Why V=5 is enough in V-fold cross-validation

Sylvain Arlot, Matthieu Lerasle

Why $V = 5$ is enough in V-fold cross-validation

Sylvain Arlot[1,*] **and Matthieu Lerasle**[2,**]

[1] *CNRS; Sierra Project-Team*
Departement d'Informatique de l'Ecole Normale Superieure (DI/ENS)
(CNRS/ENS/INRIA UMR 8548)
23 avenue d'Italie - CS 81321
75214 PARIS Cedex 13 - France e-mail: [*] sylvain.arlot@ens.fr

[2] *Univ. Nice Sophia Antipolis LJAD CNRS UMR 7351*
06100 Nice France e-mail: [**] mlerasle@unice.fr

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 V-fold penalization is asymptotically optimal. 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 these variances depend on V like $1 + 4/(V-1)$ (at least in some particular cases), suggesting the performance increases much from $V = 2$ to $V = 5$ or 10, and then is almost constant. Overall, this explains the common advice to take $V = 5$ —at least in our setting and when the computational power is limited—, as confirmed by some simulation experiments.

Keywords and phrases: V-fold cross-validation, leave-one-out, leave-p-out, resampling penalties, density estimation, model selection, penalization.

1. Introduction

Cross-validation methods are widely used in statistics, for estimating the risk of a given statistical estimator [Sto74, All74, Gei75] 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. We refer to [AC10] for more references about cross-validation for model selection.

Then, a natural question arises: which cross-validation method should be used for minimizing the risk of the selected estimator? For instance, a popular family of cross-validation methods is V-fold cross-validation [Gei75, 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 [AC10]. In short, it is proved in least-squares regression that at first order, V-fold cross-validation is suboptimal for model selection if V stays bounded, because V-fold cross-validation is biased [Arl08]. When correcting for the bias [Bur89, Arl08], we recover asymptotic optimality whatever V, but without any theoretical result distinguishing among values of V in second order terms in the risk bounds [Arl08].

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 confirmed by some simulation experiments [Arl08, for instance]. But variance computations for unbiased V-fold methods have only been made in a very specific regression setting, and they are asymptotic [Bur89].

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 [HTF09, 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 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 1) is an oracle inequality with leading constant $1 + \varepsilon_n$ with $\varepsilon_n \to 0$ as $n \to \infty$ (for unbiased V-fold methods) that holds for any value of V. To the best of our knowledge, Theorem 1 is the first non asymptotic oracle inequality for V-fold methods enjoying such properties: the leading constant $1 + o(1)$ is new in density estimation, and the fact that it holds whatever the value of V had never been obtained in any framework. Theorem 1 relies on a new concentration inequality for the V-fold penalty (Proposition 2). The second main contribution of the paper (Theorem 2) 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 [Bur89, BG05, Cel08, Cel12, CR08], which is not sufficient for our purpose, as explained in Section 4; see also the remarks after Theorem 2. In our setting, we can explain theoretically why taking, say, $V > 10$ is not necessary for getting a performance close to the optimum, as confirmed by experiments on synthetic data in Section 6.

Notation For any integer $k \geq 1$, $[\![k]\!]$ denotes $\{1, \dots, k\}$, and for any $B \subset [\![n]\!]$, ξ_B denotes $\{\xi_i, i \in B\}$, $|B|$ its cardinality and $B^c = [\![n]\!] \setminus B$. 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 $o(\cdot)$ or $O(\cdot)$ are for the regime when the number n of observations tends to infinity.

All references to the supplementary material (e.g., to sections, equations or figures) are of the form S.x, and references to the appendix are of the form A.x.

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 $s \in L^\infty(\mu)$ so $s \in L^2(\mu)$. The goal is to estimate s from $\xi_{[\![n]\!]} = (\xi_1, \dots, \xi_n)$, that is, to build an estimator $\widehat{s} = \widehat{s}(\xi_{[\![n]\!]}) \in L^2(\mu)$ such that its loss $\|\widehat{s} - s\|^2$ is as small as possible, where for any $t \in L^2(\mu)$, $\|t\|^2 := \int_{\mathcal{X}} t^2 d\mu$.

Projection estimators are among the most classical estimators in this framework, see for example [DL93, Mas07]. 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

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

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

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

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 [Mas07], 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 $\tilde{s} := \widehat{s}_{\widehat{m}}$ has a quadratic loss as small as possible, that is, comparable to the oracle loss $\inf_{m \in \mathcal{M}_n} \|\widehat{s}_m - s\|^2$. More precisely, we aim at proving an oracle inequality of the form

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

with large probability. The procedure $\widehat{m}$ is called asymptotically optimal when R_n is negligible in front of the oracle loss and $C_n \to 1$.

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

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

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

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

Penalization is a popular way of designing a model selection criterion [BBM99, Mas07]:

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

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

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

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

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 $\mathbb{B}_m = \{ t \in S_m \text{ s.t. } \|t\| \leq 1 \}$ and $(\psi_\lambda)_{\lambda \in \Lambda_m}$ denotes an orthonormal basis of S_m in $L^2(\mu)$, then,

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

where the last equality follows from Eq. (A.29).

2.3. V-fold cross validation

A standard approach for model selection is cross-validation. We refer the reader to [AC10] 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

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

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

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

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

Let $V \leq n$ 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 $\mathbb{E}\left[\mathrm{crit}_{\mathrm{id}}(m)\right]$, Burman [Bur89] proposed the bias-corrected V-fold cross-validation criterion

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

This criterion is studied in [Mas07, Sec. 7.2.1, p.204–205] in the particular case where $V = n$ under the name cross-validation estimator (see in Lemma 1).

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 [Efr83] with the bootstrap and recently generalized to all resampling schemes [Arl09]. Let $W \sim \mathcal{W}$ be some random vector of $\mathbb{R}^n$ independent from $\xi_{[\![n]\!]}$ with $n^{-1}\sum_{i=1}^n W_i = 1$, 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

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

where $\widehat{s}_m^W \in \mathrm{argmin}_{t\in S_m}\left\{P_n^W\gamma(t)\right\}$, $\mathbb{E}_W[\cdot]$ 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 [Ler12], assuming W is exchangeable, i.e., 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 in [Arl08], inspired by V-fold cross-validation: given some partition $\mathcal{B} = \mathcal{B}_{[\![V]\!]}$ of $[\![n]\!]$, the weight vector W is defined by $W_i = (1-\mathrm{Card}(\mathcal{B}_J)/n)^{-1}\mathbf{1}_{i\notin\mathcal{B}_J}$ 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

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

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

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 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 $\mathcal{B}_{\mathrm{LOO}} = \{\{1\},\dots,\{n\}\}$, 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, as proved in [Ler12].

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

$$V = |\mathcal{B}| \text{ divides } n \quad \text{and} \quad \forall K \in [\![V]\!]\ ,\ |\mathcal{B}_K| = \frac{n}{V}\ . \qquad \textbf{(Reg)}$$

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

Lemma 1. *In least-squares density estimation, under assumption* (**Reg**),

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

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

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

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

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

and the leave-p-out penalty is defined by

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

Lemma 1 is proved in Section A.1.

Remark 1. Eq. (7) was first proved in [Arl08] in a general framework that includes least-squares density estimation, assuming only (**Reg**). Eq. (10) follows from Lemma 6.11 in [Ler12] since $\mathrm{pen}_{\mathrm{LPO}}$ belongs to the family of exchangeable resampling penalties, with weights $W_i := (1-p/n)^{-1}\mathbf{1}_{i\notin A}$ and A is randomly chosen uniformly over $\mathcal{E}_p$. It can also be deduced from [Cel12, Proposition 3.1], see Section A.1.

Remark 2. It is worth mentioning the cross-validation estimators given in [Mas07, Chapter 7]. First, the unbiased cross-validation criterion defined in [Mas07, Section 7.2.1] is exactly $\mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}_{\mathrm{LOO}})$. Then, the penalized estimator of [Mas07, Theorem 7.6] is the estimator selected by the penalty

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

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

As a conclusion of this section, in the least-squares density estimation framework and assuming only (**Reg**), Lemma 1 shows 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 $[\![n]\!]$, taking $x = C(V-1)$, the V-fold penalization criterion with is denoted by

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

A key quantity in our results is the bias $\mathbb{E}[\mathcal{C}_{(C,B)}(m) - \mathrm{crit}_{\mathrm{id}}(m)]$. From Lemma A.4 in Section A.2, we have

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}(m,\mathcal{B},V-1)\right] = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = 2\mathbb{E}\left[\left\|\widehat{s}_m - s_m\right\|^2\right] \quad , \tag{12}$$

so that for any $C > 0$,

$$\mathbb{E}\left[\mathcal{C}_{(C,\mathcal{B})}(m) - \mathrm{crit}_{\mathrm{id}}(m)\right] = 2(C-1)\mathbb{E}\left[\left\|\widehat{s}_m - s_m\right\|^2\right] \quad . \tag{13}$$

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

$$\widehat{m} = \widehat{m}(\mathcal{C}_{(C,\mathcal{B})}) = \operatorname*{argmin}_{m\in\mathcal{M}_n}\left\{\mathcal{C}_{(C,\mathcal{B})}(m)\right\} \quad , \tag{14}$$

for all values of V and $C > 1/2$. Additional results on hold-out (penalization) are given in Section 8.1 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 $V \in [\![n]\!]$, 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 $C \to 1$. 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 concentration results for V-fold penalties.

Proposition 2. *Let $\xi_{[\![n]\!]}$ be i.i.d. real-valued random variables with density $s \in L^\infty(\mu)$, $\mathcal{B}$ some partition of $[\![n]\!]$ 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 $\mathbb{B}_m = \{\, t \in S_m \text{ s.t. } \|t\| \leq 1 \,\}$, $\Psi_m = \sum_{\lambda\in\Lambda_m} \psi_\lambda^2 = \sup_{t\in\mathbb{B}_m} t^2$, $b_m := \left\|\sqrt{\Psi_m}\right\|_\infty$,*

$$\mathcal{D}_m := P\Psi_m - \|s_m\|^2 = n\mathbb{E}\left[\|s_m - \widehat{s}_m\|^2\right]$$

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

$$\rho_1(m, \epsilon, s, x, n) := \frac{\|s\|_\infty x}{\epsilon n} + \frac{(b_m^2 + \|s\|^2)x^2}{\epsilon^3 n^2} \quad .$$

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

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

$$\left|\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1) - 2\|s_m - \widehat{s}_m\|^2\right| \leq \epsilon\frac{\mathcal{D}_m}{n} + \kappa\rho_1(m, \epsilon, s, x, n) \quad . \tag{16}$$

Proposition 2 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 2 does not depend on V and cannot therefore help to discriminate between different values of this parameter.

3.2. Example: histogram models

Histograms on $\mathbb{R}$ are a classical example 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 $(\mathbf{1}_\lambda)_{\lambda \in m}$. More precisely, we illustrate our results with the following examples.

Example 3. [Regular histograms on $\mathcal{X} = \mathbb{R}$]

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

In Example 3, 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 2 shows $\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1)$ is asymptotically equivalent to $\mathrm{pen}_{\dim}(m) := 2d_m/n$ when $d_m \to \infty$. Penalties of the form of $\mathrm{pen}_{\dim}$ are classical and have been studied for instance by [BBM99].

Example 4. [k-rupture points on $\mathcal{X} = [0,1]$]

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

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

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

In Example 4, the function Ψ_m is constant on each interval $[x_{i-1}, x_i)$, equal to h_i, therefore,

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

3.3. Oracle inequality for V-fold procedures

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

- *for all* $m \in \mathcal{M}_n$,

$$b_m^2 := \sup_{t \in \mathbb{B}_m} \|t\|_\infty^2 \leq n \quad \text{where} \quad \mathbb{B}_m := \{ t \in S_m, \|t\| \leq 1 \} \ , \tag{H1}$$

- *the family of the projections of* s *is uniformly bounded*: for some $a > 0$,

$$\forall m \in \mathcal{M}_n \,, \qquad \|s_m\|_\infty \leq a \ , \tag{H2}$$

- *the collection of models is nested*

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

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

Theorem 1. *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 (**Reg**) 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 $C \in (1/2, 2]$, $\delta := 2(C-1)$ and, for any $x, \epsilon > 0$,*

$$\rho_2(\epsilon, s, x, n) := \frac{(\|s\|_\infty + A)x}{\epsilon n} + \left(1 + \frac{\|s\|^2}{n}\right)\frac{x^2}{\epsilon^3 n} \quad .$$

For every $m \in \mathcal{M}_n$, let $\widehat{s}_m$ be the estimator defined by Eq. (1), and $\tilde{s} = \widehat{s}_{\widehat{m}}$ where $\widehat{m} = \widehat{m}(\mathcal{C}_{(C,\mathcal{B})})$ is defined by Eq. (14). Then, an absolute constant κ exists such that, for any $x > 0$, with probability at least $1 - e^{-x}$, for any $\epsilon \in (0, 1]$,

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

where $x_n = x + \log|\mathcal{M}_n|$.

Theorem 1 is proved in Section A.3.

Taking $\epsilon > 0$ small enough in Eq. (17), Theorem 1 proves V-fold model selection procedures satisfy an oracle inequality with large probability. The remainder term can be bounded under the following classical assumption:

$$\exists a' > 0\,,\ \forall n \in \mathbb{N}^\star,\ |\mathcal{M}_n| \leq n^{a'} \quad . \tag{**A3**}$$

For instance, (**A3**) holds in Example 3 with $a' = 1$ and in Example 4 with $a' = k$. Under (**A3**), the remainder term in Eq. (17) is bounded by $L(\log n)^2/(\epsilon^3 n)$ for some $L > 0$.

The leading constant in the oracle inequality (17) is $(1+\delta_+)/(1-\delta_-) + o(1)$ by choosing $\epsilon = o(1)$, 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 $\delta = o(1)$. 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 no asymptotic optimality result can be obtained in general when $\delta \neq o(1)$, see for instance [Sha97]. Therefore, even if the leading constant $(1+\delta_+)/(1-\delta_-)$ is only an upper bound, we conjecture it cannot be taken as small as $1 + o(1)$ unless $\delta = o(1)$; such a result can be proved in our setting using similar arguments and assumptions as in [Arl08] for instance.

For bias-corrected V-fold cross-validation, that is, $C = 1$ hence $\delta = 0$, Theorem 1 shows a first-order optimal non-asymptotic oracle inequality, since the leading constant $(1+\epsilon)/(1-\epsilon)$ can be taken equal to $1 + o(1)$, and the remainder term is small enough under assumption (**A3**), for instance. Such a result valid with no upper bound on V had never been obtained before in any setting.

Regular V-fold cross-validation is also analyzed by Theorem 1, 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 [Arl08]. On the contrary, if $\mathcal{B} = \mathcal{B}_n$ has V_n blocs, with $V_n \to \infty$, Theorem 1 implies under assumption (**A3**) the asymptotic optimality of V_n-fold cross-validation as soon as the oracle loss is much larger than $(\log n)^2/n$.

The bound obtained in Theorem 1 can be integrated and we get

$$\frac{1-\delta_- -\epsilon}{1+\delta_+ +\epsilon}\mathbb{E}\left[\left\|\tilde{s}-s\right\|^2\right] \leq \mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\|\widehat{s}_m - s\right\|^2\right] + \kappa'\rho_2(\epsilon, s, \log(|\mathcal{M}_n|))$$

for some absolute constant $\kappa' > 0$.
Assuming $C > 1/2$ is necessary according to minimal penalty results proved in [Ler12]. Assuming $C \leq 2$ only simplifies the presentation; if $C > 2$, the same proof shows Theorem 1 holds with κ replaced by $C\kappa$.
An oracle inequality similar to Theorem 1 holds in a more general setting, as proved in a previous version of this paper [AL12]; 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,

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

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, see [AL12].

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 [AC10]. In the least-squares density estimation framework, oracle inequalities were proved by [vdLDK04] 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 1 considers the strongest possible oracle, that is, trained with n data. Optimal oracle inequalities were proved by [Cel12] for leave-p-out estimators with $p \ll n$, a case also treated in Theorem 1 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 $\delta = o(1)$ and we recover the result of [Cel12]. Concerning V-fold penalization, previous results were either valid for $V = n$ only ([Mas07, Theorem 7.6] and [Ler12] for least-squares density estimation, [Arl09] for regressogram estimators), or for V bounded when n tends to infinity (for regressogram estimators [Arl08]). In comparison, Theorem 1 provides a result valid for all V, except for the assumption that V divides n, which can be removed, see [AL12]. In particular, the loss bound in [Arl08] deteriorates when V grows, while it remains stable in our result. The latter corresponds to the typical behavior of the loss ratio $\|s - \tilde{s}\|^2 / \inf_{m\in\mathcal{M}_n}\|s - \widehat{s}_m\|^2$ of V-fold penalization as a function of V in simulation experiments, see Section 6 and [Arl08], for instance.

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

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

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 $\mathcal{C}_1$ is a better method than $\mathcal{C}_2$ in some setting, we would like to prove that

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

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

Previous works and their limits. 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)

$$\left\| \widehat{s}_{\widehat{m}(\mathcal{C})} - s \right\|^2 - \inf_{m \in \mathcal{M}_n} \left\{ \|\widehat{s}_m - s\|^2 \right\} \quad \text{or} \quad \mathfrak{R}_n(\mathcal{C}) := \frac{\left\| \widehat{s}_{\widehat{m}(\mathcal{C})} - s \right\|^2}{\inf_{m \in \mathcal{M}_n} \left\{ \|\widehat{s}_m - s\|^2 \right\}} \quad .$$

Alternatively, asymptotic results show that when n tends to infinity, $\mathfrak{R}_n(\mathcal{C}) \to 1$ (asymptotic optimality of $\mathcal{C}$) or that $\mathfrak{R}_n(\mathcal{C}_1) \sim \mathfrak{R}_n(\mathcal{C}_2)$ (asymptotic equivalence of $\mathcal{C}_1$ and $\mathcal{C}_2$). A review of such results can be found in [AC10, Section 6]. 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 $\mathfrak{R}_n(\mathcal{C}^{\mathrm{VF}}) \geq \kappa(V) > 1$ with a large probability [Arl08]. The major limitation of all these previous results is 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 [AC10, 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 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 the second-order terms of $\mathfrak{R}_n(\mathcal{C}_1)$ and $\mathfrak{R}_n(\mathcal{C}_2)$. Since we focus on second-order terms, from now on, we assume $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same bias, that is,

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

In least-squares density estimation, given Lemma 1, this means 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 $\mathbb{E}[\,\|\widehat{s}_m - s\|^2\,]$ 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. (24) below, but it is not sufficient to solve our problem, since it is known the best cross-validation estimator of the risk does not necessarily yield the best model selection procedure [BS92]. 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 $\mathrm{var}(\mathcal{C}(m))$ and can make it as large as desired. For model selection, what really matters is that

$$\mathrm{sign}\,(\mathcal{C}(m_1) - \mathcal{C}(m_2)) = \mathrm{sign}\left(\|\widehat{s}_{m_1} - s\|^2 - \|\widehat{s}_{m_2} - s\|^2\right) \tag{20}$$

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 $\|\widehat{s}_{m_1} - s\|^2 - \|\widehat{s}_{m_2} - s\|^2$ is small, so that $\|\widehat{s}_{\widehat{m}(\mathcal{C})} - s\|^2$ cannot be much larger than $\inf_{m \in \mathcal{M}_n}\{\,\|\widehat{s}_m - s\|^2\,\}$.

