This result holds for any divisor $V \geqslant 2$ of n, any constant $x = C(V-1)$ in front of the penalty with $C > 1/2$, and provides an asymptotically optimal oracle inequality for the selected estimator when $C \to 1$ (assuming the setting is non parametric). In addition, as proved by Section 2.5, it implies oracle inequalities satisfied by leave-p-out procedures for all p.

3.1 Concentration of V-Fold Penalties

Concentration is the key property to establish oracle inequalities. Let us start with some new concentration results for V-fold penalties.

Proposition 4 *Let $\xi_{\llbracket n \rrbracket}$ be i.i.d. real-valued random variables with density $s \in L^\infty(\mu)$, $\mathcal{B}$ some partition of $\llbracket n \rrbracket$ into V pieces satisfying* (**Reg**), *S_m a separable linear space of measurable functions and $(\psi_\lambda)_{\lambda \in \Lambda_m}$ an orthonormal basis of S_m. Define*

$$\mathbb{B}_m = \bigl\{t \in S_m \text{ s.t. } \lVert t \rVert \leqslant 1\bigr\} \qquad \Psi_m = \sum_{\lambda \in \Lambda_m} \psi_\lambda^2 = \sup_{t \in \mathbb{B}_m} t^2 \qquad b_m := \lVert \sqrt{\Psi_m} \rVert_\infty$$

$$\mathcal{D}_m := P(\Psi_m) - \lVert s_m \rVert^2 = n\mathbb{E}\Bigl[\lVert s_m - \widehat{s}_m \rVert^2\Bigr] \enspace ,$$

where $\widehat{s}_m$ is defined by Eq. (1), *and for any $x, \epsilon > 0$,*

$$\rho_1(m, \epsilon, s, x, n) := \frac{\lVert s \rVert_\infty x}{\epsilon n} + \frac{\bigl(b_m^2 + \lVert s \rVert^2\bigr)x^2}{\epsilon^3 n^2} \enspace .$$

Then, an absolute constant κ exists such that for any $x \geqslant 0$, with probability at least $1 - 8\mathrm{e}^{-x}$, for any $\epsilon \in (0,1]$, the following two inequalities hold true

$$\left| \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1) - \frac{2\mathcal{D}_m}{n} \right| \leqslant \epsilon \frac{\mathcal{D}_m}{n} + \kappa \rho_1(m, \epsilon, s, x, n) \tag{15}$$

$$\bigl|\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1) - 2\lVert s_m - \widehat{s}_m \rVert^2\bigr| \leqslant \epsilon \frac{\mathcal{D}_m}{n} + \kappa \rho_1(m, \epsilon, s, x, n) \enspace . \tag{16}$$

Proposition 4 is proved in Section A.2. Eq. (15) gives the concentration of the V-fold penalty around its expectation $2\mathcal{D}_m/n = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$, see Eq. (12). Eq. (16) gives the concentration of the V-fold penalty around the ideal penalty, see Eq. (2). Optimizing over ϵ, the first order of the deviations of $\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1)$ around $\mathrm{pen}_{\mathrm{id}}(m)$ is driven by $\sqrt{\mathcal{D}_m}/n$. The deviation term in Proposition 4 does not depend on V and cannot therefore help to discriminate between different values of this parameter.

3.2 Example: Histogram Models

Histograms on $\mathbb{R}$ provide some classical examples of collections of models. Let $\mathcal{X}$ be a measurable subset of $\mathbb{R}$, μ denote the Lebesgue measure on $\mathcal{X}$ and m be some countable partition of $\mathcal{X}$ such that $\mu(\lambda) > 0$ for any $\lambda \in m$. The histogram space S_m based on m is the linear span of the functions $(\psi_\lambda)_{\lambda \in \Lambda_m}$ where $\Lambda_m = m$ and for every $\lambda \in m$, $\psi_\lambda = \mu(\lambda)^{-1/2}\mathbb{1}_\lambda$. More precisely, we illustrate our results with the following examples.

Example 1 (Regular histograms on $\mathcal{X} = \mathbb{R}$)

$$\mathcal{M}_n = \left\{ m_h, h \in [\![n]\!] \right\} \qquad \textit{where} \qquad \forall h \in [\![n]\!] \, , \quad m_h = \left\{ \left[\frac{\lambda}{h}, \frac{\lambda+1}{h} \right), \lambda \in \mathbb{Z} \right\} \ .$$

In Example 1, defining $d_{m_h} = h$ for every $h \in [\![n]\!]$, for every $m \in \mathcal{M}_n$, $\mathcal{D}_m = d_m - \|s_m\|^2$ since Ψ_m is constant and equal to d_m. Therefore, Proposition 4 shows that $\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1)$ is asymptotically equivalent to $\mathrm{pen}_{\mathrm{dim}}(m) := 2d_m/n$ when $d_m \to \infty$. Penalties of the form of $\mathrm{pen}_{\mathrm{dim}}$ are classical and have been studied for instance by Barron et al. (1999).

Example 2 (k-rupture points on $\mathcal{X} = [0,1]$)

$$\mathcal{M}_n = \left\{ m_{h_{[\![k+1]\!]}, x_{[\![k]\!]}} \ \textit{s.t.}\ x_1 < \cdots < x_k \in [\![n-1]\!] \text{ and} \, \forall i \in [\![k+1]\!] \, , h_i \in [\![x_i - x_{i-1}]\!] \right\} \ ,$$

where $x_0 = 0$, $x_{k+1} = n$ and for any $x_1, \ldots, x_k \in [\![n-1]\!]$ such that $x_1 < \cdots < x_k$ and any $h_{[\![k+1]\!]} \in \mathbb{N}^{k+1}$, $m_{h_{[\![k+1]\!]}, x_{[\![k]\!]}}$ is defined as the union

$$\bigcup_{i \in [\![k]\!]} \left\{ \left[\frac{x_{i-1}}{n} + \frac{(x_i - x_{i-1})(\lambda - 1)}{nh_i}, \frac{x_{i-1}}{n} + \frac{(x_i - x_{i-1})\lambda}{nh_i} \right), \lambda \in [\![h_i]\!] \right\} \ .$$

In other words, $m_{h_{[\![k+1]\!]}, x_{[\![k]\!]}}$ splits $[0,1]$ into $k+1$ pieces (at the x_i), and then splits the i-th piece into h_i pieces of equal size.

In Example 2, the function Ψ_m is constant on each interval $[x_{i-1}, x_i)$, equal to h_i, therefore,

$$\mathcal{D}_m = \sum_{i=1}^{k+1} h_i \mathbb{P}\big(\xi \in [x_{i-1}, x_i)\big) - \|s_m\|^2 \ .$$

3.3 Oracle Inequality for V-Fold Procedures

In order to state the main result, we introduce the following hypotheses:

- *A uniform bound on the L^∞ norm of the L^2 ball of the models*

$$\forall m \in \mathcal{M}_n , \qquad b_m \leqslant \sqrt{n} \tag{$\mathbf{H1}$}$$

  where we recall that $b_m := \sup_{t \in \mathbb{B}_m} \|t\|_\infty$ and $\mathbb{B}_m := \{ t \in S_m , \|t\| \leqslant 1 \}$.

- *The family of the projections of s is uniformly bounded.*

$$\exists a > 0 , \quad \forall m \in \mathcal{M}_n , \qquad \|s_m\|_\infty \leqslant a \enspace , \tag{$\mathbf{H2}$}$$

- *The collection of models is nested.*

$$\forall (m, m') \in \mathcal{M}_n^2 , \qquad S_m \cup S_{m'} \in \{ S_m , S_{m'} \} \tag{$\mathbf{H2'}$}$$

Hereafter, we define $A := a \vee \|s\|_\infty$ when ($\mathbf{H2}$) holds and $A := \|s\|_\infty$ when ($\mathbf{H2'}$) holds. On histogram spaces, ($\mathbf{H1}$) holds if and only if $\inf_{m \in \mathcal{M}_n} \inf_{\lambda \in m} \mu(\lambda) \geqslant n^{-1}$, and ($\mathbf{H2}$) holds with $a = \|s\|_\infty$.

Theorem 5 *Let $\xi_{[\![n]\!]}$ be i.i.d. real-valued random variables with common density $s \in L^\infty(\mu)$, $\mathcal{B}$ some partition of $[\![n]\!]$ into V pieces satisfying ($\mathbf{Reg}$) and $(S_m)_{m \in \mathcal{M}_n}$ be a collection of separable linear spaces satisfying ($\mathbf{H1}$). Assume that either ($\mathbf{H2}$) or ($\mathbf{H2'}$) holds true. Let $C \in (1/2, 2]$, $\delta := 2(C-1)$ and, for any $x, \epsilon > 0$,*

$$\rho_2(\epsilon, s, x, n) := \frac{Ax}{\epsilon n} + \left(1 + \frac{\|s\|^2}{n} \right) \frac{x^2}{\epsilon^3 n} \qquad \text{and} \qquad x_n = x + \log |\mathcal{M}_n| \enspace .$$

For every $m \in \mathcal{M}_n$, let $\widehat{s}_m$ be the estimator defined by Eq. (1) and $\widetilde{s} = \widehat{s}_{\widehat{m}}$ where

$$\widehat{m} = \widehat{m}\bigl(\mathcal{C}_{(C, \mathcal{B})} \bigr)$$

is defined by Eq. (14). Then, an absolute constant κ exists such that, for any $x > 0$, with probability at least $1 - \mathrm{e}^{-x}$, for any $\epsilon \in (0, 1]$,

$$\frac{1 - \delta_- - \epsilon}{1 + \delta_+ + \epsilon} \|\widetilde{s} - s\|^2 \leqslant \inf_{m \in \mathcal{M}_n} \|\widehat{s}_m - s\|^2 + \kappa \rho_2(\epsilon, s, x_n, n) \enspace . \tag{17}$$

Theorem 5 is proved in Section A.3.

Taking $\epsilon > 0$ small enough in Eq. (17), Theorem 5 proves that V-fold model selection procedures satisfy an oracle inequality with large probability. The remainder term can be bounded under the following classical hypothesis

$$\exists a' > 0, \forall n \in \mathbb{N}^\star , \qquad |\mathcal{M}_n| \leqslant n^{a'} \enspace . \tag{$\mathbf{H3}$}$$

For instance, ($\mathbf{H3}$) holds in Example 1 with $a' = 1$ and in Example 2 with $a' = k$. Under ($\mathbf{H3}$), the remainder term in Eq. (17) is bounded by $L (\log n)^2 / (\epsilon^3 n)$ for some $L > 0$, which is much smaller than the oracle loss in the nonparametric case.

The leading constant in the oracle inequality (17) is $(1 + \delta_+)/(1 - \delta_-) + \mathrm{o}(1)$ by choosing $\epsilon = \mathrm{o}(1)$, so the first-order behaviour of the upper bound on the loss is driven by δ. An asymptotic optimality result can be derived from Eq. (17) only if $\delta = \mathrm{o}(1)$. The meaning of

$\delta = 2(C-1)$ is the amount of bias of the V-fold penalization criterion, as shown by Eq. (13). Given this interpretation of δ, the model selection literature suggests that no asymptotic optimality result can be obtained in general when $\delta \neq \mathrm{o}(1)$ in the nonparametric case (see, for instance, Shao, 1997). Therefore, even if the leading constant $(1+\delta_+)/(1-\delta_-)$ is only an upper bound, we conjecture that it cannot be taken as small as $1+\mathrm{o}(1)$ unless $\delta = \mathrm{o}(1)$; such a result can be proved in our setting using similar arguments and assumptions as the ones of Arlot (2008) for instance.

For bias-corrected V-fold cross-validation, that is, $C = 1$ hence $\delta = 0$, Theorem 5 shows a first-order optimal non-asymptotic oracle inequality, since the leading constant $(1+\epsilon)/(1-\epsilon)$ can be taken equal to $1+\mathrm{o}(1)$, and the remainder term is small enough in the nonparametric case, under assumption (**H3**), for instance. Such a result valid with no upper bound on V had never been obtained before in any setting.

V-fold cross-validation is also analyzed by Theorem 5, since by Lemma 1 it corresponds to $C = 1 + 1/(2(V-1))$, hence $\delta = 1/(V-1)$. When V is fixed, the oracle inequality is asymptotically sub-optimal, which is consistent with the result proved in regression by Arlot (2008). On the contrary, if $\mathcal{B} = \mathcal{B}_n$ has V_n blocs, with $V_n \to \infty$, Theorem 5 implies under assumption (**H3**) the asymptotic optimality of V_n-fold cross-validation in the nonparametric case.

The bound obtained in Theorem 5 can be integrated and we get

$$\frac{1-\delta_- - \epsilon}{1+\delta_+ + \epsilon}\mathbb{E}\Big[\lVert \widetilde{s} - s \rVert^2\Big] \leqslant \mathbb{E}\Big[\inf_{m \in \mathcal{M}_n} \lVert \widehat{s}_m - s \rVert^2\Big] + \kappa' \rho_2\Big(\epsilon, s, \log\big(\lvert \mathcal{M}_n \rvert\big)\Big)$$

for some absolute constant $\kappa' > 0$.

Assuming $C > 1/2$ is necessary, according to minimal penalty results proved by Lerasle (2012). Assuming $C \leqslant 2$ only simplifies the presentation; if $C > 2$, the same proof shows that Theorem 5 holds with κ replaced by $C\kappa$.

An oracle inequality similar to Theorem 5 holds in a more general setting, as proved in a previous version of this paper (Arlot and Lerasle, 2012, Theorem 1); we state a less general result here for simplifying the exposition, since it does not change the message of the paper. First, assumption (**Reg**) can be relaxed into assuming the partition $\mathcal{B}$ is close to regular, that is,

$$\mathcal{B} \text{ is a partition of } [\![n]\!] \text{ of size } V \text{ and } \sup_{k \in [\![V]\!]} \left\lvert \mathrm{Card}(\mathcal{B}_k) - \frac{n}{V} \right\rvert \leqslant 1 \ , \qquad (\mathbf{Reg}')$$

which can hold for any $V \in [\![n]\!]$. Second, data $\xi_1, \ldots, \xi_n$ can belong to a general Polish space $\mathcal{X}$, at the price of some additional technical assumption.

3.4 Comparison with Previous Works on V-Fold Procedures

Few non-asymptotic oracle inequalities have been proved for V-fold penalization or cross-validation procedures.

Concerning cross-validation, previous oracle inequalities are listed in the survey by Arlot and Celisse (2010). In the least-squares density estimation framework, oracle inequalities were proved by van der Laan et al. (2004) in the V-fold case, but compared the risk of the selected estimator with the risk of an oracle trained with $n(V-1)/V$ data. In comparison, Theorem 5 considers the strongest possible oracle, that is, trained with n data. Optimal oracle inequalities were proved by Celisse (2014) for leave-p-out estimators with $p \ll n$, a case also treated in Theorem 5 by taking $V = n$ and $C = (n/p - 1/2)/(n/p - 1)$ as shown by Lemma 1. If $p \ll n$, $C \sim 1$, hence $\delta = \mathrm{o}(1)$ and we recover the result of Celisse (2014).

Concerning V-fold penalization, previous results were either valid for $V = n$ only—by Massart (2007, Theorem 7.6) and Lerasle (2012) for least-squares density estimation, by Arlot (2009) for regressogram estimators—, or for V bounded when n tends to infinity—by Arlot (2008) for regressogram estimators. In comparison, Theorem 5 provides a result valid for all V, except for the assumption that V divides n, which can be removed (Arlot and Lerasle, 2012). In particular, the loss bound by Arlot (2008) deteriorates when V grows, while it remains stable in our result. Our result is therefore much closer to the typical behavior of the loss ratio $\|\widetilde{s} - s\|^2 / \inf_{m \in \mathcal{M}_n} \|\widehat{s}_m - s\|^2$ of V-fold penalization, which usually decreases as a function of V in simulation experiments, see Section 6 and the experiments by Arlot (2008), for instance.

Theorem 5 may not satisfactorily address the parametric setting, that is, when the collection $(S_m)_{m \in \mathcal{M}_n}$ contains some fixed true model. In such a case, the usual way to obtain asymptotic optimality is to use a model selection procedure targetting identification, that is, taking $C \to +\infty$ when $n \to +\infty$. For instance, Celisse (2014, Theorem 3.3) shows that $\log(n) \ll C \ll n$ is a sufficient condition for such a result.

4. How to Compare Theoretically the Performances of Model Selection Procedures for Estimation?

The main goal of the paper is to compare the model selection performances of several (V-fold) cross-validation methods, when the goal is estimation, that is, minimizing the loss $\|\widehat{s}_{\widehat{m}} - s\|^2$ of the final estimator. In this section, we discuss how such a comparison can be made on theoretical grounds, in a general setting.

For some data-driven function $\mathcal{C} : \mathcal{M}_n \to \mathbb{R}$, the goal is to understand how $\|\widehat{s}_{\widehat{m}(\mathcal{C})} - s\|^2$ depends on $\mathcal{C}$ when the selected model is

$$\widehat{m}(\mathcal{C}) \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \bigl\{ \mathcal{C}(m) \bigr\} \enspace . \tag{18}$$

From now on, in this section, $\mathcal{C}$ is assumed to be a cross-validation estimator of the risk, but the heuristic developed here applies to the general case.

Ideal comparison. Ideally, for proving that $\mathcal{C}_1$ is a better method than $\mathcal{C}_2$ in some setting, we would like to prove that

$$\bigl\| \widehat{s}_{\widehat{m}(\mathcal{C}_1)} - s \bigr\|^2 < (1 - \varepsilon_n) \bigl\| \widehat{s}_{\widehat{m}(\mathcal{C}_2)} - s \bigr\|^2 \tag{19}$$

with a large probability, for some $\varepsilon_n \geqslant 0$.

Previous works and their limits. When the goal is estimation, the classical way to analyze the performance of a model selection procedure is to prove an oracle inequality, that is, to *upper bound* (with a large probability or in expectation)

$$\bigl\| \widehat{s}_{\widehat{m}(\mathcal{C})} - s \bigr\|^2 - \inf_{m \in \mathcal{M}_n} \Bigl\{ \|\widehat{s}_m - s\|^2 \Bigr\} \qquad \text{or} \qquad \mathfrak{R}_n(\mathcal{C}) := \frac{\bigl\| \widehat{s}_{\widehat{m}(\mathcal{C})} - s \bigr\|^2}{\inf_{m \in \mathcal{M}_n} \bigl\{ \|\widehat{s}_m - s\|^2 \bigr\}} \enspace .$$

Alternatively, asymptotic results show that when n tends to infinity, $\mathfrak{R}_n(\mathcal{C}) \to 1$ (asymptotic optimality of $\mathcal{C}$) or $\mathfrak{R}_n(\mathcal{C}_1) \sim \mathfrak{R}_n(\mathcal{C}_2)$ (asymptotic equivalence of $\mathcal{C}_1$ and $\mathcal{C}_2$); see Arlot and Celisse (2010, Section 6) for a review of such results. Nevertheless, proving Eq. (19) requires a lower bound on $\mathfrak{R}_n(\mathcal{C})$ (asymptotic or not), which has been done only once for some cross-validation method, to the best of our knowledge. In some least-squares regression setting, V-fold cross-validation ($\mathcal{C}^{\mathrm{VF}}$) performs (asymptotically) worse than all asymptotically optimal model selection procedures since $\mathfrak{R}_n(\mathcal{C}^{\mathrm{VF}}) \geqslant \kappa(V) > 1$ with a large probability (Arlot, 2008).

The major limitation of all these previous results is that they can only compare $\mathcal{C}_1$ to $\mathcal{C}_2$ at first order, that is, according to $\lim_{n\to\infty} \mathfrak{R}_n(\mathcal{C}_1)/\mathfrak{R}_n(\mathcal{C}_2)$, which only depends on the bias of $\mathcal{C}_i(m)$ $(i = 1, 2)$ as an estimator of $\mathbb{E}[\|\widehat{s}_m - s\|^2]$, hence, on the asymptotic ratio between the training set size and the sample size (Arlot and Celisse, 2010, Section 6). For instance, the leave-p-out and the hold-out with a training set of size $(n - p)$ cannot be distinguished at first order, while the leave-p-out performs much better in practice, certainly because its "variance" is much smaller.

