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

Online Appendix to the Article "Choice of V for V-Fold Cross-Validation in Least-Squares Density Estimation"

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

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

Editor: Xiaotong Shen

This appendix is organized as follows. The first section (called Section B, for consistency with the numbering of the article) gives complementary computations of variances. Then, results concerning hold-out penalization are detailed in Section D, with the proof of the oracle inequality stated in Section 8.2 (Theorem 12) and an exact computation of the variance. Section E 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 8. A useful concentration inequality is recalled in Section F. Finally, some simulation results are detailed in Section G, as a supplement to the ones of Section 6.

Appendix B. Additional Variance Computations

Proposition 17 *Let $(\psi_\lambda)_{\lambda \in \Lambda_{m_1}}$ and $(\psi_\lambda)_{\lambda \in \Lambda_{m_2}}$ be two finite orthonormal families of vectors of $L^4(\mu)$. Assume that $\mathcal{B}$ satisfies* (**Reg**) *and, for any $m \in \{m_1, m_2\}$, let*

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

Then, with the notation of Theorem 6,

$$\mathrm{Var}\big(\mathcal{C}_{id}(m_1)\big) = \frac{2(n-1)}{n^3}\beta(m_1, m_1) + \frac{2}{n}\mathrm{Var}\left(\left(1 - \frac{1}{n}\right)s_{m_1}(\xi) + \frac{1}{2n}\Psi_{m_1}(\xi)\right) \enspace .$$

We also have

$$\begin{aligned}\mathrm{Var}\big(\mathcal{C}_{id}(m_1) - \mathcal{C}_{id}(m_2)\big) &= \frac{2(n-1)}{n^3}\mathbf{B}(m_1, m_2) \\ &\quad + \frac{2}{n}\mathrm{Var}\left(\left(1 - \frac{1}{n}\right)\big(s_{m_1}(\xi) - s_{m_2}(\xi)\big) + \frac{1}{2n}\big(\Psi_{m_1}(\xi) - \Psi_{m_2}(\xi)\big)\right) \enspace .\end{aligned}$$

©2016 Sylvain Arlot and Matthieu Lerasle.

Proof Simply notice that

$$\operatorname{Var}\bigl(\mathcal{C}_{id}(m_1)\bigr) = \operatorname{Var}\bigl(P_n \gamma(\widehat{s}_{m_1})\bigr) \enspace .$$

Therefore, from (57), the variance of $\mathcal{C}_{id}(m_1)$ is the one of

$$-\frac{1}{n^2} \sum_{1 \leqslant i,j \leqslant n} U_{m_1}(\xi_i, \xi_j) - \sum_{i=1}^{n} \frac{2 s_{m_1}(\xi_i)}{n} \enspace .$$

so that, by Lemma 16,

$$\begin{aligned}
\operatorname{Var}\bigl(\mathcal{C}_{id}(m_1)\bigr) &= \frac{2(n-1)}{n^3} \beta(m_1, m_1) + \frac{1}{n^3} \operatorname{Var}\bigl(\Psi_{m_1}(\xi) - 2 s_{m_1}(\xi)\bigr) \\
&\quad + \frac{4}{n^2} \sum_{i=1}^{n} \operatorname{Cov}\bigl(\Psi_{m_1}(\xi) - 2 s_{m_1}(\xi), s_{m_1}(\xi)\bigr) + \frac{4}{n} \operatorname{Var}\bigl(s_{m_1}(\xi)\bigr) \\
&= \frac{2(n-1)}{n^3} \beta(m_1, m_1) + \frac{2}{n} \operatorname{Var}\left(\left(1 - \frac{1}{n}\right) s_{m_1}(\xi) + \frac{1}{n} \Psi_{m_1}(\xi) \right) \enspace .
\end{aligned}$$

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

B.1 Evaluation of the Terms in the Variance Formula

The following proposition gives a formula for the terms appearing in Theorem 6 and Proposition 17 which does not depend on the basis $(\psi_\lambda)_{\lambda \in \Lambda_m}$.

Proposition 18 *For any $m_1, m_2 \in \mathcal{M}_n$, we have*

$$\begin{aligned}
\beta(m_1, m_2) &= n \operatorname{Cov}\bigl(\widehat{s}_{m_1}(\xi), \widehat{s}_{m_2}(\xi)\bigr) - (n+1) \operatorname{Cov}\bigl(s_{m_1}(\xi), s_{m_2}(\xi)\bigr) \\
\mathbf{B}(m_1, m_2) &= n \operatorname{Var}\bigl((\widehat{s}_{m_1} - \widehat{s}_{m_2})(\xi)\bigr) - (n+1) \operatorname{Var}\bigl((s_{m_1} - s_{m_2})(\xi)\bigr) \enspace ,
\end{aligned} \tag{59}$$

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

Proof By definition, we have

$$\begin{aligned}
\beta(m_1, m_2) &= \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \operatorname{Cov}(\psi_\lambda(\xi_1), \psi_{\lambda'}(\xi_1))^2 \\
&= \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \bigl(P(\psi_\lambda \psi_{\lambda'}) - P\psi_\lambda P\psi_{\lambda'}\bigr)^2 \\
&= \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \bigl(P(\psi_\lambda \psi_{\lambda'})\bigr)^2 - 2 \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} P\psi_\lambda P\psi_{\lambda'} P(\psi_\lambda \psi_{\lambda'}) \\
&\quad + \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} (P\psi_\lambda P\psi_{\lambda'})^2 \\
&= \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \bigl(P(\psi_\lambda \psi_{\lambda'})\bigr)^2 - 2P(s_{m_1} s_{m_2}) + \|s_{m_1}\|^2 \|s_{m_2}\|^2.
\end{aligned}$$

Now, by Eq. (31), we have

$$
\begin{aligned}
&\operatorname{Cov}\bigl(\widehat{s}_{m_1}(\xi), \widehat{s}_{m_2}(\xi)\bigr) \\
&= \frac{1}{n^2} \sum_{1 \leqslant i,j \leqslant n} \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \operatorname{Cov}\bigl(\psi_\lambda(\xi_i)\psi_\lambda(\xi), \psi_{\lambda'}(\xi_j)\psi_{\lambda'}(\xi)\bigr) \\
&= \frac{1}{n} \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \bigl(P(\psi_\lambda \psi_{\lambda'})\bigr)^2 - (P\psi_\lambda P\psi_{\lambda'})^2 \\
&\qquad + \frac{n-1}{n} \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \bigl(P(\psi_\lambda \psi_{\lambda'}) - P\psi_\lambda P\psi_{\lambda'}\bigr) P\psi_\lambda P\psi_{\lambda'} \\
&= \frac{1}{n} \sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \bigl(P(\psi_\lambda \psi_{\lambda'})\bigr)^2 - \frac{1}{n} \lVert s_{m_1} \rVert^2 \lVert s_{m_2} \rVert^2 + \frac{n-1}{n} \operatorname{Cov}\bigl(s_{m_1}(\xi), s_{m_2}(\xi)\bigr) \enspace .
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\sum_{\lambda \in \Lambda_{m_1}} \sum_{\lambda' \in \Lambda_{m_2}} \bigl(P(\psi_\lambda \psi_{\lambda'})\bigr)^2 &= n \operatorname{Cov}\bigl(\widehat{s}_{m_1}(\xi), \widehat{s}_{m_2}(\xi)\bigr) + \lVert s_{m_1} \rVert^2 \lVert s_{m_2} \rVert^2 \\
&\qquad - (n-1) \operatorname{Cov}\bigl(s_{m_1}(\xi), s_{m_2}(\xi)\bigr) \enspace .
\end{aligned}
$$

Thus,

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

Eq. (59) follows. ■

B.2 Evaluation of the Variance in the Regular Histogram Case

The following lemma gives the value of the terms appearing in Theorem 6 for two nested regular histogram models.