Heuristic. The heuristic we propose goes as follows. Assume for simplicity $m^\star = \mathrm{argmin}_{m \in \mathcal{M}_n} \mathbb{E}[\,\|\widehat{s}_m - s\|^2\,]$ is uniquely defined. For any $\mathcal{C}$, the smallest is $\mathbb{P}(m = \widehat{m}(\mathcal{C}))$ for all $m \neq m^\star$, the better should be the performance of $\widehat{s}_{\widehat{m}(\mathcal{C})}$. 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 $\Delta_{\mathcal{C}}(m, m') := \mathcal{C}(m) - \mathcal{C}(m')$, ξ some standard Gaussian random variable and $\overline{\Phi}(t) = \mathbb{P}(\xi > t)$ for all $t \in \mathbb{R}$. Then, for every $m \in \mathcal{M}_n$,

$$\begin{aligned} &\mathbb{P}(\widehat{m}(\mathcal{C}) = m) = \mathbb{P}(\forall m' \neq m\,,\ \Delta_{\mathcal{C}}(m, m') < 0) \\ &\asymp \min_{m' \neq m} \mathbb{P}(\Delta_{\mathcal{C}}(m, m') < 0) \end{aligned} \tag{21}$$

$$\approx \min_{m' \neq m} \mathbb{P}\left(\mathbb{E}\left[\Delta_{\mathcal{C}}(m, m') \right] + \xi \sqrt{\operatorname{var}\left(\Delta_{\mathcal{C}}(m, m') \right)} < 0 \right) \tag{22}$$

$$= \overline{\Phi}\left(\mathrm{SR}_{\mathcal{C}}(m) \right) \quad \text{where} \quad \mathrm{SR}_{\mathcal{C}}(m) := \max_{m' \neq m} \frac{\mathbb{E}\left[\Delta_{\mathcal{C}}(m, m') \right]}{\sqrt{\operatorname{var}\left(\Delta_{\mathcal{C}}(m, m') \right)}} \ .$$

So, if $\mathrm{SR}_{\mathcal{C}_1}(m) > \mathrm{SR}_{\mathcal{C}_2}(m)$ for all $m \neq m^\star$, $\mathcal{C}_1$ should be better than $\mathcal{C}_2$. Under assumption (**SameBias**), this leads to the following heuristic:

$$\forall m \neq m' \, , \, \operatorname{var}\left(\Delta_{\mathcal{C}_1}(m, m') \right) < \operatorname{var}\left(\Delta_{\mathcal{C}_2}(m, m') \right) \Rightarrow \mathcal{C}_1 \text{ better than } \mathcal{C}_2 \ . \tag{23}$$

Let us make some remarks.

- The quantity $\Delta_{\mathcal{C}}(m, m')$ appears in relative bounds [Cat07, Section 1.4] which can be used as a tool for model selection [Aud04].
- Approximation (21) is the strongest one. Clearly, inequality $\leq$ holds true. The equality case is for a very particular dependence setting, when the events $\left(\left\{ \Delta_{\mathcal{C}}(m, m') < 0 \right\} \right)_{m' \in \mathcal{M}}$ are nested. In general, the left-hand side is significantly smaller than the right-hand side; we claim they vary similarly as a function of $\mathcal{C}$.
- The gaussian approximation (22) for $\Delta_{\mathcal{C}}(m, m')$ does not hold exactly, but it seems reasonable to make it, at first order at least.

In the heuristic (23), 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 S.9 and S.25. Second, 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 $\overline{\Phi}\left(\mathrm{SR}_{\mathcal{C}}(m) \right)$ is much too small, changing it even by an order of magnitude will not affect the performance of $\widehat{m}(\mathcal{C})$ significantly; hence, all m such that, say, $\mathrm{SR}_{\mathcal{C}}(m) \gg (\log(n))^\alpha$ for all $\alpha > 0$, can also be discarded. Overall, pairs (m, m') that really matter in (23) 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 classical 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 2. *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 $m \in \{m_1, m_2\}$, S_m denotes the linear span of $(\psi_\lambda)_{\lambda \in \Lambda_m}$, $\mathbb{B}_m = \{t \in S_m \, / \, \|t\| \leq 1\}$ and $\Psi_m := \sup_{t \in \mathbb{B}_m} t^2$.*

For any $\Lambda, \Lambda' \in \{\Lambda_{m_1}, \Lambda_{m_2}\}$,

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

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

Then, for every $C > 0$,

$$\mathrm{Var}\left(\mathcal{C}_{(C,\mathcal{B})}(m_1)\right) = \frac{2}{n^2}\left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n}\right)\beta(\Lambda_{m_1}, \Lambda_{m_1}) \quad (24)$$
$$+ \frac{4}{n}\mathrm{Var}\left(\left(1 + \frac{2C-1}{n}\right)s_{m_1}(\xi_1) - \frac{2C-1}{2n}\Psi_{m_1}(\xi_1)\right)$$

and $\mathrm{Var}\left(\mathcal{C}_{(C,\mathcal{B})}(m_1) - \mathcal{C}_{(C,\mathcal{B})}(m_2)\right) = \frac{2}{n^2}\left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n}\right)\mathbf{B}(m_1, m_2)$ (25)
$$+ \frac{4}{n}\mathrm{Var}\left(\left(1 + \frac{2C-1}{n}\right)(s_{m_1} - s_{m_2})(\xi_1) - \frac{2C-1}{2n}(\Psi_{m_1} - \Psi_{m_2})(\xi_1)\right) .$$

Theorem 2 is proved in Section A.4.

In the unbiased case, that is, $C = 1$,

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

for some $a, b \geq 0$ depending on n, m_1, m_2 but not on V. Given the heuristic arguments of Section 4, this shows 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 $b \ll a$, the improvement from $V = 2$ to 5 or 10 is much larger than from $V = 10$ to $V = n$, which justifies the commonly used principle that taking $V = 5$ or $V = 10$ is large enough. Assuming in addition S_{m_1} and S_{m_2} are regular histogram models (Example 3 in Section 3.2) and d_{m_1} divides d_{m_2}, then, by Lemma S.10 in Section S.1,

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

When d_{m_2}/n is at least as large as $\|s_{m_1} - s_{m_2}\|^2$, we obtain that V drives the constant of the first order term in the variance through the multiplicative factor $1 + 4/(V-1)$ in front of a. Let $\mathcal{C}_{\mathrm{id}}(m) := P_n\gamma(\widehat{s}_m) + \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ be the criterion we could use if we knew the expectation of the ideal penalty. From Proposition S.8 in Section S.1,

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

which easily compares to formula (25) 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.

Regular V-fold cross-validation and the leave-p-out are also covered by Theorem 2, according to Lemma 1, respectively with $C = 1 + 1/(2(V-1))$ and with $V = n$ and $C = 1 + 1/(2(n/p - 1))$. The conclusion is similar: increasing V decreases the variance, and 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. (24). In the leave-p-out case, we recover formulas obtained in [Cel12, CR08], with a different grouping of the variance components; Eq. (24) clearly emphasizes the influence of the bias—through $(C-1)$—on the variance. For V-fold cross-validation, we believe Eq. (24) shows in a simpler way how the variance depends on V, compared to the result of [CR08] which was focusing on the difference between V-fold cross-validation and the leave-(n/V)-out; here the difference can be written

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

A major novelty in Eq. (24) is also to cover a larger set of criteria, such as corrected-V-fold cross-validation. Note that $\operatorname{var}(\mathcal{C}_{(C,\mathcal{B})}(m_1))$ is generally much larger than $\operatorname{var}(\mathcal{C}_{(C,\mathcal{B})}(m_1) - \mathcal{C}_{(C,\mathcal{B})}(m_2))$, 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. (24) and (25), 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 hold-out criteria and their increments $\Delta_{\mathcal{C}^{\mathrm{ho}}}(m_1, m_2)$ are also computed in Proposition S.13 in Section S.2.2.

Remark 5. The term $\mathbf{B}\left(m_1, m_2\right)$ does not depend on the choice of particular bases of S_{m_1} and S_{m_2}: as proved in Proposition S.9 in Section S.1,

$$\mathbf{B}\left(m_1, m_2\right) = \operatorname{Var}\left(\left(\widehat{s}_{m_1} - \widehat{s}_{m_2}\right)(\xi)\right) - \frac{n+1}{n}\operatorname{Var}\left(\left(s_{m_1} - s_{m_2}\right)(\xi)\right) \ , \qquad (26)$$

where ξ denotes a copy of ξ_1, independent of $\xi_{[\![n]\!]}$.

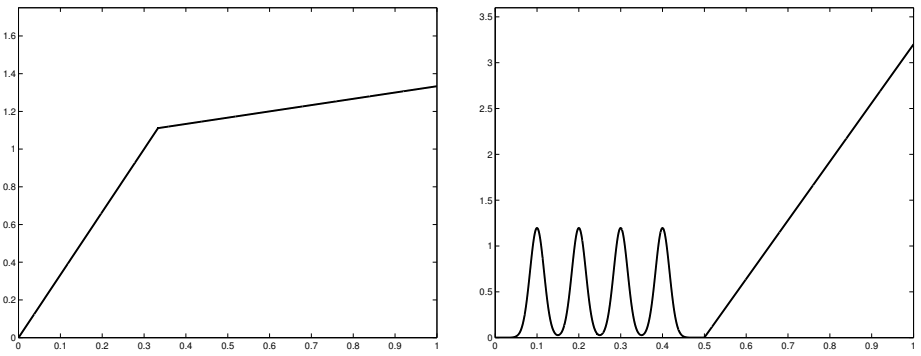

FIGURE 1. *The two densities considered. Left: setting L. Right: setting S.*

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 consider $\mathcal{X} = [0,1]$ 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: $s(x) = \frac{10x}{3}\mathbf{1}_{0\le x<1/3} + (1+\frac{x}{3})\mathbf{1}_{1\ge x\ge 1/3}$.
- Setting S: s is the mixture of the piecewise linear density $x \mapsto (8x-4)\mathbf{1}_{1\ge x\ge 1/2}$ (with weight 0.8) and four truncated gaussians with means $(k/10)_{k=1,\dots,4}$ 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: $\mathcal{M}_n = \{1,\dots,n\}$ 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: $\mathcal{M}_n = \bigcup_{k\in\{1,\dots,\widetilde{n}\}} \{k\}\times\{0,\dots,\lfloor\log_2(k)\rfloor\}\times\{0,\dots,\lfloor\log_2(\widetilde{n}-k)\rfloor\}$ 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, where the corresponding oracle models have been plotted in setting S. 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

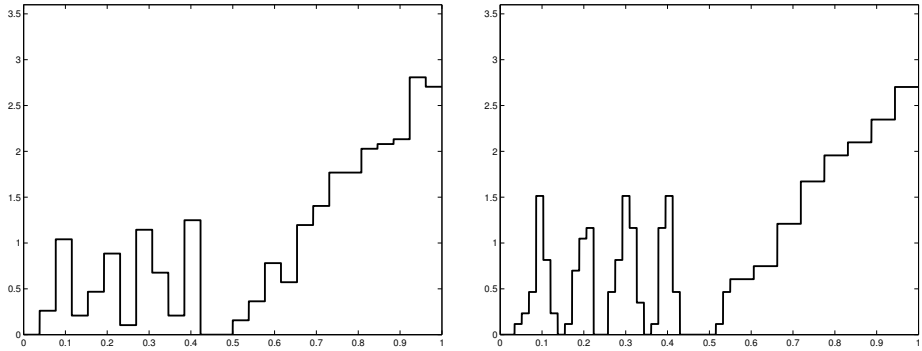

FIGURE 2. *Oracle model for one sample of size $n = 500$, in setting S. Left: Regu. Right: Dya2.*

TABLE 1

Comparison of Regu and Dya2: quadratic risks $\mathbb{E}[\,\|s - \widehat{s}_{\widehat{m}}\|^2\,]$ 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 $\mathbb{E}[\,\|s - \widehat{s}_{\widehat{m}}\|^2\,]$ among all the data-driven procedures we considered in our experiments (see Section 6.2). "Oracle" means that $\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \|s - \widehat{s}_m\|^2$ is the oracle model for each sample.

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

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:

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

Procedure	L–Dya2	S–Dya2
$\mathrm{pen}_{\mathrm{dim}}$	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
$\mathbb{E}[\,\mathrm{pen}_{\mathrm{id}}\,]$	6.52 ± 0.05	2.07 ± 0.01

- $\mathrm{pen}_{\mathrm{dim}}$ [BBM99]: 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 [Arl08], we consider all penalties above multiplied by a factor chosen among $\{1, 1.25, 1.5, 2\}$. Complete results can be found in Table S.3 in Section S.5.

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

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

which 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 [Ler12], so V-fold methods cannot improve significantly its performance, with a larger computational cost. Complete results (including Regu, with $n = 100$ and $n = 500$) are given in Tables S.3 and S.4 in Section S.5, showing 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 confirm our interpretation of Theorem 2 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 $V \in \{5, 10\}$. 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 [Arl08].

Other comments Table 2 confirms in the least-squares density estimation framework several facts previouly observed in least-squares regression [Arl08]:

- $\mathrm{pen}_{\mathrm{dim}}$ 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 S.3), but V-fold penalization then performs as well.
- V-fold penalization and $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$ perform much better when multiplying the penalty by some $C > 1$. The best overpenalization factor is $C = 2$ for L-Dya2 and $C = 1.5$ for S-Dya2, see Table S.3. Such a phenomenon can also be observed in regression [Arl08] and can explain in part the behaviour of V-fold cross-validation.
- Once V-fold penalties are multiplied by a well-chosen C, they perform significantly better than V-fold cross-validation, except for the fact that 2-fold cross-validation coincides with 2-fold penalization multiplied by 1.5 as shown by Lemma 1. Nevertheless, making a bad choice for C (which depends on the setting) can lead to worse performance with V-fold penalization, especially when $V = 2$, see Table S.3.

In other settings considered in a preliminary phase of our experiments, 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 $sd\mu$, 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 3 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 2: 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 $\mathcal{C}_{\mathrm{id}}$. On Figure 3, we can remark that for $m > m^\star$,

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

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 2, 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 the same behaviour holds when $n = 500$ with very close values for K_3 and K_4 (see Figure S.15), 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 S.10 and S.20). The fact that K_4 is close to 4 in both settings confirms the term $1 + 4/(V-1)$ appearing Theorem 2 indeed drives how $\mathrm{var}(\Delta_{\mathcal{C}_V}(m, m^\star))$ depends on V.

Figures 4 and 5 respectively show $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ and its proxy $\overline{\Phi}(\mathrm{SR}_{\mathcal{C}}(m))$ 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, remark both quantities behave similarly as a function of m and $\mathcal{C}$—see also Figure S.7—confirming empirically the heuristic of Section 4. The decrease of the variance observed on Figure 3 when V increases here translates into a better concentration of the distribution of $\widehat{m}(\mathcal{C}_V)$ around $m^\star$, which explains the performance improvement observed in Section 6.3. Figures 4–5 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 S.5, 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

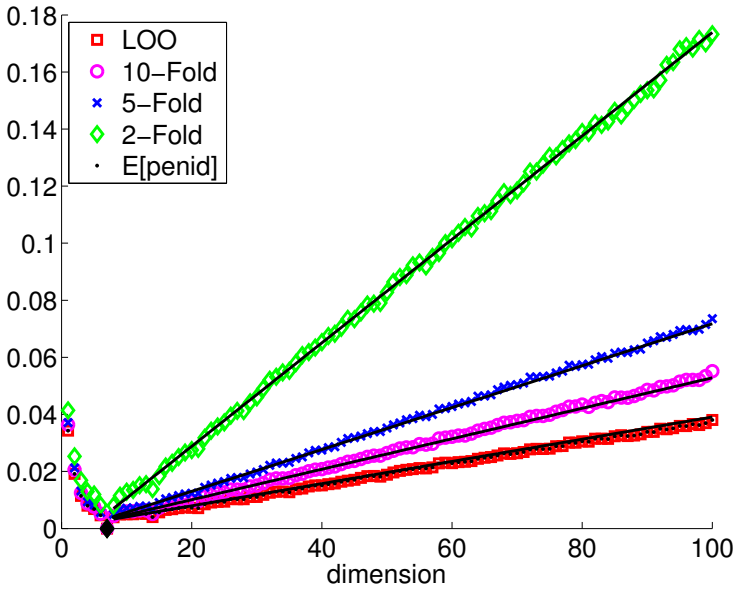

FIGURE 3. *Illustration of the variance heuristic:* $\mathrm{var}(\Delta_{\mathcal{C}}(m, m^\star))$ *as a function of* m *for five different* $\mathcal{C}$*. Setting S-Regu,* $n = 100$*. The black diamond shows* $m^\star = 7$*. The black lines show the linear approximation* $n^{-2}[29(1 + \frac{0.81}{V-1}) + 3.7(1 + \frac{3.8}{V-1})(m - m^\star)]$ *for* $m > m^\star$*.*

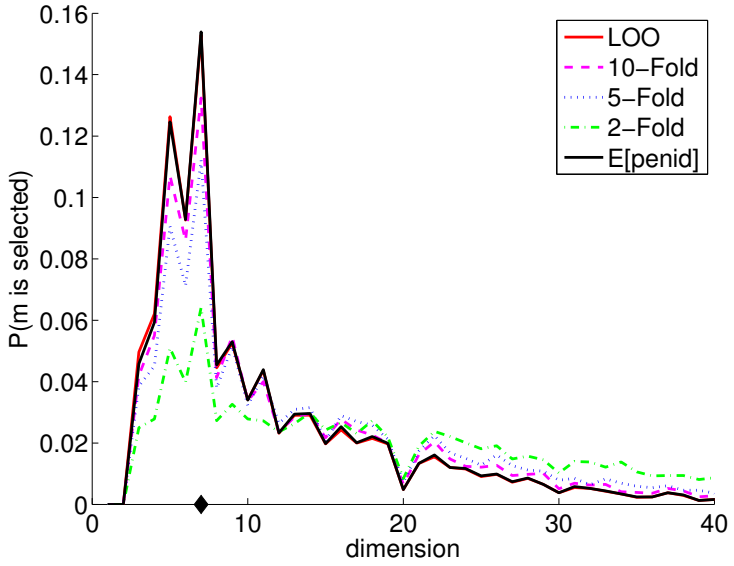

FIGURE 4. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ *as a function of* m *for five different* $\mathcal{C}$*. Setting S-Regu,* $n = 100$*. The black diamond shows* $m^\star = 7$*.*

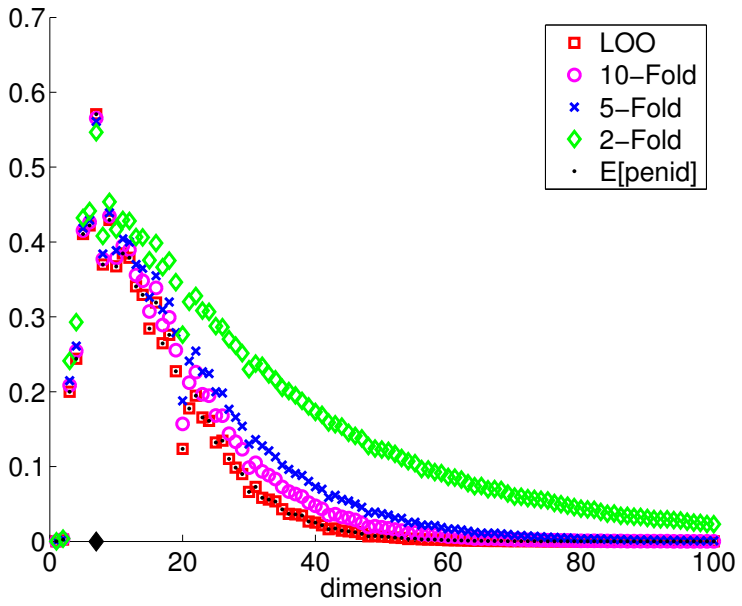

FIGURE 5. *Illustration of the variance heuristic:* $\overline{\Phi}(\mathrm{SR}_{\mathcal{C}}(m))$ *as a function of* m *for five different* $\mathcal{C}$. *Setting S-Regu,* $n = 100$. *The black diamond shows* $m^\star = 7$.

framework, two approaches are possible: a naive one—valid for all frameworks—and a faster one—specific to least-squares density estimation. For clarifying the exposition, we assume in this section (**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 $j = 1, \dots, V$ 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 S.3.1, 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 $\{1, ..., n\}$ satisfying (**Reg**), $\xi_1, \dots, \xi_n \in \mathcal{X}$ and $(\psi_\lambda)_{\lambda \in \Lambda_m}$ a finite orthonormal family of $L^2(\mu)$.