Beyond first-order. So, we must go beyond the first-order of $\mathfrak{R}_n(\mathcal{C})$ and take into account the variance of $\mathcal{C}(m)$. Nevertheless, proving a lower bound on $\mathfrak{R}_n(\mathcal{C})$ is already challenging at first order—probably the reason why only one has been proved up to now, in a specific setting only—so the challenge of computing a precise lower bound on the second order term of $\mathfrak{R}_n(\mathcal{C})$ seems too high for the present paper. We propose instead a heuristic showing that the variances of some quantities—depending on $(\mathcal{C}_i)_{i=1,2}$ and on $\mathcal{M}_n$—can be used as a proxy to a proper comparison of $\mathfrak{R}_n(\mathcal{C}_1)$ and $\mathfrak{R}_n(\mathcal{C}_2)$ at second order. Since we focus on second-order terms, from now on, we assume that $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same bias, that is,

$$\forall m \in \mathcal{M}_n, \quad \mathbb{E}\big[\mathcal{C}_1(m)\big] = \mathbb{E}\big[\mathcal{C}_2(m)\big] \ . \qquad (\textbf{SameBias})$$

In least-squares density estimation, given Lemma 1, this means that for $i \in \{1, 2\}$,

$$\mathcal{C}_i = \mathcal{C}_{(C,\mathcal{B}_i)}$$

as defined by Eq. (11), with different partitions $\mathcal{B}_i$ satisfying (**Reg**) with different $V = V_i$, but the same constant $C > 0$; $C = 1$ corresponds to the unbiased case.

The variance of the cross-validation criteria is not the correct quantity to look at. If we were only comparing cross-validation methods $\mathcal{C}_1, \mathcal{C}_2$ as estimators of $\mathbb{E}\big[\|\widehat{s}_m - s\|^2\big]$ for every single $m \in \mathcal{M}_n$, we could naturally compare them through their mean squared errors. Under assumption (**SameBias**), this would mean to compare their variances. This can be done from Eq. (23) below, but it is not sufficient to solve our problem, since it is known that the best cross-validation estimator of the risk does not necessarily yield the best model selection procedure (Breiman and Spector, 1992). More precisely, the selected model $\widehat{m}(\mathcal{C})$ defined by Eq. (18) is unchanged when $\mathcal{C}(m)$ is translated by any random quantity, but such a translation does change $\operatorname{Var}(\mathcal{C}(m))$ and can make it as large as desired. For model selection, what really matters is that

$$\operatorname{sign}\big(\mathcal{C}(m_1) - \mathcal{C}(m_2)\big) = \operatorname{sign}\Big(\|\widehat{s}_{m_1} - s\|^2 - \|\widehat{s}_{m_2} - s\|^2\Big)$$

as often as possible for every $(m_1, m_2) \in \mathcal{M}_n^2$, and that most mistakes in the ranking of models occur when $\lVert \widehat{s}_{m_1} - s \rVert^2 - \lVert \widehat{s}_{m_2} - s \rVert^2$ is small, so that $\lVert \widehat{s}_{\widehat{m}(\mathcal{C})} - s \rVert^2$ cannot be much larger than $\inf_{m \in \mathcal{M}_n} \{ \lVert \widehat{s}_m - s \rVert^2 \}$.

Heuristic. The heuristic we propose goes as follows. For simplicity, we assume that $m^\star = \operatorname{argmin}_{m \in \mathcal{M}_n} \mathbb{E}[\lVert \widehat{s}_m - s \rVert^2]$ is uniquely defined. If the goal was identification, we could directly state that for any $\mathcal{C}$, the smaller is $\mathbb{P}(m = \widehat{m}(\mathcal{C}))$ for all $m \neq m^\star$, the better should be the performance of $\widehat{m}(\mathcal{C})$. In this paper, our goal is estimation, but a similar claim can be conjectured by considering "all $m \in \mathcal{M}_n$ sufficiently far from $m^\star$ in terms of risk", that is, all $m \in \mathcal{M}_n$ such that $\mathbb{E}[\lVert \widehat{s}_m - s \rVert^2]$ is significantly worse than $\mathbb{E}[\lVert \widehat{s}_{m^\star} - s \rVert^2]$. Indeed, for any m "close to $m^\star$" in terms of risk, selecting m instead of $m^\star$ does not significantly change the performance of $\widehat{m}(\mathcal{C})$; on the contrary, for any m "far from $m^\star$" in terms of risk, selecting m instead of $m^\star$ does increase significantly the risk $\mathbb{E}[\lVert \widehat{s}_{\widehat{m}(\mathcal{C})} - s \rVert^2]$.

Then, our idea is to find a proxy for $\mathbb{P}(m = \widehat{m}(\mathcal{C}))$, that is, a quantity that should behave similarly as a function of $\mathcal{C}$ and its "variance" properties. For all $m, m' \in \mathcal{M}_n$, let $\Delta_{\mathcal{C}}(m, m') := \mathcal{C}(m) - \mathcal{C}(m')$, $\mathcal{N}$ some standard Gaussian random variable and, for all $t \in \mathbb{R}$, $\overline{\Phi}(t) = \mathbb{P}(\mathcal{N} > t)$. Then, for every $m \in \mathcal{M}_n$

$$\begin{aligned}
&\mathbb{P}\bigl(\widehat{m}(\mathcal{C}) = m\bigr) = \mathbb{P}\bigl(\forall m' \neq m,\ \Delta_{\mathcal{C}}(m, m') < 0\bigr) \\
&\asymp \min_{m' \neq m} \mathbb{P}\bigl(\Delta_{\mathcal{C}}(m, m') < 0\bigr) && (20) \\
&\approx \min_{m' \neq m} \mathbb{P}\Bigl(\mathbb{E}\bigl[\Delta_{\mathcal{C}}(m, m')\bigr] + \mathcal{N} \sqrt{\operatorname{Var}(\Delta_{\mathcal{C}}(m, m'))} < 0\Bigr) && (21) \\
&= \overline{\Phi}\bigl(\mathrm{SNR}_{\mathcal{C}}(m)\bigr) \quad \text{where} \quad \mathrm{SNR}_{\mathcal{C}}(m) := \max_{m' \neq m} \frac{\mathbb{E}\bigl[\Delta_{\mathcal{C}}(m, m')\bigr]}{\sqrt{\operatorname{Var}\bigl(\Delta_{\mathcal{C}}(m, m')\bigr)}} \enspace .
\end{aligned}$$

So, if $\mathrm{SNR}_{\mathcal{C}_1}(m) > \mathrm{SNR}_{\mathcal{C}_2}(m)$ for all m "sufficiently far from $m^\star$", $\mathcal{C}_1$ should be better than $\mathcal{C}_2$. Assuming (**SameBias**) holds true and that

$$\{m^\star\} = \operatorname*{argmin}_{m \in \mathcal{M}_n} \mathbb{E}\bigl[\mathcal{C}_1(m)\bigr] = \operatorname*{argmin}_{m \in \mathcal{M}_n} \mathbb{E}\bigl[\mathcal{C}_2(m)\bigr] \enspace , \qquad (\textbf{SameMin})$$

this leads to the following heuristic

$$\forall m \neq m', \qquad \operatorname{Var}\bigl(\Delta_{\mathcal{C}_1}(m, m')\bigr) < \operatorname{Var}\bigl(\Delta_{\mathcal{C}_2}(m, m')\bigr) \Rightarrow \mathcal{C}_1 \text{ better than } \mathcal{C}_2 \enspace . \qquad (22)$$

Indeed, for every $m \neq m'$, assumption (**SameMin**) implies that $\mathrm{SNR}_{\mathcal{C}_i}(m) > 0$ for $i = 1, 2$, hence we can restrict the max in the definition of $\mathrm{SNR}_{\mathcal{C}_i}$ to all m' such that $\mathbb{E}[\Delta_{\mathcal{C}_i}(m, m')]$ is positive. By assumption (**SameBias**), the numerator in the definition of $\mathrm{SNR}_{\mathcal{C}_i}$ does not depend on i, hence the ratio is maximal when the denominator is minimal, which leads to Eq. (22). Let us make some remarks.

- The quantity $\Delta_{\mathcal{C}}(m, m')$ appears in relative bounds (Catoni, 2007, Section 1.4) which can be used as a tool for model selection (Audibert, 2004).

- Assumptions (**SameBias**) and (**SameMin**) hold true in particular in the unbiased case, that is, when $\mathbb{E}[\mathcal{C}_i(m)] = \mathbb{E}[\lVert \widehat{s}_m - s \rVert^2]$ for all $m \in \mathcal{M}_n$ and $i \in \{1, 2\}$.

- Assumption (**SameMin**) is necessary: Figure 3 shows an example where a larger variance corresponds to better performance under assumption (**SameBias**) alone.
- As noticed above, the heuristic (22) should apply when the goal is estimation *and* when the goal is identification, provided that (**SameBias**) and (**SameMin**) hold true. What should depend on the goal is the suitable amount of bias for $\mathcal{C}_i(m)$ as an estimator of the risk $\mathbb{E}[\|\widehat{s}_m - s\|^2]$.
- Approximation (20) is the strongest one. Clearly, inequality $\leqslant$ holds true. The equality case occurs is for a very particular dependence setting, that is, when one among the events $(\{\Delta_{\mathcal{C}}(m,m') < 0\})$, $m' \in \mathcal{M}_n$, is included into all the others. In general, the left-hand side is significantly smaller than the right-hand side; we conjecture that they vary similarly as a function of $\mathcal{C}$.
- The Gaussian approximation (21) for $\Delta_{\mathcal{C}}(m,m')$ does not hold exactly, but it seems reasonable to make it, at first order at least.
- The validity of approximations (20) and (21) is supported by the numerical experiments of Section 6.

In the heuristic (22), all (m,m') do not matter equally for explaining a quantitative difference in the performances of $\mathcal{C}$. First, we can fix $m' = m^\star$, since intuitively, the strongest candidate against any $m \neq m^\star$ is $m^\star$, which clearly holds in all our experiments, see Figures 18 and 24 in Section G of the Online Appendix. Second, as mentioned above, if m and $m^\star$ are very close, that is, $\|\widehat{s}_m - s\|^2 / \|\widehat{s}_{m^\star} - s\|^2$ is smaller than the minimal order of magnitude we can expect for $\mathfrak{R}_n(\mathcal{C})$ with a data-driven $\mathcal{C}$, taking m instead of $m^\star$ does not decrease the performance significantly. Third, if $\overline{\Phi}(\mathrm{SNR}_{\mathcal{C}}(m))$ is very small, increasing it even by an order of magnitude will not affect the performance of $\widehat{m}(\mathcal{C})$ significantly; hence, all m such that, say, $\mathrm{SNR}_{\mathcal{C}}(m) \gg (\log(n))^\alpha$ for all $\alpha > 0$, can also be discarded. Overall, pairs (m,m') that really matter in (22) are pairs $(m,m^\star)$ that are at a "moderate distance", in terms of $\mathbb{E}[\|\widehat{s}_m - s\|^2 - \|\widehat{s}_{m^\star} - s\|^2]$.

5. Dependence on V of V-Fold Penalization and Cross-Validation

Let us now come back to the least-squares density estimation setting. Our goal is to compare the performance of cross-validation methods having the same bias, that is, according to Section 2.5, $\widehat{m}(\mathcal{C}_{(C,\mathcal{B})})$ with the same constant C but different partitions $\mathcal{B}$, where $\widehat{m}(\mathcal{C}_{(C,\mathcal{B})})$ is defined by Eq. (14).

Theorem 6 *Let $\xi_{[\![n]\!]}$ be i.i.d. random variables with common density $s \in L^\infty(\mu)$, $\mathcal{B}$ some partition of $[\![n]\!]$ into V pieces satisfying* (**Reg**), *and $(\psi_\lambda)_{\lambda\in\Lambda_{m_1}}$, $(\psi_\lambda)_{\lambda\in\Lambda_{m_2}}$ two orthonormal families in $L^2(\mu)$. For any $m,m' \in \{m_1,m_2\}$, we define S_m the linear span of $(\psi_\lambda)_{\lambda\in\Lambda_m}$, s_m the orthogonal projection of s onto S_m in $L^2(\mu)$, $\Psi_m := \sup_{t\in S_m \text{ s.t. } \|t\|\leqslant 1} t^2$,*

$$\beta(m,m') := \sum_{\lambda\in\Lambda_m} \sum_{\lambda'\in\Lambda_{m'}} \left(\mathbb{E}\Big[\big(\psi_\lambda(\xi_1) - P\psi_\lambda\big)\big(\psi_{\lambda'}(\xi_1) - P\psi_{\lambda'}\big) \Big] \right)^2$$

$$\text{and} \quad \mathbf{B}(m_1,m_2) := \beta(m_1,m_1) + \beta(m_2,m_2) - 2\beta(m_1,m_2) \ .$$

Then, for every $C > 0$,

$$\operatorname{Var}\bigl(\mathcal{C}_{(C,\mathcal{B})}(m_1)\bigr) = \frac{2}{n^2}\left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n}\right)\beta(m_1,m_1) \qquad (23)$$
$$+ \frac{4}{n}\operatorname{Var}\left(\left(1 + \frac{2C-1}{n}\right)s_{m_1}(\xi_1) - \frac{2C-1}{2n}\Psi_{m_1}(\xi_1)\right)$$

and $$\operatorname{Var}\bigl(\mathcal{C}_{(C,\mathcal{B})}(m_1) - \mathcal{C}_{(C,\mathcal{B})}(m_2)\bigr) = \frac{2}{n^2}\left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n}\right)\mathbf{B}(m_1,m_2) \qquad (24)$$
$$+ \frac{4}{n}\operatorname{Var}\left(\left(1 + \frac{2C-1}{n}\right)(s_{m_1} - s_{m_2})(\xi_1) - \frac{2C-1}{2n}(\Psi_{m_1} - \Psi_{m_2})(\xi_1)\right)$$

where $\mathcal{C}_{(C,\mathcal{B})}$ is defined by Eq. (11)*.*

Theorem 6 is proved in Section A.4.

Unbiased case. When $C = 1$, Theorem 6 shows that

$$\operatorname{Var}\bigl(\mathcal{C}_{(1,\mathcal{B})}(m_1) - \mathcal{C}_{(1,\mathcal{B})}(m_2)\bigr) = a + \left(1 + \frac{4}{V-1} - \frac{1}{n}\right)b$$

for some $a, b \geqslant 0$ depending on n, m_1, m_2 but not on V. If we admit that the heuristic (22) holds true, this implies that the model selection performance of bias-corrected V-fold cross-validation improves when V increases, but the improvement is at most in a second order term as soon as V is large. In particular, even if $a \ll b$, the improvement from $V = 2$ to 5 or 10 is much larger than from $V = 10$ to $V = n$, which can justify the commonly used principle that taking $V = 5$ or $V = 10$ is large enough.

Assuming in addition that S_{m_1} and S_{m_2} are regular histogram models (Example 1 in Section 3.2) with d_{m_1} that divides d_{m_2}, then, by Lemma 19 in Section B.2 of the Online Appendix,

$$a = \frac{4}{n}\left(1 + \frac{1}{n}\right)^2 \operatorname{Var}\bigl(s_{m_1}(\xi) - s_{m_2}(\xi)\bigr) \approx \mathcal{O}\left(\frac{1}{n}\|s_{m_1} - s_{m_2}\|^2\right)$$
$$\text{and} \quad b = \frac{2}{n^2}\mathbf{B}(m_1,m_2) \asymp \|s_{m_2}\|^2 \frac{d_{m_2}}{n^2} \enspace .$$

When d_{m_2}/n is at least as large as $\|s_{m_1} - s_{m_2}\|^2$, we obtain that the first-order term in the variance is of the form $\alpha + \beta/(V-1)$ where $\alpha, \beta > 0$ do not depend on V and are of the same order of magnitude, as supported by the numerical experiments of Section 6. Then, increasing V from 2 to n does reduce significantly the variance, by a constant multiplicative factor.

Let $\mathcal{C}_{\mathrm{id}}(m) := P_n\gamma(\widehat{s}_m) + \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ be the criterion we could use if we knew the expectation of the ideal penalty. From Proposition 17 in Section B of the Online Appendix,

$$\operatorname{Var}\bigl(\mathcal{C}_{\mathrm{id}}(m_1) - \mathcal{C}_{\mathrm{id}}(m_2)\bigr) = \frac{2}{n^2}\left(1 - \frac{1}{n}\right)\mathbf{B}(m_1,m_2)$$
$$+ \frac{4}{n}\operatorname{Var}\left(\left(1 - \frac{1}{n}\right)(s_{m_1} - s_{m_2})(\xi_1) + \frac{1}{2n}(\Psi_{m_1} - \Psi_{m_2})(\xi_1)\right)$$

which easily compares to formula (24) obtained for the V-fold criterion when $C = 1$. Up to smaller order terms, the difference lies in the first term, where $(1 + 4/(V-1) - 1/n)$ is replaced by $(1 - 1/n)$ when using the expectation of the ideal penalty instead of a V-fold penalty. In other words, the leave-one-out penalty—that is, taking $V = n$—behaves like the expectation of the ideal penalty.

We can also compare Eq. (23) with the asymptotic results obtained by Burman (1989), which imply that for any fixed model m_1

$$\operatorname{Var}\bigl(\mathcal{C}_{(1,\mathcal{B})}(m_1) - P\gamma(\widehat{s}_{m_1})\bigr) = \frac{\gamma_0}{n} + \left(\frac{V}{V-1}\gamma_1 + \gamma_2\right)\frac{1}{n^2} + \mathrm{o}\left(\frac{1}{n^2}\right)$$

with $\gamma_0, \gamma_1, \gamma_2$ that depend on m_1 and $\gamma_1 > 0$. Here, putting $C = 1$ in Eq. (23) yields a result with a similar flavour, valid for all $n \geqslant 1$, even if Eq. (23) computes the variance of a slightly different quantity.

Cross-validation criteria. V-fold cross-validation and the leave-p-out are also covered by Theorem 6, according to Lemma 1, respectively with $C = 1 + 1/(2(V-1))$ and with $V = n$ and $C = 1 + 1/(2(n/p - 1))$. As in the unbiased case, increasing V decreases the variance, and if we admit that the heuristic (22) holds true, V-fold cross-validation performs almost as well as the leave-(n/V)-out as soon as V is larger than 5 or 10.

Similarly, the variances of the V-fold cross-validation and leave-p-out criteria, for instance, can be derived from Eq. (23). In the leave-p-out case, we recover formulas obtained by Celisse (2014) and Celisse and Robin (2008), with a different grouping of the variance components; Eq. (23) clearly emphasizes the influence of the bias—through $(C-1)$—on the variance. For V-fold cross-validation, we believe that Eq. (23) shows in a simpler way how the variance depends on V, compared to the result of Celisse and Robin (2008) which was focusing on the difference between V-fold cross-validation and the leave-(n/V)-out; here the difference can be written

$$\frac{8}{n^2}\left(\frac{1}{V-1} - \frac{1}{n-1}\right)\left(1 + \frac{1}{2(V-1)}\right)^2 \beta(m_1, m_1) \enspace .$$

A major novelty in Eq. (23) is also to cover a larger set of criteria, such as bias-corrected V-fold cross-validation. Note that $\operatorname{Var}(\mathcal{C}_{(C,\mathcal{B})}(m_1))$ is generally much larger than

$$\operatorname{Var}\bigl(\mathcal{C}_{(C,\mathcal{B})}(m_1) - \mathcal{C}_{(C,\mathcal{B})}(m_2)\bigr) \enspace ,$$

which illustrates again why computing the former quantity might not help for understanding the model selection properties of $\mathcal{C}_{(C,\mathcal{B})}$, as explained in Section 4. For instance, comparing Eq. (23) and (24), changing s_{m_1} into $s_{m_1} - s_{m_2}$ in the second term can reduce dramatically the variance when s_{m_1} and s_{m_2} are close, which happens for the pairs (m_1, m_2) that matter for model selection according to Section 4.

The variance of other criteria and their increments are computed in subsequent sections of the paper and in the Online Appendix: Monte-Carlo cross-validation (Theorem 10 in Section 8.1 and Theorem 24 in Section C.4) and hold-out penalization (Proposition 28 in Section D.2).