Lemma 19 *Let $m_1 = \Lambda_{m_1}$ and $m_2 = \Lambda_{m_2}$ be two regular partitions of $\mathbb{R}$, as defined by Example 1 in Section 3.2, so that for $i \in \{1, 2\}$, for any $\lambda \in m_i$, $\mu(\lambda) = d_{m_i}^{-1}$. We assume that m_2 is a subpartition of m_1, that is, any element of m_2 is a subset of an element of m_1. For any $m^\star \in \{m_1, m_2\}$, we define*

$$
T_{m^\star}(x) = \sum_{\lambda \in m^\star} (\psi_\lambda(x) - P\psi_\lambda)^2 = \sup_{t \in \mathbb{B}_{m^\star}} \bigl(t(x) - Pt\bigr)^2
$$

where we recall that $\mathbb{B}_{m^\star} = \{t \in S_{m^\star} \,/\, \lVert t \rVert \leqslant 1\}$ and for any $\lambda \in m_1 \cup m_2$, $\psi_\lambda = (\mu(\lambda))^{-1/2} \mathbb{1}_\lambda$. Then, we have

$$
\beta(m_1, m_2) = d_{m_1} \lVert s_{m_2} \rVert^2 - 2P(s_{m_1} s_{m_2}) + \lVert s_{m_1} \rVert^2 \lVert s_{m_2} \rVert^2 = P(T_{m_1} s_{m_2}) \tag{60}
$$

and

$$
\begin{aligned}
\mathbf{B}(m_1, m_2) &= P\bigl(T_{m_1}(s_{m_1} - s_{m_2}) + (T_{m_2} - T_{m_1}) s_{m_2}\bigr) \\
&= (d_{m_2} - d_{m_1}) \lVert s_{m_2} \rVert^2 + d_{m_1} \lVert s_{m_1} - s_{m_2} \rVert^2 \\
&\qquad - 2 \operatorname{Var}_P(s_{m_1} - s_{m_2}) - \lVert s_{m_1} - s_{m_2} \rVert^4 \enspace .
\end{aligned}
$$

Proof On the one hand, by definition,

$$\begin{aligned}
&\beta(m_1,m_2)\\
&= \sum_{\lambda\in m_1}\sum_{\lambda'\in m_2}\left(\mathbb{E}\Big[\big(\psi_\lambda(\xi_1)-P\psi_\lambda\big)\big(\psi_{\lambda'}(\xi_1)-P\psi_{\lambda'}\big)\Big]\right)^2\\
&= \sum_{\lambda\in m_1}\sum_{\lambda'\in m_2}\left(\big[P(\psi_\lambda\psi_{\lambda'})\big]^2 - 2P(\psi_\lambda\psi_{\lambda'})P\psi_\lambda P\psi_{\lambda'} + (P\psi_\lambda)^2(P\psi_{\lambda'})^2\right)\\
&= \sum_{\lambda\in m_1}\sum_{\lambda'\in m_2}\big[P(\psi_\lambda\psi_{\lambda'})\big]^2 - 2P\Bigg(\underbrace{\Big(\sum_{\lambda\in m_1}(P\psi_\lambda)\psi_\lambda\Big)}_{=s_{m_1}}\underbrace{\Big(\sum_{\lambda\in m_2}(P\psi_\lambda)\psi_\lambda\Big)}_{=s_{m_2}}\Bigg)\\
&\qquad + \underbrace{\sum_{\lambda\in m_1}(P\psi_\lambda)^2}_{=\|s_{m_1}\|^2}\underbrace{\sum_{\lambda\in m_2}(P\psi_\lambda)^2}_{=\|s_{m_2}\|^2} \enspace .
\end{aligned}$$

For computing the first term, we use that $\psi_\lambda\psi_{\lambda'}=0$ if $\lambda\cap\lambda'=\emptyset$ and m_2 is a subpartition of m_1, so that

$$\begin{aligned}
\sum_{\lambda\in m_1}\sum_{\lambda'\in m_2}\big[P(\psi_\lambda\psi_{\lambda'})\big]^2 &= \sum_{\lambda\in m_1}\sum_{\substack{\lambda'\in m_2\\ \lambda'\subset\lambda}}\big[P(\psi_\lambda\psi_{\lambda'})\big]^2\\
&= \sum_{\lambda\in m_1}\frac{1}{\mu(\lambda)}\sum_{\substack{\lambda'\in m_2\\ \lambda'\subset\lambda}}(P\psi_{\lambda'})^2 = d_{m_1}\sum_{\lambda'\in m_2}(P\psi_{\lambda'})^2 = d_{m_1}\|s_{m_2}\|^2
\end{aligned}$$

hence

$$\beta(m_1,m_2) = d_{m_1}\|s_{m_2}\|^2 - 2P(s_{m_1}s_{m_2}) + \|s_{m_1}\|^2\|s_{m_2}\|^2 \enspace .$$

On the other hand, by definition of T_m,

$$\begin{aligned}
P(T_{m_1}s_{m_2}) &= \sum_{\lambda\in m_1}\sum_{\lambda'\in m_2}P\big((\psi_\lambda-P\psi_\lambda)^2\psi_{\lambda'}P(\psi_{\lambda'})\big)\\
&= \sum_{\lambda\in m_1}\sum_{\lambda'\in m_2}\big(P(\psi_\lambda^2\psi_{\lambda'})(P\psi_{\lambda'}) - 2P(\psi_\lambda\psi_{\lambda'})(P\psi_\lambda)(P\psi_{\lambda'}) + (P\psi_\lambda)^2(P\psi_{\lambda'})^2\big)\\
&= P\Bigg(\underbrace{\sum_{\lambda\in m_1}\psi_\lambda^2}_{=d_{m_1}}\underbrace{\sum_{\lambda'\in m_2}(P\psi_{\lambda'})\psi_{\lambda'}}_{=s_{m_2}}\Bigg) - 2P(s_{m_1}s_{m_2}) + \|s_{m_1}\|^2\|s_{m_2}\|^2
\end{aligned}$$

which proves Eq. (60) since $P(s_{m_2}) = \|s_{m_2}\|^2$.

Now, we remark that Eq. (60) also gives formulas for $\beta(m_i,m_i)$, $i\in\{1,2\}$, since m_i is a subpartition of itself. So, the second formula for $\beta(m_i,m_j)$ in Eq. (60) yields

$$\mathbf{B}(m_1,m_2) = P\big(T_{m_1}s_{m_1} + T_{m_2}s_{m_2} - 2T_{m_1}s_{m_2}\big)$$

$$= P\big(T_{m_1}(s_{m_1} - s_{m_2}) + (T_{m_2} - T_{m_1})s_{m_2}\big) \ .$$

Similarly, the first formula for $\beta(m_i, m_j)$ in Eq. (60) gives

$$\begin{aligned}
&\mathbf{B}(m_1, m_2) \\
&= d_{m_1}\big(\|s_{m_1}\|^2 - \|s_{m_2}\|^2\big) + (d_{m_2} - d_{m_1})\|s_{m_2}\|^2 - 2P\big((s_{m_1} - s_{m_2})^2\big) + \big(\|s_{m_1}\|^2 - \|s_{m_2}\|^2\big)^2 \\
&= (d_{m_2} - d_{m_1})\|s_{m_2}\|^2 + d_{m_1}\|s_{m_1} - s_{m_2}\|^2 - 2\operatorname{Var}_P(s_{m_1} - s_{m_2}) - \|s_{m_1} - s_{m_2}\|^4 \ ,
\end{aligned}$$

where we used that $P(s_m) = \|s_m\|^2$ and $\|s_{m_1} - s_{m_2}\|^2 = \|s_{m_1}\|^2 - \|s_{m_2}\|^2$. ■

Appendix C. Results on MCCV and Some Other Cross-Validation Criteria

We prove here the results stated in Section 8.1. Note that we here prove slightly more general results (Theorems 23 and 24), from which Theorems 9 and 10 are corollaries. In particular, we do not always restrict to MCCV criteria: we always assume (**SameSize**) and (**Ind**) hold true, but we sometimes do not need to have (**MCCV**) satisfied.

C.1 Preliminary Computations

Our proofs rely on a simple closed-form formula for cross-validation criteria. Let us start by the hold-out criterion. Let $T \subset [\![n]\!]$ with $|T| = n - p$, independent from D_n. Then,

$$\begin{aligned}
\operatorname{crit}_{\mathrm{HO}}(m, T) &= P_n^{(T^c)}\gamma\Big(\widehat{s}_m^{(T)}\Big) \\
&= \big\|\widehat{s}_m^{(T)}\big\|^2 - 2P_n^{(T^c)}\Big(\widehat{s}_m^{(T)}\Big) \\
&= \big\|\widehat{s}_m^{(T)} - s_m\big\|^2 + \|s_m\|^2 + 2\Big\langle \widehat{s}_m^{(T)} - s_m, \, s_m \Big\rangle \\
&\qquad - 2\Big(P_n^{(T^c)} - P\Big)\Big(\widehat{s}_m^{(T)} - s_m\Big) - 2P\Big(\widehat{s}_m^{(T)} - s_m\Big) - 2P_n^{(T^c)}(s_m) \\
&= \big\|\widehat{s}_m^{(T)} - s_m\big\|^2 - 2\Big(P_n^{(T^c)} - P\Big)\Big(\widehat{s}_m^{(T)} - s_m\Big) - 2P_n^{(T^c)}(s_m) + \|s_m\|^2 \qquad (61)
\end{aligned}$$

where the last equality uses that

$$P\Big(\widehat{s}_m^{(T)} - s_m\Big) = \Big\langle \widehat{s}_m^{(T)} - s_m, \, s \Big\rangle = \Big\langle \widehat{s}_m^{(T)} - s_m, \, s_m \Big\rangle$$

since s_m is the orthogonal projection in $L^2(\mu)$ of s_m onto S_m and $\widehat{s}_m^{(T)} - s_m \in S_m$.