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

Output:

Empirical risk: $P_n \gamma(\widehat{s}_m) = -\mathcal{S}/V^2$

V-fold cross-validation criterion: $\mathrm{crit}_{\mathrm{VFCV}}(m) = \frac{\mathcal{T}}{V(V-1)} - \frac{\mathcal{S}-\mathcal{T}}{(V-1)^2}$

V-fold penalty: $\mathrm{pen}_{\mathrm{VF}}(m) = (\mathrm{crit}_{\mathrm{VFCV}}(m) - P_n\gamma(\widehat{s}_m))\frac{V-1/2}{V-1}$.

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 3. *Algorithm 1 is correct and has a computational complexity of order* $(n + V^2)\,\mathrm{Card}(\Lambda_m)$.

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

Proposition 3 is proved in Section S.3.2. Note that closed-form formulas are available for the leave-p-out criterion in least-squares density estimation [Cel12], allowing to compute it with a complexity of order $n\,\mathrm{Card}(\Lambda_m)$ 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, we state an additional result and we discuss the model selection literature in least-squares density estimation.

8.1. Hold-out penalization

Our analysis of V-fold procedures for model selection can be extended to hold-out methods, that is, when data are split only once. Similarly to the definition of the hold-out criterion in Eq. (4), the hold-out penalty is defined as

$$\forall x \geq 0\,, \quad \mathrm{pen}_{\mathrm{HO}}(m, T, x) := 2x\left(P_n^{(T)} - P_n\right)\left(\widehat{s}_m^{(T)} - \widehat{s}_m\right) \quad , \tag{27}$$

that is, the hold-out estimator of the expectation of the ideal penalty, written as $\mathbb{E}[2(P_n - P)(\widehat{s}_m - s_m)]$, 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. (27)— 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 $\tau = |T|/n$, it comes from Lemma S.11 that

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{HO}}(m, T, x)\right] = x\frac{1-\tau}{\tau}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \quad .$$

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

Remark 6. Since $P_n = \tau P_n^{(T)} + (1-\tau) P_n^{(T^c)}$, by linearity of the estimator $\widehat{s}_m$,

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

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:

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

Theorem 3. *Let $\xi_{[\![n]\!]}$ be i.i.d real-valued random variables with common density $s \in L^\infty(\mu)$, $T \subset [\![n]\!]$ with $\tau = |T|/n \in (0,1)$ 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 $\tilde{s}_{HO} = \widehat{s}_{\widehat{m}_{\mathrm{HO}}}$ where*

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

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

$$\frac{1 - \delta_- - \epsilon}{1 + \delta_+ + \epsilon} \left\| \tilde{s}_{HO} - s \right\|^2 \leq \inf_{m \in \mathcal{M}_n} \left\| \widehat{s}_m - s \right\|^2 + \kappa \left(\frac{A x_n}{\epsilon n} + \frac{\tau^2 + (1-\tau)^2}{\tau(1-\tau)} \frac{x_n^2}{\epsilon^3 n} \right) \ . \quad (28)$$

Theorem 3 is proved in Section S.2.1.

Theorem 3 extends Theorem 1 to hold-out penalties, under similar assumptions. As in Theorem 1, δ quantifies the bias of the hold-out penalized criterion, and plays the same role in the leading constant of the oracle inequality (28).

The main difference between Theorems 1 and 3 lies in the remainder term. For making proper comparisons, let V be some divisor of n and $T \subset [\![n]\!]$ such that $|T| = n - n/V$. Then, the remainder term in Eq. (28) is larger than the one of Eq. (17) in Theorem 1 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.

Similarly to Theorem 2, 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. See Proposition S.13 in Section S.2.2 for details.

8.2. Other model selection procedures for 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, [Mas07, Section 7.2] prove an oracle inequality for some penalization procedures, but they 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 in [Rig06] and some T-estimators in [Bir13]. Note the models considered in [Bir13] are more general than ours, but the corresponding estimators are not computable in practice, and the oracle inequality in [Bir13] also has a suboptimal constant C_n. Some aggregation procedures also satisfy oracle inequalities, as proved for instance in [RT07, BTWB10]. Overall, under our assumptions, none of these results imply strictly better bounds than ours.

Let us finally mention [BR06] proposed 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 S.3 where we presented the performances of $\mathrm{pen}_{\mathrm{dim}}$ with different over-penalizing constants.

8.3. Conclusion on the choice of V

Overall, choosing V requires a trade-off between:

- Computational complexity, usually proportional to V, slightly different in the least-squares density estimation setting since it can be reduced to $(n + V^2)\,\mathrm{Card}(\Lambda_m)$ or even to $n + V^2\,\mathrm{Card}(\Lambda_m)$, see Section 7.
- Statistical performance in terms of risk, which is better when the bias and the variance are small. The bias decreases as V increases for V-fold cross-validation, but it can be removed completely or fixed to any desired value by using V-fold penalization instead, see Lemma 1. The variance decreases as V increases, but it almost reaches its minimal value by taking, say, $V = 5$ or $V = 10$, as shown by theoretical and empirical arguments in Sections 5 and 6.

The most common advices for choosing V in the literature [for instance HTF09, Section 7.10.1] are between $V = 5$ and $V = 10$. This article provides clear evidence why taking V larger does not reduce the variance significantly. Concerning the bias, Lemma 1 shows 5-fold (resp. 10-fold) cross-validation corresponds to overpenalization by a factor $1 + 1/8$ (resp. $1 + 1/18$), which is likely to be a good amount in many cases; in our simulation experiments, the best overpenalization factor is even larger, see also [Arl08].

Note however our results are only valid for some least-squares algorithms, and it is reported in the literature [AC10] that V-fold cross-validation behaves differently as a function of V in other settings.

Finally, we would like to address the question of choosing between V-fold cross-validation and penalization. The answer is rather simple—at least in least-squares density estimation—since Lemma 1 shows V-fold cross-validation is a

particular instance of V-fold penalization, with $C = 1+1/(2(V-1))$. So, if one wants to overpenalize by a factor $1+1/(2(V-1))$, V-fold cross-validation is definitely the good choice. Otherwise, the best choice would be V-fold penalization with another value for C, depending on how much one wants to overpenalize.

Acknowledgments 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. The authors also 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), and the first author acknowledges the support of the GARGANTUA project funded by the Mastodons program of CNRS.

References

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

[AL12] Sylvain Arlot and Matthieu Lerasle. V-fold cross-validation and V-fold penalization in least-squares density estimation, October 2012. arXiv:1210.5830v1.

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

[Arl08] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

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

[Aud04] Jean-Yves Audibert. A better variance control for pac-bayesian classification. Technical Report 905b, Laboratoire de Probabilités et Modèles Aléatoires, 2004. Available online at http://imagine.enpc.fr/publications/papers/04PMA-905Bis.pdf.

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

[BG05] 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 Anniv. Ser.*, pages 75–95. Springer, New York, 2005.

[Bir13] Lucien Birgé. Model selection for density estimation with $\mathbb{L}_2$-loss. *Probability Theory and Related Fields*, pages 1–42, 2013.

[BR06] Lucien Birgé and Yves Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Statist.*, 10, 2006.

[BS92] Leo Breiman and Philip Spector. Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, 60(3):291–319, 1992.

[BTWB10] Florentina Bunea, Alexandre B. Tsybakov, Marten H. Wegkamp, and Adrian Barbu. Spades and mixture models. *Ann. Statist.*, 38(4):2525–2558, 2010.

[Bur89] 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.

[Cat07] Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Inst. Math. Statist., 2007.

[Cel08] Alain Celisse. *Model Selection Via Cross-Validation in Density Estimation, Regression and Change-Points Detection*. PhD thesis, University Paris-Sud 11, December 2008. http://tel.archives-ouvertes.fr/tel-00346320/.

[Cel12] Alain Celisse. Optimal cross-validation in density estimation. Technical report, arXiv, 2012. arXiv:0811.0802v3.

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

[DL93] 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.

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

[Gei75] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

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

[HTF09] 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.

[Ler11] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877, 2011.

[Ler12] Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(3):884–908, 2012.

[Mas07] 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, With a foreword by Jean

Picard.

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

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

[Sha97] Jun Shao. An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2):221–264, 1997. With comments and a rejoinder by the author.

[Sto74] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974.

[vdLDK04] Mark J. van der Laan, Sandrine Dudoit, and Sunduz Keles. Asymptotic optimality of likelihood-based cross-validation. *Stat. Appl. Genet. Mol. Biol.*, 3:Art. 4, 27 pp. (electronic), 2004.

Appendix A: Proofs

Before proving the main results stated in the paper, let us recall a simple result that we use repeatedly in the proofs: 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

$$\sup_{\sum_{\lambda\in\Lambda_m} a_\lambda^2 \leq 1} \left(\sum_{\lambda\in\Lambda_m} a_\lambda b_\lambda \right)^2 = \sum_{\lambda\in\Lambda_m} b_\lambda^2 \ . \tag{A.29}$$

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 the converse inequality holds true.

A.1. Proof of Lemma 1

Let us first recall here the proof of Eq. (7) (coming from [Arl08]) for the sake of completeness. By (**Reg**), $P_n - P_n^{(\mathcal{B}_K^c)} = V^{-1}(P_n^{(\mathcal{B}_K)} - P_n^{(\mathcal{B}_K^c)})$ and $P_n^{(\mathcal{B}_K)} - P_n = (V-1)V^{-1}(P_n^{(\mathcal{B}_K)} - P_n^{(\mathcal{B}_K^c)})$, so that

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

Eq. (8) and (9) follow simultaneously from Eq. (A.33) below. Let $\mathcal{E}$ be a set of subsets of $[\![n]\!]$ such that

$$\forall A \in \mathcal{E}\,, \quad |A| = p \quad \text{and} \quad \frac{1}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} P_n^{(A^c)} = P_n \ . \tag{A.30}$$

Let us consider the associated penalty

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

and the associated cross-validation criterion

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

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. (A.30) holds true with $p = n/V$ under assumption (**Reg**). When $\mathcal{E} = \mathcal{E}_p := \{\, A \subset [\![n]\!] \text{ s.t. } |A| = p \,\}$, Eq. (A.30) 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. (A.30), we get

$$\begin{aligned}
\mathrm{pen}_{\mathcal{E}}(m, C) &= \frac{2C}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} (P_n^{(A^c)} - P_n) \left(\widehat{s}_m^{(A^c)}\right) \\
&= \frac{2C}{|\mathcal{E}|} \sum_{A \in \mathcal{E}} \sum_{\lambda \in \Lambda_m} \left[\left(P_n^{(A^c)}(\psi_\lambda) - P_n(\psi_\lambda) \right) P_n^{(A^c)}(\psi_\lambda) \right] \\
&= \frac{2C}{|\mathcal{E}|} \sum_{\lambda \in \Lambda_m} \left[\sum_{A \in \mathcal{E}} \left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - P_n(\psi_\lambda) \sum_{A \in \mathcal{E}} P_n^{(A^c)}(\psi_\lambda) \right] \\
&= \frac{2C}{|\mathcal{E}|} \sum_{\lambda \in \Lambda_m} \sum_{A \in \mathcal{E}} \left[\left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - \left(P_n(\psi_\lambda) \right)^2 \right] \ .
\end{aligned} \tag{A.31}$$

On the other hand, using that $P_n^{(A)} = \frac{n}{p} P_n - \frac{n-p}{p} P_n^{(A^c)}$ by (A.30),

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

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

$$= \left(\frac{2n}{p} - 1 \right) \frac{1}{|\mathcal{E}|} \sum_{\lambda \in \Lambda_m} \sum_{A \in \mathcal{E}} \left[\left(P_n^{(A^c)}(\psi_\lambda) \right)^2 - \left(P_n(\psi_\lambda) \right)^2 \right] \quad , \tag{A.32}$$

where we used again Eq. (A.30). Comparing Eq. (A.31) and (A.32) gives

$$\mathrm{crit}_{\mathcal{E}}(m) = P_n \gamma \left(\widehat{s}_m \right) + \mathrm{pen}_{\mathcal{E}} \left(m, \frac{n}{p} - \frac{1}{2} \right) \tag{A.33}$$

which implies Eq. (8) and (9). Eq. (10) follows by [Ler12]. □

We now prove the statements made in Remarks 1–2 below Lemma 1. Eq. (10) can also be deduced from [Cel12, Proposition 2.1], which proves that

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

Elementary algebraic computations show then that

$$\mathrm{crit}_{\mathrm{LPO}}(m, p) - P_n \gamma(\widehat{s}_m) = \frac{2n-p}{n^2(n-p)} \sum_{\lambda \in \Lambda_m} \left(\sum_{i=1}^{n} \psi_\lambda(\xi_i)^2 - \frac{1}{n-1} \sum_{1 \le i \ne j \le n} \psi_\lambda(\xi_i) \psi_\lambda(\xi_j) \right) \tag{A.34}$$

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

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

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

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

For Remark 2, note first the CV estimator in [Mas07, Sec. 7.2.1, p.204–205] is defined as the minimizer of

$$\left\| \widehat{s}_m \right\|^2 - \frac{2}{n(n-1)} \sum_{1 \le i \ne j \le n} \sum_{\lambda \in \Lambda_m} \psi_\lambda(\xi_i) \psi_\lambda(\xi_j)$$

$$= P_n\gamma\left(\widehat{s}_m\right) + \frac{2}{n^2}\sum_{\lambda\in\Lambda_m}\left(\sum_{i=1}^{n}\psi_\lambda(\xi_i)^2 - \frac{1}{n-1}\sum_{1\le i\ne j\le n}\psi_\lambda(\xi_i)\psi_\lambda(\xi_j)\right) \ . \tag{A.35}$$

On the other hand, from Eq. (A.34) and (9) with $p = 1$, we have

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

Hence, from Eq. (A.35), the CV estimator is the minimizer of $\mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}_{\mathrm{LOO}})$. [Mas07, Theorem 7.6] studies the minimizers of the criterion

$$P_n\gamma(\widehat{s}_m) + \frac{C}{n^2}\sum_{i=1}^{n}\sum_{\lambda\in\Lambda_m}\psi_\lambda(\xi_i)^2 \ , \tag{A.36}$$

where $C = (1+\epsilon)^6$ for any $\epsilon > 0$. Let $\alpha = C/n$, so that $\alpha = (C-\alpha)/(n-1)$. Then, the criterion (A.36) is equal to

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

A.2. Proof of Proposition 2

Note the two formulas given for Ψ_m in the statement of Proposition 2 coincide by Eq. (A.29). The proof is decomposed into 3 lemmas.

Lemma A.4. *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*

$$U(m) := \frac{1}{n^2}\sum_{1\le k\ne k'\le V}\ \sum_{\substack{i\in\mathcal{B}_k\\ j\in\mathcal{B}_{k'}}}\ \sum_{\lambda\in\Lambda_m}\left(\psi_\lambda(\xi_i) - P\psi_\lambda\right)\left(\psi_\lambda(\xi_j) - P\psi_\lambda\right) \ . \tag{A.37}$$

Then, the V-fold penalty is equal to

$$\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) = \frac{2C}{V-1}\left\|s_m - \widehat{s}_m\right\|^2 - \frac{2VC}{(V-1)^2}U(m) \tag{A.38}$$

and $$\mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}\left(m, \mathcal{B}, \frac{V-1}{2}\right)\right] = \mathbb{E}\left[\left\|s_m - \widehat{s}_m\right\|^2\right] = \frac{\mathcal{D}_m}{2n} \quad . \tag{A.39}$$

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

$$\begin{aligned}
\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) &= C\mathbb{E}_W\left[(P_n - P_n^W)(\gamma(\widehat{s}_m^W))\right] \\
&= 2C\mathbb{E}_W\left[(P_n^W - P_n)(\widehat{s}_m^W)\right] \\
&= 2C\mathbb{E}_W\left[(P_n^W - P_n)(\widehat{s}_m^W - \widehat{s}_m)\right] \qquad \text{by (\textbf{Reg})} \\
&= 2C\sum_{\lambda \in \Lambda_m} \mathbb{E}_W\left[\left[(P_n^W - P_n)(\psi_\lambda)\right]^2\right] \\
&= 2C\sum_{\lambda \in \Lambda_m} \mathbb{E}_W\left[\left[(P_n^W - P_n)(\psi_\lambda - P\psi_\lambda)\right]^2\right] \\
&= \frac{2C}{n^2}\sum_{\lambda \in \Lambda_m}\sum_{1 \le i,j \le n} E_{i,j}^{(\mathrm{VF})}(\psi_\lambda(\xi_i) - P\psi_\lambda)(\psi_\lambda(\xi_j) - P\psi_\lambda)
\end{aligned} \tag{A.40}$$

where $E_{i,j}^{(\mathrm{VF})} := \mathbb{E}\left[(W_i - 1)(W_j - 1)\right]$. Since $\mathbb{E}[W_i] = 1$ by (**Reg**) and $W_i W_j = (V/(V-1))^2 \mathbf{1}_{J \notin \{J_0, J_1\}}$ if $i \in \mathcal{B}_{J_0}$ and $j \in \mathcal{B}_{J_1}$, 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,

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

and Eq. (A.38) follows by Eq. (3). Eq. (A.39) directly follows from Eq. (A.38). □

Lemma A.5. *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 $\mathbb{B}_m = \{t \in S_m \text{ s.t. } \|t\| \le 1\}$, $\mathcal{D}_m = \sum_{\lambda \in \Lambda_m} P\left(\psi_\lambda^2\right) - \|s_m\|^2$ 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 $1 - 2e^{-x}$, we have for every $\epsilon > 0$,*

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

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

$$\mathrm{Var}\left(t(\xi_1)\right) \le \int_{\mathbb{R}} t^2 s d\mu \le \|s\|_\infty \|t\|^2 \le \|s\|_\infty \quad , \tag{A.41}$$

which gives the conclusion thanks to Proposition S.14. □

Lemma A.6. *Assume that $\xi_{[\![n]\!]}$ is a sequence of i.i.d. real-valued random variables with common density $s \in L^\infty(\mu)$ and $\mathcal{B}_{[\![V]\!]}$ is some partition of $[\![n]\!]$ satisfying* (**Reg**)*. Let S_m denote a separable subspace of $L^2(\mu)$ with orthonormal basis $(\psi_\lambda)_{\lambda\in\Lambda_m}$ such that $b_m := \sup_{t\in S_m, \|t\|\leq 1} \|t\|_\infty < +\infty$. Let $U(m)$ be the U-statistics defined by Eq.* (A.37)*. Using the notations of Lemma A.5, an absolute constant κ exists such that, with probability larger than $1 - 6e^{-x}$,*

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

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

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

Proof. For any $x, y \in \mathbb{R}$ and $i, j \in [\![n]\!]$, let us define

$$g_{i,j}(x,y) = \begin{cases} 0 & \text{if} \quad \exists k \in [\![V]\!] \text{ s.t. } \{i,j\} \subset \mathcal{B}_k \\ \sum_{\lambda\in\Lambda_m}(\psi_\lambda(x) - P\psi_\lambda)(\psi_\lambda(y) - P\psi_\lambda) & \text{otherwise.} \end{cases}$$

Then,

$$U(m) = \frac{2}{n^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} g_{i,j}(\xi_i, \xi_j) .$$

From Theorem 3.4 in [HRB03], an absolute constant κ exists such that, for any $x > 0$ and $\epsilon \in (0,1]$,

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

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

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

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

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

$$\begin{aligned}
\overline{A}^2 &= \sum_{k=2}^{V}\sum_{k'=1}^{k-1}\sum_{i\in\mathcal{B}_k, j\in\mathcal{B}_{k'}}\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_m}\mathbb{E}\left[(\psi_\lambda(\xi_1)-P\psi_\lambda)(\psi_{\lambda'}(\xi_1)-P\psi_{\lambda'})\right]^2 \\
&= \frac{n^2(V-1)}{2V}\sum_{\lambda\in\Lambda_m}\left(\sup_{\sum_{\lambda'\in\Lambda_m}a_{\lambda'}^2\le 1}\mathbb{E}\left[(\psi_\lambda(\xi_1)-P\psi_\lambda)\sum_{\lambda'\in\Lambda_m}a_{\lambda'}(\psi_{\lambda'}(\xi_1)-P\psi_{\lambda'})\right]\right)^2 \\
&= \frac{n^2(V-1)}{2V}\sum_{\lambda\in\Lambda_m}\left(\sup_{t\in\mathbb{B}_m}\mathbb{E}\left[(\psi_\lambda(\xi_1)-P\psi_\lambda)(t(\xi_1)-Pt)\right]\right)^2 \\
&\le \frac{n^2(V-1)}{2V}\mathcal{D}_m\sup_{t\in\mathbb{B}_m}\mathbb{E}\left[\left(t(\xi_1)-P(t)\right)^2\right] \\
&\le \frac{n^2(V-1)}{2V}\left\|s\right\|_\infty\mathcal{D}_m \qquad \text{by Eq. (A.41)} \quad .
\end{aligned}$$