Remark 7 *The term $\mathbf{B}(m_1, m_2)$ does not depend on the choice of particular bases of S_{m_1} and S_{m_2}: as proved by Proposition 18 in Section B of the Online Appendix*

$$\mathbf{B}(m_1, m_2) = n\operatorname{Var}\bigl((\widehat{s}_{m_1} - \widehat{s}_{m_2})(\xi)\bigr) - (n+1)\operatorname{Var}\bigl((s_{m_1} - s_{m_2})(\xi)\bigr) \enspace .$$

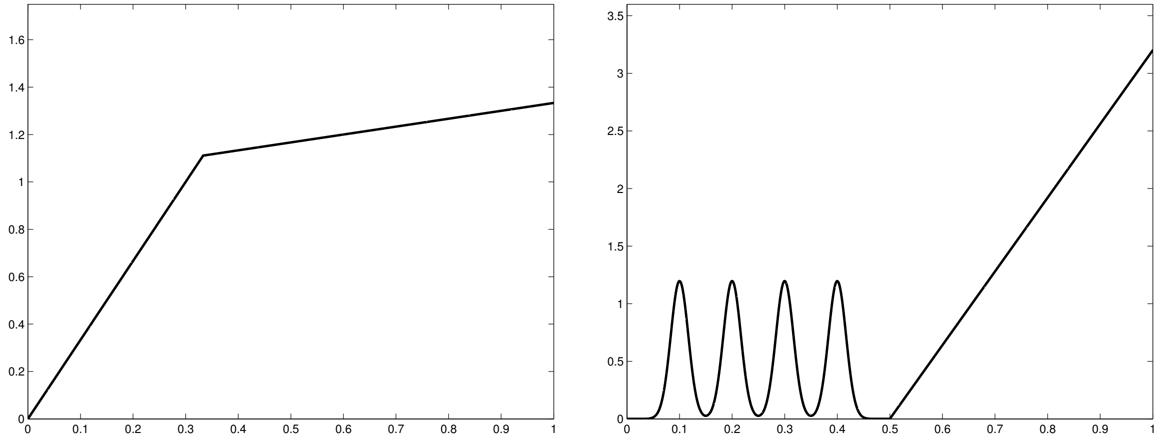

Figure 1: The two densities considered. Left: setting L. Right: setting S.

6. Simulation Study

This section illustrates the main theoretical results of the paper with some experiments on synthetic data.

6.1 Setting

In this section, we take $\mathcal{X} = [0,1]$ and μ is the Lebesgue measure on $\mathcal{X}$. Two examples are considered for the target density s and for the collection of models $(S_m)_{m \in \mathcal{M}_n}$.

Two density functions s are considered, see Figure 1:

- Setting L: $s(x) = \frac{10x}{3} \mathbb{1}_{0 \leqslant x < 1/3} + (1 + \frac{x}{3}) \mathbb{1}_{1 \geqslant x \geqslant 1/3}$.
- Setting S: s is the mixture of the piecewise linear density $x \mapsto (8x-4)\mathbb{1}_{1 \geqslant x \geqslant 1/2}$ (with weight 0.8) and four truncated Gaussian densities with means $(k/10)_{k=1,\ldots,4}$ and standard deviation $1/60$ (each with weight 0.05).

Two collections of models are considered, both leading to histogram estimators: for every $m \in \mathcal{M}_n$, S_m is the set of piecewise constant functions on some partition Λ_m of $\mathcal{X}$.

- "Regu" for regular histograms: $\mathcal{M}_n = \{1, \ldots, n\}$ where for every $m \in \mathcal{M}_n$, Λ_m is the regular partition of $[0,1]$ into m bins.
- "Dya2" for dyadic regular histograms with two bin sizes and a variable change-point:

$$\mathcal{M}_n = \bigcup_{k \in \{1,\ldots,\widetilde{n}\}} \{k\} \times \left\{0, \ldots, \lfloor \log_2(k) \rfloor \right\} \times \left\{0, \ldots, \lfloor \log_2(\widetilde{n} - k) \rfloor \right\}$$

  where $\widetilde{n} = \lfloor n / \log(n) \rfloor$ and for every $(k,i,j) \in \mathcal{M}_n$, $\Lambda_{(k,i,j)}$ is the union of the regular partition of $[0, k/\widetilde{n})$ into 2^i pieces and the regular partition of $[k/\widetilde{n}, 1]$ into 2^j pieces.

The difference between "Regu" and "Dya2" can be visualized on Figure 2, on which the corresponding oracle estimators $\widehat{s}_{\widehat{m}^\star}$ have been plotted for one sample in setting S, where

$$\widehat{m}^\star \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \left\| \widehat{s}_m - s \right\|^2 \enspace .$$

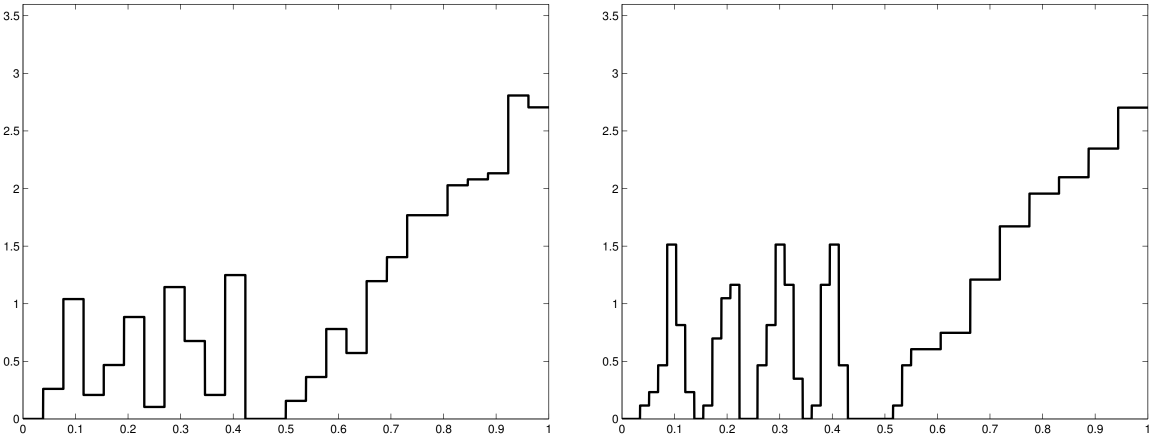

Figure 2: Oracle estimator for one sample of size $n = 500$, in setting S. Left: Regu. Right: Dya2.

Setting	Oracle(Regu)	Oracle(Dya2)	Best(Regu)	Best(Dya2)
L	13.4 ± 0.1	5.46 ± 0.02	25.8 ± 0.1	19.4 ± 0.1
S	62.4 ± 0.1	43.9 ± 0.1	100.9 ± 0.2	83.4 ± 0.2

Table 1: Comparison of Regu and Dya2: quadratic risks $\mathbb{E}[\|\widehat{s}_{\widehat{m}} - s\|^2]$ of "Oracle" and "Best" estimators (multiplied by 10^3) with the two collections of models. "Best" means that $\widehat{m}$ is the data-driven procedure minimizing $\mathbb{E}[\|\widehat{s}_{\widehat{m}} - s\|^2]$ among all the data-driven procedures we considered in our experiments (see Section 6.2). "Oracle" means that $\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \|\widehat{s}_m - s\|^2$ is the oracle model for each sample.

While "Regu" is one of the simplest and most classical collections for density estimation, the flexibility of "Dya2" allows to adapt to the variability of the smoothness of s. Intuitively, in settings L and S, the optimal bin size is smaller on $[0, 1/2]$ (where s is varying fastly) than on $[1/2, 1]$ (where $|s'|$ is much smaller).

Another point of comparison of Regu and Dya2 is given by Table 1, that reports values of the quadratic risks obtained depending on the collection of models considered. Table 1 shows that in settings L and S, the collection Dya2 helps reducing the quadratic risk by approximately 20% (when comparing the best data-driven procedures of our experiment), and even more when comparing oracle estimators (30% in setting S, 59% in setting L). Therefore, in settings L and S, it is worth considering more complex collections of models (such as Dya2) than regular histograms.

Let us finally remark that Dya2 does not reduce the quadratic risk in all settings as significantly as in settings L and S. We performed similar experiments with a few other density functions, sometimes leading to less important differences between Regu and Dya2 in terms of risk (results not shown). The oracle model was always better with Dya2, but in two cases, the risk of the best data-driven procedure with Dya2 was larger than with Regu by 6 to 8%.

6.2 Procedures Compared

In each setting, we consider the following model selection procedures:

- $\mathrm{pen}_{\mathrm{dim}}$ (Barron et al., 1999): penalization with $\mathrm{pen}(m) = 2\,\mathrm{Card}(\Lambda_m)/n$.
- V-fold cross-validation with $V \in \{2, 5, 10, n\}$, see Section 2.3.
- V-fold penalties (with leading constant $x = V - 1$, that is, bias-corrected V-fold cross-validation), for $V \in \{2, 5, 10, n\}$, see Section 2.4.
- for comparison, penalization with $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$, that is, $\widehat{m}(\mathcal{C}_{\mathrm{id}})$.

Since it is often suggested to multiply the usual penalties by some factor larger than one (Arlot, 2008), we consider all penalties above multiplied by a factor $C \in [0, 10]$. Complete results can be found in Section G of the Online Appendix.

6.3 Model Selection Performances

In each setting, all procedures are compared on $N = 10\,000$ independent synthetic data sets of size $n = 500$. For measuring their respective model selection performances, for each procedure $\widehat{m}(\mathcal{C})$ we estimate

$$C_{\mathrm{or}}(\mathcal{C}) := \mathbb{E}\big[\mathfrak{R}_n(\mathcal{C})\big] = \mathbb{E}\left[\frac{\big\|\widehat{s}_{\widehat{m}(\mathcal{C})} - s\big\|^2}{\inf_{m \in \mathcal{M}_n} \big\|\widehat{s}_m - s\big\|^2}\right]$$

by the corresponding average over the N simulated data sets; $C_{\mathrm{or}}(\mathcal{C})$ represents the constant that would appear in front of an oracle inequality. The uncertainty of estimation of $C_{\mathrm{or}}(\mathcal{C})$ is measured by the empirical standard deviation of $\mathfrak{R}_n(\mathcal{C})$ divided by $\sqrt{N}$. The results are reported in Table 2 for settings L and S, with the collection Dya2.

Results for Regu are not reported here since dimensionality-based penalties are already known to work well with Regu (Lerasle, 2012), so V-fold methods cannot improve significantly their performance, with a larger computational cost. Complete results (including Regu, with $n = 100$ and $n = 500$) are given in Tables 3 and 4 in Section G of the Online Appendix, showing that the performances of $\mathrm{pen}_{\mathrm{dim}}$ and V-fold methods indeed are very close.

Performance as a function of V***.*** Let us first consider V-fold penalization. In both settings L and S, as suggested by our theoretical results, C_{or} decreases when V increases. The improvement is large when V goes from 2 to 5 (27% for L, 10% for S) and small when V goes from 5 to 10 and when V goes from 10 to $n = 500$ (each time, 8% for L, 2% for S). Since the main influence of V is on the variance of the V-fold penalty, these experiments support our interpretation of Theorem 6 in Section 5: increasing V helps much more from 2 to 5 or 10 than from 10 to n.

The picture is less clear for V-fold cross-validation, for which almost no difference is observed among $V \in \{2, 5, 10, n\}$—less than 2%—, and C_{or} is minimized for $V \in \{5, 10\}$.

Procedure	L–Dya2	S–Dya2
pen_{dim}	8.27 ± 0.07	3.21 ± 0.01
pen2F	10.21 ± 0.08	2.39 ± 0.01
pen5F	7.47 ± 0.06	2.16 ± 0.01
pen10F	6.89 ± 0.06	2.11 ± 0.01
penLOO	6.35 ± 0.05	2.06 ± 0.01
2FCV	6.41 ± 0.05	2.05 ± 0.01
5FCV	6.27 ± 0.05	2.05 ± 0.01
10FCV	6.24 ± 0.05	2.05 ± 0.01
LOO	6.34 ± 0.05	2.06 ± 0.01
$\mathbb{E}[\text{pen}_{\text{id}}]$	6.52 ± 0.05	2.07 ± 0.01

Table 2: Estimated model selection performances, see text. ‘LOO’ is a shortcut for ‘leave-one-out’, that is, V-fold with $V = n = 500$.

Indeed, increasing V simultaneously decreases the bias and the variance of the V-fold cross-validation criterion, leading to various possible behaviours of C_{or} as a function of V, depending on the setting. The same phenomenon has been observed in regression (Arlot, 2008).

Overpenalization. In all settings considered in this paper, V-fold penalization performs much better when multiplying the penalty by $C > 1$, as illustrated by Figure 3. In particular, the best overpenalization factor for pen_{LOO} is $C_n^\star \approx 2.5$ for L-Dya2 and $C_n^\star \approx 1.4$ for S-Dya2, when $n = 500$. Such a phenomenon, which can also be observed in regression (Arlot, 2008), is related to the fact that some nonparametric model selection problems are “practically parametric”, using the terminology of Liu and Yang (2011), that is, BIC beats AIC and the optimal C is closer to $\log(n)/2$ than to 1. For instance, Figure 3 shows that L-Dya2 is practically parametric, while S-Dya2 is practically nonparametric since AIC beats BIC and the optimal C is close to 1.

Given an overpenalization factor C close to its optimal value $C_n^\star$, V-fold penalization performs significantly better than V-fold cross-validation in settings S-Dya2 and L-Dya2 (Figure 3). Since V-fold cross-validation corresponds to taking

$$C = C^{\text{VF}}(V) := 1 + \frac{1}{2(V-1)}$$

according to Lemma 1, this mostly means that $C^{\text{VF}}(V)$ is not close to $C_n^\star$ in these settings. In addition, when $C \approx C_n^\star$ is fixed, increasing V always improves the performance of V-fold penalization, as predicted by the heuristic of Section 4 and the theoretical results of Section 5. Let us emphasize that this fact does not depend on the parametricness of the setting: although the value of $C_n^\star$ is quite different for S-Dya2 and L-Dya2, in both cases, we observe qualitatively the same relationship between V and the performance of the procedure.

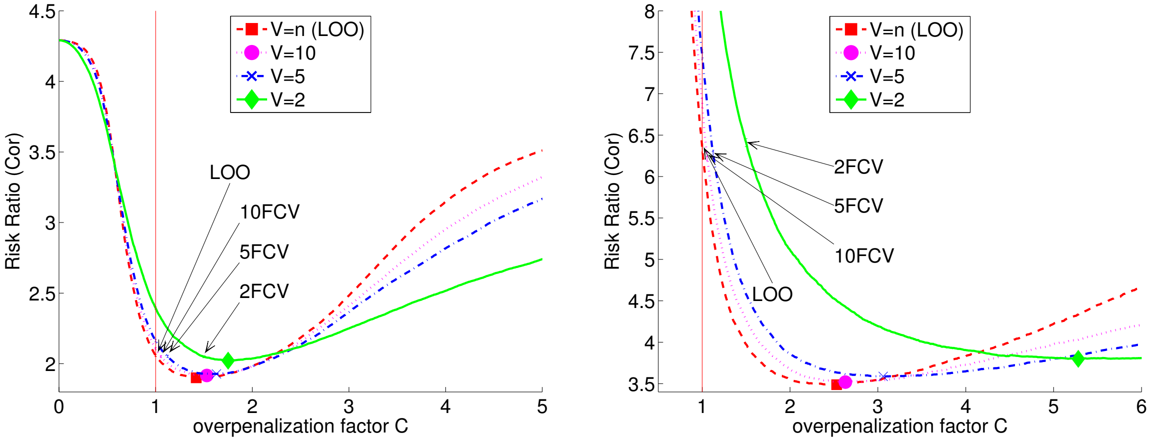

Figure 3: Overpenalization in settings S-Dya2 (left) and L-Dya2 (right), with $n = 500$ in both cases. Each plot represents the estimated model selection performance $C_{\mathrm{or}}(\mathcal{C}_{(C,\mathcal{B})})$ of several penalization procedures, as a function of the overpenalization constant C; unbiased risk estimation ($C = 1$) is materialized by a vertical red line. For each value of V, the estimated optimal value of C is shown on the graph; some arrows also show the performance of V-fold cross-validation, that is, $C = 1 + 1/[2(V-1)]$. Error bars are not shown for clarity; Table 2 shows their order of magnitude, which is smaller than visible differences in the above graph. The performance obtained with the penalty $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ (not shown on the graph) is almost the same as with the leave-one-out penalty.

The results reported in Section G of the Online Appendix lead to similar conclusions in several other settings, as well as unshown results in a truly parametric setting, with a true model of dimension 2. Although a wider simulation study would be necessary to get general conclusions, this suggests at least that the heuristic of Section 4 and the theoretical results of Section 5 can be applied to both parametric and nonparametric settings.

Figure 3 also helps understanding how the performance of V-fold cross-validation depends on V in Table 2. Indeed, the performance of V-fold cross-validation for each value of V can be visualized on Figure 3 by taking the point of abscissa $C = C^{\mathrm{VF}}(V)$ on the curve associated with V-fold penalization. Two phenomena are coupled when $C \leqslant C_n^\star$, which always holds in our simulations for V-fold cross-validation since $\max_V C^{\mathrm{VF}}(V) = 1.5$ and the estimated value of $C_n^\star$ is always larger. (i) The performance improves when V is fixed and C gets closer to $C_n^\star$. (ii) The performance improves when C is fixed and V increases. Even if both phenomena (i) and (ii) seem quite universal, their coupling can result in various behaviours for V-fold cross-validation as a function of V, as shown by Table 3 in Section G of the Online Appendix for instance.

Other comments.

- $\mathrm{pen}_{\mathrm{dim}}$ performs much worse than V-fold penalization (except $V = 2$ in setting L) with the collection Dya2. On the contrary, $\mathrm{pen}_{\mathrm{dim}}$ does well with Regu (see Table 3 in Section G of the Online Appendix), but V-fold penalization then performs as well.

- In other settings considered in a preliminary phase of our experiments, for V-fold penalization, differences between $V = 2$ and $V = 5$ were sometimes smaller or not significant, but always with the same ordering (that is, the worse performance for $V = 2$ when C is fixed). In a few settings, for which the "change-point" in the smoothness of s was close to the median of $s\mathrm{d}\mu$, we found $\mathrm{pen}_{\mathrm{dim}}$ among the best procedures with collection Dya2; then, V-fold penalization and cross-validation always had a performance very close to $\mathrm{pen}_{\mathrm{dim}}$. Both phenomena lead us to discard all settings for which there were no significant difference to comment.

6.4 Variance as a Function of V

We now illustrate the results of Section 5 about the variance of V-fold penalization and the heuristic of Section 4 about its influence on model selection. We focus on the unbiased case, that is, criteria $\mathcal{C}_{(1,\mathcal{B})}$ with partitions $\mathcal{B}$ satisfying (**Reg**). Since the distribution of $(\mathcal{C}_{(1,\mathcal{B})}(m))_{m\in\mathcal{M}_n}$ then only depends on $V = |\mathcal{B}|$, we write $\mathcal{C}_V$ instead of $\mathcal{C}_{(1,\mathcal{B})}$ by abuse of notation. All results presented in this subsection have been obtained from $N = 10\,000$ independent samples in setting S with a sample size $n = 100$ and the collection Regu—for which models are naturally indexed by their dimension.