The last two terms in the right-hand side of Eq. (61) can be rewritten as

$$\begin{aligned}
-2P_n^{(T^c)}(s_m) + \|s_m\|^2 &= -2\Big(P_n^{(T^c)} - P\Big)(s_m) - 2P(s_m) + \|s_m\|^2 \\
&= -2\Big(P_n^{(T^c)} - P\Big)(s_m) - \|s_m\|^2
\end{aligned}$$

since $\|s_m\|^2 = P(s_m)$. For the first two terms, we write that

$$\|\widehat{s}_m^{(T)} - s_m\|^2 - 2\Big(P_n^{(T^c)} - P\Big)\Big(\widehat{s}_m^{(T)} - s_m\Big)$$

$$= \sum_{\lambda\in\Lambda_m} \Big[\big((P_n^{(T)} - P)(\psi_\lambda)\big)^2 - 2\Big(P_n^{(T^c)} - P\Big)(\psi_\lambda)\Big(P_n^{(T)} - P\Big)(\psi_\lambda)\Big]$$

$$= \sum_{\lambda\in\Lambda_m} \Bigg[\frac{1}{(n-p)^2} \sum_{1\leqslant i,j\leqslant n} \mathbb{1}_{i\in T,\, j\in T} \big(\psi_\lambda(\xi_i) - P\psi_\lambda\big)\big(\psi_\lambda(\xi_j) - P\psi_\lambda\big)$$

$$- \frac{2}{p(n-p)} \sum_{1\leqslant i,j\leqslant n} \mathbb{1}_{i\in T^c,\, j\in T} \big(\psi_\lambda(\xi_i) - P\psi_\lambda\big)\big(\psi_\lambda(\xi_j) - P\psi_\lambda\big)\Bigg]$$

$$= \sum_{1\leqslant i,j\leqslant n} \left[\frac{\mathbb{1}_{j\in T}}{n-p}\left(\frac{\mathbb{1}_{i\in T}}{n-p} - \frac{2\mathbb{1}_{i\in T^c}}{p}\right) U_m(\xi_i,\xi_j)\right]$$

where we recall that for any $x, y \in \mathcal{X}$,

$$U_m(x,y) = \sum_{\lambda\in\Lambda_m} \big(\psi_\lambda(x) - P\psi_\lambda\big)\big(\psi_\lambda(y) - P\psi_\lambda\big) = \sum_{\lambda\in\Lambda_m} \psi_\lambda(x)\psi_\lambda(y) - s_m(x) - s_m(y) + \|s_m\|^2$$

is defined by Eq. (56), and that $U_m(x,x) = \Psi_m(x) - 2s_m(x) + \|s_m\|^2$.

Therefore, Eq. (61) can be rewritten as

$$\mathrm{crit}_{\mathrm{HO}}(m,T) = \sum_{1\leqslant i,j\leqslant n} \left[\frac{\mathbb{1}_{j\in T}}{n-p}\left(\frac{\mathbb{1}_{i\in T}}{n-p} - \frac{2\mathbb{1}_{i\in T^c}}{p}\right) U_m(\xi_i,\xi_j)\right] - 2\big(P_n^{(T^c)} - P\big)(s_m) - \|s_m\|^2$$

$$= \sum_{1\leqslant i,j\leqslant n} \omega_{i,j}^{\mathrm{HO}}(T) U_m(\xi_i,\xi_j) + \sum_{i=1}^n \sigma_i^{\mathrm{HO}}(T)\big(s_m(\xi_i) - P(s_m)\big) - \|s_m\|^2 \tag{62}$$

with

$$\omega_{i,j}^{\mathrm{HO}}(T) = \frac{\mathbb{1}_{j\in T}}{n-p}\left(\frac{\mathbb{1}_{i\in T}}{n-p} - \frac{2\mathbb{1}_{i\in T^c}}{p}\right)$$

$$\sigma_i^{\mathrm{HO}}(T) = \frac{-2}{p}\mathbb{1}_{i\in T^c} \quad .$$

As a consequence, under assumption (**SameSize**),

$$\mathrm{crit}_{\mathrm{CV}}\big(m, (T_j)_{1\leqslant j\leqslant K}\big) = \sum_{1\leqslant i,j\leqslant n} \omega_{i,j} U_m(\xi_i,\xi_j) + \sum_{i=1}^n \sigma_i\big(s_m(\xi_i) - P(s_m)\big) - \|s_m\|^2 \tag{63}$$

with

$$\omega_{i,j} = \frac{1}{B}\sum_{K=1}^B \left[\frac{\mathbb{1}_{j\in T_K}}{n-p}\left(\frac{\mathbb{1}_{i\in T_K}}{n-p} - \frac{2\mathbb{1}_{i\in T_K^c}}{p}\right)\right]$$

$$\sigma_i = \frac{-2}{pB}\sum_{K=1}^B \mathbb{1}_{i\in T_K^c} \quad .$$

Note that Eq. (63) is consistent with previously obtained formulas. For V-fold cross-validation, under assumption (**Reg**), Eq. (63) holds with

$$\omega_{i,j} = \omega_{i,j}^{\mathrm{VF}} := \frac{1}{n^2}\begin{cases} \frac{V}{V-1} & \text{if } i \text{ and } j \text{ belong to the same block} \\ -\left(\frac{V}{V-1}\right)^2 & \text{otherwise} \end{cases}$$

$$\sigma_i = \sigma_i^{\mathrm{VF}} := \frac{-2}{n} \enspace ,$$

which can also be obtained from the combination of Eq. (8) in Lemma 1 and Eq. (58). For the leave-p-out, Eq. (63) holds with

$$\omega_{i,j} = \omega_{i,j}^{\mathrm{LPO}} := \begin{cases} \frac{1}{n(n-p)} & \text{if } i \neq j \\ \frac{-(n-p+1)}{n(n-1)(n-p)} & \text{otherwise} \end{cases}$$
$$\sigma_i = \sigma_i^{\mathrm{LPO}} := \frac{-2}{n} \enspace ,$$

which can also be obtained from Eq. (10) in Lemma 1 and Eq. (58).

Using that $U_m(x,x) = \Psi_m(x) - 2s_m(x) + \|s_m\|^2$, Eq. (63) can be rewritten as

$$\begin{aligned} \mathrm{crit}_{\mathcal{C}}\bigl(m, (T_j)_{1 \leqslant j \leqslant K}\bigr) &= \left(\sum_{i=1}^n \omega_{i,i} \right) \bigl(\mathcal{D}_m - \|s_m\|^2 \bigr) - \|s_m\|^2 + \sum_{i=1}^n \omega_{i,i} \bigl(\Psi_m(\xi_i) - \mathcal{D}_m \bigr) \\ &\quad + \sum_{i=1}^n (-2\omega_{i,i} + \sigma_i) \bigl(s_m(\xi_i) - P s_m \bigr) + \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} U_m(\xi_i, \xi_j) \enspace . \end{aligned}$$

Using (**SameSize**) we have

$$\sum_{i=1}^n \omega_{i,i} = \frac{1}{B(n-p)^2} \sum_{K=1}^B \sum_{i=1}^n \mathbb{1}_{i \in T_K} = \frac{1}{n-p} \enspace ,$$

and we get

$$\begin{aligned} \mathrm{crit}_{\mathcal{C}}\bigl(m, (T_j)_{1 \leqslant j \leqslant K}\bigr) &= \frac{\mathcal{D}_m - \|s_m\|^2}{n-p} - \|s_m\|^2 + \sum_{i=1}^n \omega_{i,i} \bigl(\Psi_m(\xi_i) - \mathcal{D}_m \bigr) \\ &\quad + \sum_{i=1}^n (-2\omega_{i,i} + \sigma_i) \bigl(s_m(\xi_i) - P s_m \bigr) + \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} U_m(\xi_i, \xi_j) \enspace . \end{aligned} \tag{64}$$

C.2 Concentration Inequalities

In the proof of Theorem 9 in Section C.3, given formula (64) for the cross-validation criterion, we need concentration inequalities for the three random sums appearing in Eq. (64). These are stated and proved in three lemmas below.

Concentration of $\sum_{i=1}^n \omega_{i,i}(\Psi_m(\xi_i) - \mathcal{D}_m)$.

Lemma 20 *Assume that* (**SameSize**), (**Ind**) *and* (**H1**) *hold true. Then, for any $x > 0$, an event of probability at least $1 - 2\mathrm{e}^{-x}$ exists on which the following holds true: for any $\epsilon \in (0,1]$,*

$$\left| \sum_{i=1}^n \omega_{i,i} \bigl(\Psi_m(\xi_i) - \mathcal{D}_m \bigr) \right| \leqslant \epsilon \frac{\mathcal{D}_m}{n-p} + \frac{5x(n+A)}{3\epsilon(n-p)^2} \enspace .$$

Proof By (**Ind**), conditionally to $(\omega_{i,i})_{1 \leqslant i \leqslant n}$, $\sum_{i=1}^n \omega_{i,i} \bigl(\Psi_m(\xi_i) - \mathcal{D}_m \bigr)$ is a sum of independent real-valued random variables. So, we can apply Bernstein's inequality.