Let now $a_1,\dots,a_n,b_1,\dots,b_n$ be functions in $L^2(\mu)$ such that

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

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

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=2}^n\sum_{j=1}^{i-1}a_i(\xi_i)b_j(\xi_j)g_{i,j}(\xi_i,\xi_j)\right] \\
&= \sum_{k=2}^{V}\sum_{k'=1}^{V-1}\sum_{i\in\mathcal{B}_k,j\in\mathcal{B}_{k'}}\sum_{\lambda\in\Lambda_m}\mathbb{E}\left[a_i(\xi_i)(\psi_\lambda(\xi_i)-P\psi_\lambda)\right]\mathbb{E}\left[b_j(\xi_j)(\psi_\lambda(\xi_j)-P\psi_\lambda)\right] \\
&\le \sum_{k=2}^{V}\sum_{k'=1}^{V-1}\sum_{i\in\mathcal{B}_k,j\in\mathcal{B}_{k'}}\sum_{\lambda\in\Lambda_m}\frac{\mathbb{E}\left[a_i(\xi_i)(\psi_\lambda(\xi_i)-P\psi_\lambda)\right]^2+\mathbb{E}\left[b_j(\xi_j)(\psi_\lambda(\xi_j)-P\psi_\lambda)\right]^2}{2} \ .
\end{aligned} \tag{A.42}$$

Now, we have, for every $i\in[\![n]\!]$, using Eq. (A.29), Cauchy-Schwarz inequality and the fact that for every $t\in L^2(\mu)$, $\mathrm{Var}\left(t(\xi_1)\right)\le\left\|s\right\|_\infty\left\|t\right\|^2$,

$$\begin{aligned}
\sum_{\lambda\in\Lambda_m}\mathbb{E}\left[a_i(\xi_i)(\psi_\lambda(\xi_i)-P\psi_\lambda)\right]^2 &= \sup_{\sum_{\lambda\in\Lambda_m}t_\lambda^2\le 1}\left(\mathbb{E}\left[a_i(\xi_i)\sum_{\lambda\in\Lambda_m}t_\lambda\psi_\lambda(\xi_i)-P(t_\lambda\psi_\lambda)\right]\right)^2 \\
&= \sup_{t\in\mathbb{B}_m}\left(\mathbb{E}\left[a_i(\xi_i)(t(\xi_i)-Pt\right]\right)^2 \\
&\le \mathbb{E}\left[a_i(\xi_i)^2\right]\sup_{t\in\mathbb{B}_m}\mathrm{Var}\left(t(\xi_1)\right)\le\mathbb{E}\left[a_i(\xi_i)^2\right]\left\|s\right\|_\infty \ .
\end{aligned}$$

Plugging this bound in (A.42) yields

$$\mathbb{E}\left[\sum_{i=2}^{n}\sum_{j=1}^{i-1} a_i(\xi_i)b_j(\xi_j)g_{i,j}(\xi_i,\xi_j)\right] \leq \|s\|_\infty \sum_{k=2}^{V}\sum_{k'=1}^{V-1}\sum_{i\in\mathcal{B}_k, j\in\mathcal{B}_{k'}} \frac{\mathbb{E}\left[a_i(\xi_i)^2\right]+\mathbb{E}\left[b_j(\xi_j)^2\right]}{2}$$
$$\leq n\|s\|_\infty \ .$$

Hence,

$$\overline{B} \leq n\|s\|_\infty \ . \tag{A.43}$$

Now, for any $x \in \mathbb{R}$, using (**Reg**),

$$\sum_{i=2}^{n}\mathbb{E}\left[g_{i,1}(\xi_i,x)^2\right] = \frac{n(V-1)}{V}\mathbb{E}\left[\left(\sum_{\lambda\in\Lambda_m}(\psi_\lambda(x)-P\psi_\lambda)(\psi_\lambda(\xi_1)-P\psi_\lambda)\right)^2\right] \ .$$

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

$$\|g_x\|^2 = \sum_{\lambda\in\Lambda_m}(\psi_\lambda(x)-P\psi_\lambda)^2 \leq 2\sum_{\lambda\in\Lambda_m}(\psi_\lambda(x))^2 + 2\sum_{\lambda\in\Lambda_m}(P\psi_\lambda)^2$$
$$= 2\Psi_m(x)^2 + 2\|s_m\|^2 \leq 2(b_m^2+\|s_m\|^2) \ .$$

Hence,

$$\sum_{i=2}^{n}\mathbb{E}\left[g_{i,1}(\xi_i,x)^2\right] = \frac{n(V-1)}{V}\mathrm{Var}\left(g_x(\xi_1)\right) \leq \frac{n(V-1)\|g_x\|^2\|s\|_\infty}{V}$$
$$\leq \frac{2n(V-1)}{V}\left(b_m^2+\|s_m\|^2\right)\|s\|_\infty$$

which proves

$$\overline{C}^2 \leq \frac{2n(V-1)}{V}\left(b_m^2+\|s_m\|^2\right)\|s\|_\infty \ . \tag{A.44}$$

Finally, from Cauchy-Schwarz inequality,

$$\sum_{\lambda\in\Lambda_m}(\psi_\lambda(x)-P\psi_\lambda)(\psi_\lambda(y)-P\psi_\lambda) \leq \sup_{x\in\mathbb{R}}\sum_{\lambda\in\Lambda_m}\left(\psi_\lambda(x)-P\psi_\lambda\right)^2 \leq 2(b_m^2+\|s_m\|^2) \ .$$

Hence,

$$\overline{D} \leq 2\left(b_m^2+\|s_m\|^2\right) \ . \tag{A.45}$$

□

Let us conclude the proof of Proposition 2. From Lemmas A.4 and A.6, an absolute constant κ exists such that, with probability larger than $1-6e^{-x}$, for every $\epsilon \in (0,1]$,

$$\left|\mathrm{pen}_{\mathrm{VF}}(m,V,V-1) - 2\|s_m-\widehat{s}_m\|^2\right|$$

$$= \frac{2V}{V-1}\,|U(m)| \leq \epsilon\frac{\mathcal{D}_m}{n} + \kappa\left(\frac{\|s\|_\infty\, x}{\epsilon n} + \frac{(b_m^2 + \|s\|^2)x^2}{n^2}\right) \quad . \tag{A.46}$$

Using in addition Lemma A.5, we get that an absolute constant κ' exists such that with probability larger than $1 - 8e^{-x}$, for every $\epsilon \in (0,1]$, Eq. (A.46) holds true and

$$\left|\mathrm{pen}_{\mathrm{VF}}(m,V,V-1) - \frac{2\mathcal{D}_m}{n}\right| \leq \epsilon\frac{\mathcal{D}_m}{n} + \kappa\left(\frac{\|s\|_\infty\, x}{\epsilon n} + \frac{(b_m^2\epsilon^{-3} + \|s\|^2)x^2}{n^2}\right) \quad ,$$

which implies Eq. (15) and (16). □

A.3. Proof of Theorem 1

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

$$\begin{aligned}\|\widehat{s}_{\widehat{m}} - s\|^2 - \left(\mathrm{pen}_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}_{\mathrm{VF}}(\widehat{m},V,C(V-1))\right)& \\ \leq \|\widehat{s}_m - s\|^2 + \left(\mathrm{pen}_{\mathrm{VF}}(m,V,C(V-1)) - \mathrm{pen}_{\mathrm{id}}(m)\right) \quad .&\end{aligned} \tag{A.47}$$

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

$$\begin{aligned}\|\widehat{s}_{\widehat{m}} - s\|^2 &\leq \|\widehat{s}_m - s\|^2 + \left(\mathrm{pen}_{\mathrm{VF}}(m,V,C(V-1)) - 2\|s_m - \widehat{s}_m\|^2\right) \\ &\quad - \left(\mathrm{pen}_{\mathrm{VF}}(\widehat{m},V,C(V-1)) - 2\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\|^2\right) + 2(P_n - P)(s_m - s_{\widehat{m}}) \\ &= \|\widehat{s}_m - s\|^2 + \left(\mathrm{pen}_{\mathrm{VF}}(m,V,C(V-1)) - 2C\|s_m - \widehat{s}_m\|^2\right) \\ &\quad - \left(\mathrm{pen}_{\mathrm{VF}}(\widehat{m},V,C(V-1)) - 2C\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\|^2\right) + 2(P_n - P)(s_m - s_{\widehat{m}}) \\ &\quad + 2(C-1)\left[\|\widehat{s}_m - s_m\|^2 - \|\widehat{s}_{\widehat{m}} - s_{\widehat{m}}\|^2\right] \quad .\end{aligned} \tag{A.48}$$

Let $x > 0$ and $x_n = \log(|\mathcal{M}_n|) + x$. A union bound in Proposition 2 gives

$$\begin{aligned}&\mathbb{P}\left(\exists m \in \mathcal{M}_n,\, \epsilon \in (0,1] \text{ s.t. } \left|\mathrm{pen}_{\mathrm{VF}}(m,V,V-1) - 2\|s_m - \widehat{s}_m\|^2\right| > \epsilon\frac{\mathcal{D}_m}{n} + \kappa\rho_1(m,\epsilon,s,x_n,n)\right) \\ &\leq 8\sum_{m\in\mathcal{M}_n} e^{-x_n} = 8e^{-x}\sum_{m\in\mathcal{M}_n}\frac{1}{|\mathcal{M}_n|} = 8e^{-x}\end{aligned} \tag{A.49}$$

and a union bound in Lemma A.5 gives

$$\begin{aligned}&\mathbb{P}\left(\exists m \in \mathcal{M}_n,\, \epsilon \in (0,1] \text{ s.t. } \left|\|\widehat{s}_m - s_m\|^2 - \frac{\mathcal{D}_m}{n}\right| > \epsilon\frac{\mathcal{D}_m}{n} + \kappa\rho_1(m,\epsilon,s,x_n,n)\right) \\ &\leq \sum_{m\in\mathcal{M}_n} e^{-x_n} = e^{-x} \quad .\end{aligned} \tag{A.50}$$

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 $s \in L^\infty(\mu)$,

$$\mathrm{Var}\left((s_m - s_{m'})(\xi_1)\right) \leq \|s\|_\infty \|s_m - s_{m'}\|^2 \quad .$$

Under Assumption (**H2**),

$$\|s_m - s_{m'}\|_\infty \leq \|s_m\|_\infty + \|s_{m'}\|_\infty \leq 2a \quad .$$

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

$$\|s_m - s_{m'}\|_\infty \leq b_{m''} \|s_m - s_{m'}\| \leq \sqrt{n} \|s_m - s_{m'}\| \quad .$$

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

$$\begin{aligned}(P_n - P)(s_m - s_{m'}) &\leq \sqrt{\frac{2x\mathrm{Var}\left((s_m - s_{m'})(\xi_1)\right)}{n}} + \frac{\|s_m - s_{m'}\|_\infty x}{3n} \\ &\leq \epsilon \|s_m - s_{m'}\|^2 + \frac{\kappa(Ax + x^2)}{\epsilon n} \quad .\end{aligned}$$

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 $\alpha, \beta, \epsilon > 0$, $\alpha\beta \leq \epsilon\alpha^2 + (\beta^2)/(4\epsilon)$. A union bound gives that for any $x > 0$, with probability at least $1 - e^{-x}$, for every $m, m' \in \mathcal{M}_n$ and every $\epsilon \in (0, 1]$,

$$(P_n - P)(s_m - s_{m'}) \leq \epsilon \|s_m - s_{m'}\|^2 + \frac{\kappa(Ax_n + x_n^2)}{\epsilon n} \quad . \tag{A.51}$$

Plugging Eq. (A.49), (A.50) and (A.51) into Eq. (A.48) and using that $C \in (1/2, 2]$ yields that, with probability $1 - 10e^{-x}$, for any $\epsilon \in (0, 1/2]$,

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

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

$$2 \sup_{m \in \mathcal{M}_n} \rho_1(m, \epsilon, s, x_n, n) + \frac{Ax_n + x_n^2}{\epsilon n} \leq \frac{(2\|s\|_\infty + A)x_n}{\epsilon n} + \left(3 + \frac{2\|s\|^2}{n} \right) \frac{x_n^2}{\epsilon^3 n}$$

for every $\epsilon \in (0, 1]$. Hence, with probability larger than $1 - 10e^{-x}$, for any $\epsilon \in (0, 1]$,

$$\frac{1 - \delta_- - \epsilon}{1 + \delta_+ + \epsilon} \|\widehat{s}_{\widehat{m}} - s\|^2 \leq \|\widehat{s}_m - s\|^2 + \kappa \left[\frac{(\|s\|_\infty + A)x_n}{\epsilon n} + \left(1 + \frac{\|s\|^2}{n} \right) \frac{x_n^2}{\epsilon^3 n} \right]$$

for some absolute constant $\kappa > 0$, which implies the result. □

A.4. Proof of Theorem 2

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

$$\begin{aligned} U_m(x,y) &:= \sum_{\lambda \in \Lambda_m} (\psi_\lambda(x) - P\psi_\lambda)(\psi_\lambda(y) - P\psi_\lambda) \\ &= K_m(x,y) - s_m(x) - s_m(y) + \|s_m\|^2 \quad . \end{aligned} \tag{A.52}$$

For every $x \in \mathcal{X}$, $K_m(x,x) = \Psi_m(x)$ by Eq. (A.29), $U_m(x,x) = \Psi_m(x) - 2s_m(x) + \|s_m\|^2$ and, by independence, for every $m, m' \in \{m_1, m_2\}$

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

hence, $\operatorname{Var}\left(U_{m_1}(\xi_1,\xi_2) - U_{m_2}(\xi_1,\xi_2)\right) = \mathbf{B}\left(m_1, m_2\right)$. For every $m \in \{m_1, m_2\}$, by Eq. (A.52),

$$P_n\gamma(\widehat{s}_m) = -\sum_{\lambda \in \Lambda_m} (P_n\psi_\lambda)^2 = -\frac{1}{n^2} \sum_{1 \le i,j \le n} K_m(\xi_i, \xi_j) \tag{A.53}$$

$$= -\frac{1}{n^2} \sum_{1 \le i,j \le n} U_m(\xi_i,\xi_j) - \frac{2}{n}\sum_{i=1}^n s_m(\xi_i) + \|s_m\|^2 \quad . \tag{A.54}$$

Moreover, by Eq. (A.40) in the proof of Lemma A.4,

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

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

It follows that

$$\mathcal{C}_{C,\mathcal{B}}(m) = \sum_{1 \le i,j \le n} \frac{2CE_{i,j}^{(\mathrm{VF})} - 1}{n^2} U_m(\xi_i,\xi_j) + \sum_{i=1}^n \frac{-2s_m(\xi_i)}{n} + \|s_m\|^2 \quad .$$

Hence, up to the deterministic term $\|s_m\|^2$, $\mathcal{C}_{C,\mathcal{B}}(m)$ has the form of a function $\mathcal{C}_m$ defined in Lemma A.7 below with

$$\omega_{i,j} = \frac{2CE_{i,j}^{(\mathrm{VF})} - 1}{n^2}, \qquad f_m = \frac{-2s_m}{n} \quad .$$

It remains to evaluate the quantities appearing in Lemma A.7 for these weights and function. First,

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

Second, by (**Reg**),

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

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

It follows that

$$\sum_{1 \le i \le n} \omega_{i,i}^2 = \frac{(2C-1)^2}{n^3} \ , \qquad \sum_{i=1}^{n} \omega_{i,i} = \frac{2C-1}{n}$$

$$\text{and} \quad \sum_{1 \le i \ne j \le n} \omega_{i,j} \omega_{j,i} = \sum_{1 \le i \ne j \le n} \omega_{i,j}^2 = \frac{1}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \ .$$

Hence, from Lemma A.7, for every $m, m' \in \{m_1, m_2\}$,

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

Therefore,

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

$$\operatorname{Var} \left(\mathcal{C}_{C,\mathcal{B}}(m_1) - \mathcal{C}_{C,\mathcal{B}}(m_2) \right) = \frac{2}{n^2} \left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n} \right) \mathbf{B} \left(m_1, m_2 \right)$$

$$
\begin{aligned}
&+ \frac{1}{n}\mathrm{Var}\left(2(s_{m_1} - s_{m_2})(\xi) - \frac{2C-1}{n}\left(U_{m_1}(\xi,\xi) - U_{m_2}(\xi,\xi)\right)\right) \\
&= \frac{2}{n^2}\left(1 + \frac{4C^2}{V-1} - \frac{(2C-1)^2}{n}\right)\mathrm{Var}\left(U_{m_1}(\xi,\xi) - U_{m_2}(\xi_1,\xi_2)\right) \\
&+ \frac{4}{n}\mathrm{Var}\left(\left(1 + \frac{2C-1}{n}\right)(s_{m_1} - s_{m_2})(\xi) - \frac{2C-1}{2n}\left(\Psi_{m_1}(\xi) - \Psi_{m_2}(\xi)\right)\right) \quad ,
\end{aligned}
$$

which concludes the proof. □

Lemma A.7. *Let* $\mathcal{C}_m = \sum_{1\le i,j\le n}\omega_{i,j}U_m(\xi_i,\xi_j) + \sum_{i=1}^n f_m(\xi_i)$, *where* U_m *is defined by Eq. (A.52) and* $f_m \in L^2(\mu)$. *We have*

$$
\begin{aligned}
\mathrm{cov}\left(\mathcal{C}_m, \mathcal{C}_{m'}\right) = &\left(\sum_{1\le i\ne j\le n}\omega_{i,j}^2 + \omega_{i,j}\omega_{j,i}\right)\mathrm{cov}\left(U_m(\xi_1,\xi_2), U_{m'}(\xi_1,\xi_2)\right) \\
&+ \left(\sum_{i=1}^n \omega_{i,i}^2\right)\mathrm{cov}\left(U_m(\xi_1,\xi_1), U_{m'}(\xi_1,\xi_1)\right) \\
&+ \left(\sum_{i=1}^n \omega_{i,i}\right)\left[\mathrm{cov}\left(U_m(\xi_1,\xi_1), f_{m'}(\xi_1)\right) + \mathrm{cov}\left(U_{m'}(\xi_1,\xi_1), f_m(\xi_1)\right)\right] \\
&+ n\,\mathrm{cov}\left(f_m(\xi_1), f_{m'}(\xi_1)\right) \quad .
\end{aligned}
$$

Proof. We develop the covariance to get

$$
\begin{aligned}
\mathrm{cov}\left(\mathcal{C}_m, \mathcal{C}_{m'}\right) = &\sum_{1\le i,j,k,\ell\le n}\omega_{i,j}\omega_{k,\ell}\,\mathrm{cov}\left(U_m(\xi_i,\xi_j), U_{m'}(\xi_k,\xi_\ell)\right) \\
&+ \sum_{1\le i,j,k\le n}\omega_{i,j}\,\mathrm{cov}\left(U_m(\xi_i,\xi_j), f_{m'}(\xi_k)\right) \\
&+ \sum_{1\le i,j,k\le n}\omega_{i,j}\,\mathrm{cov}\left(U_{m'}(\xi_i,\xi_j), f_m(\xi_k)\right) \\
&+ \sum_{1\le i,j\le n}\mathrm{cov}\left(f_m(\xi_i), f_{m'}(\xi_j)\right) \quad .
\end{aligned}
$$

The proof is then concluded with the following remarks, that follow by independence of the random variables $\xi_{[\![n]\!]}$.