First, Figure 4 shows the variance of $\Delta_{\mathcal{C}_V}(m, m^\star) = \mathcal{C}_V(m) - \mathcal{C}_V(m^\star)$ as a function of the dimension m of S_m, illustrating the conclusions of Theorem 6: the variance decreases when V increases. More precisely, the variance decrease is significant between $V = 2$ and $V = 5$, an order of magnitude smaller between $V = 5$ and $V = 10$ and between $V = 10$ and $V = n$, while the leave-one-out $\mathcal{C}_n$ is hard to distinguish from the ideal penalized criterion

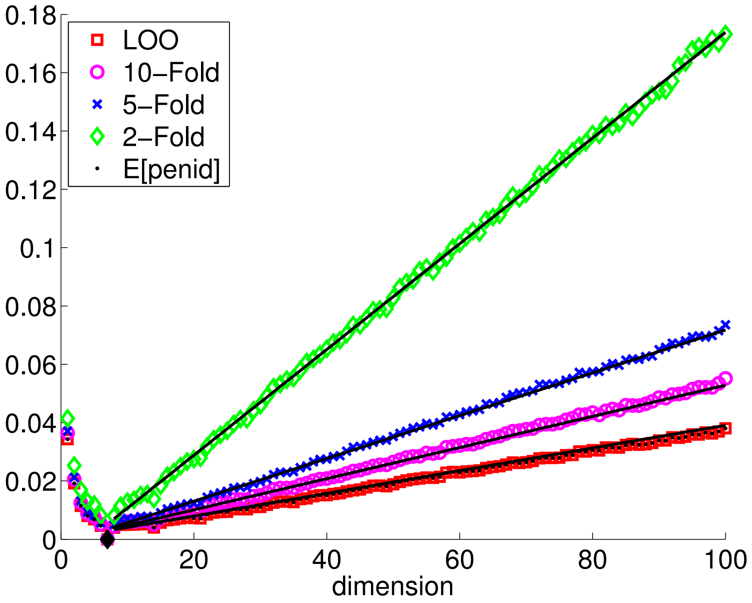

Figure 4: Illustration of the variance heuristic: $\mathrm{Var}(\Delta_{\mathcal{C}}(m, m^\star))$ as a function of m for five different $\mathcal{C}$. Setting S-Regu, $n = 100$. The black diamond shows $m^\star = 7$. The black lines show the linear approximation $n^{-2}[29(1+\frac{0.81}{V-1})+3.7(1+\frac{3.8}{V-1})(m-m^\star)]$ for $m > m^\star$.

$\mathcal{C}_{\mathrm{id}}$. On Figure 4, we can remark that for $m > m^\star$

$$\operatorname{Var}(\Delta_{\mathcal{C}_V}(m, m^\star)) \approx \frac{1}{n^2}\left[K_1\left(1 + \frac{K_2}{V-1}\right) + K_3\left(1 + \frac{K_4}{V-1}\right)(m - m^\star)\right]$$

with $K_1 \approx 29$, $K_2 \approx 0.81$, $K_3 \approx 3.7$ and $K_4 \approx 3.8$. The shape of the dependence on V already appears in Theorem 6, the above formula clarifies the relative importance of the terms called a and b in Section 5, and their dependence on the dimension m of S_m. Remark that the same behaviour holds when $n = 500$ with very close values for K_3 and K_4 (see Figure 25 in Section G of the Online Appendix), as well as in setting L with $n = 100$ or $n = 500$ with $K_3 \approx 2.1$ and $K_4 \approx 4.2$ (see Figures 19 and 30 in Section G of the Online Appendix). The fact that K_4 is close to 4 in both settings supports that the term $1 + 4/(V-1)$ appearing Theorem 6 indeed drives how $\operatorname{Var}(\Delta_{\mathcal{C}_V}(m, m^\star))$ depends on V.

Figures 5 and 6 respectively show $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ and its proxy $\overline{\Phi}(\mathrm{SNR}_{\mathcal{C}}(m))$ as a function of m for $\mathcal{C} = \mathcal{C}_V$ with $V \in \{2, 5, 10, n\}$ and for $\mathcal{C} = \mathcal{C}_{\mathrm{id}}$. First, we remark that both quantities behave similarly as a function of m and $\mathcal{C}$—see also Figure 16 in Section G of the Online Appendix—supporting empirically the heuristic of Section 4. The decrease of the variance observed on Figure 4 when V increases here translates into a better concentration of the distribution of $\widehat{m}(\mathcal{C}_V)$ around $m^\star$, which can explain the performance improvement observed in Section 6.3. Figures 5–6 actually show how the decrease of the variance quantitatively influences the distribution of $\widehat{m}(\mathcal{C}_V)$: $\widehat{m}(\mathcal{C}_5)$ is significantly more concentrated than $\widehat{m}(\mathcal{C}_2)$, while the difference between $V = 10$ and $V = 5$ is much smaller and comparable to the difference between $V = n$ and $V = 10$; $\mathcal{C}_n$ is hard to distinguish from $\mathcal{C}_{\mathrm{id}}$. Similar experiments with $n = 500$ and in setting L are reported in Section G of the Online Appendix, leading to similar conclusions.

7. Fast Algorithm for Computing V-Fold Penalties for Least-Squares Density Estimation

Since the use of V-fold algorithms is motivated by computational reasons, it is important to discuss the actual computational cost of V-fold penalization and cross-validation as a function of V. In the least-squares density estimation framework, two approaches are possible: a naive one—valid for all other frameworks—, and a faster one—specific to least-squares density estimation. For clarifying the exposition, we assume in this section that (**Reg**) holds true—so, V divides n. The general algorithm for computing the V-fold penalized criterion and/or the V-fold cross-validation criterion consists in training the estimator with data sets $(\xi_i)_{i \notin \mathcal{B}_j}$ for $j = 1, \ldots, V$ and then testing each trained estimator on the data sets $(\xi_i)_{i \in \mathcal{B}_j}$ and/or $(\xi_i)_{i \notin \mathcal{B}_j}$. In the least-squares density estimation framework, for any model S_m given through an orthonormal family $(\psi_\lambda)_{\lambda \in \Lambda_m}$ of elements of $L^2(\mu)$, we get the "naive" algorithm described and analysed more precisely in Section E.1 of the Online Appendix, whose complexity is of order $nV \operatorname{Card}(\Lambda_m)$.

Several simplifications occur in the least-squares density estimation framework, that allow to avoid a significant part of the computations made in the naive algorithm.

Algorithm 1

Input: *$\mathcal{B}$ some partition of $\{1, \ldots, n\}$ satisfying* (**Reg**)*, $\xi_1, \ldots, \xi_n \in \mathcal{X}$ and $(\psi_\lambda)_{\lambda \in \Lambda_m}$ a finite orthonormal family of $L^2(\mu)$.*

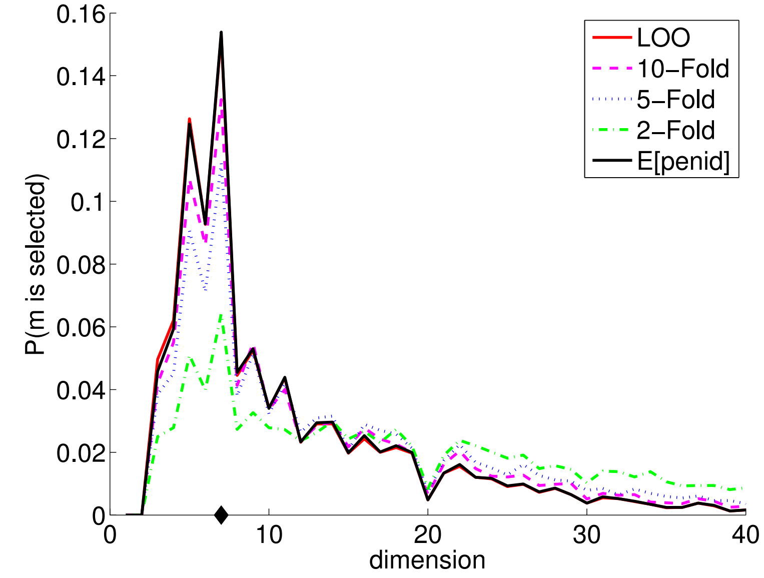

Figure 5: $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ as a function of m for five different $\mathcal{C}$. Setting S-Regu, $n = 100$. The black diamond shows $m^\star = 7$.

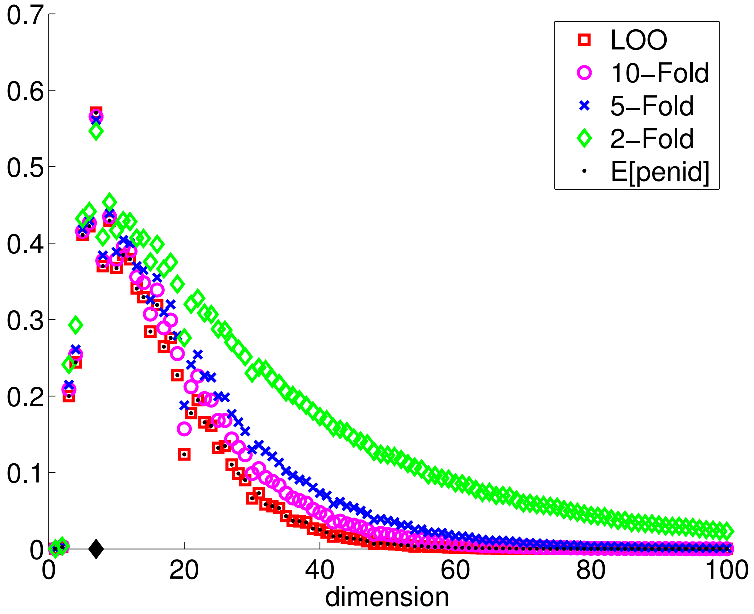

Figure 6: Illustration of the variance heuristic: $\overline{\Phi}(\mathrm{SNR}_{\mathcal{C}}(m))$ as a function of m for five different $\mathcal{C}$. Setting S-Regu, $n = 100$. The black diamond shows $m^\star = 7$.

1. *For $i \in \{1, \ldots, V\}$ and $\lambda \in \Lambda_m$, compute $A_{i,\lambda} := \frac{V}{n} \sum_{j \in B_i} \psi_\lambda(\xi_j)$.*
2. *For $i, j \in \{1, \ldots, V\}$, compute $C_{i,j} := \sum_{\lambda \in \Lambda_m} A_{i,\lambda} A_{j,\lambda}$.*
3. *Compute $\mathcal{S} := \sum_{1 \leqslant i,j \leqslant V} C_{i,j}$ and $\mathcal{T} := \operatorname{tr}(C)$.*

***Output*:**
Empirical risk: $P_n \gamma(\widehat{s}_m) = \dfrac{-\mathcal{S}}{V^2}$;
V-fold cross-validation criterion: $\operatorname{crit}_{\mathrm{VFCV}}(m) = \dfrac{\mathcal{T}}{V(V-1)} - \dfrac{\mathcal{S} - \mathcal{T}}{(V-1)^2}$;
V-fold penalty: $\operatorname{pen}_{\mathrm{VF}}(m) = \bigl(\operatorname{crit}_{\mathrm{VFCV}}(m) - P_n \gamma(\widehat{s}_m)\bigr) \dfrac{V - 1/2}{V - 1}$.

To the best of our knowledge, Algorithm 1 is new, even for computing the V-fold cross-validation criterion. Its correctness and complexity are analyzed with the following proposition.

Proposition 8 *Algorithm 1 is correct and has a computational complexity of order*

$$\bigl(n + V^2\bigr) \operatorname{Card}(\Lambda_m) \enspace .$$

In the histogram case, that is, when Λ_m is a partition of $\mathcal{X}$ and $\forall \lambda \in \Lambda_m$, $\psi_\lambda = \mu(\lambda)^{-1/2} \mathbb{1}_\lambda$, the computational complexity of Algorithm 1 can be reduced to the order of $n + V^2 \operatorname{Card}(\Lambda_m)$.

Proposition 8 is proved in Section E.2 of the Online Appendix. It shows that Algorithm 1 is significantly faster than the “naive” algorithm, by a factor of order

$$\frac{nV}{n + V^2} = \left(\frac{1}{V} + \frac{V}{n} \right)^{-1} \ll 1 \qquad \text{if} \qquad 1 \ll V \ll n \, .$$

Note that closed-form formulas are available for the leave-p-out criterion in least-squares density estimation (Celisse, 2014), allowing to compute it with a complexity of order $n \operatorname{Card}(\Lambda_m)$ in general, and smaller in some particular cases—for instance, n for histograms.

8. Discussion

Before discussing how to choose V when using V-fold methods for model selection—or more generally for choosing among a given family of estimators—, we state some additional results and we discuss the model selection literature in least-squares density estimation.

8.1 Monte-Carlo Cross-Validation

Our analysis of V-fold procedures for model selection can be extended to some other cross-validation procedures. We here present results for Monte-Carlo cross-validation (MCCV, Picard and Cook, 1984), also known as repeated cross-validation, where B training samples of the same size $n - p$ are chosen independently and uniformly (see also Arlot and Celisse, 2010, Section 4.3.2). Formally, we consider the criterion

$$\operatorname{crit}_{\mathrm{CV}}\bigl(m, (T_K)_{1 \leqslant K \leqslant B}\bigr) := \frac{1}{B} \sum_{K=1}^{B} \operatorname{crit}_{\mathrm{HO}}(m, T_K) \enspace , \tag{25}$$

where $T_1, \dots, T_B$ are subsets of $\llbracket n \rrbracket$ and we recall that the hold-out criterion is defined by Eq. (4). We make the following three assumptions throughout this subsection

$$\exists p \in \llbracket n-1 \rrbracket \, , \quad \forall j \in \llbracket B \rrbracket \, , \qquad |T_j| = n - p = n\tau_n \enspace , \tag{\textbf{SameSize}}$$

$$(T_K)_{1 \leqslant K \leqslant B} \quad \text{is independent from} \quad D_n \enspace , \tag{\textbf{Ind}}$$

$$T_1, \dots, T_B \quad \text{are independent with uniform distribution over} \quad \mathcal{E}_{n-p} \enspace , \tag{\textbf{MCCV}}$$

where we recall that $\mathcal{E}_{n-p} = \{ A \subset \llbracket n \rrbracket \text{ s.t. } |A| = n-p \}$. Under these assumptions, we write $\mathcal{C}^{\mathrm{MCCV}}(m)$ as a shortcut for $\mathrm{crit}_{\mathrm{CV}}(m, (T_K)_{1 \leqslant K \leqslant B})$.

Similarly to Theorem 5, we prove in Section C.3 of the Online Appendix the following oracle inequality for MCCV.

Theorem 9 *Let $\xi_{\llbracket n \rrbracket}$ be i.i.d. real-valued random variables with common density $s \in L^\infty(\mu)$, $(T_K)_{1 \leqslant K \leqslant B}$ some sequence of subsets of $\llbracket n \rrbracket$ satisfying* (**SameSize**)*,* (**Ind**) *and* (**MCCV**) *and $(S_m)_{m \in \mathcal{M}_n}$ be a collection of separable linear spaces satisfying* (**H1**)*. Assume that either* (**H2**) *or* (**H2′**) *holds true. For every $m \in \mathcal{M}_n$, let $\widehat{s}_m$ be the estimator defined by Eq.* (1)*, and $\widetilde{s} = \widehat{s}_{\widehat{m}}$ where*

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \Bigl\{ \mathrm{crit}_{\mathrm{CV}} \bigl(m, (T_K)_{1 \leqslant K \leqslant B} \bigr) \Bigr\}$$

and $\mathrm{crit}_{\mathrm{CV}}$ is defined by Eq. (25)*. Let us define, for any $x, y, \epsilon > 0$, $x_n = x + \log |\mathcal{M}_n|$ and*

$$\rho_3(\epsilon, x, y, n, \tau_n, B, A) := \frac{1}{n\tau_n^2} \left(1 + \frac{B \wedge (\log n + y)}{B(1-\tau_n)} \right)^\alpha \left(\frac{Ax}{\tau_n \epsilon} + \frac{(A \vee 1)x^2}{\epsilon^3} \right)$$

with $\alpha = 1$ under assumption (**H2**) *and $\alpha = 2$ under assumption* (**H2′**)*. Then, an absolute constant $\kappa > 0$ exists such that, for any $x, y \geqslant 0$, with probability at least $1 - \mathrm{e}^{-x} - \mathrm{e}^{-y}$, for any $\epsilon \in (0, \kappa^{-1})$,*

$$\left(1 - \frac{\epsilon}{\tau_n} \right) \| \widetilde{s} - s \|^2 \leqslant \frac{1+\epsilon}{\tau_n} \inf_{m \in \mathcal{M}_n} \Bigl\{ \| \widehat{s}_m - s \|^2 \Bigr\} + \kappa \rho_3(\epsilon, x_n, y, n, \tau_n, B, A) \enspace . \tag{26}$$

Theorem 9 actually is a corollary of a more general result (Theorem 23 in Section C.3 of the Online Appendix), which is valid without assumption (**MCCV**) and extends therefore our previous results on V-fold cross-validation).

Very few results exist in the literature about the model selection performance of MCCV with an estimation goal. Some asymptotic optimality result has been obtained by Burman (1990) for spline regression, and some oracle inequalities comparing the risk of the selected estimator with the risk of an oracle trained with $\tau_n n < n$ data have been proved by van der Laan and Dudoit (2003) in a general framework and by van der Laan et al. (2004) for density estimation with the Kullback-Leibler loss. In comparison, Theorem 9 provides a precise non-asymptotic comparison to an oracle trained with n data.

As in Theorem 5, the leading constant of the oracle inequality (26) is directly related to the bias, which is here quantified by $\tau_n^{-1} - 1 \geqslant 0$ instead of δ. The remainder term ρ_3 is also comparable to ρ_2 in Theorem 5: they differ by a factor between τ_n^{-2} (when B is large enough) and $\tau_n^{-2}(1-\tau_n)^{-\alpha}$ (when B is small). In particular, let $V \geqslant 2$ and assume that

$p = n/V$ in Theorem 9, hence $\tau_n = 1 - V^{-1} \in [1/2, 1)$. Then, for the hold-out ($B = 1$), ρ_3 is larger than ρ_2 by a factor V^α with $\alpha \in \{1, 2\}$. For $B = V$, MCCV with $\tau_n = 1 - V^{-1}$ can be called "Monte-Carlo V-fold" (MCVF); then, with $y \approx \log n$, we loose a factor at most $\log n$ for MCVF compared to V-fold cross-validation. Finally, when B is large enough, that is, larger than $V \log n$, ρ_3 and ρ_2 are of the same order.

The above comparison of remainder terms suggests a hierarchy between several cross-validation methods with a common training sample size $n - p = n\tau_n$: from the (presumably) worse to the (presumably) best procedure, the hold-out, Monte-Carlo CV with $B = V$, V-fold CV, Monte-Carlo CV with B large and the leave-p-out. Nevertheless, upper bounds comparison can be misleading, so, following the heuristics (22) presented in Section 4, we compute below the variance of $\Delta_{\mathcal{C}}(m, m')$ when $\mathcal{C}$ is a Monte-Carlo CV criterion.

Theorem 10 *We consider the setting and notation of Theorem 6, and we assume that* (**SameSize**)*,* (**MCCV**) *and* (**Ind**) *hold true. We recall that $\mathcal{C}^{\mathrm{MCCV}}(m)$ is defined above at the beginning of Section 8.1. Then, for regular histogram models m_1, m_2* (*Example 1 in Section 3.2*)*, we have*

$$\begin{aligned}\operatorname{Var}\bigl(\mathcal{C}^{\mathrm{MCCV}}(m_1) - \mathcal{C}^{\mathrm{MCCV}}(m_2)\bigr) = C_1^{\mathrm{MC}}(B, n, \tau_n) \frac{2}{n^2} \mathbf{B}(m_1, m_2) \\ + C_2^{\mathrm{MC}}(B, n, \tau_n) \frac{4}{n} \operatorname{Var}\bigl(s_{m_1}(\xi_1) - s_{m_2}(\xi_1)\bigr)\end{aligned} \tag{27}$$

where

$$C_1^{\mathrm{MC}}(B, n, \tau_n) = \frac{1}{B}\left(\frac{1}{\tau_n^2} + \frac{2}{\tau_n(1-\tau_n)} - \frac{1}{n\tau_n^3}\right) + \left(1 - \frac{1}{B}\right)\left[1 + \frac{1}{n-1}\left(\frac{1}{\tau_n} + 1\right)^2 - \frac{1}{n\tau_n^2}\right]$$

$$C_2^{\mathrm{MC}}(B, n, \tau_n) = \frac{1}{B}\left(\frac{1}{n^2\tau_n^3} + \frac{1}{1-\tau_n}\right) + \left(1 - \frac{1}{B}\right)\left(1 + \frac{1}{n\tau_n}\right)^2$$

and we recall that $\tau_n = |T_K|/n = 1 - (p/n)$.

Theorem 10 is proved in Section C.4 of the Online Appendix, as a corollary of a more general result, called Theorem 24, which holds for all models m_1, m_2—not only regular histograms—and provides a formula for the variance of the criterion itself—not its increments. Let us make a few comments.