First, for any $i \in [\![n]\!]$, using (**SameSize**),

$$\omega_{i,i} = \frac{1}{B} \sum_{K=1}^{B} \frac{\mathbb{1}_{i \in T_K}}{(n-p)^2} \leqslant \frac{1}{(n-p)^2}$$

and using Eq. (49),

$$\lVert \Psi_m \rVert_\infty \leqslant \lVert U_m \rVert_\infty \leqslant 2 \Bigl(b_m^2 + \lVert s_m \rVert^2 \Bigr) \enspace ,$$

so that

$$\omega_{i,i} \Psi_m(\xi_i) \leqslant \max_{1 \leqslant i \leqslant n} \omega_{i,i} \times \lVert \Psi_m \rVert_\infty \leqslant \frac{2 \Bigl(b_m^2 + \lVert s_m \rVert^2 \Bigr)}{(n-p)^2}$$

almost surely.

Second, using (**SameSize**), we have

$$\sum_{i=1}^n \omega_{i,i}^2 \leqslant \max_{1 \leqslant i \leqslant n} \omega_{i,i} \times \sum_{i=1}^n \omega_{i,i} \leqslant \frac{1}{(n-p)^3}$$

and using Eq. (49) again,

$$\mathbb{E}\bigl[\Psi_m(\xi_i)^2 \bigr] \leqslant \lVert \Psi_m \rVert_\infty \times P(\Psi_m) = \lVert \Psi_m \rVert_\infty \times \mathcal{D}_m \leqslant 2 \Bigl(b_m^2 + \lVert s_m \rVert^2 \Bigr) \mathcal{D}_m \enspace ,$$

so that

$$\sum_{i=1}^n \omega_{i,i}^2 \mathbb{E}\bigl[\Psi_m(\xi_i)^2 \bigr] \leqslant \frac{2 \Bigl(b_m^2 + \lVert s_m \rVert^2 \Bigr) \mathcal{D}_m}{(n-p)^3} \enspace .$$

Then, by Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10), conditionally to $(\omega_{i,i})_{1 \leqslant i \leqslant n}$, an event of probability at least $1 - 2\mathrm{e}^{-x}$ exists on which

$$\begin{aligned}
\left| \sum_{i=1}^n \omega_{i,i} \bigl(\Psi_m(\xi_i) - \mathcal{D}_m \bigr) \right| &\leqslant 2 \sqrt{ \frac{x \Bigl(b_m^2 + \lVert s_m \rVert^2 \Bigr) \mathcal{D}_m}{(n-p)^3} } + \frac{2 \Bigl(b_m^2 + \lVert s_m \rVert^2 \Bigr)}{(n-p)^2} \frac{x}{3} \\
&\leqslant \epsilon \frac{\mathcal{D}_m}{n-p} + \left(\frac{2}{3} + \frac{1}{\epsilon} \right) \frac{x \Bigl(b_m^2 + \lVert s_m \rVert^2 \Bigr)}{(n-p)^2} \\
&\leqslant \epsilon \frac{\mathcal{D}_m}{n-p} + \frac{5}{3\epsilon} \frac{x(n+A)}{(n-p)^2}
\end{aligned}$$

for any $\epsilon \in (0,1]$, where we used that $b_m^2 \leqslant n$ by (**H1**), and that $\lVert s_m \rVert^2 \leqslant \lVert s \rVert^2 \leqslant \lVert s \rVert_\infty \leqslant A$. The result follows by integrating this conditional concentration inequality with respect to $(\omega_{i,i})_{1 \leqslant i \leqslant n}$. ■

Concentration of $\sum_{i=1}^n (-2\omega_{i,i} + \sigma_i)(s_m(\xi_i) - Ps_m)$**.**

Lemma 21 *Assume that* (**SameSize**) *and* (**Ind**) *hold true. Then, for any $x > 0$, an event of probability at least $1 - \mathrm{e}^{-x}$ exists on which the following holds true: for any $\epsilon \in (0, 1]$,*

$$\begin{aligned} &\sum_{i=1}^n (-2\omega_{i,i} + \sigma_i)\big(s_m(\xi_i) - Ps_m - s_{m'}(\xi_i) - Ps_{m'}\big) \\ &\qquad\qquad \leqslant \epsilon \| s_m - s_{m'} \|^2 + R_n^{21}(x, \epsilon, \pi^*, A) \end{aligned} \tag{65}$$

where the remainder term depends on the additional assumption that we make. If (**H2**) *holds true, then*

$$R_n^{21}(x, \epsilon, \pi^*, A) := \frac{16Ax}{3\epsilon}\left(\frac{1}{(n-p)^2} + \frac{\pi^\star}{p} \right) \enspace .$$

If (**H1**) *and* (**H2′**) *hold true, then, some numerical constant $\kappa > 0$ exists such that*

$$R_n^{21}(x, \epsilon, \pi^*, A) := \frac{\kappa}{\epsilon}\left[Ax\left(\frac{1}{(n-p)^2} + \frac{\pi^\star}{p} \right) + x^2 n \left(\frac{1}{(n-p)^2} + \frac{\pi^\star}{p} \right)^2 \right] \enspace .$$

Before proving Lemma 21, let us introduce some useful notation: given a sequence $T_1, \ldots, T_B$ of subsets of $[\![n]\!]$, for every $i, j \in [\![n]\!]$, we define

$$\pi_i = \frac{1}{B}\sum_{K=1}^B \mathbb{1}_{i \in T_K^c} \qquad \pi_{i,j} = \frac{1}{B}\sum_{K=1}^B \mathbb{1}_{i \in T_K^c} \mathbb{1}_{j \in T_K^c} \qquad \text{and} \qquad \pi^* = \max_{i=1,\ldots,n} \pi_i \enspace .$$

Note that, assuming (**SameSize**), we have

$$\begin{gathered} 0 \leqslant \pi_{i,j} \leqslant \min(\pi_i, \pi_j) \leqslant \pi^* \leqslant 1 \qquad \sum_{i=1}^n \pi_i = p \\ \sum_{i=1}^n \pi_{i,j} = p\pi_j \leqslant p\pi^* \qquad \text{and} \qquad \sum_{1 \leqslant i,j \leqslant n} \pi_{i,j} = p^2 \enspace . \end{gathered} \tag{66}$$

Proof of Lemma 21 By (**Ind**), conditionally to $(-2\omega_{i,i} + \sigma_i)_{1 \leqslant i \leqslant n}$,

$$\sum_{i=1}^n \omega_{i,i}\big(\Psi_m(\xi_i) - \mathcal{D}_m\big)$$

is a sum of independent real-valued random variables. So, we can apply Bernstein's inequality.

First, we notice that for every $i \in [\![n]\!]$,

$$-2\omega_{i,i} + \sigma_i = \frac{1}{B}\sum_{K=1}^B \left(\frac{-2}{(n-p)^2}\mathbb{1}_{i \in T_K} - \frac{2}{p}\mathbb{1}_{i \notin T_K} \right) = -2\left(\frac{1}{(n-p)^2}(1 - \pi_i) + \frac{\pi_i}{p} \right)$$

hence

$$|-2\omega_{i,i} + \sigma_i| = 2\left(\frac{1}{(n-p)^2}(1 - \pi_i) + \frac{\pi_i}{p} \right) \leqslant 2\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p} \right)$$

since $0 \leqslant \pi_i \leqslant \pi^* \leqslant 1$. So, for every $i \in [\![n]\!]$,

$$(-2\omega_{i,i} + \sigma_i)\bigl(s_m(\xi_i) - s_{m'}(\xi_i)\bigr) \leqslant 2\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right)\|s_m - s_{m'}\|_\infty$$

almost surely. Second,

$$\begin{aligned}\sum_{i=1}^n (-2\omega_{i,i} + \sigma_i)^2 &\leqslant 2\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right)\sum_{i=1}^n |-2\omega_{i,i} + \sigma_i| \\ &= 2\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right) 2\left(\frac{1}{n-p} + 1\right) \\ &\leqslant 8\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right)\end{aligned}$$

and

$$\mathbb{E}\left[\bigl(s_m(\xi) - s_{m'}(\xi)\bigr)^2\right] \leqslant \|s\|_\infty \|s_m - s_{m'}\|^2 \leqslant A\|s_m - s_{m'}\|^2$$

so that

$$\sum_{i=1}^n \mathbb{E}\left[\Bigl((-2\omega_{i,i} + \sigma_i)\bigl(s_m(\xi) - s_{m'}(\xi)\bigr)\Bigr)^2\right] \leqslant 8A\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right)\|s_m - s_{m'}\|^2 \enspace .$$

Then, by Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10), conditionally to $(-2\omega_{i,i} + \sigma_i)_{1 \leqslant i \leqslant n}$, an event of probability at least $1 - \mathrm{e}^{-x}$ exists on which

$$\sum_{i=1}^n (-2\omega_{i,i} + \sigma_i)\bigl(s_m(\xi_i) - Ps_m - s_{m'}(\xi_i) - Ps_{m'}\bigr) \leqslant R^0(m, m')$$

$$R^0(m, m') := \sqrt{16xA\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right)\|s_m - s_{m'}\|^2} + \frac{2x\|s_m - s_{m'}\|_\infty}{3}\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right) \enspace .$$

Since $1 - 2\mathrm{e}^{-x}$ is deterministic, the same inequality holds unconditionally on an event of probability at least $1 - 2\mathrm{e}^{-x}$.