1. $\mathrm{cov}\left(f_m(\xi_i), f_{m'}(\xi_j)\right) \ne 0$ only when $\xi_i = \xi_j$, therefore

$$
\sum_{1\le i,j\le n}\mathrm{cov}\left(f_m(\xi_i), f_{m'}(\xi_j)\right) = \sum_{i=1}^n \mathrm{cov}\left(f_m(\xi_i), f_{m'}(\xi_i)\right) = n\,\mathrm{cov}\left(f_m(\xi_1), f_{m'}(\xi_1)\right) \quad .
$$

2. By definition (A.52) of U_m, $\mathrm{cov}\left(U_m(\xi_i,\xi_j), f_{m'}(\xi_k)\right) \ne 0$ only when $i = j = k$, hence

$$\sum_{1\leq i,j,k\leq n} \omega_{i,j} \operatorname{cov}\left(U_m(\xi_i,\xi_j), f_{m'}(\xi_k)\right) = \left(\sum_{i=1}^{n} \omega_{i,i}\right) \operatorname{cov}\left(U_m(\xi_1,\xi_1), f_{m'}(\xi_1)\right) \quad .$$

3. By definition (A.52) of U_m, $\operatorname{cov}\left(U_m(\xi_i,\xi_j), U_m(\xi_k,\xi_l)\right) \neq 0$ only when $i=j=k=\ell$ or $i=k\neq j=\ell$ or $i=\ell\neq j=k$. It follows that

$$\begin{aligned}
&\sum_{1\leq i,j,k,\ell\leq n} \omega_{i,j}\omega_{k,\ell} \operatorname{cov}\left(U_m(\xi_i,\xi_j), U_{m'}(\xi_k,\xi_\ell)\right) \\
&= \left(\sum_{1\leq i\neq j\leq n} \omega_{i,j}^2 + \omega_{i,j}\omega_{j,i}\right) \operatorname{cov}\left(U_m(\xi_1,\xi_2), U_{m'}(\xi_1,\xi_2)\right) \\
&\quad + \left(\sum_{i=1}^{n} \omega_{i,i}^2\right) \operatorname{cov}\left(U_m(\xi_1,\xi_1), U_{m'}(\xi_1,\xi_1)\right) \quad .
\end{aligned}$$

□

Appendix S: Supplementary material

The supplementary material is organized as follows. Section S.1 gives complementary computations of variances. Then, results concerning hold-out penalization are detailed in Section S.2, with the proof of the oracle inequality stated in Section 8.1 (Theorem 3) and an exact computation of the variance. Section S.3 provides complements on the computational aspects stated in Section 7. In particular, we state and analyse the basic algorithm for computing the V-fold criteria and we give the proof of Proposition 3. A useful concentration inequality is recalled in Section S.4. Finally, some simulation results are detailed in Section S.5, as a supplement to the ones of Section 6.

S.1. Additional variance computations

Proposition S.8. *Let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ and $(\psi_\lambda)_{\lambda\in\Lambda_{m'}}$ be two finite orthonormal families of vectors of $L^4(\mu)$, $C > 0$ some constant. Assume that $\mathcal{B}$ satisfies* (**Reg**) *and, for any $m \in \mathcal{M}_n$, let*

$$\mathcal{C}_{id}(m) = P_n\gamma(\widehat{s}_m) + \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \quad .$$

Then, with the notation of Theorem 2, for every $m \in \mathcal{M}_n$,

$$\mathrm{Var}\left(\mathcal{C}_{id}(m)\right) = \frac{2(n-1)}{n^3}\beta\left(\Lambda_m, \Lambda_m\right) + \frac{2}{n}\mathrm{Var}\left(\left(1-\frac{1}{n}\right)s_m(\xi) + \frac{1}{2n}\Psi_m(\xi)\right) \quad .$$

For any (m, m') in $\mathcal{M}_n$, we also have

$$\begin{aligned}\mathrm{Var}\left(\mathcal{C}_{id}(m) - \mathcal{C}_{id}(m')\right) &= \frac{2(n-1)}{n^3}\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right) \\ &\quad + \frac{2}{n}\mathrm{Var}\left(\left(1-\frac{1}{n}\right)\left(s_m(\xi) - s_{m'}(\xi)\right) + \frac{1}{2n}\left(\Psi_m(\xi) - \Psi_{m'}(\xi)\right)\right) \quad .\end{aligned}$$

Proof of Proposition S.8. Simply notice that

$$\mathrm{Var}\left(\mathcal{C}_{id}(m)\right) = \mathrm{Var}\left(P_n\gamma(\widehat{s}_m)\right) \quad .$$

Therefore, from (A.53), the variance of $\mathcal{C}_{id}(m)$ is the one of

$$-\frac{1}{n^2}\sum_{1\le i,j\le n} U_m(\xi_i, \xi_j) - \sum_{i=1}^{n}\frac{2s_m(\xi_i)}{n} \quad .$$

so that, by Lemma A.7,

$$\begin{aligned}\mathrm{Var}\left(\mathcal{C}_{id}(m)\right) &= \frac{2(n-1)}{n^3}\beta\left(\Lambda_m, \Lambda_m\right) + \frac{1}{n^3}\mathrm{Var}\left(\Psi_m(\xi) - 2s_m(\xi)\right) \\ &\quad + \frac{4}{n^2}\sum_{i=1}^{n}\mathrm{cov}\left(\Psi_m(\xi) - 2s_m(\xi), s_m(\xi)\right) + \frac{4}{n}\mathrm{Var}\left(s_m(\xi)\right) \\ &= \frac{2(n-1)}{n^3}\beta\left(\Lambda_m, \Lambda_m\right) + \frac{2}{n}\mathrm{Var}\left(\left(1-\frac{1}{n}\right)s_m(\xi) + \frac{1}{n}\Psi_m(\xi)\right) \quad .\end{aligned}$$

The variance of the increments follows from the same computations. □

Evaluation of the terms in the variance term The following proposition gives a formulation of the terms appearing in Theorem 2 and Proposition S.8 that does not depend of the basis $(\psi_\lambda)_{\lambda\in\Lambda_m}$.

Proposition S.9. *For any $m, m' \in \mathcal{M}_n$, we have*

$$\begin{aligned}\beta(\Lambda_m,\Lambda_{m'}) &= n\operatorname{cov}(\widehat{s}_m(\xi),\widehat{s}_{m'}(\xi)) - (n+1)\operatorname{cov}(s_m(\xi),s_{m'}(\xi))\\ \mathbf{B}(\Lambda_m,\Lambda_{m'}) &= n\mathrm{Var}((\widehat{s}_m-\widehat{s}_{m'})(\xi)) - (n+1)\mathrm{Var}((s_m-s_{m'})(\xi))\end{aligned} \tag{S.55}$$

Proof of Proposition S.9. By definition, we have

$$\begin{aligned}\beta(\Lambda_m,\Lambda_{m'}) &:= \sum_{\lambda\in\Lambda_m,\,\lambda'\in\Lambda_{m'}}\left(C^{(1,1)}_{\lambda,\lambda'}\right)^2 = \sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\operatorname{cov}(\psi_\lambda,\psi_{\lambda'})^2\\ &= \sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P(\psi_\lambda\psi_{\lambda'}) - P\psi_\lambda P\psi_{\lambda'}\right)^2\\ &= \sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P(\psi_\lambda\psi_{\lambda'})\right)^2 - 2\sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}P\psi_\lambda P\psi_{\lambda'}P(\psi_\lambda\psi_{\lambda'}) + \sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P\psi_\lambda P\psi_{\lambda'}\right)^2\\ &= \sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P(\psi_\lambda\psi_{\lambda'})\right)^2 - 2P[s_m s_{m'}] + \|s_m\|^2\|s_{m'}\|^2.\end{aligned}$$

Now, let ξ denote a copy of ξ_1, independent of $\xi_{[\![n]\!]}$. We have

$$\begin{aligned}\operatorname{cov}(\widehat{s}_m(\xi),\widehat{s}_{m'}(\xi)) &= \frac{1}{n^2}\sum_{1\le i,j\le n}\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}}\operatorname{cov}(\psi_\lambda(\xi_i)\psi_\lambda(\xi),\psi_{\lambda'}(\xi_j)\psi_{\lambda'}(\xi))\\ &= \frac{1}{n}\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}}\left(P(\psi_\lambda\psi_{\lambda'})\right)^2 - \left(P\psi_\lambda P\psi_{\lambda'}\right)^2\\ &\quad+\frac{n-1}{n}\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}}\left(P(\psi_\lambda\psi_{\lambda'}) - P\psi_\lambda P\psi_{\lambda'}\right)P\psi_\lambda P\psi_{\lambda'}\\ &= \frac{1}{n}\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}}\left(P(\psi_\lambda\psi_{\lambda'})\right)^2\\ &\quad-\frac{1}{n}\|s_m\|^2\|s_{m'}\|^2+\frac{n-1}{n}\operatorname{cov}(s_m(\xi),s_{m'}(\xi))\end{aligned}$$

It follows that

$$\begin{aligned}\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}}\left(P(\psi_\lambda\psi_{\lambda'})\right)^2 &= n\operatorname{cov}(\widehat{s}_m(\xi),\widehat{s}_{m'}(\xi)) + \|s_m\|^2\|s_{m'}\|^2\\ &\quad-(n-1)\operatorname{cov}(s_m(\xi),s_{m'}(\xi)).\end{aligned} \tag{S.56}$$

Thus,

$$\beta(\Lambda_m,\Lambda_{m'}) = n\operatorname{cov}(\widehat{s}_m(\xi),\widehat{s}_{m'}(\xi)) - (n+1)\operatorname{cov}(s_m(\xi),s_{m'}(\xi)).$$

□

Evaluation of the variance in the regular histogram case The following lemma gives the value of the terms appearing in Theorem 2 in the histogram case.

Lemma S.10. *For any $i \in [\![n]\!]$ and $\lambda \in \Lambda$, let $\xi_{\lambda,i} = \psi_\lambda(\xi_i) - P\psi_\lambda$ and for $m^\star \in \{m, m'\}$, let $T_{m^\star}(x) = \sum_{\lambda \in m^\star} (\psi_\lambda(x) - P\psi_\lambda)^2 = \sup_{t \in \mathbb{B}_{m^\star}} (t(x) - Pt)^2$. The random variables $(\xi_{\lambda,1})_{\lambda\in\Lambda}, \dots, (\xi_{\lambda,n})_{\lambda\in\Lambda}$ are independent and, if m is a regular partition of $\mathbb{R}$, with $\mu(\lambda) = d_m^{-1}$ for any $\lambda \in m$ and if m' is a subpartition of m,*

$$\beta(\Lambda_m, \Lambda_{m'}) = \sum_{\lambda\in m, \lambda'\in m'} \mathbb{E}[\xi_{1,\lambda}\xi_{1,\lambda'}]^2 = P(T_m s_{m'})$$

and

$$\begin{aligned}\mathbf{B}(\Lambda_m, \Lambda_{m'}) &= P((T_m(s_m - s_{m'}) + (T_{m'} - T_m)s_{m'}) \\ &= d_{m'}\|s_{m'}\|^2 - d_m\|s_m\|^2 - 2\operatorname{var}_P(s_m - s_{m'}) - \|s_m - s_{m'}\|^4 \quad .\end{aligned}$$

Proof.

$$\begin{aligned}&\sum_{\lambda\in m, \lambda'\in m'} \mathbb{E}[\xi_{1,\lambda}\xi_{1,\lambda'}]^2 \\ &= \sum_{\lambda\in m, \lambda'\in m'} \Big([P(\psi_\lambda\psi_{\lambda'})]^2 - 2P(P\psi_\lambda\psi_\lambda P\psi_{\lambda'}\psi_{\lambda'}) + (P\psi_\lambda)^2 (P\psi_{\lambda'})^2 \Big) \\ &= \sum_{\lambda\in m, \lambda'\in m'} [P(\psi_\lambda\psi_{\lambda'})]^2 - 2P(s_m s_{m'}) + \|s_m\|^2 \|s_{m'}\|^2 \quad .\end{aligned}$$

Moreover,

$$\begin{aligned}\sum_{\lambda\in m, \lambda'\in m'} [P(\psi_\lambda\psi_{\lambda'})]^2 &= \sum_{\lambda\in m} \sum_{\lambda'\subset\lambda} [P(\psi_\lambda\psi_{\lambda'})]^2 \\ &= \sum_{\lambda\in m} \frac{1}{\mu(\lambda)} \sum_{\lambda'\subset\lambda} (P\psi_{\lambda'})^2 = d_m \sum_{\lambda'\in m'} (P\psi_{\lambda'})^2 \quad .\end{aligned}$$

Hence,

$$\sum_{\lambda\in m, \lambda'\in m'} \mathbb{E}[\xi_{1,\lambda}\xi_{1,\lambda'}]^2 = d_m\|s_{m'}\|^2 - 2P(s_m s_{m'}) + \|s_m\|^2\|s_{m'}\|^2 \quad ,$$

and

$$P(T_m s_{m'}) = d_m\|s_{m'}\|^2 - 2P(s_m s_{m'}) + \|s_m\|^2\|s_{m'}\|^2 \quad .$$

It follows that

$$\begin{aligned}&\mathbf{B}(\Lambda_m, \Lambda_{m'}) = P((T_m(s_m - s_{m'}) + (T_{m'} - T_m)s_{m'}) \\ &= d_m(\|s_m\|^2 - \|s_{m'}\|^2) + (d_{m'} - d_m)\|s_{m'}\|^2 - 2P((s_m - s_{m'})^2) + (\|s_m\|^2 - \|s_{m'}\|^2)^2 \\ &= (d_{m'} - d_m)\|s_{m'}\|^2 + d_m\|s_m - s_{m'}\|^2 - 2\operatorname{var}_P(s_m - s_{m'}) - \|s_m - s_{m'}\|^4 \quad .\end{aligned}$$

□

S.2. Results on hold-out penalization

This section gathers the proof of Theorem 3 (oracle inequality for hold-out penalization) and the variance computations we can make for hold-penalization.

S.2.1. Proof of Theorem 3

The hold-out penalty is equal to

$$\begin{aligned}\mathrm{pen}_{\mathrm{HO}}(m,T,x) &= 2x(1-\tau)^2(P_n^{(T)}-P_n^{(T^c)})(\widehat{s}_m^{(T)}-\widehat{s}_m^{(T^c)}) \\ &= 2x(1-\tau)^2\sum_{\lambda\in\Lambda_m}\left[\left(P_n^{(T)}-P_n^{(T^c)}\right)\psi_\lambda\right]^2 \quad . \quad \text{(S.57)}\end{aligned}$$

As for Theorem 1, the oracle inequality is based on a concentration result for $\mathrm{pen}_{\mathrm{HO}}(m,T,x)$. Let start with an exact formula for the hold-out penalty (Lemma S.11, analogous to Lemma A.4).

Lemma S.11. *For all $m\in\mathcal{M}_n$, we have*

$$\begin{aligned}&\mathrm{pen}_{\mathrm{HO}}(m,T,x) \\ &= 2x(1-\tau)^2\left[\left\|\widehat{s}_m^{(T)}-s_m\right\|^2+\left\|\widehat{s}_m^{(T^c)}-s_m\right\|^2-2(P_n^{(T)}-P)\left(\widehat{s}_m^{(T^c)}-s_m\right)\right] \quad .\end{aligned}$$

In particular, for $\tau=|T|/n$,

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{HO}}(m,T,x)\right] = 2x\frac{1-\tau}{\tau}\frac{\mathcal{D}_m}{n} \quad .$$

Proof of Lemma S.11. By definition,

$$\begin{aligned}\mathrm{pen}_{\mathrm{HO}}(m,T,x) &= 2x(1-\tau)^2 \sum_{\lambda\in\Lambda_m}\left\{\left((P_n^{(T^c)}-P)\psi_\lambda\right)^2+\left((P_n^{(T)}-P)\psi_\lambda\right)^2\right\} \\ &\quad -2x(1-\tau)^2 \sum_{\lambda\in\Lambda_m}\left\{2\left((P_n^{(T^c)}-P)\psi_\lambda\right)\left((P_n^{(T)}-P)\psi_\lambda\right)\right\} \\ &= 2x(1-\tau)^2 \left[\left\|\widehat{s}_m^{(T^c)}-s_m\right\|^2+\left\|\widehat{s}_m^{(T)}-s_m\right\|^2\right. \\ &\quad \left.-2(P_n^{(T)}-P)\left(\sum_{\lambda\in\Lambda_m}\left((P_n^{(T^c)}-P)\psi_\lambda\right)\psi_\lambda\right)\right].\end{aligned}$$

□

Lemma S.12. *For all $m\in\mathcal{M}_n$ and $x>0$, with probability larger than $1-2e^{-x}$, for all $\eta>0$,*

$$\left|(P_n^{(T)}-P)\left(\widehat{s}_m^{(T^c)}-s_m\right)\right| \le \frac{\eta}{2}\left\|\widehat{s}_m^{(T^c)}-s_m\right\|^2+\frac{2\|s\|_\infty x}{\eta\tau n}+\frac{b_m^2x^2}{9\eta(\tau n)^2} \quad . \quad \text{(S.58)}$$

Proof of Lemma S.12. Let us apply Bernstein's inequality conditionally to $(\xi_i)_{i\notin T}$ to the function $t = \left(\widehat{s}_m^{(T^c)} - s_m\right)$. Recall that $v_m^2 \leq \|s\|_\infty$, hence,

$$\left\|\widehat{s}_m^{(T^c)} - s_m\right\|_\infty \leq \left\|\widehat{s}_m^{(T^c)} - s_m\right\| b_m \ ,$$
$$\operatorname{var}\left(\widehat{s}_m^{(T^c)}(\xi) - s_m(\xi)\,\middle|\,(\xi_i)_{i\notin T}\right) \leq \left\|\widehat{s}_m^{(T^c)} - s_m\right\|^2 v_m^2 \leq \left\|\widehat{s}_m^{(T^c)} - s_m\right\|^2 \|s\|_\infty \ .$$

Hence, for all $x > 0$, with probability larger than $1 - 2e^{-x}$, conditionally to $(\xi_i)_{i\notin T}$,

$$\begin{aligned}\left|(P_n^{(T)} - P)\left(\widehat{s}_m^{(T^c)} - s_m\right)\right| &\leq \left\|\widehat{s}_m^{(T^c)} - s_m\right\| \left(\sqrt{\frac{2\|s\|_\infty x}{\tau n}} + \frac{b_m x}{3\tau n}\right) \\ &\leq \frac{\eta}{2}\left\|\widehat{s}_m^{(T^c)} - s_m\right\|^2 + \frac{1}{\eta}\left(\frac{2\|s\|_\infty x}{\tau n} + \frac{b_m^2 x^2}{9(\tau n)^2}\right) \ .\end{aligned}$$

As the bound on the probability does not depend on $(\xi_i)_{i\notin T}$, the same inequality holds unconditionally. □

Proof of Theorem 3. From Theorem 4.1 in [Ler11] (recalled in Proposition S.14), Lemma S.11 and Lemma S.12 that there exists an absolute constant κ such that, for all $x > 0$, with probability larger than $1 - 8e^{-x}$, for all $\epsilon \in (0, 1]$,

$$\begin{aligned}\forall m \in \mathcal{M}_n,\ &\left|\mathrm{pen}_{\mathrm{HO}}\left(m, T, \frac{\tau}{1-\tau}\right) - \|\widehat{s}_m - s_m\|^2\right| \\ &\leq \epsilon\|\widehat{s}_m - s_m\|^2 + \kappa\left(\frac{\|s\|_\infty x_n}{\epsilon n} + \frac{b_m^2 x_n^2}{\epsilon^3 n^2}\frac{\tau^2 + (1-\tau)^2}{\tau(1-\tau)}\right) \ .\end{aligned}$$

We can then conclude the proof as in Theorem 1.