Eq. (27) is similar to the formula obtained for bias-corrected V-fold and V-fold penalization, see Eq. (24) in Theorem 6. In the particular case of regular histogram models, Eq. (24) even fits the general form of Eq. (27), with constants $C_i^{\mathrm{penVF}}(V, n, C)$ instead of $C_i^{\mathrm{MC}}(B, n, \tau_n)$.

Assuming the heuristics of Section 4 is valid, for m_1, m_2 which matter for model selection, the two terms $2n^{-2}\mathbf{B}(m_1, m_2)$ and $4n^{-1}\operatorname{Var}(s_{m_1}(\xi_1) - s_{m_2}(\xi_1))$ are of the same order of magnitude (see Section 5). Then, we can compare model selection performance of several cross-validation methods by comparing the values of the constants C_i only.

In order to get a variance of the same order of magnitude as the one of bias-corrected V-fold CV—that is, constants C_i of order 1—, MCCV requires to take τ_n far enough from 0 and 1, hence training and sample sets of comparable sizes, unless B is large enough.

Eq. (27) allows to compare the hold-out ($B = 1$) with the leave-p-out ($B \to +\infty$), for a given value $n\tau_n = n - p$ of the training sample size. Let us assume for simplicity that $n \to +\infty$ and $\tau_n \gg n^{-1/2}$. Then,

$$C_1^{\mathrm{MC}}(1, n, \tau_n) \sim \frac{1}{\tau_n^2} + \frac{2}{\tau_n(1-\tau_n)} > 11 \qquad \text{and} \qquad C_2^{\mathrm{MC}}(1, n, \tau_n) \sim \frac{1}{1-\tau_n} \geqslant 1$$

$$\text{whereas} \qquad C_1^{\mathrm{MC}}(\infty, n, \tau_n) \to 1 \qquad \text{and} \qquad C_2^{\mathrm{MC}}(\infty, n, \tau_n) \to 1$$

which shows an improvement at least by a constant factor in general. When τ_n tends to zero—leave-most-out—or 1—such as for the leave-one-out—, the improvement is by an order of magnitude. The fact that the leave-p-out has a smaller variance than the hold-out is not surprising at all—it holds in full generality, as a consequence of Jensen's inequality—, but the exact quantification of the improvement given by Theorem 10 is new and can be useful in practice for choosing the number of splits B when using Monte-Carlo cross-validation.

Eq. (27) also allows to compare V-fold cross-validation, given by Theorem 6 with

$$C = 1 + \frac{1}{2(V-1)} \enspace ,$$

with MCCV with $B = V$ and $\tau_n = (V-1)/V$, which can be named "Monte-Carlo V-fold" cross-validation. The only difference between the two methods is that the V splits are chosen independently for "Monte-Carlo V-fold", whereas the usual V-fold makes a balanced use of each observation—putting it exactly $(V-1)$ times in the training set. Let us assume for simplicity that $n \to +\infty$ while $V = V_n$ can vary with n. Then, we have

$$C_1^{\mathrm{MCVF}}(V_n, n) := C_1^{\mathrm{MC}}\left(V_n, n, \frac{V_n - 1}{V_n}\right) \sim 3 + \frac{2V_n + 1}{V_n(V_n-1)} + \frac{1}{(V_n-1)^2}$$

$$C_1^{\mathrm{VF}}(V_n, n) := C_1^{\mathrm{penVF}}\left(V_n, n, 1 + \frac{1}{2(V_n-1)}\right) \sim 1 + \frac{4}{V_n - 1} + \frac{4}{(V_n-1)^2} + \frac{1}{(V_n-1)^3}$$

$$\text{hence} \qquad \frac{C_1^{\mathrm{MCVF}}(V_n, \infty)}{C_1^{\mathrm{VF}}(V_n, \infty)} > 1 \text{ if } V_n \geqslant 3 \enspace , \qquad \frac{C_1^{\mathrm{MCVF}}(V_n, n)}{C_1^{\mathrm{VF}}(V_n, n)} \xrightarrow[n, V_n \to +\infty]{} 3 \enspace ,$$

$$C_2^{\mathrm{MCVF}}(V_n, n) := C_2^{\mathrm{MC}}\left(V_n, n, \frac{V_n - 1}{V_n}\right) \sim 2 - \frac{1}{V_n} \in \left[\frac{3}{2}, 2\right]$$

$$\text{and} \qquad C_2^{\mathrm{VF}}(V_n, n) := C_2^{\mathrm{penVF}}\left(V_n, n, 1 + \frac{1}{2(V_n-1)}\right) \to 1 \enspace .$$

Overall, we get that V-fold cross-validation has a smaller variance than "Monte-Carlo V-fold" for $V \geqslant 3$, at least for n large enough, and that the improvement is by a constant factor between $3/2$ and 3. Since increasing V cannot decrease the variance of (bias-corrected) VFCV by more than a small constant factor, the above difference between two methods with the same computational complexity is quite important. This supports strongly the use of V-fold CV methods instead of "Monte-Carlo V-fold". Such an improvement was previously noticed in the asymptotic computations of Burman (1989); here we show that it holds in a non-asymptotic framework, where the models m_1, m_2 can depend on n.

8.2 Hold-Out Criteria

Our analysis of cross-validation procedures for model selection can also be extended to hold-out criteria. First, let us emphasize that the hold-out criterion defined by Eq. (4) corresponds to taking $B = 1$ in the results of Section 8.1, since choosing T uniformly over $\mathcal{E}_{n-p}$, independently from D_n, is equivalent to choosing some arbitrary T of size $n-p$ before seeing the data D_n.

Second, similarly to the definition of the hold-out criterion in Eq. (4), we can define the hold-out penalty by

$$\forall x \geqslant 0, \quad \mathrm{pen}_{\mathrm{HO}}(m, T, x) := 2x\Bigl(P_n^{(T)} - P_n\Bigr)\Bigl(\widehat{s}_m^{(T)} - \widehat{s}_m\Bigr) \ , \tag{28}$$

that is, the hold-out estimator of $\mathbb{E}[2(P_n - P)(\widehat{s}_m - s_m)]$ which is equal to the expectation of the ideal penalty, see Eq. (2). We do not define $\mathrm{pen}_{\mathrm{HO}}$ by Eq. (6) with $V = 1$ and $T = \mathcal{B}_1^c$—that is, the hold-out estimator of $\mathbb{E}[(P - P_n)\gamma(\widehat{s}_m)]$, which amounts to removing the centering term $-\widehat{s}_m$ in Eq. (28)—because this would dramatically increase its variability. Note that adding such a term $-\widehat{s}_m$ in Eq. (6) does not change the value of the V-fold penalty under (**Reg**) since $\sum_{K=1}^{V}(P_n^{(\mathcal{B}_K^c)} - P_n) = 0$.

Denoting by $\tau_n = |T|/n$ as in Section 8.1, it comes from Lemma 26 in Section D.1 of the Online Appendix that

$$\mathbb{E}\bigl[\mathrm{pen}_{\mathrm{HO}}(m, T, x)\bigr] = x\frac{1-\tau_n}{\tau_n}\mathbb{E}\bigl[\mathrm{pen}_{\mathrm{id}}(m)\bigr] \ .$$

In the following, we choose $x = C\tau_n/(1-\tau_n)$ so that $C = 1$ corresponds to the unbiased case, as in the previous sections for the V-fold penalty.

Remark 11 *Since $P_n = \tau_n P_n^{(T)} + (1-\tau_n)P_n^{(T^c)}$, by linearity of the estimator $\widehat{s}_m$,*

$$\mathrm{pen}_{\mathrm{HO}}(m, T, x) := 2x(1-\tau_n)^2\Bigl(P_n^{(T)} - P_n^{(T^c)}\Bigr)\Bigl(\widehat{s}_m^{(T)} - \widehat{s}_m^{(T^c)}\Bigr)$$

which is symmetric in T and T^c, hence $\mathrm{pen}_{\mathrm{HO}}(m, T^c, x) = \mathrm{pen}_{\mathrm{HO}}(m, T, x)$. In particular, if $|T| = n/2$, the 2-fold penalty computed on the partition $\mathcal{B} = \{T, T^c\}$ and the hold-out penalty coincide

$$\forall x > 0, \quad \mathrm{pen}_{\mathrm{VF}}\bigl(m, \{T, T^c\}, x\bigr) = \mathrm{pen}_{\mathrm{HO}}(m, T, x) \ .$$

Theorem 12 *Let $\xi_{[\![n]\!]}$ be i.i.d. real-valued random variables, $s \in L^\infty(\mu)$ their common density, $T \subset [\![n]\!]$ with $\tau_n = |T|/n \in (0,1)$ and $(S_m)_{m\in\mathcal{M}_n}$ be a collection of separable linear spaces satisfying* (**H1**)*. Assume that either* (**H2**) *or* (**H2′**) *holds true. Let $C \in (1/2, 2]$ and $\delta := 2(C-1)$. For every $m \in \mathcal{M}_n$, let $\widehat{s}_m$ be the projection estimator onto S_m defined by Eq. (1), and $\widetilde{s}_{\mathrm{HO}} = \widehat{s}_{\widehat{m}_{\mathrm{HO}}}$ where*

$$\widehat{m}_{\mathrm{HO}} = \operatorname*{argmin}_{m\in\mathcal{M}_n}\biggl\{P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{HO}}\biggl(m, T, \frac{C\tau_n}{1-\tau_n}\biggr)\biggr\} \ .$$

Then, an absolute constant κ exists such that, for any $x > 0$, defining $x_n = x + \log|\mathcal{M}_n|$, with probability at least $1 - \mathrm{e}^{-x}$, for any $\epsilon \in (0,1]$,

$$\frac{1-\delta_- - \epsilon}{1+\delta_+ + \epsilon}\|\widetilde{s}_{\mathrm{HO}} - s\|^2 \leqslant \inf_{m\in\mathcal{M}_n}\|\widehat{s}_m - s\|^2 + \kappa\biggl(\frac{Ax_n}{\epsilon n} + \frac{\tau_n^2 + (1-\tau_n)^2}{\tau_n(1-\tau_n)}\frac{x_n^2}{\epsilon^3 n}\biggr) \ . \tag{29}$$

Theorem 12 is proved in Section D.1 of the Online Appendix.

Theorem 12 extends Theorem 5 to hold-out penalties, under similar assumptions. As in Theorem 5, δ quantifies the bias of the hold-out penalized criterion, and plays the same role in the leading constant of the oracle inequality (29).

We can compare the results obtained for hold-out and V-fold penalization in Theorems 5 and 12. For this comparison, let V be some divisor of n, $T \subset [\![n]\!]$ such that $|T| = n - n/V$ and choose the same C so that both criteria have the same bias δ. Then, the only difference lies in the remainder term, the one in Eq. (29) is larger than the one of Eq. (17) in Theorem 5 by a factor of order V when V is large. These only are upper bounds, but at least they are consistent with the common intuition about the stabilizing effect of averaging over V folds. We can also compare the results obtained for hold-out penalization in Theorem 12 and for the hold-out criterion in Theorem 9. First, hold-out penalization gives a flexibility to choose an unbiased criterion and therefore to obtain asymptotically optimal oracle inequalities while hold-out criteria are always biased for fixed τ_n, hence a leading constant $\tau_n^{-1} > 1$ in the oracle inequality. The loss in the remainder term is also smaller in Eq. (29) than in Eq. (26) by a factor of order $\tau_n^{-1}(1-\tau_n)^{-1}$ under assumption (**H2′**).

Similarly to Theorems 6 and 10, the variance terms can be computed for the hold-out penalty in order to understand separately the roles of the training sample size and of averaging over the V splits, in the V-fold criteria. Detailed results are given by Proposition 28 in Section D.2 of the Online Appendix.

8.3 Other Oracle Inequalities for Least-Squares Density Estimation

Although the primary topic of the paper is the study of V-fold procedures, let us compare briefly our results to other oracle inequalities that have been proved in the least-squares density estimation setting. For projection estimators, Massart (2007, Section 7.2) proves an oracle inequality for some penalization procedures, which are suboptimal since the leading constant C_n does not tend to 1 as n goes to $+\infty$. Oracle inequalities have also been proved for other estimators: blockwise Stein estimators (Rigollet, 2006), linear estimators (Goldenshluger and Lepski, 2011) and some T-estimators (Birgé, 2013). The models considered by Birgé (2013) are more general than ours, but the corresponding estimators are not computable in practice, and the oracle inequality by Birgé (2013) also has a suboptimal constant C_n. Some aggregation procedures also satisfy oracle inequalities (Rigollet and Tsybakov, 2007; Bunea et al., 2010). Overall, under our assumptions, none of these results imply strictly better bounds than ours.

Let us finally mention that Birgé and Rozenholc (2006) propose a precise evaluation of the penalty term in the case of regular histogram models and the log-likelihood contrast. Their final penalty is a function of the dimension, only slightly modified compared to $\mathrm{pen}_{\mathrm{dim}}$, performing very well on regular histograms. These performances are likely to become much worse on the collection Dya2 presented in Section 6. This can be seen, for example, in Table 3 in Section G of the Online Appendix, where we present the performances of $\mathrm{pen}_{\mathrm{dim}}$ with different over-penalizing constants.

8.4 Conclusion on the Choice of V

This section summarizes the results of the paper in order to address the main question we would like to answer: How to choose a V-fold procedure for model selection?

Generality of the results. The results of the paper only hold for projection estimators in least-squares density estimation, but we conjecture that most of the statements below are valid much more generally. At least, they have been observed experimentally for projection estimators in least-squares regression (Arlot, 2008) and they are supported by theoretical results for kernel density estimators (Magalhães, 2015, Chapters 3–4). Nevertheless, it is reported in the literature that V-fold cross-validation can behave differently in other settings (Arlot and Celisse, 2010), so we must keep in mind that the statements below may not be universal.

Let us also recall that we focus here on model selection with an estimation goal, that is, minimizing the risk of the final estimator; see Yang (2006, 2007) and Celisse (2014) for results when the goal is identification.

Choice of a model selection procedure. Choosing among procedures of the form $\widehat{m}(\mathcal{C})$, as defined by Eq. (18), requires to take into account three quantities:

- *the bias* of $\mathcal{C}(m)$ as an estimator of the risk of $\widehat{s}_m$ for every $m \in \mathcal{M}_n$, or equivalently, the *overpenalization factor* C, which usually drives the performance at first order when $n \to +\infty$, as in Theorem 5. The simulation experiments of Section 6 also show that varying C can strongly change the performance of the procedure. In all settings considered in the paper, some $C_n^\star$ exists (the optimal overpenalization constant) such that the performance decreases for $C \in [0, C_n^\star]$ and increases for $C > C_n^\star$ (Figure 3).

  Note that $C_n^\star$ strongly depends on the setting, and can also vary with V when using V-fold penalization (in particular from $V = 2$ to $V \geqslant 5$). In the nonparametric case, when $n \to +\infty$, Theorem 5 shows that $C_n^\star \sim 1$. On the contrary, in the parametric case, when $n \to +\infty$, it is known that a BIC-type penalty performs better, hence $C_n^\star \to +\infty$. For a finite sample size, Section 6 and Liu and Yang (2011) show that some nonparametric settings can be "practically parametric", that is, $C_n^\star$ can be much larger than 1.

- *the variance* of increments $\mathcal{C}(m) - \mathcal{C}(m')$ drives the performance $\widehat{m}(\mathcal{C})$ at second order, according to the heuristic of Section 4, which suggests that this variance should be minimized, at least for a given "good enough" value of the overpenalization factor C.

- *the computational complexity* of the procedure $\widehat{m}(\mathcal{C})$, that we want to minimize—for a given statistical performance—, or on which some upper bound is given—fixed budget.

V-fold cross-validation. The paper analyzes how the aboves three terms depend on V when $\mathcal{C} = \mathcal{C}_V^{\mathrm{VFCV}}$ is a V-fold cross-validation procedure, under assumption (**Reg**). First, by Lemma 1, its overpenalization factor is $C^{\mathrm{VF}}(V) = 1 + 1/[2(V-1)] \in [1, 3/2]$, which decreases to 1 as V increases to $+\infty$. Second, by Theorem 6, its variance decreases as V increases. Theoretical and empirical arguments in Sections 5 and 6 show that the variance almost reaches its minimal value by taking, say, $V = 5$ or $V = 10$. Third, by Section 7, its computational complexity is proportional to V in general; in the least-squares density estimation setting, it can be reduced to $(n + V^2)\,\mathrm{Card}(\Lambda_m)$.

These three results can explain why the most common advices for choosing V in the literature (for instance Breiman and Spector, 1992; Hastie et al., 2009, Section 7.10.1) are between $V = 5$ and $V = 10$. Indeed, taking V larger does not reduce the variance significantly—with almost no impact on the risk of the final estimator—, and it reduces the overpenalization factor although $C_n^\star$ is often larger than $C^{\mathrm{VF}}(10) = 19/18$ or $C^{\mathrm{VF}}(5) = 9/8$. So, if $C_n^\star$ is not much larger than $1 + 1/8$, which is likely to occur in many nonparametric settings, taking $V = 5$ or 10 can be close to be optimal.

Nevertheless, other situations can occur, for instance in (practically) parametric settings where $C_n^\star$ is much larger, possibly leading to the failure of the heuristic "$5 \leqslant V \leqslant 10$ is almost optimal". More generally, understanding precisely how $\mathcal{C}_V^{\mathrm{VFCV}}$ performs as a function of V seems to be a difficult question: V influences the performance in two opposite directions simultaneously, through the bias and the variance, so that various behaviours can result from this coupling of bias and variance, as shown in the simulation experiments.

***V*-fold penalization.** Lemma 1 shows that a natural way to solve this difficulty is to consider instead a V-fold penalization procedure $\mathcal{C}_{(C,V)}^{\mathrm{pen}_{\mathrm{VF}}}$, with overpenalization factor $C > 0$. The value $C = C^{\mathrm{VF}}(V)$ corresponds to V-fold cross-validation, but any other value of C can also be considered, making it easier to understand. Indeed, the overpenalization factor is directly given by C, while the variance and computational complexity of $\mathcal{C}_{(C,V)}^{\mathrm{pen}_{\mathrm{VF}}}$ vary with V—independently from C—exactly as for V-fold cross-validation. So, V should be taken as large as possible—depending on the maximal computational budget available—, while C should be taken as close as possible to $C_n^\star$.

Compared to V-fold cross-validation, another interest of V-fold penalization is the improvement of the performance for a given computational cost, that is, a given value of V, because it is then possible to take C closer to $C_n^\star$ than $C^{\mathrm{VF}}(V)$. This is especially true in (practically) parametric settings for which $C_n^\star > 3/2 \geqslant C^{\mathrm{VF}}(V)$ for all $V \geqslant 2$.

Data-driven overpenalization factor C**.** Although the paper shows that choosing well C is a key practical problem, making an optimal data-driven choice of C remains an open question which deserves to be studied, even independently from the analysis of cross-validation procedures. We postpone such a study to future works, but we can already make two suggestions. First, an external cross-validation loop can be used for choosing C, if the computational power is not a limitation. Second, a procedure built for choosing between AIC and BIC can be used in order to detect whether C should be close to 1 or significantly larger (see, for instance, Liu and Yang, 2011 and references therein).

Acknowledgments

The authors thank the two referees for their comments that allowed us to improve the paper. The authors thank gratefully Yannick Baraud and Guillaume Obozinski for precious comments on an earlier version of the paper, and Nelo Magalhães for his careful reading and helpful remarks on this earlier version. Let us also emphasize that this paper has changed a lot since its first version (Arlot and Lerasle, 2012), and that the most recent

results (Section 8.1, Section C of the Online Appendix) and some aspects of the proofs (for instance, the systematic use of $U_m(x,y)$ and $K_m(x,y)$) have been strongly influenced by our ongoing collaboration with Nelo Magalhães (Chapters 3–4 Magalhães, 2015).
This work was done while the first author was financed by CNRS and member of the Sierra team in the Departement d'Informatique de l'Ecole normale superieure (CNRS/ENS/INRIA UMR 8548), 45 rue d'Ulm, F-75230 Paris Cedex 05, France. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-09-JCJC-0027-01 (Detect project) and ANR 2011 BS01 010 01 (projet Calibration). The first author also acknowledges the support of the GARGANTUA project funded by the Mastodons program of CNRS. The first author was partly supported by Institut des Hautes Études Scientifiques (IHES, Le Bois-Marie, 35, route de Chartres, 91440 Bures-Sur-Yvette, France) during the last days of writing of this paper.