We now upperbound $R^0(m, m')$, differently depending on the assumption we make. On the one hand, if (**H2**) holds true,

$$\|s_m - s_{m'}\|_\infty \leqslant \|s_m\|_\infty + \|s_{m'}\|_\infty \leqslant 2A$$

and we get

$$\begin{aligned}R^0(m, m') &\leqslant \sqrt{16Ax\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right)\|s_m - s_{m'}\|^2} + \frac{4Ax}{3}\left(\frac{1}{(n-p)^2} + \frac{\pi^*}{p}\right) \\ &\leqslant \epsilon\|s_m - s_{m'}\|^2 + \frac{16Ax}{3\epsilon}\left(\frac{1}{(n-p)^2} + \frac{\pi^\star}{p}\right)\end{aligned}$$

for any $\epsilon \in (0, 1]$, which proves Eq. (65). On the other hand, if (**H1**) and (**H2′**) hold true, $s_m - s_{m'} \in S_{m''}$ with $m'' \in \{m, m'\}$, so that

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

and we get

$$R^0(m,m') \leqslant \sqrt{16xA\left(\frac{1}{(n-p)^2}+\frac{\pi^*}{p}\right)\|s_m - s_{m'}\|^2} + \frac{2x\sqrt{n}\|s_m - s_{m'}\|}{3}\left(\frac{1}{(n-p)^2}+\frac{\pi^*}{p}\right)$$

$$\leqslant \epsilon\|s_m - s_{m'}\|^2 + \frac{1}{\epsilon}\left[8Ax\left(\frac{1}{(n-p)^2}+\frac{\pi^\star}{p}\right) + \frac{2}{9}x^2 n\left(\frac{1}{(n-p)^2}+\frac{\pi^\star}{p}\right)^2\right]$$

for any $\epsilon \in (0,1]$, which proves Eq. (65) with $\kappa = 8$. ■

Concentration of $\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} U_m(\xi_i,\xi_j)$***.***

Lemma 22 *Suppose that assumptions* (**SameSize**), (**Ind**) *and* (**H1**) *hold true. Then, an absolute constant* $\kappa > 0$ *exists such that, for any* $x > 1$*, with probability larger than* $1 - 6\mathrm{e}^{-x}$*, for any* $\epsilon \in (0,1]$*,*

$$\left|\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} U_m(\xi_i,\xi_j)\right| \leqslant \frac{\epsilon \mathcal{D}_m}{n-p} + \frac{\kappa n}{(n-p)^2}\left(1+\frac{n\pi^*}{p}\right)\left[\frac{nAx}{(n-p)\epsilon} + \left(1+\frac{A}{n}\right)x^2\right] . \qquad (67)$$

Proof We start with the following symmetrization trick

$$\begin{aligned}\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} U_m(\xi_i,\xi_j) &= \sum_{1\leqslant i<j\leqslant n} \omega_{i,j} U_m(\xi_i,\xi_j) + \omega_{j,i} U_m(\xi_j,\xi_i) \\ &= \sum_{1\leqslant i<j\leqslant n} (\omega_{i,j} + \omega_{j,i}) U_m(\xi_i,\xi_j) \\ &= \sum_{1\leqslant i\neq j\leqslant n} \omega'_{i,j} U_m(\xi_i,\xi_j) \ ,\end{aligned}$$

where

$$\begin{aligned}\omega'_{i,j} = \frac{\omega_{i,j}+\omega_{j,i}}{2} &= \frac{1}{(n-p)^2}\left[1 - (\pi_i+\pi_j)\frac{n}{p} + \left(\frac{2n}{p}-1\right)\pi_{i,j}\right] \\ &= \frac{1}{(n-p)^2}\left[(1-\pi_{i,j}) + \frac{n}{p}(\pi_{i,j}-\pi_i) + \frac{n}{p}(\pi_{i,j}-\pi_j)\right] .\end{aligned}$$

From the last formula for $\omega'_{i,j}$, using Eq. (66), we get that

$$(\omega'_{i,j})^2 \leqslant \frac{1}{(n-p)^4}\left[1 + \frac{n^2}{p^2}(\pi_i+\pi_j)^2\right] \qquad (68)$$

$$\text{and} \qquad \max_{i,j\in[\![n]\!]} |\omega'_{i,j}| \leqslant \frac{1}{(n-p)^2}\left(1 + \frac{2n}{p}\pi^*\right) . \qquad (69)$$

The concentration of the U-statistics follows from Houdré and Reynaud-Bouret (2003, Theorem 3.4), that is Eq. (44) with $g_{i,j}(\xi_i,\xi_j) = \omega'_{i,j} U_m(\xi_i,\xi_j)$. To apply this result, it remains to compute the terms $\overline{A}$, $\overline{B}$, $\overline{C}$, $\overline{D}$. First,

$$2\overline{A}^2 = \sum_{1\leqslant i\neq j\leqslant n} (\omega'_{i,j})^2 \mathbb{E}\left[U_m(\xi_i,\xi_j)^2\right] \leqslant \|s\|_\infty \mathcal{D}_m \sum_{1\leqslant i\neq j\leqslant n} (\omega'_{i,j})^2$$

by Eq (45). Algebraic computations and Eq. (68) and (66) show that

$$\begin{aligned}\sum_{1\leqslant i\neq j\leqslant n}(\omega'_{i,j})^2 &\leqslant \frac{1}{(n-p)^4}\sum_{1\leqslant i\neq j\leqslant n}\left[1+\frac{n^2}{p^2}(\pi_i+\pi_j)^2\right]\\ &\leqslant \frac{1}{(n-p)^4}\sum_{1\leqslant i,j\leqslant n}\left[1+\frac{n^2}{p^2}(\pi^*\pi_i+\pi^*\pi_j+2\pi_i\pi_j)\right]\\ &= \frac{n^2}{(n-p)^4}\left(3+\frac{2\pi^* n}{p}\right)\end{aligned}$$

Hence,

$$\overline{A}\leqslant \frac{n}{(n-p)^2}\sqrt{\left(\frac{3}{2}+\frac{\pi^* n}{p}\right)\|s\|_\infty \mathcal{D}_m}\enspace .$$

Second, let a_i and b_j be functions such that $\sum_{i=1}^n \mathbb{E}\left[a_i(\xi)^2\right]\leqslant 1$ and $\sum_{i=1}^n \mathbb{E}\left[b_i(\xi)^2\right]\leqslant 1$. Eq (47) shows that

$$\left|\mathbb{E}\left[a_i(\xi)b_j(\xi')U_m(\xi,\xi')\right]\right|\leqslant \frac{\|s\|_\infty}{2}\left(\mathbb{E}\left[a_i(\xi)^2\right]+\mathbb{E}\left[b_j(\xi)^2\right]\right)\enspace ,$$

hence, using Eq. (69),

$$\begin{aligned}\overline{B} &= \sum_{1\leqslant i\neq j\leqslant n}\omega'_{i,j}\mathbb{E}\left[a_i(\xi)b_j(\xi')U_m(\xi,\xi')\right]\\ &\leqslant \max_{1\leqslant i\neq j\leqslant n}\left|\omega'_{i,j}\right|\frac{\|s\|_\infty}{2}\sum_{1\leqslant i\neq j\leqslant n}\left(\mathbb{E}\left[a_i(\xi)^2\right]+\mathbb{E}\left[b_j(\xi)^2\right]\right)\\ &\leqslant \frac{n\|s\|_\infty}{(n-p)^2}\left(1+\frac{2n}{p}\pi^*\right)\enspace .\end{aligned}$$

Third, Eq (48) shows that, for any $x>0$,

$$\mathbb{E}\left[U_m(\xi,x)^2\right]\leqslant 2\left(b_m^2+\|s_m\|^2\right)\|s\|_\infty$$

and by Eq. (68) we have

$$\sum_{i=2}^n(\omega'_{i,1})^2\leqslant \frac{1}{(n-p)^4}\sum_{i=2}^n\left(1+(\pi_i+\pi_1)^2\frac{n^2}{p^2}\right)\leqslant \frac{n}{(n-p)^4}\left(1+2\pi^*\frac{n}{p}\right)^2\enspace .$$