□

S.2.2. Variance

Proposition S.13. *Assume that $|T| \in [\![n-1]\!]$ and denote, for any $m \in \mathcal{M}$,*

$$\mathcal{C}^{\mathrm{ho}}_{(C,T)}(m) = P_n\gamma(\widehat{s}) + \mathrm{pen}_{\mathrm{HO}}\left(m, T, C\tau/(1-\tau)\right).$$

Then, with the notations introduced in Theorem 2, for every $m, m' \in \mathcal{M}_n$,

$$\begin{aligned}\operatorname{Var}\left(\mathcal{C}^{\mathrm{ho}}_{(C,T)}(m)\right) &= \frac{4}{n}\operatorname{Var}\left(\left(1 + \frac{2C-1}{n}\right)s_m(\xi) - \frac{2C-1}{2n}\Psi_m(\xi)\right) \\ &\quad + \frac{2}{n^2}\left[1 + 4C^2 - \frac{(2C-1)^2}{n}\right]\beta(\Lambda_m, \Lambda_m) \qquad \text{(S.59)} \\ &\quad + \frac{4C^2}{n^3}\frac{(1-2\tau)^2}{\tau(1-\tau)}\left(\operatorname{Var}\left(\Psi_m(\xi) - 2s_m(\xi)\right) - 2\beta\left(\Lambda_m, \Lambda_m\right)\right) \ .\end{aligned}$$

$$
\begin{aligned}
&\mathrm{Var}\left(\mathcal{C}^{\mathrm{ho}}_{(C,T)}(m) - \mathcal{C}^{\mathrm{ho}}_{(C,T)}(m')\right) \\
&= \frac{4}{n}\mathrm{Var}\left(\left(1 + \frac{2C-1}{n}\right)(s_m(\xi) - s_{m'}(\xi)) - \frac{2C-1}{2n}(\Psi_m(\xi) - \Psi_{m'}(\xi))\right) \\
&+ \frac{2}{n^2}\left[1 + 4C^2 - \frac{(2C-1)^2}{n}\right]\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right) \qquad \text{(S.60)} \\
&+ \frac{4C^2}{n^3}\frac{(1-2\tau)^2}{\tau(1-\tau)}\left(\mathrm{Var}\left((\Psi_m(\xi) - \Psi_{m'}(\xi)) - 2(s_m(\xi) - s_{m'}(\xi))\right) - 2\mathbf{B}\left(\Lambda_m, \Lambda_m\right)\right) \ .
\end{aligned}
$$

Proof of Proposition S.13. By definition,

$$
\begin{aligned}
\mathrm{pen}_{\mathrm{HO}}(m, T, x) &= 2x \sum_{\lambda \in \Lambda_m}\left[\left(P_n^{(T)} - P_n\right)\psi_\lambda\right]^2 \\
&= \frac{2x}{n^2}\sum_{\lambda \in \Lambda_m}\left(\sum_{i=1}^n \left(\frac{1}{\tau}\mathbf{1}_{i \in T} - 1\right)\psi_\lambda(\xi_i)\right)^2 \\
&= \frac{2x}{n^2}\sum_{i,j=1}^n E^{(\mathrm{HO})}_{i,j} U_m(\xi_i, \xi_j) \ , \qquad \text{(S.61)}
\end{aligned}
$$

where, for all $i, j \in \{1 \dots, n\}$,

$$
\begin{aligned}
U_m(\xi_i, \xi_j) &= \sum_{\lambda \in \Lambda_m}(\psi_\lambda(\xi_i) - P\psi_\lambda)(\psi_\lambda(\xi_j) - P\psi_\lambda) \ , \\
E^{(\mathrm{HO})}_{i,j} &= \left(\frac{1}{\tau}\mathbf{1}_{i \in T} - 1\right)\left(\frac{1}{\tau}\mathbf{1}_{j \in T} - 1\right) \ ,
\end{aligned}
$$

Therefore, from (A.53), if $x = C\tau/(1-\tau)$,

$$
\begin{aligned}
\mathcal{C}^{\mathrm{ho}}_{(C,T)}(m) &:= P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{HO}}(m, T, x) \\
&= \sum_{1 \le i,j \le n}\frac{2xE^{(\mathrm{HO})}_{i,j} - 1}{n^2}U_m(\xi_i, \xi_j) - \sum_{i=1}^n \frac{2}{n}s_m(\xi_i) + \|s_m\|^2 \ .
\end{aligned}
$$

By definition

$$
E^{(\mathrm{HO})}_{i,j} = \left(\frac{1-\tau}{\tau}\right)^2 \mathbf{1}_{i,j \in T} - \frac{1-\tau}{\tau}\mathbf{1}_{i \in T,\, j \notin T} - \frac{1-\tau}{\tau}\mathbf{1}_{i \notin T,\, j \in T} + \mathbf{1}_{i,j \notin T} \ ,
$$

Therefore,

$$
\begin{aligned}
\sum_{i=1}^n E^{(\mathrm{HO})}_{i,i} &= n\left(\tau\left(\frac{1-\tau}{\tau}\right)^2 + 1 - \tau\right) = n\frac{1-\tau}{\tau} \ , \qquad \text{(S.62)} \\
\sum_{i=1}^n \left(E^{(\mathrm{HO})}_{i,i}\right)^2 &= n\left(\tau\left(\frac{1-\tau}{\tau}\right)^4 + 1 - \tau\right) = n(1-\tau)\frac{(1-\tau)^3 + \tau^3}{\tau^3} \ . \qquad \text{(S.63)}
\end{aligned}
$$

Moreover, $E_{i,j}^{(\mathrm{HO})}$ satisfy

$$\sum_{1\le i,j\le n} E_{i,j}^{(\mathrm{HO})} = \mathbb{E}\left[\left(\sum_{i=1}^{n}\left(\frac{1}{\tau}\mathbf{1}_{i\in T}-1\right)\right)^2\right] = 0 \ , \tag{S.64}$$

so Eq. (S.62) implies that

$$\sum_{1\le i\ne j\le n}\left(E_{i,j}^{(\mathrm{HO})}\right) = -\sum_{i=1}^{n} E_{i,i}^{(\mathrm{HO})} = -n\frac{1-\tau}{\tau} \ . \tag{S.65}$$

In addition,

$$\begin{aligned}\sum_{1\le i\ne j\le n}\left(E_{i,j}^{(\mathrm{HO})}\right)^2 &= 2n^2\tau(1-\tau)\left(\frac{1-\tau}{\tau}\right)^2 + n\tau(n\tau-1)\left(\frac{1-\tau}{\tau}\right)^4 \\ &\quad + n(1-\tau)(n(1-\tau)-1) \\ &= n^2(1-\tau)^2\left(2\frac{1-\tau}{\tau}+\left(\frac{1-\tau}{\tau}\right)^2+1\right) \\ &\quad - n(1-\tau)\left(\left(\frac{1-\tau}{\tau}\right)^3+1\right) \\ &= n^2\left(\frac{1-\tau}{\tau}\right)^2 - n(1-\tau)\frac{(1-\tau)^3+\tau^3}{\tau^3} \ . \end{aligned} \tag{S.66}$$

According to (S.61) and (A.53), $\mathrm{Var}\left(\mathcal{C}^{\mathrm{ho}}_{(C,T)}(m)\right)$ can be computed using Lemma A.7 with

$$\forall i,j\in\{1,\dots,n\}\,, \quad \omega_{i,j} = \frac{1}{n^2}\left(2xE_{i,j}^{(\mathrm{HO})}-1\right) \qquad \text{and} \qquad f_m = \frac{-2s_m}{n} \ .$$

So, using Eq. (S.62), (S.63), (S.65) and (S.66), we have

$$\begin{aligned}\sum_{i=1}^{n}\omega_{i,i}^2 &= \frac{1}{n^4}\left[4x^2\sum_{i=1}^{n}\left(E_{i,i}^{(\mathrm{HO})}\right)^2 - 4x\sum_{i=1}^{n}E_{i,i}^{(\mathrm{HO})}+n\right] \\ &= \frac{1}{n^3}\left[4x^2(1-\tau)\frac{(1-\tau)^3+\tau^3}{\tau^3} - 4x\frac{1-\tau}{\tau}+1\right] \end{aligned} \tag{S.67}$$

$$\begin{aligned}\sum_{1\le i\ne j\le n}\omega_{i,j}^2 &= \frac{1}{n^4}\left[4x^2\sum_{1\le i\ne j\le n}\left(E_{i,j}^{(\mathrm{HO})}\right)^2 - 4x\sum_{1\le i\ne j\le n}E_{i,j}^{(\mathrm{HO})}+n(n-1)\right] \\ &= \frac{1}{n^4}\left[4x^2\left(n^2\left(\frac{1-\tau}{\tau}\right)^2 - n(1-\tau)\frac{(1-\tau)^3+\tau^3}{\tau^3}\right)+4xn\frac{1-\tau}{\tau}+n(n-1)\right] \end{aligned} \tag{S.68}$$

$$\sum_{i=1}^{n} \omega_{i,i} = \frac{1}{n}\left(2x\frac{1-\tau}{\tau} - 1\right) \quad . \tag{S.69}$$

Therefore, by Lemma A.7 with $m' = m$, we deduce

$$\begin{aligned}
\mathrm{Var}\left(\mathcal{C}^{\mathrm{ho}}_{(C,T)}(m)\right) &= \frac{1}{n^3}\left[4C^2\frac{(1-2\tau)^2}{\tau(1-\tau)} + (2C-1)^2\right]\mathrm{Var}\left(\Psi_m(\xi) - 2s_m(\xi)\right) \\
&\quad + \frac{2}{n^2}\left[1 + 4C^2 - \frac{1}{n}\left(4C^2\frac{(1-2\tau)^2}{\tau(1-\tau)} + (2C-1)^2\right)\right]\beta\left(\Lambda_m, \Lambda_m\right) \\
&\quad - \frac{4}{n^2}\left(2C-1\right)\mathrm{cov}\left(\Psi_m(\xi) - 2s_m(\xi), s_m(\xi)\right) + \frac{4}{n}\mathrm{Var}\left(s_m(\xi)\right) \\
&= \frac{4}{n}\mathrm{Var}\left(\left(1 + \frac{2C-1}{n}\right)s_m(\xi) - \frac{2C-1}{2n}\Psi_m(\xi)\right) \\
&\quad + \frac{2}{n^2}\left[1 + 4C^2 - \frac{(2C-1)^2}{n}\right]\beta(\Lambda_m, \Lambda_m) \\
&\quad + \frac{4C^2}{n^3}\frac{(1-2\tau)^2}{\tau(1-\tau)}\left(\mathrm{Var}\left(\Psi_m(\xi) - 2s_m(\xi)\right) - 2\beta\left(\Lambda_m, \Lambda_m\right)\right) \quad .
\end{aligned}$$

Eq (S.60) follows from the same computations. □

S.3. Additional comments on computational issues

S.3.1. Naive implementation

Algorithm S.2.

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

1. For $j \in \{1, \ldots, V\}$,
   (a) train $\widehat{s}_m(\cdot)$ with the data set $(\xi_i)_{i\notin\mathcal{B}_j}$, that is, for all $\lambda \in \Lambda_m$, compute $\alpha_{\lambda,j} := P_n^{(-\mathcal{B}_j)}(\psi_\lambda) = \frac{V}{(V-1)n}\sum_{i\notin\mathcal{B}_j}\psi_\lambda(\xi_i)$ so that $\widehat{s}_m^{(-\mathcal{B}_j)} = \sum_{\lambda\in\Lambda_m}\alpha_{\lambda,j}\psi_\lambda$
   (b) compute the norm of $\widehat{s}_m^{(-\mathcal{B}_j)}$: $N_j := \sum_{\lambda\in\Lambda_m}\alpha_{\lambda,j}^2$
   (c) compute $Q_j := P_n^{(\mathcal{B}_j)}\left(\widehat{s}_m^{(-\mathcal{B}_j)}\right) = \frac{V}{n}\sum_{\lambda\in\Lambda_m}\sum_{i\in\mathcal{B}_j}\alpha_{\lambda,j}\psi_\lambda(\xi_i)$
   (d) compute $R_j := P_n^{(-\mathcal{B}_j)}\left(\widehat{s}_m^{(-\mathcal{B}_j)}\right) = \frac{V}{n(V-1)}\sum_{\lambda\in\Lambda_m}\sum_{i\notin\mathcal{B}_j}\alpha_{\lambda,j}\psi_\lambda(\xi_i)$
2. Compute the V-fold cross-validation criterion: $\mathcal{C} = V^{-1}\sum_{j=1}^{V}(N_j - 2Q_j)$
3. Empirical risk:
   (a) Train $\widehat{s}_m(\cdot)$ with the data set $(\xi_i)_{1\le i\le n}$, that is, for all $\lambda \in \Lambda_m$, compute $\alpha_\lambda := P_n(\psi_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(\xi_i)$ so that $\widehat{s}_m = \sum_{\lambda\in\Lambda_m}\alpha_\lambda\psi_\lambda$
   (b) compute the norm of $\widehat{s}_m$: $N := \sum_{\lambda\in\Lambda_m}\alpha_\lambda^2$

(c) compute $R := \frac{1}{n} \sum_{\lambda \in \Lambda_m} \sum_{i=1}^{n} \alpha_\lambda \psi_\lambda(\xi_i)$

4. Compute the V-fold penalty: $\mathcal{D} := 2(V-1)V^{-2} \sum_{j=1}^{V} (Q_j - R_j)$
   Output:
   Empirical risk: $N - 2R$
   V-fold cross-validation estimator of the risk of $\widehat{s}_m$: $\mathrm{crit}_{\mathrm{VFCV}}(m) = \mathcal{C}$
   V-fold penalty: $\mathrm{pen}_{\mathrm{VF}}(m) = \mathcal{D}$.

Assuming the computational cost of evaluation ψ_λ at some point $\xi \in \Xi$ is of order 1, the computational cost of this naive algorithm S.2 is as follows: $n(V-1)d_m$ for step 1, V for steps 2 and 4, nd_m for step 3. So the overall cost of computing the V-fold penalization criterion for m is of order nVd_m.

S.3.2. Proof of Proposition 3

Let us first note that for every $i \in \{1, \dots, V\}$ and $\lambda \in \Lambda_m$, $A_{i,\lambda} = P_n^{(\mathcal{B}_i)}(\psi_\lambda)$. So, at step 2, for every $i, j \in \{1, \dots, V\}$,

$$C_{i,j} = \sum_{\lambda \in \Lambda_m} P_n^{(\mathcal{B}_i)}(\psi_\lambda) P_n^{(\mathcal{B}_j)}(\psi_\lambda) = P_n^{(\mathcal{B}_i)} \left(\sum_{\lambda \in \Lambda_m} P_n^{(\mathcal{B}_j)}(\psi_\lambda) \psi_\lambda \right) = P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_j)})$$

and by symmetry $C_{i,j} = C_{j,i} = P_n^{(\mathcal{B}_j)}(\widehat{s}_m^{(\mathcal{B}_i)})$.

Correctness of Algorithm 1 By assumption (**Reg**), we have

$$P_n = \frac{1}{V} \sum_{j=1}^{V} P_n^{(\mathcal{B}_j)} \quad , \qquad \widehat{s}_m = \frac{1}{V} \sum_{j=1}^{V} \widehat{s}_m^{(\mathcal{B}_j)} \quad ,$$

$$P_n^{(-\mathcal{B}_i)} = \frac{1}{V-1} \sum_{\substack{1 \le j \le V \\ j \ne i}} P_n^{(\mathcal{B}_j)} \quad \text{and} \quad \widehat{s}_m^{(-\mathcal{B}_i)} = \frac{1}{V-1} \sum_{\substack{1 \le j \le V \\ j \ne i}} \widehat{s}_m^{(\mathcal{B}_j)} \quad .$$

Therefore,

$$\|\widehat{s}_m\|^2 = -P_n \gamma(\widehat{s}_m) = P_n(\widehat{s}_m) = \frac{1}{V^2} \sum_{1 \le i,j \le V} P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_j)}) = \frac{1}{V^2} \mathcal{S}$$

and

$$\begin{aligned}
\mathrm{crit}_{\mathrm{VFCV}}(m) &= \frac{1}{V} \sum_{j=1}^{V} P_n^{(\mathcal{B}_j)} \gamma\left(\widehat{s}_m^{(-\mathcal{B}_j)} \right) \\
&= \frac{1}{V} \sum_{j=1}^{V} \left(\left\| \widehat{s}_m^{(-\mathcal{B}_j)} \right\|^2 - 2 P_n^{(\mathcal{B}_j)} \left(\widehat{s}_m^{(-\mathcal{B}_j)} \right) \right) \\
&= \frac{1}{V} \sum_{j=1}^{V} \left(\frac{1}{(V-1)^2} \sum_{\substack{1 \le i,\ell \le V \\ i,\ell \ne j}} P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)}) - \frac{2}{V-1} \sum_{i \ne j} P_n^{(\mathcal{B}_j)} \left(\widehat{s}_m^{(\mathcal{B}_i)} \right) \right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{V(V-1)^2} \sum_{1\le i,\ell\le V} \left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)}) \sum_{j=1}^{V} \mathbf{1}_{i\neq j\,,\,\ell\neq j} \right) - \frac{2}{V(V-1)} \sum_{1\le i\neq j\le V} P_n^{(\mathcal{B}_j)}\left(\widehat{s}_m^{(\mathcal{B}_i)}\right) \\
&= \frac{1}{V(V-1)^2} \sum_{1\le i,\ell\le V} \left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)})(V-1-\mathbf{1}_{i\neq\ell}) \right) - \frac{2}{V(V-1)} (\mathcal{S}-\mathcal{T}) \\
&= \frac{1}{V(V-1)} \sum_{1\le i\le V} \left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_i)}) \right) + \frac{V-2}{V(V-1)^2} \sum_{1\le i\neq\ell\le V} \left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)}) \right) \\
&\quad - \frac{2}{V(V-1)} (\mathcal{S}-\mathcal{T}) \\
&= \frac{1}{V(V-1)} \mathcal{T} + \frac{V-2}{V(V-1)^2} (\mathcal{S}-\mathcal{T}) - \frac{2}{V(V-1)} (\mathcal{S}-\mathcal{T}) \\
&= \frac{1}{V(V-1)} \mathcal{T} - \frac{1}{(V-1)^2} (\mathcal{S}-\mathcal{T}) \quad ,
\end{aligned}
$$

so the formula for $\mathrm{crit}_{\mathrm{VFCV}}$ is correct. Lemma 1 implies the formula for $\mathrm{pen}_{\mathrm{VF}}$ is also correct.

Computational cost of Algorithm 1 Step 1 has a cost of order $V \times \mathrm{Card}(\Lambda_m) \times (n/V) = n\,\mathrm{Card}(\Lambda_m)$. Step 2 has a cost of order $V^2\,\mathrm{Card}(\Lambda_m)$. Step 3 has a cost of order V^2. Summing the three steps yields the result.

Computational cost for histograms In the histogram case, step 1 can be performed with a cost of order $V\,\mathrm{Card}(\Lambda_m) + n$. Indeed, one can initialize the $V \times \mathrm{Card}(\Lambda_m)$ matrix A with zeros (cost: $V\,\mathrm{Card}(\Lambda_m)$), and then go sequentially through the data set: for $j = 1, \dots, n$, find the unique $i(j) \in \{1, \dots, V\}$ such that $j \in \mathcal{B}_{i(j)}$, the unique $\lambda(j) \in \Lambda_m$ such that $\xi_j \in \lambda(j)$, and add $(V/n)\psi_\lambda(\xi_j)$ to $A_{(i(j),\lambda(j))}$. Since the partitions $\mathcal{B}$ and Λ_m can be coded so that finding $i(j)$ and $\lambda(j)$ has a cost of order 1, the resulting cost of step 1 is $V\,\mathrm{Card}(\Lambda_m) + n$, hence the overall cost is of order $V^2\,\mathrm{Card}(\Lambda_m) + n$. □

S.4. Probabilistic Tool

Proposition S.14 ([Ler11])**.** *Let $\xi_{[\![N]\!]}$ be iid random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$, with common distribution P. Let S be a symmetric class of functions bounded by b. For all $t \in S$, let $P_N t = N^{-1} \sum_{i=1}^N t(\xi_i)$, $v^2 = \sup_{t\in S} P[(t - Pt)^2]$, $Z = \sup_{t\in S}(P_N - P)t$, $D = N\mathbb{E}(Z^2)$. There exists an absolute constant κ such that, for all $x > 0$, with probability larger than $1 - 2e^{-x}$, for all $\epsilon \in (0, 1]$,*

$$
\left| Z^2 - \frac{D}{N} \right| \le \epsilon \frac{D}{N} + \kappa \left(\frac{v^2 x}{\epsilon N} + \frac{b^2 x^2}{\epsilon^3 N^2} \right) \quad .
$$

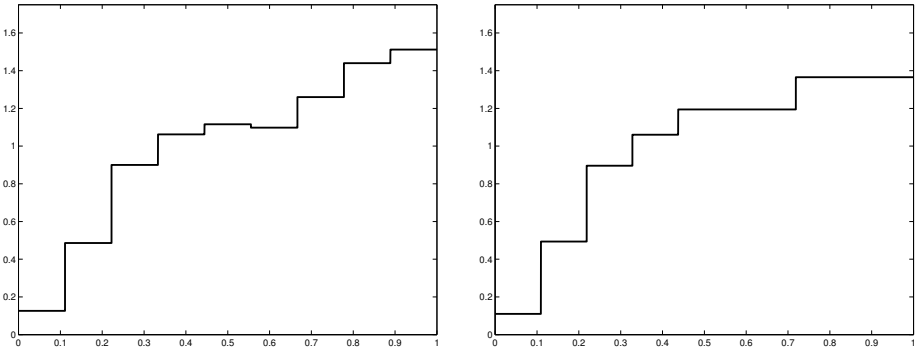

FIGURE S.6. *Oracle model for some sample of size $n = 500$, in setting L. Left: Regu. Right: Dya2.*

S.5. Additional simulation results

This section provides simulation results in addition to the ones of Section 6. Figure S.6 is an analogous of Figure 2 in setting L, that illustrates the difference between the model collections Regu and Dya2. Table S.3 is an extended version of Table 2, with more procedures compared and two additional settings (L-Regu and S-Regu). Table S.4 provides a similar comparison of model selection performances with a reduced sample size $n = 100$, again from $N = 10\,000$ independent samples.