Appendix A. Proofs

Before proving the main results stated in the paper, let us recall two simple results that we use repeatedly in the paper. First, if $(b_\lambda)_{\lambda\in\Lambda_m}$ is a family of real numbers such that $\sum_{\lambda\in\Lambda_m} b_\lambda^2 < \infty$, then

$$\sup_{\sum_{\lambda\in\Lambda_m} a_\lambda^2 \leqslant 1} \left(\sum_{\lambda\in\Lambda_m} a_\lambda b_\lambda \right)^2 = \sum_{\lambda\in\Lambda_m} b_\lambda^2 \enspace . \tag{30}$$

The left-hand side is smaller than the right-hand side by Cauchy-Schwarz inequality, and considering $a_\lambda = b_\lambda / (\sum_{\lambda'\in\Lambda_m} b_{\lambda'}^2)^{1/2}$ shows that the converse inequality holds true. Second, for any probability distribution Q on $\mathcal{X}$,

$$\sum_{\lambda\in\Lambda_m} (Q\psi_\lambda)\psi_\lambda \in \operatorname*{argmin}_{t\in S_m} \bigl\{ Q\gamma(t) \bigr\} \enspace , \tag{31}$$

a result which provides in particular a formula for $\widehat{s}_m$ and for s_m, by taking $Q = P_n$ and $Q = P$, respectively.

A.1 Proof of Lemma 1

Let us first recall here the proof of Eq. (7)—coming from Arlot (2008)—for the sake of completeness. By (**Reg**),

$$P_n - P_n^{(\mathcal{B}_K^c)} = \frac{1}{V} \Bigl(P_n^{(\mathcal{B}_K)} - P_n^{(\mathcal{B}_K^c)} \Bigr) \qquad \text{and} \qquad P_n^{(\mathcal{B}_K)} - P_n = \frac{V-1}{V} \Bigl(P_n^{(\mathcal{B}_K)} - P_n^{(\mathcal{B}_K^c)} \Bigr) \enspace ,$$

so that

$$\begin{aligned} \mathcal{C}_{1,\mathcal{B}}(m) &:= P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1) \\ &= P_n\gamma(\widehat{s}_m) + \frac{V-1}{V^2} \sum_{K=1}^{V} \Bigl[\Bigl(P_n^{(\mathcal{B}_K)} - P_n^{(\mathcal{B}_K^c)} \Bigr) \gamma\Bigl(\widehat{s}_m^{(\mathcal{B}_K^c)} \Bigr) \Bigr] \\ &= P_n\gamma(\widehat{s}_m) + \frac{1}{V} \sum_{K=1}^{V} \Bigl[\Bigl(P_n^{(\mathcal{B}_K)} - P_n \Bigr) \gamma\Bigl(\widehat{s}_m^{(\mathcal{B}_K^c)} \Bigr) \Bigr] \end{aligned}$$

$$= \mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}) \enspace .$$

Eq. (8) and (9) follow simultaneously from Eq. (35) below. Let $\mathcal{E}$ be a set of subsets of $\llbracket n \rrbracket$ such that

$$\forall A \in \mathcal{E}, \quad |A| = p \quad \text{and} \quad \frac{1}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} P_n^{(A^c)} = P_n \enspace . \tag{32}$$

Let us consider the associated penalty

$$\mathrm{pen}_{\mathcal{E}}(m, C) = \frac{C}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \left(P_n - P_n^{(A^c)} \right) \gamma\left(\widehat{s}_m^{(A^c)} \right) = \frac{2C}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \left(P_n^{(A^c)} - P_n \right) \left(\widehat{s}_m^{(A^c)} \right)$$

and the associated cross-validation criterion

$$\mathrm{crit}_{\mathcal{E}}(m) = \frac{1}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} P_n^{(A)} \gamma\left(\widehat{s}_m^{(A^c)} \right) \enspace .$$

When $\mathcal{E} = \mathcal{B}$, we get the V-fold penalty $\mathrm{pen}_{\mathrm{VF}} = \mathrm{pen}_{\mathcal{E}}$ and the V-fold cross-validation criterion $\mathrm{crit}_{\mathrm{VFCV}} = \mathrm{crit}_{\mathcal{E}}$, and Eq. (32) holds true with $p = n/V$ under assumption (**Reg**). When $\mathcal{E} = \mathcal{E}_p := \{ A \subset \llbracket n \rrbracket \text{ s.t. } |A| = p \}$, Eq. (32) always holds true and we get the leave-p-out penalty $\mathrm{pen}_{\mathrm{LPO}} = \mathrm{pen}_{\mathcal{E}}$ and the leave-p-out cross-validation criterion $\mathrm{crit}_{\mathrm{LPO}} = \mathrm{crit}_{\mathcal{E}}$.

Let $(\psi_\lambda)_{\lambda \in \Lambda_m}$ be some orthonormal basis of S_m in $L^2(\mu)$. On the one hand, using Eq. (32), we get

$$\begin{aligned}
\mathrm{pen}_{\mathcal{E}}(m, C) &= \frac{2C}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \left(P_n^{(A^c)} - P_n \right) \left(\widehat{s}_m^{(A^c)} \right) \\
&= \frac{2C}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \sum_{\lambda \in \Lambda_m} \left[\left(P_n^{(A^c)}(\psi_\lambda) - P_n(\psi_\lambda) \right) P_n^{(A^c)}(\psi_\lambda) \right] \\
&= \frac{2C}{|\mathcal{E}|} \sum_{\lambda \in \Lambda_m} \left[\sum_{A \in \mathcal{E}} \left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - P_n(\psi_\lambda) \sum_{A \in \mathcal{E}} P_n^{(A^c)}(\psi_\lambda) \right] \\
&= \frac{2C}{|\mathcal{E}|} \sum_{\lambda \in \Lambda_m} \sum_{A \in \mathcal{E}} \left[\left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - (P_n(\psi_\lambda))^2 \right] \enspace .
\end{aligned} \tag{33}$$

On the other hand, using that $P_n^{(A)} = \frac{n}{p} P_n - \frac{n-p}{p} P_n^{(A^c)}$ by Eq. (32),

$$\begin{aligned}
&\mathrm{crit}_{\mathcal{E}}(m) - P_n \gamma(\widehat{s}_m) \\
&= \frac{1}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \left[P_n^{(A)} \gamma\left(\widehat{s}_m^{(A^c)} \right) - P_n \gamma(\widehat{s}_m) \right] \\
&= \frac{1}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \left[\left\| \widehat{s}_m^{(A^c)} \right\|^2 - 2 P_n^{(A)} \left(\widehat{s}_m^{(A^c)} \right) - \| \widehat{s}_m \|^2 + 2 P_n(\widehat{s}_m) \right] \\
&= \frac{1}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \sum_{\lambda \in \Lambda_m} \left[\left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - 2 P_n^{(A)}(\psi_\lambda) P_n^{(A^c)}(\psi_\lambda) + \left(P_n(\psi_\lambda) \right)^2 \right] \\
&= \frac{1}{|\mathcal{E}|} \sum_{\lambda \in \Lambda_m} \sum_{A \in \mathcal{E}} \left[\left(\frac{2n}{p} - 1 \right) \left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - \frac{2n}{p} P_n(\psi_\lambda) P_n^{(A^c)}(\psi_\lambda) + \left(P_n(\psi_\lambda) \right)^2 \right]
\end{aligned}$$

$$= \left(\frac{2n}{p} - 1 \right) \frac{1}{|\mathcal{E}|} \sum_{\lambda \in \Lambda_m} \sum_{A \in \mathcal{E}} \left[\left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - \left(P_n(\psi_\lambda) \right)^2 \right] \ , \tag{34}$$

where we used again Eq. (32). Comparing Eq. (33) and (34) gives

$$\mathrm{crit}_{\mathcal{E}}(m) = P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathcal{E}}\left(m, \frac{n}{p} - \frac{1}{2} \right) \tag{35}$$

which implies Eq. (8) and (9). Eq. (10) follows by Lemma A.11 of Lerasle (2012). ■

We now prove the statements made in Remarks 2–3 below Lemma 1.

Proof of Remark 2 We first note that Eq. (10) can also be deduced from Celisse (2014, Proposition 2.1), which proves

$$\mathrm{crit}_{\mathrm{LPO}}(m,p) = \frac{1}{n(n-p)} \sum_{\lambda \in \Lambda_m} \left(\sum_{i=1}^n \psi_\lambda(\xi_i)^2 - \frac{n-p+1}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} \psi_\lambda(\xi_i)\psi_\lambda(\xi_j) \right) \ .$$

Elementary algebraic computations then show that

$$\begin{aligned} &\mathrm{crit}_{\mathrm{LPO}}(m,p) - P_n\gamma(\widehat{s}_m) \\ &\qquad = \frac{2n-p}{n^2(n-p)} \sum_{\lambda \in \Lambda_m} \left(\sum_{i=1}^n \psi_\lambda(\xi_i)^2 - \frac{1}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} \psi_\lambda(\xi_i)\psi_\lambda(\xi_j) \right) \end{aligned} \tag{36}$$

hence for any $p, p' \in [\![n]\!]$,

$$\frac{n/p - 1}{n/p - 1/2} \big(\mathrm{crit}_{\mathrm{LPO}}(m,p) - P_n\gamma(\widehat{s}_m) \big) = \frac{n/p' - 1}{n/p' - 1/2} \big(\mathrm{crit}_{\mathrm{LPO}}(m,p') - P_n\gamma(\widehat{s}_m) \big) \ .$$

In particular, when $p' = 1$, from Eq. (9), since $\mathrm{pen}_{\mathrm{LPO}}(m,1,C) = \mathrm{pen}_{\mathrm{LOO}}(m,C)$,

$$\begin{aligned} \mathrm{pen}_{\mathrm{LPO}}\left(m, p, \frac{n}{p} - \frac{1}{2} \right) &= \frac{n/p - 1/2}{n/p - 1} \frac{n-1}{n-1/2} \mathrm{pen}_{\mathrm{LPO}}\left(m, 1, n - \frac{1}{2} \right) \\ &= \mathrm{pen}_{\mathrm{LOO}}\left(m, (n-1)\frac{n/p - 1/2}{n/p - 1} \right) \ . \end{aligned}$$

■

Proof of Remark 3 Note first that the CV estimator of Massart (2007, Sec. 7.2.1, p. 204–205) is defined as the minimizer of

$$\begin{aligned} &\|\widehat{s}_m\|^2 - \frac{2}{n(n-1)} \sum_{1 \leqslant i \neq j \leqslant n} \sum_{\lambda \in \Lambda_m} \psi_\lambda(\xi_i)\psi_\lambda(\xi_j) \\ &\qquad = P_n\gamma(\widehat{s}_m) + \frac{2}{n^2} \sum_{\lambda \in \Lambda_m} \left(\sum_{i=1}^n \psi_\lambda(\xi_i)^2 - \frac{1}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} \psi_\lambda(\xi_i)\psi_\lambda(\xi_j) \right) \ . \end{aligned} \tag{37}$$

On the other hand, from Eq. (36) and (9) with $p = 1$, we have

$$\mathrm{pen}_{\mathrm{LOO}}(m, n-1) = \frac{2}{n^2} \sum_{\lambda \in \Lambda_m} \left(\sum_{i=1}^{n} \psi_\lambda(\xi_i)^2 - \frac{1}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} \psi_\lambda(\xi_i)\psi_\lambda(\xi_j) \right) \ .$$

Hence, from Eq. (37), the CV estimator is the minimizer of $\mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}_{\mathrm{LOO}})$. Massart (2007, Theorem 7.6) studies the minimizers of the criterion

$$P_n\gamma(\widehat{s}_m) + \frac{C}{n^2}\sum_{i=1}^{n}\sum_{\lambda \in \Lambda_m} \psi_\lambda(\xi_i)^2 \ , \tag{38}$$

where $C = (1+\epsilon)^6$ for any $\epsilon > 0$. Let $\alpha = C/n$, so that $\alpha = (C-\alpha)/(n-1)$. Then, the criterion (38) is equal to

$$\begin{aligned}
&(1-\alpha)P_n\gamma(\widehat{s}_m) + \frac{C-\alpha}{n^2}\sum_{\lambda \in \Lambda_m}\sum_{i=1}^{n}\psi_\lambda(\xi_i)^2 - \frac{\alpha}{n^2}\sum_{\lambda \in \Lambda_m}\sum_{1 \leqslant i \neq j \leqslant n}\psi_\lambda(\xi_i)\psi_\lambda(\xi_j) \\
&= (1-\alpha)P_n\gamma(\widehat{s}_m) + \frac{C-\alpha}{n^2}\sum_{\lambda \in \Lambda_m}\left(\sum_{i=1}^{n}\psi_\lambda(\xi_i)^2 - \frac{1}{n-1}\sum_{\lambda \in \Lambda_m}\sum_{1 \leqslant i \neq j \leqslant n}\psi_\lambda(\xi_i)\psi_\lambda(\xi_j)\right) \\
&= (1-\alpha)\left[P_n\gamma(\widehat{s}_m) + \frac{C-\alpha}{2(1-\alpha)}\,\mathrm{pen}_{\mathrm{LOO}}(m, n-1)\right] \\
&= (1-\alpha)\left[P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{LOO}}\left(m, \frac{C(n-1)^2}{2(n-C)}\right)\right] \ .
\end{aligned}$$

■

A.2 Proof of Proposition 4

Note that the two formulas given for Ψ_m in the statement of Proposition 4 coincide by Eq. (30). The proof is decomposed into 3 lemmas.

Lemma 13 *Let $\xi_{[\![n]\!]}$ denote i.i.d. random variables taking value in a Polish space $\mathcal{X}$, $\mathcal{B}_{[\![V]\!]}$ some partition of $[\![n]\!]$ satisfying* (**Reg**)*, S_m some separable linear subspace of $L^2(\mu)$ with orthonormal basis $(\psi_\lambda)_{\lambda \in \Lambda_m}$ and*

$$U(m) := \frac{1}{n^2}\sum_{1 \leqslant k \neq k' \leqslant V}\sum_{i \in \mathcal{B}_k}\sum_{j \in \mathcal{B}_{k'}}\sum_{\lambda \in \Lambda_m}\big(\psi_\lambda(\xi_i) - P\psi_\lambda\big)\big(\psi_\lambda(\xi_j) - P\psi_\lambda\big) \ . \tag{39}$$

Then, the V-fold penalty is equal to

$$\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) = \frac{2C}{V-1}\|s_m - \widehat{s}_m\|^2 - \frac{2VC}{(V-1)^2}U(m) \tag{40}$$

$$\textit{and} \quad \mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}\left(m, \mathcal{B}, \frac{V-1}{2}\right)\right] = \mathbb{E}\big[\|s_m - \widehat{s}_m\|^2\big] = \frac{\mathcal{D}_m}{2n} \ . \tag{41}$$

Proof Let $W_i = \frac{V}{V-1} \mathbb{1}_{i \notin \mathcal{B}_J}$ and use the formulation (5) of the V-fold penalty as a resampling penalty. Then,

$$\begin{aligned}
\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) &= C \, \mathbb{E}_W \Bigl[\bigl(P_n - P_n^W \bigr) \Bigl(\gamma \bigl(\widehat{s}_m^W \bigr) \Bigr) \Bigr] \\
&= 2C \, \mathbb{E}_W \Bigl[\bigl(P_n^W - P_n \bigr) \bigl(\widehat{s}_m^W \bigr) \Bigr] \\
&= 2C \, \mathbb{E}_W \Bigl[\bigl(P_n^W - P_n \bigr) \bigl(\widehat{s}_m^W - \widehat{s}_m \bigr) \Bigr] \qquad \text{by } (\mathbf{Reg}) \\
&= 2C \sum_{\lambda \in \Lambda_m} \mathbb{E}_W \Bigl[\Bigl(\bigl(P_n^W - P_n \bigr) (\psi_\lambda) \Bigr)^2 \Bigr] \\
&= 2C \sum_{\lambda \in \Lambda_m} \mathbb{E}_W \Bigl[\Bigl(\bigl(P_n^W - P_n \bigr) (\psi_\lambda - P\psi_\lambda) \Bigr)^2 \Bigr] \\
&= \frac{2C}{n^2} \sum_{\lambda \in \Lambda_m} \sum_{1 \leqslant i,j \leqslant n} e_{i,j}^{(\mathrm{VF})} \bigl(\psi_\lambda(\xi_i) - P\psi_\lambda \bigr) \bigl(\psi_\lambda(\xi_j) - P\psi_\lambda \bigr) \qquad (42)
\end{aligned}$$

where $e_{i,j}^{(\mathrm{VF})} := \mathbb{E}[(W_i - 1)(W_j - 1)]$. Since $\mathbb{E}[W_i] = 1$ by (**Reg**) and

$$W_i W_j = \left(\frac{V}{V-1} \right)^2 \mathbb{1}_{J \notin \{J_0, J_1\}} \qquad \text{if} \quad i \in \mathcal{B}_{J_0} \quad \text{and} \quad j \in \mathcal{B}_{J_1} \enspace ,$$

we get that $e_{i,j}^{(\mathrm{VF})} = (V-1)^{-1}$ if i and j belong to the same block and $e_{i,j}^{(\mathrm{VF})} = -(V-1)^{-2}$ otherwise. So,

$$\begin{aligned}
&\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) \\
&= \frac{2C}{n^2 (V-1)} \sum_{\lambda \in \Lambda_m} \sum_{k=1}^{V} \sum_{(i,j) \in \mathcal{B}_k} \bigl(\psi_\lambda(\xi_i) - P\psi_\lambda \bigr) \bigl(\psi_\lambda(\xi_j) - P\psi_\lambda \bigr) - \frac{2C}{(V-1)^2} U(m) \\
&= \frac{2C}{V-1} \sum_{\lambda \in \Lambda_m} \bigl((P_n - P) \psi_\lambda \bigr)^2 - \frac{2CV}{(V-1)^2} U(m)
\end{aligned}$$

and Eq. (40) follows by Eq. (3). Eq. (41) directly follows from Eq. (40). ■

Lemma 14 *Let $\xi_{[\![n]\!]}$ be i.i.d. random variables taking values in a Polish space $\mathcal{X}$ with common density $s \in L^\infty(\mu)$, S_m a separable linear subspace of $L^2(\mu)$ and denote by $(\psi_\lambda)_{\lambda \in \Lambda_m}$ an orthonormal basis of S_m. Let $\mathbb{B}_m = \{t \in S_m \text{ s.t. } \|t\| \leqslant 1\}$, $\mathcal{D}_m = \sum_{\lambda \in \Lambda_m} P\bigl(\psi_\lambda^2\bigr) - \|s_m\|^2$ and assume that $b_m = \sup_{t \in \mathbb{B}_m} \|t\|_\infty < \infty$. An absolute constant κ exists such that, for any $x > 0$, with probability larger than $1 - 2\mathrm{e}^{-x}$, we have for every $\epsilon > 0$,*

$$\left| \|s_m - \widehat{s}_m\|^2 - \frac{\mathcal{D}_m}{n} \right| \leqslant \epsilon \frac{\mathcal{D}_m}{n} + \kappa \left(\frac{\|s\|_\infty x}{(\epsilon \wedge 1) n} + \frac{b_m^2 x^2}{(\epsilon \wedge 1)^3 n^2} \right) \enspace .$$

Proof By Eq. (3), $\|s_m - \widehat{s}_m\|^2 = \sup_{t \in \mathbb{B}_m} [(P_n - P)(t)]^2$ has expectation $\mathcal{D}_m / n$. In addition, for any $t \in \mathbb{B}_m$,

$$\mathrm{Var}\bigl(t(\xi_1)\bigr) \leqslant \int_{\mathbb{R}} t^2 s \, \mathrm{d}\mu \leqslant \|s\|_\infty \|t\|^2 \leqslant \|s\|_\infty \enspace , \qquad (43)$$

which gives the conclusion thanks to a result by Lerasle (2011, Theorem 4.1 of the supplementary material), which is recalled in the Online Appendix (Proposition 29 in Section F). ■