So, for any $x>0$,

$$\sum_{i=2}^n(\omega'_{i,1})^2\mathbb{E}\left[U_m(\xi,x)^2\right]\leqslant 2\left(b_m^2+\|s_m\|^2\right)\|s\|_\infty\times\frac{1}{(n-p)^4}\left(1+2\pi^*\frac{n}{p}\right)^2$$

hence

$$\overline{C}\leqslant \left(1+2\pi^*\frac{n}{p}\right)\frac{n}{(n-p)^2}\sqrt{\frac{2\left(b_m^2+\|s_m\|^2\right)\|s\|_\infty}{n}}\enspace .$$

Fourth, using Eq (49) and (69),

$$\overline{D} \leqslant \max_{i,j \in [\![n]\!]} \left| \omega'_{i,j} \right| \sup_{x,y} \left| U_m(x,y) \right| \leqslant \left(1 + \frac{2n}{p} \pi^* \right) \frac{n}{(n-p)^2} \frac{2\big(b_m^2 + \|s_m\|^2\big)}{n} \enspace .$$

Now, we remark that $b_m^2 \leqslant n$ by (**H1**), and $\|s_m\|^2 \leqslant \|s\|^2 \leqslant \|s\|_\infty \leqslant A$, and we can plug this two inequalities in the upper bounds above. By (**Ind**), we can apply Houdré and Reynaud-Bouret (2003, Theorem 3.4), conditionally on the weights $\omega_{i,j}$. We obtain that an absolute constant $\kappa > 0$ exists such that, for any $x > 1$, with probability larger than $1 - 6\mathrm{e}^{-x}$, for any $\epsilon \in (0,1]$, Eq. (67) holds true. ■

C.3 Oracle Inequality (Proof of Theorem 9)

Theorem 9 actually is a corollary of the following general result.

Theorem 23 *Let $\xi_{[\![n]\!]}$ be i.i.d. real-valued random variables with common density s with respect to μ, such that $s \in L^\infty(\mu)$, $(T_K)_{1 \leqslant K \leqslant B}$ be some sequence of subsets of $[\![n]\!]$ satisfying* (**SameSize**) *and* (**Ind**), *and $(S_m)_{m \in \mathcal{M}_n}$ be a collection of separable linear spaces satisfying* (**H1**). *Assume that either* (**H2**) *or* (**H2′**) *holds true. For every $m \in \mathcal{M}_n$, let $\widehat{s}_m$ be the estimator defined by Eq. (1), and $\widetilde{s} = \widehat{s}_{\widehat{m}}$ where*

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \left\{ \mathrm{crit}_{\mathrm{CV}}\big(m, (T_K)_{1 \leqslant K \leqslant B}\big) \right\}$$

and $\mathrm{crit}_{\mathrm{CV}}$ is defined by Eq. (25). Define $\pi^* = \max_{i=1,\ldots,n} \frac{1}{B} \sum_{K=1}^{B} \mathbb{1}_{i \in T_K^c}$ and for any $x, \epsilon, \kappa > 0$,

$$\rho_4(\epsilon, x, \kappa, n, \tau_n, \pi^*, A) := \frac{\kappa}{n \tau_n^2} \left(1 + \frac{\pi^*}{1 - \tau_n} \right)^{\alpha} \left[\frac{Ax}{\tau_n \epsilon} + \frac{(A \vee 1) x^2}{\epsilon^3} \right]$$

with $\alpha = 1$ under assumption (**H2**) *and $\alpha = 2$ under assumption* (**H2′**). *Then, an absolute constant $\kappa > 0$ exists such that, for any $x \geqslant 0$, with probability at least $1 - 12|\mathcal{M}_n|^2 \mathrm{e}^{-x}$, for any $\epsilon \in (0, \kappa^{-1})$,*

$$\left(1 - \frac{\epsilon}{\tau_n} \right) \|\widetilde{s} - s\|^2 \leqslant \frac{1 + \epsilon}{\tau_n} \inf_{m \in \mathcal{M}_n} \left\{ \|\widehat{s}_m - s\|^2 \right\} + \rho_4(\epsilon, x, \kappa, n, \tau_n, \pi^*, A) \enspace .$$

The oracle inequality of Theorem 23 is similar to the one of Theorem 5, with δ replaced by $1/\tau_n - 1$ (both quantities correspond to the bias of the criterion as an estimator of the risk) and a slightly different remainder term. In addition to the remarks already made about Theorem 5, we can make the following comments.

- The remainder term ρ_4 is of order x^2/n, as in Theorem 5 under the following sufficient conditions: (i) τ_n stays away from 0, (ii) $\pi^*/(1 - \tau_n)$ is bounded.
- For V-fold criteria, $\tau_n = (V-1)/V \geqslant 1/2$ and $\pi^*/(1 - \tau_n) = 1$, so conditions (i) and (ii) are satisfied and we recover an oracle inequality for V-fold cross-validation similar to Theorem 5.

- The leading constant in front of the oracle inequality of Theorem 23 is of order $1/\tau_n$, so we can get asymptotic optimality only if $\tau_n \to 1$, that is, $p \ll n$. This is consistent with the fact that the bias of the cross-validation criterion is negligible at first order if and only if $\tau_n \to 1$.
- For hold-out criteria, $\pi^* = 1$ so the remainder term is of order $x^2/(n(1-\tau_n)^\alpha) \geqslant x^2/p$ which is large when τ_n is close to 1, that is, when p is small. Hence, for such criteria, we cannot get a leading constant close to 1 and a "small" remainder term.

Let us now explain why Theorem 9 is also a corollary of Theorem 23.

Proof of Theorem 9 We only have to prove some upper bound on π^* under assumption (**MCCV**), thanks to which Theorem 9 is a straightforward corollary of Theorem 23.

By (**SameSize**) and (**MCCV**), for any $i \in [\![n]\!]$, π_i is the empirical mean of K independent Bernoulli random variables with common parameter $\mathbb{P}(i \in T_K^c) = p/n$. Then, by Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10)

$$\forall y > 0, \forall i \in [\![n]\!] \, , \qquad \mathbb{P}\left(\pi_i - \frac{p}{n} > \sqrt{\frac{2p(n-p)y}{n^2 B}} + \frac{x}{3B} \right) \leqslant \mathrm{e}^{-y} \enspace .$$

A union bound over $i \in [\![n]\!]$ yields that for any $x > 0$,

$$\mathbb{P}\left(\pi^* \leqslant 1 \wedge \left(\frac{2p}{n} + \frac{\log n + x}{B} \right) \right) \geqslant 1 - \mathrm{e}^{-x} \enspace ,$$

where we used also that $\pi^* \leqslant 1$ almost surely. Theorem 9 follows. ∎

We finally prove Theorem 23.

Proof of Theorem 23 Throughout the proof, L denotes some positive numerical constant, whose value may change from line to line. Given Eq. (64), the proof relies on concentration inequalities that are detailed in Section C.2. Let us fix $x \geqslant 0$ and define for every $\kappa \geqslant 1$ the event $\Omega_{good}(\kappa, x)$ where all the following inequalities hold for any $m, m' \in \mathcal{M}_n$ and any $\epsilon \in (0,1]$

$$\begin{aligned}
&\left| \sum_{i=1}^n \omega_{i,i} \bigl(\Psi_m(\xi_i) - \mathcal{D}_m \bigr) \right| \leqslant \epsilon \frac{\mathcal{D}_m}{n-p} + \kappa \frac{(n+A)x}{\epsilon (n-p)^2} \\
&\sum_{i=1}^n (-2\omega_{i,i} + \sigma_i) \bigl(s_m(\xi_i) - P s_m - s_{m'}(\xi_i) - P s_{m'} \bigr) \leqslant \epsilon \| s_m - s_{m'} \|^2 + R_n^{21}(x, \epsilon, \pi^*, A) \\
&\left| \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} U_m(\xi_i, \xi_j) \right| \leqslant \epsilon \frac{\mathcal{D}_m}{n-p} + \kappa \frac{n}{(n-p)^2} \left(1 + \pi^* \frac{n}{p} \right) \left[\frac{nAx}{(n-p)\epsilon} + (n+A) \frac{x^2}{n} \right] \\
&\left| \| \widehat{s}_m - s_m \|^2 - \frac{\mathcal{D}_m}{n} \right| \leqslant \epsilon \frac{\mathcal{D}_m}{n} + \kappa \frac{A x^2}{\epsilon^3 n} \enspace .
\end{aligned}$$

It follows from Lemmas 14, 20, 21 and 22 that an absolute constant $\kappa > 0$ exists such that $\mathbb{P}(\Omega_{good}(\kappa, x)) \geqslant 1 - |\mathcal{M}_n|^2 \mathrm{e}^{-x} - 10 |\mathcal{M}_n| \mathrm{e}^{-x}$. Let us remark that we can assume

$x \geqslant \log(11) \geqslant 1$ in the following, since otherwise the above probability bound is negative. On $\Omega_{good}(\kappa, x)$, for every $m \in \mathcal{M}_n$ and $\epsilon \in (0,1)$,

$$\frac{\mathcal{D}_m}{n} \leqslant \frac{1}{1-\epsilon} \|\widehat{s}_m - s_m\|^2 + \frac{LAx^2}{\epsilon^3(1-\epsilon)n} \quad . \tag{70}$$