The study of variance of Section 6.4 (setting S with $n = 100$) is completed with Figure S.7, which tests the validity of the heuristic of Section 4, Figure S.8, which is the equivalent of Figure 4 without zooming on the smallest dimensions, and Figure S.9, which shows that

$$\forall m \neq m^\star\,, \quad \mathrm{SR}(m) \approx \frac{\mathbb{E}\left[\Delta(m, m^\star)\right]}{\sqrt{\mathrm{var}\left(\Delta(m, m^\star)\right)}}\,.$$

TABLE S.3
Simulation results: settings L and S, $n = 500$. The best procedures (up to standard-deviations) are bolded, where the data-driven procedures are considered separately from the procedures using the knowledge of $\mathbb{E}[\,\mathrm{pen}_{\mathrm{id}}\,]$.

Experiment	L–Dya2	L–Regu	S–Dya2	S–Regu
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	6.52 ± 0.05	2.33 ± 0.01	2.07 ± 0.01	1.75 ± 0.01
$1.25 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	4.81 ± 0.04	2.01 ± 0.01	1.94 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	4.12 ± 0.03	$\mathbf{1.93 \pm 0.01}$	$\mathbf{1.92 \pm 0.01}$	1.65 ± 0.003
$2 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	$\mathbf{3.61 \pm 0.02}$	1.96 ± 0.01	2.01 ± 0.01	1.84 ± 0.004
$\mathrm{pen}_{\mathrm{dim}}$	8.27 ± 0.07	2.33 ± 0.01	3.21 ± 0.01	1.75 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{dim}}$	5.95 ± 0.05	2.01 ± 0.01	3.01 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \mathrm{pen}_{\mathrm{dim}}$	4.99 ± 0.04	$\mathbf{1.94 \pm 0.01}$	3.03 ± 0.01	1.66 ± 0.003
$2 \times \mathrm{pen}_{\mathrm{dim}}$	4.38 ± 0.03	1.97 ± 0.01	3.24 ± 0.01	1.85 ± 0.004
$\mathrm{pen}_{\mathrm{LOO}}$	6.35 ± 0.05	2.33 ± 0.01	2.06 ± 0.01	1.75 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{LOO}}$	4.62 ± 0.04	2.01 ± 0.01	1.92 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \mathrm{pen}_{\mathrm{LOO}}$	3.97 ± 0.03	$\mathbf{1.94 \pm 0.01}$	$\mathbf{1.90 \pm 0.005}$	1.66 ± 0.003
$2 \times \mathrm{pen}_{\mathrm{LOO}}$	$\mathbf{3.55 \pm 0.02}$	1.97 ± 0.01	1.98 ± 0.01	1.85 ± 0.004
$\mathrm{pen}_{\mathrm{VF}}$ (V=10)	6.89 ± 0.06	2.42 ± 0.02	2.11 ± 0.01	1.77 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	5.01 ± 0.04	2.04 ± 0.01	1.95 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	4.27 ± 0.03	1.94 ± 0.01	1.92 ± 0.01	1.63 ± 0.004
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	3.68 ± 0.02	1.94 ± 0.01	1.98 ± 0.01	1.78 ± 0.004
$\mathrm{pen}_{\mathrm{VF}}$ (V=5)	7.47 ± 0.06	2.55 ± 0.02	2.16 ± 0.01	1.80 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	5.50 ± 0.04	2.10 ± 0.01	1.98 ± 0.01	1.63 ± 0.004
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	4.58 ± 0.03	1.96 ± 0.01	1.93 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	3.86 ± 0.02	$\mathbf{1.93 \pm 0.01}$	1.98 ± 0.01	1.73 ± 0.004
$\mathrm{pen}_{\mathrm{VF}}$ (V=2)	10.21 ± 0.08	3.37 ± 0.03	2.39 ± 0.01	2.01 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	7.69 ± 0.06	2.49 ± 0.02	2.15 ± 0.01	1.71 ± 0.01
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	6.41 ± 0.05	2.18 ± 0.01	2.05 ± 0.01	1.63 ± 0.004
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	5.11 ± 0.04	1.99 ± 0.01	2.04 ± 0.01	1.64 ± 0.004
LOO	6.34 ± 0.05	2.33 ± 0.01	2.06 ± 0.01	1.75 ± 0.01
10-fold CV	6.24 ± 0.05	2.29 ± 0.01	2.05 ± 0.01	1.71 ± 0.01
5-fold CV	6.27 ± 0.05	2.26 ± 0.01	2.05 ± 0.01	1.68 ± 0.01
2-fold CV	6.41 ± 0.05	2.18 ± 0.01	2.05 ± 0.01	1.63 ± 0.004
Oracle: $10^{-3}\times$	5.46 ± 0.02	13.39 ± 0.05	43.86 ± 0.09	62.37 ± 0.13
Best: $10^{-3}\times$	19.38 ± 0.10	25.77 ± 0.10	83.39 ± 0.22	100.86 ± 0.23

TABLE S.4
Simulation results: settings L and S, $n = 100$. The best procedures (up to standard-deviations) are bolded, where the data-driven procedures are considered separately from the procedures using the knowledge of $\mathbb{E}[\,\mathrm{pen}_{\mathrm{id}}\,]$.

Experiment	L–Dya2	L–Regu	S–Dya2	S–Regu
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	8.38 ± 0.08	3.29 ± 0.03	1.97 ± 0.01	2.09 ± 0.01
$1.25 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	6.53 ± 0.07	2.61 ± 0.02	$\mathbf{1.93 \pm 0.01}$	1.72 ± 0.01
$1.5 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	5.59 ± 0.06	$\mathbf{2.46 \pm 0.02}$	$\mathbf{1.92 \pm 0.01}$	1.61 ± 0.01
$2 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	$\mathbf{4.72 \pm 0.05}$	2.57 ± 0.01	1.94 ± 0.005	$\mathbf{1.60 \pm 0.004}$
$\mathrm{pen}_{\mathrm{dim}}$	9.67 ± 0.09	3.28 ± 0.03	2.17 ± 0.01	2.09 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{dim}}$	7.85 ± 0.08	2.62 ± 0.02	2.10 ± 0.01	1.72 ± 0.01
$1.5 \times \mathrm{pen}_{\mathrm{dim}}$	6.74 ± 0.07	$\mathbf{2.48 \pm 0.02}$	2.05 ± 0.01	1.62 ± 0.01
$2 \times \mathrm{pen}_{\mathrm{dim}}$	5.70 ± 0.06	2.60 ± 0.01	2.00 ± 0.01	1.61 ± 0.004
$\mathrm{pen}_{\mathrm{LOO}}$	8.10 ± 0.08	3.29 ± 0.03	1.97 ± 0.01	2.09 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{LOO}}$	6.20 ± 0.06	2.62 ± 0.02	1.92 ± 0.01	1.72 ± 0.01
$1.5 \times \mathrm{pen}_{\mathrm{LOO}}$	5.18 ± 0.05	$\mathbf{2.49 \pm 0.02}$	1.91 ± 0.01	1.62 ± 0.01
$2 \times \mathrm{pen}_{\mathrm{LOO}}$	$\mathbf{4.44 \pm 0.04}$	2.59 ± 0.01	1.94 ± 0.005	1.61 ± 0.004
$\mathrm{pen}_{\mathrm{VF}}$ (V=10)	8.61 ± 0.08	3.54 ± 0.04	1.97 ± 0.01	2.21 ± 0.01
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	6.76 ± 0.07	2.76 ± 0.02	1.92 ± 0.01	1.78 ± 0.01
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	5.77 ± 0.06	2.52 ± 0.02	$\mathbf{1.90 \pm 0.01}$	1.64 ± 0.01
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	4.81 ± 0.05	2.57 ± 0.01	1.91 ± 0.01	$\mathbf{1.60 \pm 0.004}$
$\mathrm{pen}_{\mathrm{VF}}$ (V=5)	9.14 ± 0.08	3.92 ± 0.04	1.98 ± 0.01	2.34 ± 0.02
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	7.38 ± 0.07	2.90 ± 0.03	1.93 ± 0.01	1.85 ± 0.01
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	6.31 ± 0.06	2.60 ± 0.02	$\mathbf{1.91 \pm 0.01}$	1.68 ± 0.01
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	5.21 ± 0.05	2.56 ± 0.02	$\mathbf{1.90 \pm 0.01}$	$\mathbf{1.60 \pm 0.005}$
$\mathrm{pen}_{\mathrm{VF}}$ (V=2)	11.15 ± 0.09	6.14 ± 0.08	2.01 ± 0.01	2.92 ± 0.02
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	9.61 ± 0.08	4.05 ± 0.05	1.97 ± 0.01	2.24 ± 0.01
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	8.60 ± 0.07	3.30 ± 0.03	1.94 ± 0.01	1.94 ± 0.01
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	7.30 ± 0.07	2.80 ± 0.02	1.91 ± 0.01	1.70 ± 0.01
LOO	8.04 ± 0.08	3.26 ± 0.03	1.97 ± 0.01	2.07 ± 0.01
10-fold CV	8.11 ± 0.08	3.28 ± 0.03	1.95 ± 0.01	2.06 ± 0.01
5-fold CV	8.15 ± 0.08	3.28 ± 0.03	1.95 ± 0.01	2.01 ± 0.01
2-fold CV	8.60 ± 0.07	3.30 ± 0.03	1.94 ± 0.01	1.94 ± 0.01
Oracle: $10^{-3}\times$	12.66 ± 0.05	33.58 ± 0.16	118.21 ± 0.25	133.04 ± 0.28
Best: $10^{-3}\times$	56.15 ± 0.53	83.42 ± 0.51	224.09 ± 0.63	212.84 ± 0.61

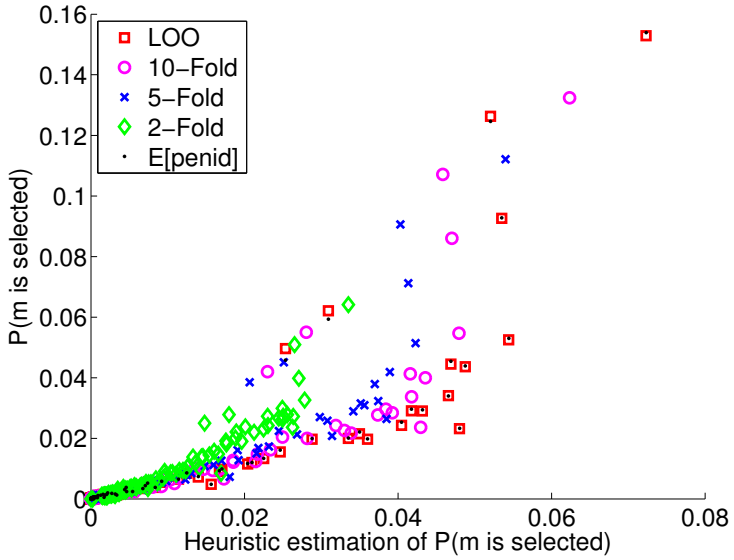

FIGURE S.7. *Illustration of the variance heuristic:* $\mathbb{P}(\widehat{m} = m)$ *as a function of* $\overline{\Phi}(\mathrm{SR}(m))$ *(renormalized to have a sum equal to one). Setting S-Regu,* $n = 100$.

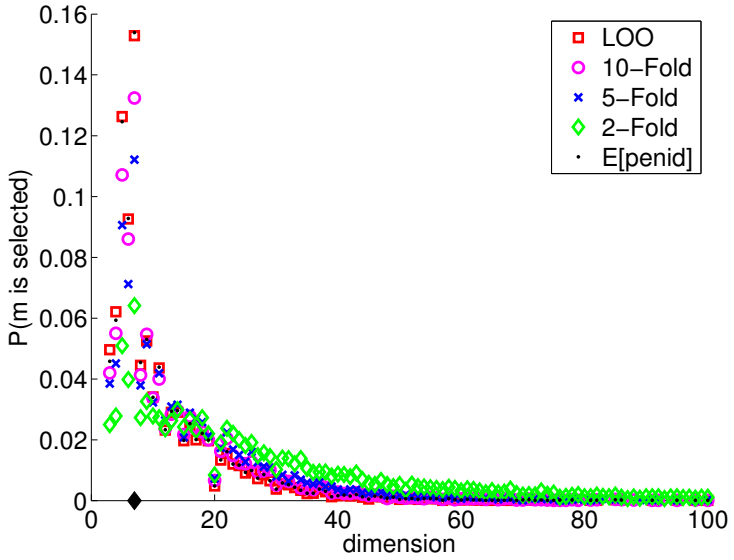

FIGURE S.8. *Setting S-Regu,* $n = 100$. $\mathbb{P}(\widehat{m} = m)$ *as a function of* m. *The black diamond shows* $m^\star = 7$.

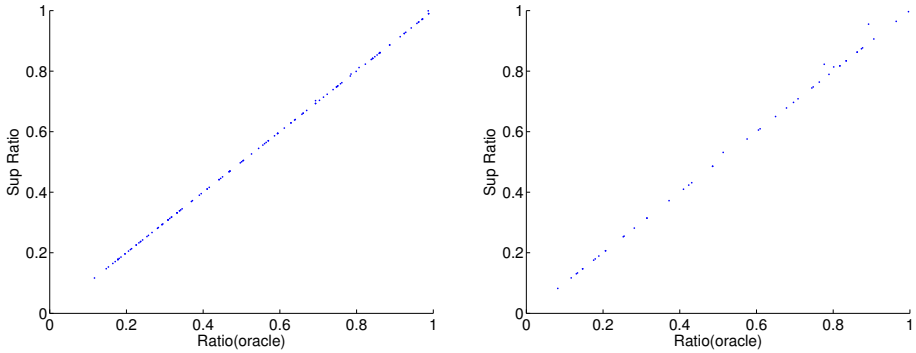

FIGURE S.9. SR(m) *as a function of the ratio at* $m' = m^\star$. $n = 100$. *Left: S-Regu. Right: L-Regu.*

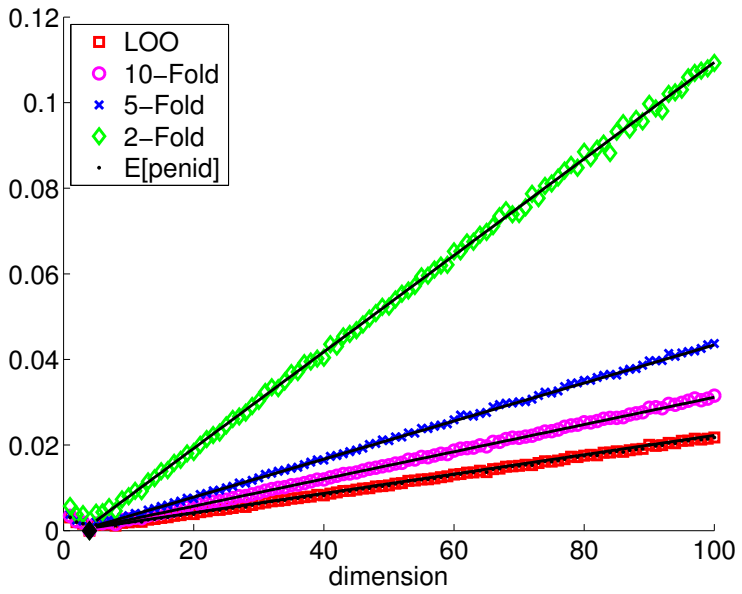

FIGURE S.10. *L-Regu,* $n = 100$. $\mathrm{var}(\Delta_{\mathcal{C}}(m, m^\star))$ *as a function of* m. *The black lines show the linear approximation* $n^{-2}[5.6(1 + \frac{1.1}{V-1}) + 2.2(1 + \frac{4.2}{V-1})(m - m^\star)]$ *for* $m > m^\star = 4$.

All graphs plotted about the variance in setting S with $n = 100$ are also provided for setting L with $n = 100$, based upon $N = 10\,000$ independent samples, see Figures S.9–S.14. The case of a sample size $n = 500$ has also been considered in both settings S and L, based upon $N = 1\,000$ independent samples, see Figures S.15–S.25.

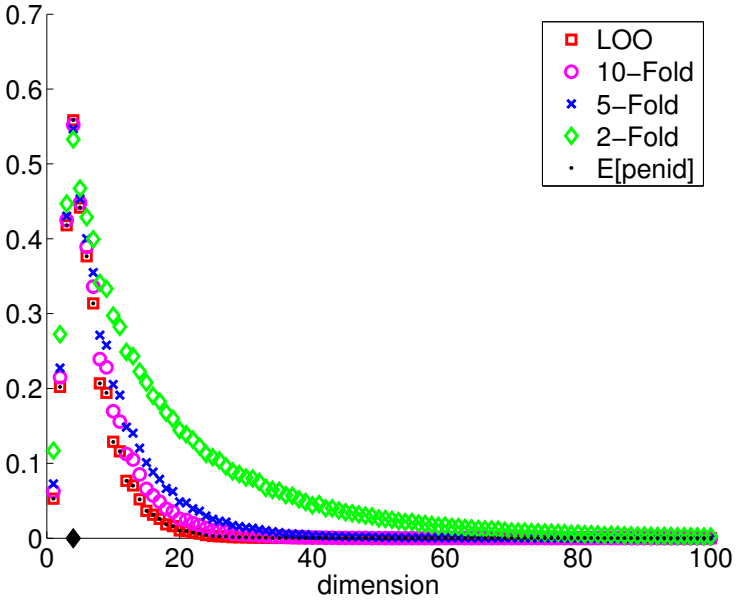

FIGURE S.11. *L-Regu*, $n = 100$. $\overline{\Phi}(\mathrm{SR}_{\mathcal{C}}(m))$ *as a function of* m.

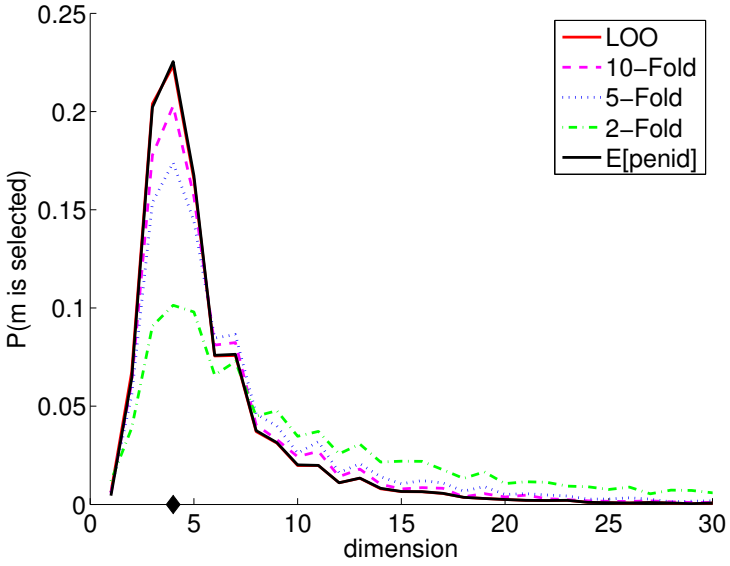

FIGURE S.12. *L-Regu*, $n = 100$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ *as a function of* m.