Lemma 15 *Assume that $\xi_{\llbracket n \rrbracket}$ is a sequence of i.i.d. real-valued random variables with common density $s \in L^{\infty}(\mu)$ and $\mathcal{B}_{\llbracket V \rrbracket}$ is some partition of $\llbracket n \rrbracket$ satisfying* (**Reg**). *Let S_m denote a separable subspace of $L^2(\mu)$ with orthonormal basis $(\psi_\lambda)_{\lambda \in \Lambda_m}$ such that*

$$b_m := \sup_{t \in S_m, \|t\| \leqslant 1} \|t\|_{\infty} < +\infty \enspace .$$

Let $U(m)$ be the U-statistics defined by Eq. (39). *Using the notations of Lemma 14, an absolute constant κ exists such that, with probability larger than $1 - 6\mathrm{e}^{-x}$,*

$$\bigl|U(m)\bigr| \leqslant \frac{3\sqrt{(V-1)\|s\|_{\infty}\mathcal{D}_m x}}{\sqrt{V}n} + \kappa\left(\frac{\|s\|_{\infty}x}{n} + \frac{(b_m^2 + \|s\|^2)x^2}{n^2}\right) \enspace .$$

Hence, an absolute constant κ' exists such that, for any $x > 0$, with probability larger than $1 - 6\mathrm{e}^{-x}$, for any $\theta \in (0,1]$,

$$\bigl|U(m)\bigr| \leqslant \theta\frac{\mathcal{D}_m}{n} + \kappa'\left(\frac{\|s\|_{\infty}x}{\theta n} + \frac{(b_m^2 + \|s\|^2)x^2}{n^2}\right) \enspace .$$

Proof For any $x, y \in \mathbb{R}$ and $i, j \in \llbracket n \rrbracket$, let us define

$$U_m(x,y) = \sum_{\lambda \in \Lambda_m} \bigl(\psi_\lambda(x) - P\psi_\lambda\bigr)\bigl(\psi_\lambda(y) - P\psi_\lambda\bigr)$$

and $$g_{i,j}(x,y) = U_m(x,y)\mathbb{1}_{\{\exists k,k' \in \llbracket V \rrbracket \text{ s.t. } k \neq k',\, i \in \mathcal{B}_k,\, j \in \mathcal{B}_{k'}\}}$$

so that $$U(m) = \frac{2}{n^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} g_{i,j}(\xi_i, \xi_j) = \frac{2}{n^2}\sum_{k=2}^{V}\sum_{k'=1}^{k-1}\sum_{i \in \mathcal{B}_k, j \in \mathcal{B}_{k'}} U_m(\xi_i, \xi_j) \enspace .$$

From Houdré and Reynaud-Bouret (2003, Theorem 3.4), an absolute constant κ exists such that, for any $x > 0$ and $\epsilon \in (0,1]$,

$$\mathbb{P}\left(|U(m)| \geqslant \frac{1}{n^2}\left[(4+\epsilon)\overline{A}\sqrt{x} + \kappa\left(\frac{\overline{B}x}{\epsilon} + \frac{\overline{C}x^{3/2}}{\epsilon^3} + \frac{\overline{D}x^2}{\epsilon^3}\right)\right]\right) \leqslant 6\mathrm{e}^{-x} \enspace . \tag{44}$$

$$\overline{A}^2 = \sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}\bigl[g_{i,j}(\xi_i,\xi_j)^2\bigr] \enspace ,$$

$$\overline{B} = \sup\left\{ \mathbb{E}\left[\sum_{i=2}^{n}\sum_{j=1}^{i-1} a_i(\xi_i)b_j(\xi_j)g_{i,j}(\xi_i,\xi_j)\right] \right.$$
$$\left. \text{such that} \quad \mathbb{E}\left[\sum_{i=1}^{n} a_i^2(\xi_i)\right] \leqslant 1 \quad \text{and} \quad \mathbb{E}\left[\sum_{i=1}^{n} b_i^2(\xi_i)\right] \leqslant 1 \right\} \enspace ,$$

$$\overline{C}^2 = \sup_{x \in \mathbb{R}} \left\{ \sum_{i=2}^{n} \mathbb{E}\bigl[g_{i,1}(\xi_i, x)^2 \bigr] \right\} \quad \text{and} \quad \overline{D} = \sup_{x,y} \bigl| g_{i,j}(x,y) \bigr| \enspace .$$

It remains to upper bound these different terms for proving the first inequality, and the second inequality follows. First,

$$\begin{aligned}
\mathbb{E}\Bigl[U_m(\xi_1, \xi_2)^2 \Bigr] &= \sum_{\lambda \in \Lambda_m, \lambda' \in \Lambda_m} \mathbb{E}\Bigl[\bigl(\psi_\lambda(\xi_1) - P\psi_\lambda \bigr) \bigl(\psi_{\lambda'}(\xi_1) - P\psi_{\lambda'} \bigr) \Bigr]^2 \\
&= \sum_{\lambda \in \Lambda_m} \left(\sup_{\sum_{\lambda' \in \Lambda_m} a_{\lambda'}^2 \leqslant 1} \mathbb{E}\left[\bigl(\psi_\lambda(\xi_1) - P\psi_\lambda \bigr) \sum_{\lambda' \in \Lambda_m} a_{\lambda'} \bigl(\psi_{\lambda'}(\xi_1) - P\psi_{\lambda'} \bigr) \right] \right)^2 \\
&= \sum_{\lambda \in \Lambda_m} \left(\sup_{t \in \mathbb{B}_m} \mathbb{E}\Bigl[\bigl(\psi_\lambda(\xi_1) - P\psi_\lambda \bigr) \bigl(t(\xi_1) - P(t) \bigr) \Bigr] \right)^2 \\
&\leqslant \mathcal{D}_m \sup_{t \in \mathbb{B}_m} \mathbb{E}\Bigl[\bigl(t(\xi_1) - P(t) \bigr)^2 \Bigr] \\
&\leqslant \lVert s \rVert_\infty \mathcal{D}_m \qquad\qquad \text{by Eq. (43)}
\end{aligned} \tag{45}$$

so that

$$\overline{A}^2 = \sum_{k=2}^{V} \sum_{k'=1}^{k-1} \sum_{i \in \mathcal{B}_k, j \in \mathcal{B}_{k'}} \mathbb{E}\Bigl[U_m(\xi_i, \xi_j)^2 \Bigr] \leqslant \frac{n^2 (V-1)}{2V} \times \lVert s \rVert_\infty \mathcal{D}_m \enspace .$$

Second, let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be functions in $L^2(\mu)$ such that

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i^2(\xi_i) \right] \leqslant 1 \qquad \text{and} \qquad \mathbb{E}\left[\sum_{i=1}^{n} b_i^2(\xi_i) \right] \leqslant 1 \enspace .$$

Using successively the independence of the ξ_i and that $\alpha\beta \leqslant (\alpha^2 + \beta^2)/2$ for every $\alpha, \beta \in \mathbb{R}$, for every $i \neq j$,

$$\begin{aligned}
&\Bigl| \mathbb{E}\bigl[a_i(\xi_i) b_j(\xi_j) U_m(\xi_i, \xi_j) \bigr] \Bigr| \\
&= \left| \sum_{\lambda \in \Lambda_m} \mathbb{E}\Bigl[a_i(\xi_i) \bigl(\psi_\lambda(\xi_i) - P\psi_\lambda \bigr) \Bigr] \mathbb{E}\Bigl[b_j(\xi_j) \bigl(\psi_\lambda(\xi_j) - P\psi_\lambda \bigr) \Bigr] \right| \\
&\leqslant \frac{1}{2} \sum_{\lambda \in \Lambda_m} \left(\mathbb{E}\Bigl[a_i(\xi_i) \bigl(\psi_\lambda(\xi_i) - P\psi_\lambda \bigr) \Bigr]^2 + \mathbb{E}\Bigl[b_j(\xi_j) \bigl(\psi_\lambda(\xi_j) - P\psi_\lambda \bigr) \Bigr]^2 \right) \enspace .
\end{aligned} \tag{46}$$

Now, we have, for every $i \in [\![n]\!]$, using Eq. (30), Cauchy-Schwarz inequality and the fact that for every $t \in L^2(\mu)$, $\operatorname{Var}(t(\xi_1)) \leqslant \lVert s \rVert_\infty \lVert t \rVert^2$,

$$\sum_{\lambda \in \Lambda_m} \mathbb{E}\Bigl[a_i(\xi_i) \bigl(\psi_\lambda(\xi_i) - P\psi_\lambda \bigr) \Bigr]^2 = \sup_{\sum_{\lambda \in \Lambda_m} t_\lambda^2 \leqslant 1} \left(\mathbb{E}\left[a_i(\xi_i) \sum_{\lambda \in \Lambda_m} t_\lambda \psi_\lambda(\xi_i) - P(t_\lambda \psi_\lambda) \right] \right)^2$$

$$\begin{aligned}
&= \sup_{t \in \mathbb{B}_m} \left(\mathbb{E}\Big[a_i(\xi_i) \big(t(\xi_i) - P(t) \big) \Big] \right)^2 \\
&\leqslant \mathbb{E}\big[a_i(\xi_i)^2 \big] \sup_{t \in \mathbb{B}_m} \operatorname{Var}\big(t(\xi_1) \big) \leqslant \mathbb{E}\big[a_i(\xi_i)^2 \big] \lVert s \rVert_\infty \enspace .
\end{aligned}$$

Plugging this bound in (46) yields

$$\Big| \mathbb{E}\big[a_i(\xi_i) b_j(\xi_j) U_m(\xi_i, \xi_j) \big] \Big| \leqslant \frac{\lVert s \rVert_\infty}{2} \Big(\mathbb{E}\big[a_i(\xi_i)^2 \big] + \mathbb{E}\big[b_j(\xi_j)^2 \big] \Big) \tag{47}$$

hence

$$\overline{B} \leqslant n \lVert s \rVert_\infty \enspace .$$

Third, for every $x, y \in \mathbb{R}$, let $g_x(y) = \sum_{\lambda \in \Lambda_m} (\psi_\lambda(x) - P\psi_\lambda) \psi_\lambda(y)$ so that

$$\begin{aligned}
\lVert g_x \rVert^2 = \sum_{\lambda \in \Lambda_m} \big(\psi_\lambda(x) - P\psi_\lambda \big)^2 &\leqslant 2 \sum_{\lambda \in \Lambda_m} \big(\psi_\lambda(x) \big)^2 + 2 \sum_{\lambda \in \Lambda_m} (P\psi_\lambda)^2 \\
&= 2\Psi_m(x)^2 + 2 \lVert s_m \rVert^2 \leqslant 2 \Big(b_m^2 + \lVert s_m \rVert^2 \Big) \enspace .
\end{aligned}$$

Then,

$$\mathbb{E}\big[U_m(\xi_i, x)^2 \big] = \operatorname{Var}\big(g_x(\xi_1) \big) \leqslant \lVert g_x \rVert^2 \lVert s \rVert_\infty \leqslant 2 \Big(b_m^2 + \lVert s_m \rVert^2 \Big) \lVert s \rVert_\infty \tag{48}$$

and, using (**Reg**), we get that

$$\overline{C}^2 \leqslant \frac{2n(V-1)}{V} \Big(b_m^2 + \lVert s_m \rVert^2 \Big) \lVert s \rVert_\infty \enspace .$$

Fourth, from Cauchy-Schwarz inequality, for every $x, y \in \mathcal{X}$,

$$U_m(x, y) \leqslant \sup_{x \in \mathbb{R}} \sum_{\lambda \in \Lambda_m} \big(\psi_\lambda(x) - P\psi_\lambda \big)^2 \leqslant 2 \Big(b_m^2 + \lVert s_m \rVert^2 \Big) \enspace . \tag{49}$$

Hence,

$$\overline{D} \leqslant 2 \Big(b_m^2 + \lVert s_m \rVert^2 \Big)$$

and we get the desired result. ■

Let us conclude the proof of Proposition 4. From Lemmas 13 and 15, an absolute constant κ exists such that, with probability larger than $1 - 6\mathrm{e}^{-x}$, for every $\epsilon \in (0, 1]$,

$$\begin{aligned}
&\Big| \operatorname{pen}_{\mathrm{VF}}(m, V, V-1) - 2 \lVert s_m - \widehat{s}_m \rVert^2 \Big| \\
&\quad = \frac{2V}{V-1} \big| U(m) \big| \leqslant \epsilon \frac{\mathcal{D}_m}{n} + \kappa \left(\frac{\lVert s \rVert_\infty x}{\epsilon n} + \frac{\big(b_m^2 + \lVert s \rVert^2 \big) x^2}{n^2} \right) \enspace .
\end{aligned} \tag{50}$$

Using in addition Lemma 14, we get that an absolute constant κ' exists such that with probability larger than $1 - 8\mathrm{e}^{-x}$, for every $\epsilon \in (0, 1]$, Eq. (50) holds true and

$$\left| \operatorname{pen}_{\mathrm{VF}}(m, V, V-1) - \frac{2\mathcal{D}_m}{n} \right| \leqslant \epsilon \frac{\mathcal{D}_m}{n} + \kappa \left(\frac{\lVert s \rVert_\infty x}{\epsilon n} + \frac{\big(b_m^2 \epsilon^{-3} + \lVert s \rVert^2 \big) x^2}{n^2} \right) \enspace ,$$

which implies Eq. (15) and (16). ■

A.3 Proof of Theorem 5

By construction, the penalized estimator satisfies, for any $m \in \mathcal{M}_n$,

$$\begin{aligned}&\left\|\widehat{s}_{\widehat{m}} - s\right\|^2 - \Bigl(\mathrm{pen}_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}_{\mathrm{VF}}\bigl(\widehat{m}, V, C(V-1)\bigr)\Bigr) \\ &\leqslant \left\|\widehat{s}_m - s\right\|^2 + \Bigl(\mathrm{pen}_{\mathrm{VF}}\bigl(m, V, C(V-1)\bigr) - \mathrm{pen}_{\mathrm{id}}(m)\Bigr) \enspace .\end{aligned}$$

Now, by Eq. (2) and (3), $\mathrm{pen}_{\mathrm{id}}(m) = 2\|\widehat{s}_m - s_m\|^2 + 2(P_n - P)(s_m)$, hence

$$\begin{aligned}\left\|\widehat{s}_{\widehat{m}} - s\right\|^2 &\leqslant \left\|\widehat{s}_m - s\right\|^2 + \Bigl[\mathrm{pen}_{\mathrm{VF}}\bigl(m, V, C(V-1)\bigr) - 2\left\|s_m - \widehat{s}_m\right\|^2\Bigr] \\ &\qquad - \Bigl[\mathrm{pen}_{\mathrm{VF}}\bigl(\widehat{m}, V, C(V-1)\bigr) - 2\left\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\right\|^2\Bigr] + 2(P_n - P)(s_m - s_{\widehat{m}}) \\ &= \left\|\widehat{s}_m - s\right\|^2 + \Bigl[\mathrm{pen}_{\mathrm{VF}}\bigl(m, V, C(V-1)\bigr) - 2C\left\|s_m - \widehat{s}_m\right\|^2\Bigr] \\ &\qquad - \Bigl[\mathrm{pen}_{\mathrm{VF}}\bigl(\widehat{m}, V, C(V-1)\bigr) - 2C\left\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\right\|^2\Bigr] + 2(P_n - P)(s_m - s_{\widehat{m}}) \\ &\qquad + 2(C-1)\Bigl(\left\|\widehat{s}_m - s_m\right\|^2 - \left\|\widehat{s}_{\widehat{m}} - s_{\widehat{m}}\right\|^2\Bigr) \enspace . \end{aligned} \tag{51}$$

Let $x > 0$ and $x_n = \log(|\mathcal{M}_n|) + x$. A union bound in Proposition 4 gives

$$\begin{aligned}&\mathbb{P}\Bigl(\exists m \in \mathcal{M}_n,\, \epsilon \in (0,1] \text{ s.t. } \bigl|\mathrm{pen}_{\mathrm{VF}}(m, V, V-1) - 2\left\|s_m - \widehat{s}_m\right\|^2\bigr| \\ &\qquad > \epsilon\frac{\mathcal{D}_m}{n} + \kappa\rho_1(m, \epsilon, s, x_n, n)\Bigr) \leqslant 8\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-x_n} = 8\mathrm{e}^{-x}\sum_{m\in\mathcal{M}_n}\frac{1}{|\mathcal{M}_n|} = 8\mathrm{e}^{-x}\end{aligned} \tag{52}$$

and a union bound in Lemma 14 gives

$$\begin{aligned}&\mathbb{P}\Biggl(\exists m \in \mathcal{M}_n,\, \epsilon \in (0,1] \text{ s.t. } \left|\left\|\widehat{s}_m - s_m\right\|^2 - \frac{\mathcal{D}_m}{n}\right| > \epsilon\frac{\mathcal{D}_m}{n} + \kappa\rho_1(m, \epsilon, s, x_n, n)\Biggr) \\ &\leqslant 2\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-x_n} = 2\mathrm{e}^{-x} \enspace .\end{aligned} \tag{53}$$

It remains to bound $2(P_n - P)(s_m - s_{m'})$ uniformly over m and m' in $\mathcal{M}_n$. In order to apply Bernstein's inequality, we first bound the variance and the sup norm of $s_m - s_{m'}$ for some $m, m' \in \mathcal{M}_n$. Since $s \in L^\infty(\mu)$,

$$\mathrm{Var}\bigl((s_m - s_{m'})(\xi_1)\bigr) \leqslant \|s\|_\infty \left\|s_m - s_{m'}\right\|^2 \enspace .$$

Under assumption (**H2**)

$$\left\|s_m - s_{m'}\right\|_\infty \leqslant \left\|s_m\right\|_\infty + \left\|s_{m'}\right\|_\infty \leqslant 2a \enspace .$$

Under assumption (**H2′**), $s_m - s_{m'} \in S_{m''}$ for some $m'' \in \{m, m'\}$, hence by (**H1**) we have

$$\left\|s_m - s_{m'}\right\|_\infty \leqslant b_{m''}\left\|s_m - s_{m'}\right\| \leqslant \sqrt{n}\left\|s_m - s_{m'}\right\| \enspace .$$