By definition of $\widehat{m}$, for every $m \in \mathcal{M}_n$,

$$\|\widehat{s}_{\widehat{m}} - s\|^2 \leqslant \|\widehat{s}_m - s\|^2 + \Big(\mathrm{crit}_{\mathrm{CV}}(m) - \|\widehat{s}_m - s\|^2\Big) - \Big(\mathrm{crit}_{\mathrm{CV}}(\widehat{m}) - \|\widehat{s}_{\widehat{m}} - s\|^2\Big) \quad . \tag{71}$$

In addition, by Eq. (64),

$$\begin{aligned}
\mathrm{crit}_{\mathrm{CV}}(m) - \|\widehat{s}_m - s\|^2 = \frac{\mathcal{D}_m - \|s_m\|^2}{n-p} &- \underbrace{\left(\|s_m\|^2 + \|s_m - s\|^2\right)}_{=\|s\|^2} - \frac{\mathcal{D}_m}{n} \\
&+ \sum_{i=1}^n \omega_{i,i}\big(\Psi_m(\xi_i) - \mathcal{D}_m\big) + \sum_{i=1}^n (-2\omega_{i,i} + \sigma_i)\big(s_m(\xi_i) - Ps_m\big) \\
&+ \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} U_m(\xi_i, \xi_j) - \left(\|\widehat{s}_m - s\|^2 - \frac{\mathcal{D}_m}{n}\right) \quad .
\end{aligned}$$

So, on $\Omega_{good}(\kappa, x)$, for every $m, m' \in \mathcal{M}_n$ and $\epsilon \in (0, 1/5)$,

$$\begin{aligned}
&\mathrm{crit}_{\mathrm{CV}}(m) - \|\widehat{s}_m - s\|^2 - \Big(\mathrm{crit}_{\mathrm{CV}}(m') - \|\widehat{s}_{m'} - s\|^2\Big) \\
&\leqslant \mathcal{D}_m \left(\frac{1+2\epsilon}{n-p} - \frac{1-\epsilon}{n}\right)_+ + \mathcal{D}_{m'} \left(\frac{1+\epsilon}{n} - \frac{1-2\epsilon}{n-p}\right)_+ + \epsilon \|s_m - s_{m'}\|^2 \\
&\quad + R_n^{21}(x, \epsilon, \pi^*, A) + \frac{Ln}{(n-p)^2}\left(1 + \pi^* \frac{n}{p}\right)\left(\frac{nAx}{(n-p)\epsilon} + \frac{(A \vee 1)x^2}{\epsilon^3}\right) + \frac{\|s_{m'}\|^2}{n-p} \\
&\leqslant \frac{n}{1-\epsilon}\left(\frac{1+2\epsilon}{n-p} - \frac{1-\epsilon}{n}\right)_+ \|\widehat{s}_m - s_m\|^2 + \frac{n}{1-\epsilon}\left(\frac{1+\epsilon}{n} - \frac{1-2\epsilon}{n-p}\right)_+ \|\widehat{s}_{m'} - s_{m'}\|^2 \\
&\quad + 2\epsilon\|s_m - s\|^2 + 2\epsilon\|s_{m'} - s\|^2 \\
&\quad + R_n^{21}(x, \epsilon, \pi^*, A) + \frac{Ln}{(n-p)^2}\left(1 + \pi^* \frac{n}{p}\right)\left(\frac{nAx}{(n-p)\epsilon} + \frac{(A \vee 1)x^2}{\epsilon^3}\right) \\
&\leqslant \max\left\{\frac{1}{1-\epsilon}\left(\frac{1}{\tau_n} - 1 + \epsilon + \frac{2\epsilon}{\tau_n}\right)_+, 2\epsilon\right\} \|\widehat{s}_m - s\|^2 \\
&\quad + \max\left\{\frac{1}{1-\epsilon}\left(1 - \frac{1}{\tau_n} + \epsilon + \frac{2\epsilon}{\tau_n}\right)_+, 2\epsilon\right\} \|\widehat{s}_{m'} - s\|^2 \\
&\quad + R_n^{21}(x, \epsilon, \pi^*, A) + \frac{L}{n\tau_n^2}\left(1 + \frac{\pi^*}{1-\tau_n}\right)\left(\frac{Ax}{\tau_n \epsilon} + \frac{(A \vee 1)x^2}{\epsilon^3}\right) \\
&\leqslant \left(\frac{1}{\tau_n} - 1 + \frac{L\epsilon}{\tau_n}\right)\|\widehat{s}_m - s\|^2 + \frac{4\epsilon}{\tau_n}\|\widehat{s}_{m'} - s\|^2 \\
&\quad + R_n^{21}(x, \epsilon, \pi^*, A) + \frac{L}{n\tau_n^2}\left(1 + \frac{\pi^*}{1-\tau_n}\right)\left(\frac{Ax}{\tau_n \epsilon} + \frac{(A \vee 1)x^2}{\epsilon^3}\right)
\end{aligned}$$

where we used Eq. (70) for the second inequality. Note also that by Lemma 21,

$$\begin{aligned}
R_n^{21}(x,\epsilon,\pi^*,A) &\leqslant \frac{\kappa}{n\epsilon}\left[Ax\left(\frac{1}{n\tau_n^2}+\frac{\pi^\star}{1-\tau_n}\right)+x^2\left(\frac{1}{n\tau_n^2}+\frac{\pi^\star}{1-\tau_n}\right)^2\right] \\
&\leqslant \frac{\kappa}{n\epsilon}\left[Ax\left(\frac{1}{\tau_n}+\frac{\pi^\star}{1-\tau_n}\right)+x^2\left(\frac{1}{\tau_n}+\frac{\pi^\star}{1-\tau_n}\right)^2\right] \\
&\leqslant \frac{\kappa}{n\epsilon}\left[\frac{Ax}{\tau_n}\left(1+\frac{\pi^\star}{1-\tau_n}\right)+\frac{x^2}{\tau_n^2}\left(1+\frac{\pi^\star}{1-\tau_n}\right)^2\right]
\end{aligned}$$

where the term $\frac{x^2}{\tau_n^2}(1+\frac{\pi^\star}{1-\tau_n})^2$ is not present in $R_n^{21}(x,\epsilon,\pi^*,A)$ under assumption (**H2**), so that $R_n^{21}(x,\epsilon,\pi^*,A)\leqslant\rho_4(\epsilon,x,\kappa,n,\tau_n,\pi^*,A)$ whatever the assumption among (**H2**) and (**H2′**).

Therefore, Eq. (71) yields that, on $\Omega_{good}(\kappa,x)$, for every $m,m'\in\mathcal{M}_n$ and $\epsilon\in(0,1/5)$,

$$\begin{aligned}
\left(1-\frac{4\epsilon}{\tau_n}\right)\|\widehat{s}_{\widehat{m}}-s\|^2 &\leqslant \frac{1+L\epsilon}{\tau_n}\|\widehat{s}_m-s\|^2+R_n^{21}(x,\epsilon,\pi^*,A) \\
&\quad+\frac{L}{n\tau_n^2}\left(1+\frac{\pi^*}{1-\tau_n}\right)\left(\frac{Ax}{\tau_n\epsilon}+\frac{(A\vee 1)x^2}{\epsilon^3}\right)
\end{aligned}$$

hence the result by changing ϵ into ϵ/L. ■

C.4 Variance (Proof of Theorem 10)

We prove in this section the variance computation of Theorem 10, which is a straightforward corollary of the following result, since Ψ_{m_1} and Ψ_{m_2} are constant for regular histogram models.