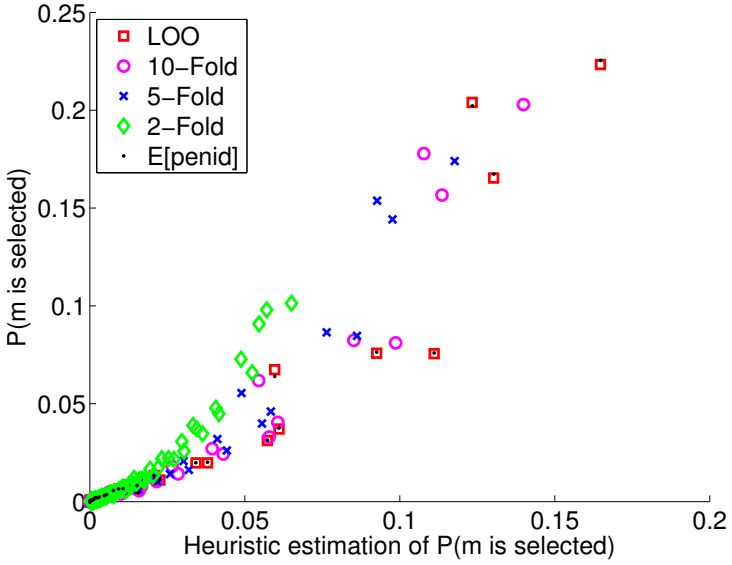

FIGURE S.13. *L-Regu*, $n = 100$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ *as a function of* $\overline{\Phi}(\mathrm{SR}_{\mathcal{C}}(m))$.

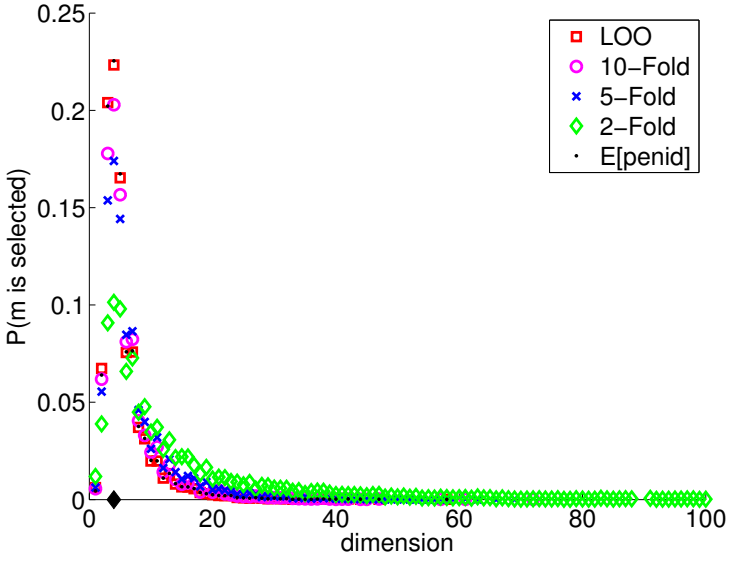

FIGURE S.14. *L-Regu*, $n = 100$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ *as a function of* m.

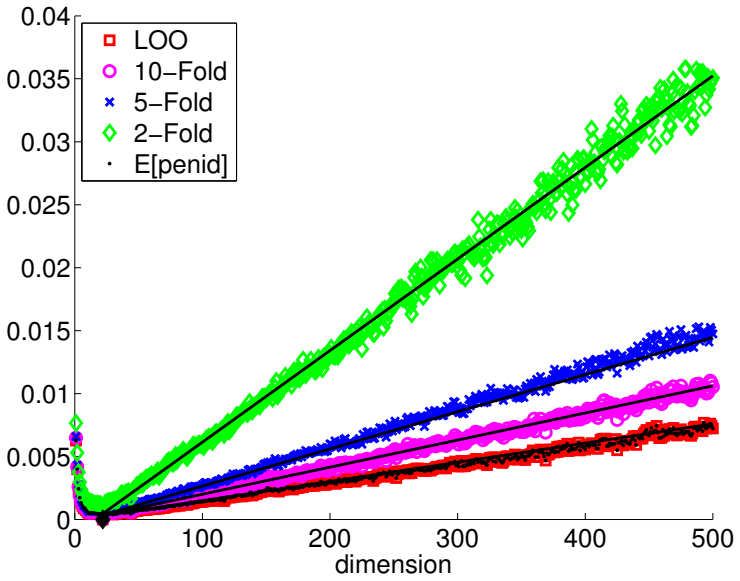

FIGURE S.15. *S-Regu,* $n = 500$. $\mathrm{var}(\Delta_{\mathcal{C}_V}(m, m^\star))$ *as a function of* m. *The black lines show the linear approximation* $n^{-2}[75(1 + \frac{0.52}{V-1}) + 3.8(1 + \frac{3.8}{V-1})(m - m^\star)]$ *for* $m > m^\star = 22$.

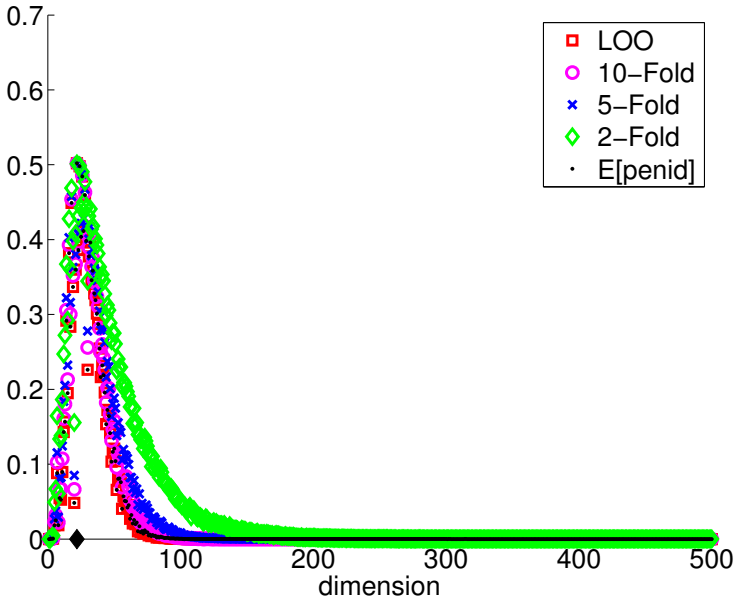

FIGURE S.16. *S-Regu,* $n = 500$. $\overline{\Phi}(\mathrm{SR}_{\mathcal{C}_V}(m))$ *as a function of* m.

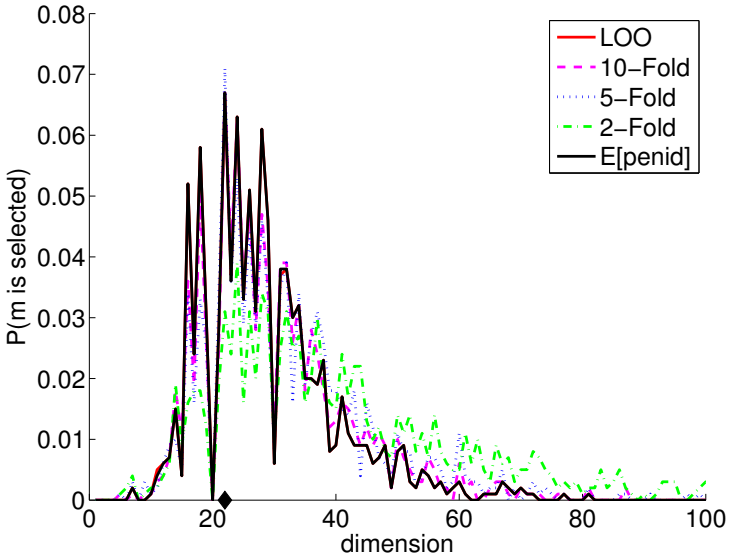

FIGURE S.17. *S-Regu*, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ *as a function of* m.

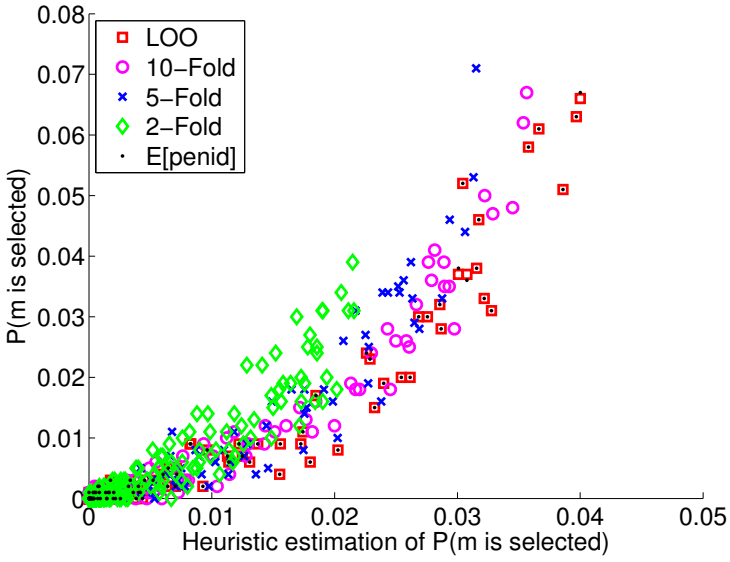

FIGURE S.18. *S-Regu*, $n = 500$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ *as a function of* $\overline{\Phi}(\mathrm{SR}(m))$.

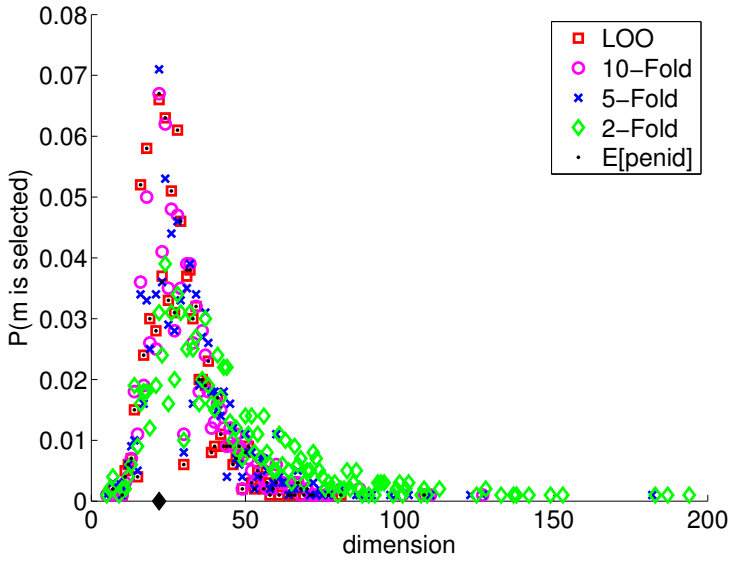

FIGURE S.19. *S-Regu*, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ *as a function of* m.

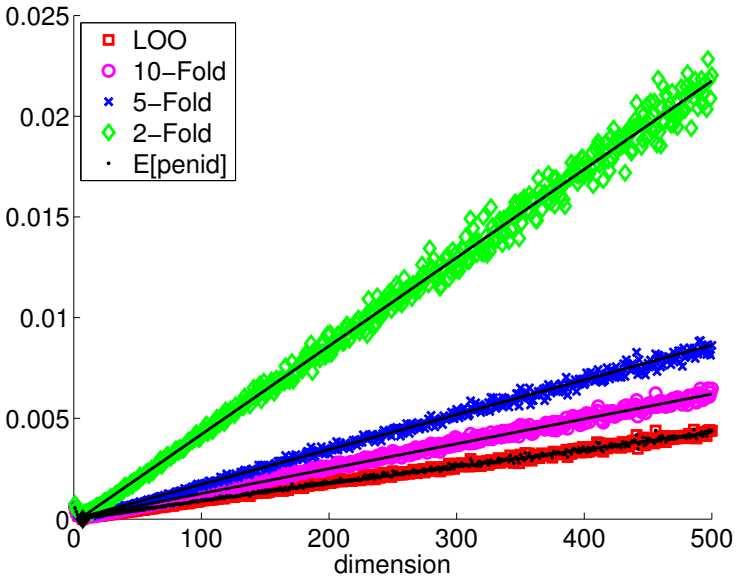

FIGURE S.20. *L-Regu*, $n = 500$. $\mathrm{var}(\Delta_{\mathcal{C}_V}(m, m^\star))$ *as a function of* m. *The black lines show the linear approximation* $n^{-2}[28(1 + \frac{0.06}{V-1}) + 2.1(1 + \frac{4.2}{V-1})(m - m^\star)]$ *for* $m > m^\star = 7$.

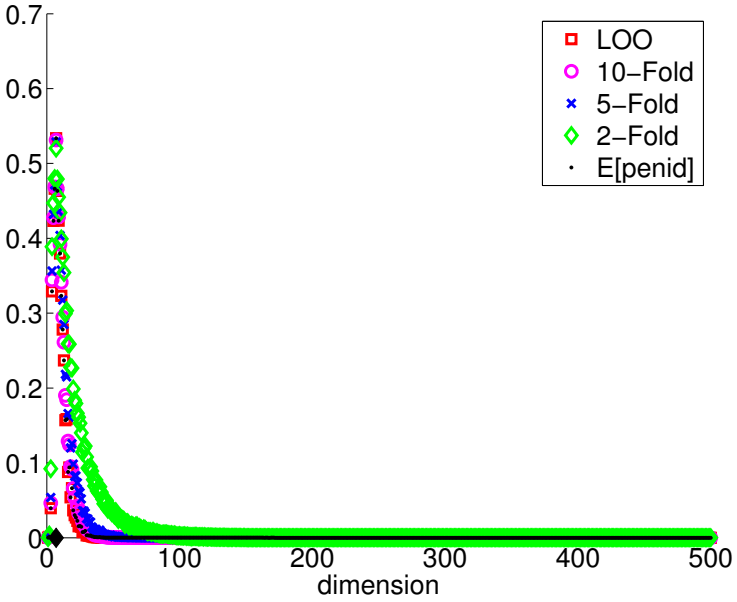

FIGURE S.21. *L-Regu*, $n = 500$. $\overline{\Phi}(\mathrm{SR}_{\mathcal{C}_V}(m))$ *as a function of* m.

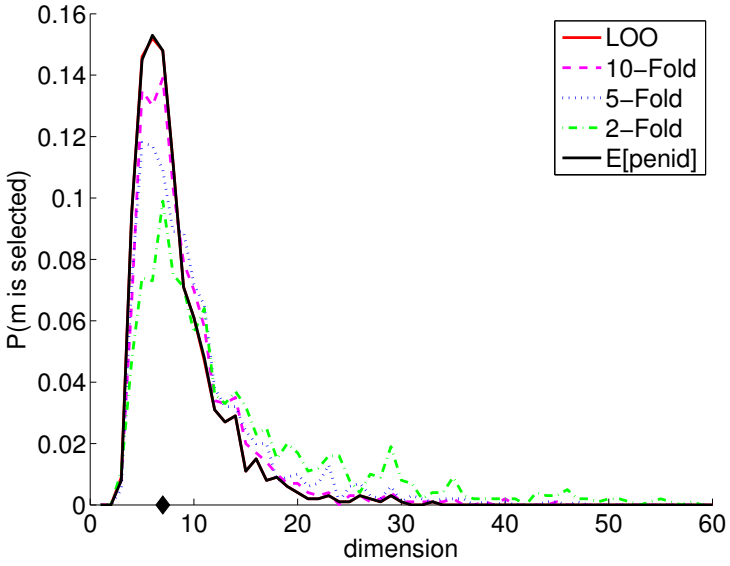

FIGURE S.22. *L-Regu*, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ *as a function of* m.

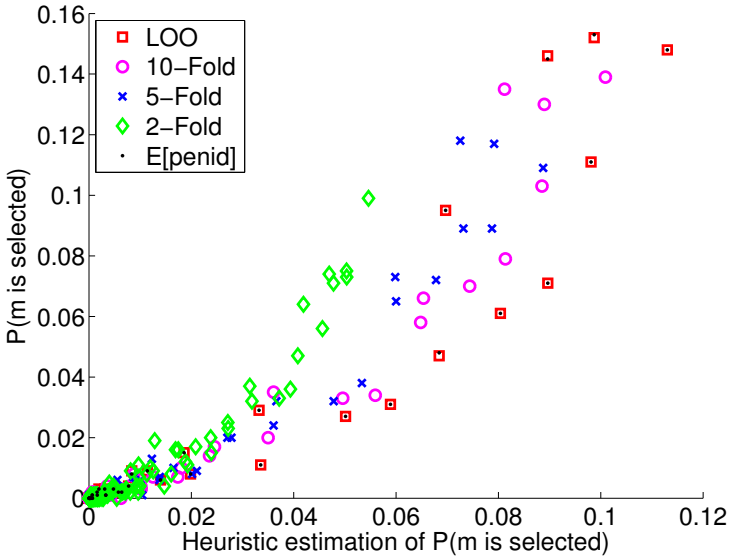

FIGURE S.23. *L-Regu*, $n = 500$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ *as a function of* $\overline{\Phi}(\mathrm{SR}(m))$.

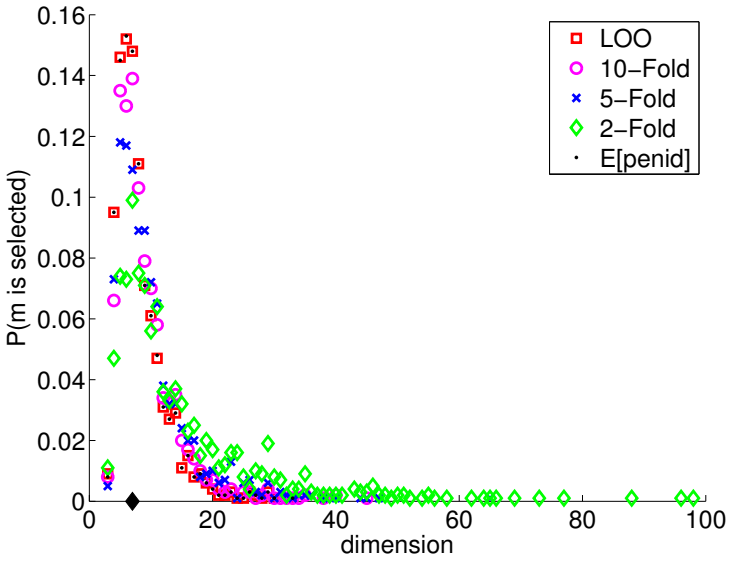

FIGURE S.24. *L-Regu*, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ *as a function of* m.

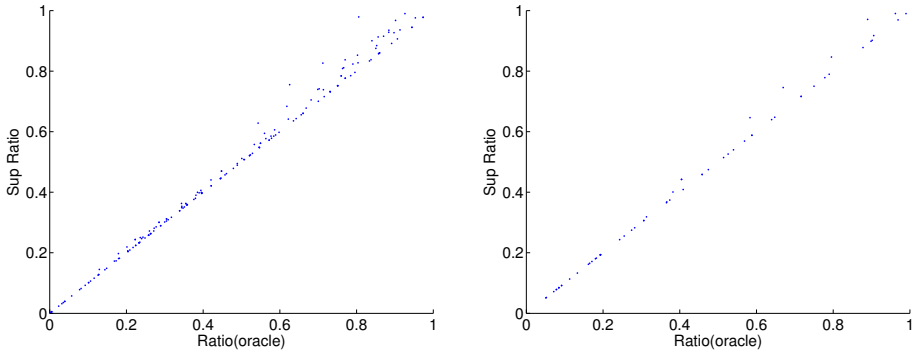

FIGURE S.25. SR(m) *as a function of the ratio at* $m' = m^\star$. $n = 500$. *Left: S-Regu. Right: L-Regu.*