Therefore, by Bernstein's inequality, for any $x > 0$, for any m, m', with probability larger than $1 - \mathrm{e}^{-x}$, for any $\epsilon \in (0,1]$,

$$\begin{aligned}(P_n - P)(s_m - s_{m'}) &\leqslant \sqrt{\frac{2x \operatorname{Var}\bigl((s_m - s_{m'})(\xi_1)\bigr)}{n}} + \frac{\|s_m - s_{m'}\|_\infty x}{3n} \\ &\leqslant \epsilon \|s_m - s_{m'}\|^2 + \frac{\kappa\bigl(Ax + x^2\bigr)}{\epsilon n} \enspace .\end{aligned}$$

for some absolute constant κ, where the last inequality is obtained by considering separately the cases (**H2**) and (**H2′**), and by using that for every $\alpha, \beta, \epsilon > 0$, $\alpha\beta \leqslant \epsilon\alpha^2 + (\beta^2)/(4\epsilon)$. A union bound gives that for any $x > 0$, with probability at least $1 - |\mathcal{M}_n|^2 \mathrm{e}^{-x}$, for every $m, m' \in \mathcal{M}_n$ and every $\epsilon \in (0,1]$,

$$(P_n - P)(s_m - s_{m'}) \leqslant \epsilon \|s_m - s_{m'}\|^2 + \frac{\kappa\bigl(Ax + x^2\bigr)}{\epsilon n} \tag{54}$$

for some absolute constant κ. Plugging Eq. (52), (53) and (54) into Eq. (51) and using that $C \in (1/2, 2]$ yields that, with probability $1 - (|\mathcal{M}_n|^2 + 10)\mathrm{e}^{-x}$, for any $\epsilon \in (0, 1/2]$,

$$\begin{aligned}(1 - 4\epsilon)\|\widehat{s}_{\widehat{m}} - s\|^2 &\leqslant (1 + 4\epsilon)\|\widehat{s}_m - s\|^2 + (\delta_+ + 4\epsilon)\frac{\mathcal{D}_m}{n} + (\delta_- + 3\epsilon)\frac{\mathcal{D}_{\widehat{m}}}{n} \\ &\qquad + \kappa\left(\rho_1(m, \epsilon, s, x, n) + \rho_1(\widehat{m}, \epsilon, s, x, n) + \frac{Ax + x^2}{\epsilon n}\right) \\ &\leqslant (1 + \delta_+ + 16\epsilon)\|\widehat{s}_m - s\|^2 + (\delta_- + 8\epsilon)\|\widehat{s}_{\widehat{m}} - s_m\|^2 \\ &\qquad + \kappa'\left(\rho_1(m, \epsilon, s, x, n) + \rho_1(\widehat{m}, \epsilon, s, x, n) + \frac{Ax + x^2}{\epsilon n}\right)\end{aligned}$$

for some absolute constants $\kappa, \kappa' > 0$. Since $b_m \leqslant \sqrt{n}$ for all $m \in \mathcal{M}_n$, we get

$$2 \sup_{m \in \mathcal{M}_n} \rho_1(m, \epsilon, s, x, n) + \frac{Ax + x^2}{\epsilon n} \leqslant \frac{\bigl(2\|s\|_\infty + A\bigr)x}{\epsilon n} + \left(3 + \frac{2\|s\|^2}{n}\right)\frac{x^2}{\epsilon^3 n}$$

for every $\epsilon \in (0,1]$. Hence, with probability larger than $1 - (|\mathcal{M}_n|^2 + 10)\mathrm{e}^{-x}$, for any $\epsilon \in (0,1]$,

$$\frac{1 - \delta_- - \epsilon}{1 + \delta_+ + \epsilon}\|\widehat{s}_{\widehat{m}} - s\|^2 \leqslant \|\widehat{s}_m - s\|^2 + \kappa\left[\frac{\bigl(\|s\|_\infty + A\bigr)x}{\epsilon n} + \left(1 + \frac{\|s\|^2}{n}\right)\frac{x^2}{\epsilon^3 n}\right] \tag{55}$$

for some absolute constant $\kappa > 0$. To conclude, we remark that Eq. (17) clearly holds true when $|\mathcal{M}_n| = 1$, so we can assume that $|\mathcal{M}_n| \geqslant 2$. Therefore, for every $x > 0$, Eq. (55) holds true with probability at least

$$1 - \Bigl(|\mathcal{M}_n|^2 + 10\Bigr)\mathrm{e}^{-x} \geqslant 1 - |\mathcal{M}_n|^4 \mathrm{e}^{-x} \geqslant 1 - \mathrm{e}^{-x + 4\log|\mathcal{M}_n|} \enspace .$$

So, if we replace x by $4x_n \geqslant x + 4\log|\mathcal{M}_n|$ in Eq. (55), we get that Eq. (17) holds true with probability at least $1 - \mathrm{e}^{-x}$ for some absolute constant $\kappa > 0$, slightly larger than the one appearing in Eq. (55). ■

A.4 Proof of Theorem 6

For every $x, y \in \mathcal{X}$ and $m \in \{m_1, m_2\}$, let $K_m(x,y) := \sum_{\lambda \in \Lambda_m} \psi_\lambda(x)\psi_\lambda(y)$ and remark that

$$\begin{aligned} U_m(x,y) &= \sum_{\lambda \in \Lambda_m} \bigl(\psi_\lambda(x) - P\psi_\lambda\bigr)\bigl(\psi_\lambda(y) - P\psi_\lambda\bigr) \\ &= K_m(x,y) - s_m(x) - s_m(y) + \lVert s_m \rVert^2 \ . \end{aligned} \tag{56}$$

For every $x \in \mathcal{X}$, $K_m(x,x) = \Psi_m(x)$ by Eq. (30), $U_m(x,x) = \Psi_m(x) - 2s_m(x) + \lVert s_m \rVert^2$ and, by independence, for every $m, m' \in \{m_1, m_2\}$

$$\begin{aligned} &\operatorname{Cov}\bigl(U_m(\xi_1,\xi_2), U_{m'}(\xi_1,\xi_2)\bigr) \\ &\quad = \sum_{\lambda \in \Lambda_m, \lambda' \in \Lambda_{m'}} \mathbb{E}\Bigl[\bigl(\psi_\lambda(\xi_1) - P\psi_\lambda\bigr)\bigl(\psi_\lambda(\xi_2) - P\psi_\lambda\bigr)\bigl(\psi_{\lambda'}(\xi_1) - P\psi_{\lambda'}\bigr)\bigl(\psi_{\lambda'}(\xi_2) - P\psi_{\lambda'}\bigr)\Bigr] \\ &\quad = \sum_{\lambda \in \Lambda_m, \lambda' \in \Lambda_{m'}} \mathbb{E}\Bigl[\bigl(\psi_\lambda(\xi_1) - P\psi_\lambda\bigr)\bigl(\psi_{\lambda'}(\xi_1) - P\psi_{\lambda'}\bigr)\Bigr]^2 = \beta\bigl(m, m'\bigr) \ , \end{aligned}$$

hence, $\operatorname{Var}(U_{m_1}(\xi_1,\xi_2) - U_{m_2}(\xi_1,\xi_2)) = \mathbf{B}(m_1, m_2)$. For every $m \in \{m_1, m_2\}$, by Eq. (56),

$$\begin{aligned} P_n\gamma(\widehat{s}_m) &= -\sum_{\lambda \in \Lambda_m} (P_n\psi_\lambda)^2 = -\frac{1}{n^2} \sum_{1 \leqslant i,j \leqslant n} K_m(\xi_i, \xi_j) \\ &= -\frac{1}{n^2} \sum_{1 \leqslant i,j \leqslant n} U_m(\xi_i,\xi_j) - \frac{2}{n} \sum_{i=1}^n s_m(\xi_i) + \lVert s_m \rVert^2 \ . \end{aligned} \tag{57}$$

Moreover, by Eq. (42) in the proof of Lemma 13,

$$\operatorname{pen}_{\mathrm{VF}}\bigl(m, \mathcal{B}, C(V-1)\bigr) = \frac{2C}{n^2} \sum_{1 \leqslant i,j \leqslant n} E_{i,j}^{(\mathrm{VF})} U_m(\xi_i, \xi_j)$$

$$\text{where} \quad \forall I, J \in \{1, \ldots, V\},\, \forall i \in B_I,\, \forall j \in B_J, \quad E_{i,j}^{(\mathrm{VF})} = 1 - \frac{V \mathbb{1}_{I \neq J}}{V-1} = (V-1) e_{i,j}^{(\mathrm{VF})} \ .$$

It follows that

$$\mathcal{C}_{C,\mathcal{B}}(m) = \sum_{1 \leqslant i,j \leqslant n} \frac{2CE_{i,j}^{(\mathrm{VF})} - 1}{n^2} U_m(\xi_i, \xi_j) + \sum_{i=1}^n \frac{-2s_m(\xi_i)}{n} + \lVert s_m \rVert^2 \ . \tag{58}$$

Hence, up to the deterministic term $\lVert s_m \rVert^2$, $\mathcal{C}_{C,\mathcal{B}}(m)$ has the form of a function $\mathcal{C}_m$ defined in Lemma 16 below with

$$\overline{\omega}_{i,j} = \frac{2CE_{i,j}^{(\mathrm{VF})} - 1}{n^2} \ , \qquad f_m = \frac{-2s_m}{n} \qquad \text{and} \qquad \overline{\sigma}_i = 1 \ .$$

It remains to evaluate the quantities appearing in Lemma 16 for these weights and function. First,

$$\sum_{i=1}^n E_{i,i}^{(\mathrm{VF})} = n \qquad \text{and} \qquad \sum_{i=1}^n \Bigl(E_{i,i}^{(\mathrm{VF})}\Bigr)^2 = n \ .$$

Second, by (**Reg**),

$$\sum_{1 \leqslant i \neq j \leqslant n} \left(E_{i,j}^{(\mathrm{VF})} \right) = n \Big(\frac{n}{V} - 1 \Big) + \frac{-1}{(V-1)} \times \frac{n^2 (V-1)}{V} = -n$$

$$\text{and} \quad \sum_{1 \leqslant i \neq j \leqslant n} \left(E_{i,j}^{(\mathrm{VF})} \right)^2 = n \left[\Big(\frac{n}{V} - 1 \Big) + \frac{n}{V(V-1)} \right] = \frac{n^2}{V-1} - n \enspace .$$

It follows that

$$\sum_{1 \leqslant i \leqslant n} \overline{\omega}_{i,i}^2 = \frac{(2C-1)^2}{n^3} \enspace , \qquad \sum_{i=1}^n \overline{\omega}_{i,i} \overline{\sigma}_i = \frac{2C-1}{n}$$

$$\text{and} \quad \sum_{1 \leqslant i \neq j \leqslant n} \overline{\omega}_{i,j} \overline{\omega}_{j,i} = \sum_{1 \leqslant i \neq j \leqslant n} \overline{\omega}_{i,j}^2 = \frac{1}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \enspace .$$

Hence, from Lemma 16, for every $m, m' \in \{m_1, m_2\}$,

$$\begin{aligned}
\operatorname{Cov}\big(\mathcal{C}_{C,\mathcal{B}}(m), \mathcal{C}_{C,\mathcal{B}}(m') \big) &= \frac{2}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \beta\big(m, m'\big) \\
&\quad + \frac{(2C-1)^2}{n^3} \operatorname{Cov}\big(U_m(\xi,\xi), U_{m'}(\xi,\xi) \big) + \frac{4}{n} \operatorname{Cov}\big(s_m(\xi), s_{m'}(\xi) \big) \\
&\quad - \frac{2(2C-1)}{n^2} \Big[\operatorname{Cov}\big(U_m(\xi,\xi), s_{m'}(\xi) \big) + \operatorname{Cov}\big(U_{m'}(\xi,\xi), s_m(\xi) \big) \Big] \\
&= \frac{2}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \beta\big(m, m'\big) \\
&\quad + \frac{1}{n} \operatorname{Cov}\left(\frac{2C-1}{n} U_m(\xi,\xi) - 2 s_m(\xi), \frac{2C-1}{n} U_{m'}(\xi,\xi) - 2 s_{m'}(\xi) \right) \enspace .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{Var}\big(\mathcal{C}_{C,\mathcal{B}}(m_1) \big) &= \frac{2}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \beta(m_1, m_1) \\
&\quad + \frac{1}{n} \operatorname{Var}\left(\frac{2C-1}{n} U_{m_1}(\xi,\xi) - 2 s_{m_1}(\xi) \right)
\end{aligned}$$

$$\begin{aligned}
\text{and} \quad \operatorname{Var}(\mathcal{C}_{C,\mathcal{B}}(m_1) - \mathcal{C}_{C,\mathcal{B}}(m_2)) &= \frac{2}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \mathbf{B}(m_1, m_2) \\
&\quad + \frac{1}{n} \operatorname{Var}\left(2(s_{m_1} - s_{m_2})(\xi) - \frac{2C-1}{n} \big(U_{m_1}(\xi,\xi) - U_{m_2}(\xi,\xi) \big) \right) \\
&= \frac{2}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \operatorname{Var}\big(U_{m_1}(\xi,\xi) - U_{m_2}(\xi_1,\xi_2) \big) \\
&\quad + \frac{4}{n} \operatorname{Var}\left(\left(1 + \frac{2C-1}{n} \right) (s_{m_1} - s_{m_2})(\xi) - \frac{2C-1}{2n} \big(\Psi_{m_1}(\xi) - \Psi_{m_2}(\xi) \big) \right) \enspace ,
\end{aligned}$$

which concludes the proof. ■

Lemma 16 *Let $\mathcal{C}_m = \sum_{1 \leqslant i,j \leqslant n} \overline{\omega}_{i,j} U_m(\xi_i, \xi_j) + \sum_{i=1}^n \overline{\sigma}_i f_m(\xi_i)$, where U_m is defined by Eq. (56) and $f_m \in L^2(\mu)$. For every m, m', we have*

$$\begin{aligned}\operatorname{Cov}(\mathcal{C}_m, \mathcal{C}_{m'}) &= \left(\sum_{1 \leqslant i \neq j \leqslant n} \overline{\omega}_{i,j}^2 + \overline{\omega}_{i,j}\overline{\omega}_{j,i} \right) \operatorname{Cov}\bigl(U_m(\xi_1, \xi_2), U_{m'}(\xi_1, \xi_2)\bigr) \\ &+ \left(\sum_{i=1}^n \overline{\omega}_{i,i}^2 \right) \operatorname{Cov}\bigl(U_m(\xi_1, \xi_1), U_{m'}(\xi_1, \xi_1)\bigr) \\ &+ \left(\sum_{i=1}^n \overline{\omega}_{i,i}\overline{\sigma}_i \right) \Bigl[\operatorname{Cov}\bigl(U_m(\xi_1, \xi_1), f_{m'}(\xi_1)\bigr) + \operatorname{Cov}\bigl(U_{m'}(\xi_1, \xi_1), f_m(\xi_1)\bigr) \Bigr] \\ &+ \left(\sum_{i=1}^n \overline{\sigma}_i^2 \right) \operatorname{Cov}\bigl(f_m(\xi_1), f_{m'}(\xi_1)\bigr) \ .\end{aligned}$$

Proof We develop the covariance to get

$$\begin{aligned}\operatorname{Cov}(\mathcal{C}_m, \mathcal{C}_{m'}) = &\sum_{1 \leqslant i,j,k,\ell \leqslant n} \overline{\omega}_{i,j}\overline{\omega}_{k,\ell} \operatorname{Cov}\bigl(U_m(\xi_i, \xi_j), U_{m'}(\xi_k, \xi_\ell)\bigr) \\ &+ \sum_{1 \leqslant i,j,k \leqslant n} \overline{\omega}_{i,j}\overline{\sigma}_k \operatorname{Cov}\bigl(U_m(\xi_i, \xi_j), f_{m'}(\xi_k)\bigr) \\ &+ \sum_{1 \leqslant i,j,k \leqslant n} \overline{\omega}_{i,j}\overline{\sigma}_k \operatorname{Cov}\bigl(U_{m'}(\xi_i, \xi_j), f_m(\xi_k)\bigr) \\ &+ \sum_{1 \leqslant i,j \leqslant n} \overline{\sigma}_i\overline{\sigma}_j \operatorname{Cov}\bigl(f_m(\xi_i), f_{m'}(\xi_j)\bigr) \ .\end{aligned}$$

The proof is then concluded with the following remarks, which rely on the fact that the random variables $\xi_{[\![n]\!]}$ are independent and identically distributed.

1. $\operatorname{Cov}\bigl(f_m(\xi_i), f_{m'}(\xi_j)\bigr) = 0$ unless $i \neq j$, therefore

$$\sum_{1 \leqslant i,j \leqslant n} \overline{\sigma}_i\overline{\sigma}_j \operatorname{Cov}\bigl(f_m(\xi_i), f_{m'}(\xi_j)\bigr) = \left(\sum_{i=1}^n \overline{\sigma}_i^2 \right) \operatorname{Cov}\bigl(f_m(\xi_1), f_{m'}(\xi_1)\bigr) \ .$$

2. By definition (56) of U_m, $\operatorname{Cov}\bigl(U_m(\xi_i, \xi_j), f_{m'}(\xi_k)\bigr) = 0$ unless $i = j = k$, hence

$$\sum_{1 \leqslant i,j,k \leqslant n} \overline{\omega}_{i,j}\overline{\sigma}_k \operatorname{Cov}\bigl(U_m(\xi_i, \xi_j), f_{m'}(\xi_k)\bigr) = \left(\sum_{i=1}^n \overline{\omega}_{i,i}\overline{\sigma}_i \right) \operatorname{Cov}\bigl(U_m(\xi_1, \xi_1), f_{m'}(\xi_1)\bigr) \ .$$

3. By definition (56) of U_m, $\operatorname{Cov}\bigl(U_m(\xi_i, \xi_j), U_m(\xi_k, \xi_l)\bigr) = 0$ unless $i = j = k = \ell$ or $i = k \neq j = \ell$ or $i = \ell \neq j = k$. It follows that

$$\sum_{1 \leqslant i,j,k,\ell \leqslant n} \overline{\omega}_{i,j}\overline{\omega}_{k,\ell} \operatorname{Cov}\bigl(U_m(\xi_i, \xi_j), U_{m'}(\xi_k, \xi_\ell)\bigr)$$

$$= \left(\sum_{1 \leqslant i \neq j \leqslant n} \overline{\omega}_{i,j}^2 + \overline{\omega}_{i,j} \overline{\omega}_{j,i} \right) \mathrm{Cov}\big(U_m(\xi_1, \xi_2), U_{m'}(\xi_1, \xi_2) \big)$$

$$+ \left(\sum_{i=1}^{n} \overline{\omega}_{i,i}^2 \right) \mathrm{Cov}\big(U_m(\xi_1, \xi_1), U_{m'}(\xi_1, \xi_1) \big) \enspace .$$

■

References

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

Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. `http://arxiv.org/pdf/0802.0566v2.pdf`.

Sylvain Arlot. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624 (electronic), 2009.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Sylvain Arlot and Matthieu Lerasle. V-fold cross-validation and V-fold penalization in least-squares density estimation, October 2012. `http://arxiv.org/pdf/1210.5830v1.pdf`.

Jean-Yves Audibert. A better variance control for pac-bayesian classification. Technical Report 905b, Laboratoire de Probabilités et Modèles Aléatoires, 2004. Available electronically at `http://imagine.enpc.fr/publications/papers/04PMA-905Bis.pdf`.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

Yoshua Bengio and Yves Grandvalet. Bias in estimating the variance of K-fold cross-validation. In *Statistical Modeling and Analysis for Complex Data Problems*, volume 1 of *GERAD 25th Anniversary Series*, pages 75–95. Springer, New York, 2005.

Lucien Birgé. Model selection for density estimation with $\mathbb{L}_2$-loss. *Probability Theory and Related Fields*, pages 1–42, 2013.

Lucien Birgé and Yves Rozenholc. How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, 10, 2006.

Leo Breiman and Philip Spector. Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, 60(3):291–319, 1992.

Florentina Bunea, Alexandre B. Tsybakov, Marten H. Wegkamp, and Adrian Barbu. Spades and mixture models. *The Annals of Statistics*, 38(4):2525–2558, 2010.

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

Prabir Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā (Statistics). Series A*, 52(3):314–345, 1990.

Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007.

Alain Celisse. *Model Selection Via Cross-Validation in Density Estimation, Regression and Change-Points Detection*. PhD thesis, University Paris-Sud 11, December 2008. Available electronically at `http://tel.archives-ouvertes.fr/tel-00346320/`.

Alain Celisse. Optimal cross-validation in density estimation with the L^2-loss. *The Annals of Statistics*, 42(5):1879–1910, 10 2014.

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

Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*]. Springer-Verlag, Berlin, 1993.

Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39 (3):1608–1632, 2011.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data Mining, Inference, and Prediction.

Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic Inequalities and Applications*, volume 56 of *Progress in Probability*, pages 55–69. Birkhäuser, Basel, 2003.

Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. *The Annals of Statistics*, 39(4):1852–1877, 2011.

Matthieu Lerasle. Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, 48(3):884–908, 2012.

Wei Liu and Yuhong Yang. Parametric or nonparametric? A parametricness index for model selection. *The Annals of Statistics*, 39(4):2074–2102, 2011.

Nelo Magalhães. *Cross-Validation and Penalization for Density Estimation.* PhD thesis, University Paris-Sud 11, May 2015. Available electronically at `http://tel.archives-ouvertes.fr/tel-01164581/`.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics.* Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

Richard R. Picard and R. Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

Philippe Rigollet. Adaptive density estimation using the blockwise Stein method. *Bernoulli*, 12(2):351–370, 2006.

Philippe Rigollet and Alexander B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

Mats Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics. Theory and Applications*, 9(2):65–78, 1982.

Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–264, 1997. With comments and a rejoinder by the author.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B. Methodological*, 36:111–147, 1974.

Mark J. van der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Working Paper 130, U.C. Berkeley Division of Biostatistics, November 2003. Available electronically at `http://www.bepress.com/ucbbiostat/paper130`.

Mark J. van der Laan, Sandrine Dudoit, and Sunduz Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3:Art. 4, 27 pp. (electronic), 2004.

Yuhong Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

Yuhong Yang. Comparing learning methods for classification. *Statistica Sinica*, 16(2): 635–657, 2006.

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