Theorem 24 *We consider the setting and notation of Theorem 6. We recall that*

$$\mathcal{C}^{\mathrm{MCCV}}(m)=\mathrm{crit}_{\mathrm{CV}}(m,(T_K)_{1\leqslant K\leqslant B})$$

for some sequence $T_1,\dots,T_B$ of subsets of $[\![n]\!]$ satisfying (**SameSize**), (**MCCV**) *and* (**Ind**), *where* $\mathrm{crit}_{\mathrm{CV}}$ *is defined by Eq. (25) Then, we have*

$$\begin{aligned}
&\mathrm{Var}\Big(\mathcal{C}^{\mathrm{MCCV}}(m_1)-\mathcal{C}^{\mathrm{MCCV}}(m_2)\Big) \\
&=C_1^{\mathrm{MC}}(B,n,\tau_n)\frac{2}{n^2}\mathbf{B}(m_1,m_2) \qquad (72)\\
&\quad+\frac{4}{Bn}\frac{1}{n^2\tau_n^3}\mathrm{Var}\Big((s_{m_1}-s_{m_2})(\xi_1)-\frac{1}{2}(\Psi_{m_1}-\Psi_{m_2})(\xi_1)\Big) \\
&\quad+\frac{4}{Bn}\frac{1}{1-\tau_n}\mathrm{Var}\big(s_{m_1}(\xi_1)-s_{m_2}(\xi_1)\big) \\
&\quad+\left(1-\frac{1}{B}\right)\frac{4}{n}\mathrm{Var}\Big(\left(1+\frac{1}{n\tau_n}\right)(s_{m_1}-s_{m_2})(\xi_1)-\frac{1}{2n\tau_n}(\Psi_{m_1}-\Psi_{m_2})(\xi_1)\Big)
\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}\Bigl(\mathcal{C}^{\mathrm{MCCV}}(m_1)\Bigr) &= C_1^{\mathrm{MC}}(B,n,\tau_n)\frac{2}{n^2}\beta(m_1,m_1) \\ &\quad + \frac{1}{B}\frac{4}{n}\left[\frac{1}{n^2\tau_n^3}\operatorname{Var}\Bigl(s_{m_1}(\xi_1) - \frac{1}{2}\Psi_{m_1}(\xi_1)\Bigr) + \frac{1}{1-\tau_n}\operatorname{Var}\bigl(s_{m_1}(\xi_1)\bigr)\right] \\ &\quad + \Bigl(1-\frac{1}{B}\Bigr)\frac{4}{n}\operatorname{Var}\biggl(\Bigl(1+\frac{1}{n\tau_n}\Bigr)s_{m_1}(\xi_1) - \frac{1}{2n\tau_n}\Psi_{m_1}(\xi_1)\biggr)\end{aligned} \tag{73}$$

where

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

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

Theorem 24 is proved below. Note that a similar argument can be used for computing the variance of Monte-Carlo penalized criteria, where the Monte-Carlo penalty is defined from hold-out penalties similarly to MCCV. Indeed, given Lemma 25, we only need to compute the variance of hold-out penalized criteria (as done by Proposition 28) and the variance of leave-p-out criteria (as done by combining Lemma 1 and Theorem 6).

Before proving Theorem 24, we state and prove a general result that relates the variance of (increments of) Monte-Carlo CV criteria to the variance of (increments of) hold-out and leave-p-out criteria.

Lemma 25 *Let* $n \geqslant p \geqslant 1$ *and* $F : \mathcal{X}^n \times \mathfrak{P}(\llbracket n \rrbracket) \to \mathbb{R}$ *be some mesurable function. Let* D_n *denote some sample of* n *independent variables with common distribution* P*. Assume that* (**SameSize**), (**Ind**) *and* (**MCCV**) *hold true, as well as*

$$\forall T \in \mathcal{E}_{n-p}, \qquad \mathbb{E}\bigl[F(D_n,T)^2\bigr] < +\infty \ . \tag{74}$$

Let $B \geqslant 1$ *and* $T_1,\ldots,T_B$ *be some random sequence of subsets of* $\llbracket n \rrbracket$*. Let us define*

$$Z_B := \frac{1}{B}\sum_{K=1}^{B} F(D_n,T_K) \qquad \text{and} \qquad F^{\mathrm{lpo}}(D_n,p) := \frac{1}{\binom{n}{p}}\sum_{T\in\mathcal{E}_{n-p}} F(D_n,T) \ .$$

Then, we have

$$\begin{aligned}\operatorname{Var}(Z_B) &= \operatorname{Var}\bigl(F^{\mathrm{lpo}}(D_n,p)\bigr) + \frac{1}{B}\mathbb{E}\Bigl[\operatorname{Var}\bigl(F(D_n,T_1)\,|\,D_n\bigr)\Bigr] &(75)\\ &= \operatorname{Var}\bigl(F^{\mathrm{lpo}}(D_n,p)\bigr) + \frac{1}{B}\Bigl[\operatorname{Var}\bigl(F(D_n,T_1)\bigr) - \operatorname{Var}\bigl(F^{\mathrm{lpo}}(D_n,p)\bigr)\Bigr] &(76)\\ &= \Bigl(1-\frac{1}{B}\Bigr)\operatorname{Var}\bigl(F^{\mathrm{lpo}}(D_n,p)\bigr) + \frac{1}{B}\operatorname{Var}\bigl(F(D_n,T_1)\bigr) \ .\end{aligned}$$

Proof of Lemma 25 By (74), Z_B admits a finite variance. Then, we can write that

$$\operatorname{Var}(Z_B) = \operatorname{Var}\bigl(\mathbb{E}[Z_B\,|\,D_n]\bigr) + \mathbb{E}\bigl[\operatorname{Var}(Z_B\,|\,D_n)\bigr] \ .$$

By (**SameSize**), (**MCCV**) and (**Ind**),

$$\mathbb{E}[Z_B \mid D_n] = F^{\mathrm{lpo}}(D_n, p) \qquad \text{and} \qquad \mathrm{Var}(Z_B \mid D_n) = \frac{1}{B} \mathrm{Var}\bigl(F(D_n, T_1) \mid D_n\bigr)$$

which proves Eq. (75). Eq. (76) follows by remarking that $Z_B = F(D_n, T_1)$ when $B = 1$. ■

We can now prove Theorem 24. The idea is to apply Lemma 25 when $F(D_n, T)$ is the hold-out estimator of the risk of $\widehat{s}_m$, so that $F^{\mathrm{lpo}}(D_n, p)$ corresponds to some leave-p-out estimator of the risk. Similarly, Lemma 25 applies when $F(D_n, T)$ is the difference between the hold-out estimators of the risks of $\widehat{s}_{m_1}$ and $\widehat{s}_{m_2}$.

Proof of Theorem 24 First note that the variance of the CV criterion at model m_1 can be deduced from the variance of the increment between the CV criterion at model m_1 and CV criterion at model m_0 with $S_{m_0} = \{0\}$ the null model. So, Eq. (73) directly follows from Eq. (72).

For proving Eq. (72), we apply Lemma 25 with

$$F(D_n, T) = F_{\mathrm{HO}, m_1, m_2}(D_n, T) := \mathrm{crit}_{\mathrm{HO}}(m_1, T) - \mathrm{crit}_{\mathrm{HO}}(m_2, T)$$

so that

$$F^{\mathrm{lpo}}_{\mathrm{HO}, m_1, m_2}(D_n, p) = \mathcal{C}_{\left(\frac{n/p - 1/2}{n/p - 1}, \mathcal{B}_{\mathrm{LOO}}\right)}(m_1) - \mathcal{C}_{\left(\frac{n/p - 1/2}{n/p - 1}, \mathcal{B}_{\mathrm{LOO}}\right)}(m_2)$$

by Eq. (10) in Lemma 1.

The variance of $F^{\mathrm{lpo}}_{\mathrm{HO}, m_1, m_2}(D_n, p)$ is given by Theorem 6: by Eq. (24) with $V = n$ and

$$C = \frac{n/p - 1/2}{n/p - 1} = 1 + \frac{p}{2(n-p)} \quad ,$$

we get

$$\begin{aligned}
&\mathrm{Var}\Bigl(F^{\mathrm{lpo}}_{\mathrm{HO}, m_1, m_2}(D_n, p)\Bigr) \\
&= \frac{2}{n^2} \left[1 + \frac{4}{n-1} \left(1 + \frac{p}{2(n-p)} \right)^2 - \frac{n}{(n-p)^2} \right] \mathbf{B}(m_1, m_2) \qquad (77) \\
&\quad + \frac{4}{n} \mathrm{Var}\left(\left(1 + \frac{1}{n-p} \right) (s_{m_1} - s_{m_2})(\xi_1) - \frac{1}{2(n-p)} (\Psi_{m_1} - \Psi_{m_2})(\xi_1) \right) \\
&= \left[1 + \frac{1}{n-1} \left(\frac{1}{\tau_n} + 1 \right)^2 - \frac{1}{n \tau_n^2} \right] \frac{2}{n^2} \mathbf{B}(m_1, m_2) \qquad (78) \\
&\quad + \frac{4}{n} \mathrm{Var}\left(\left(1 + \frac{1}{n \tau_n} \right) (s_{m_1} - s_{m_2})(\xi_1) - \frac{1}{2 n \tau_n} (\Psi_{m_1} - \Psi_{m_2})(\xi_1) \right)
\end{aligned}$$

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

It now remains to compute the variance of

$$F_{\mathrm{HO}, m_1, m_2}(D_n, T) := \mathrm{crit}_{\mathrm{HO}}(m_1, T) - \mathrm{crit}_{\mathrm{HO}}(m_2, T) \ .$$

By Eq. (62), $F_{\mathrm{HO},m_1,m_2}(D_n,T)$ has the same variance as $\mathcal{C}_{m_1}-\mathcal{C}_{m_2}$ where $\mathcal{C}_m$ is defined as in Lemma 16 with

$$\overline{\omega}_{i,j}=\omega_{i,j}^{\mathrm{HO}}(T) \qquad \overline{\sigma}_i=\sigma_i^{\mathrm{HO}}(T) \qquad \text{and} \qquad f_m=s_m \ .$$

Since $|T|=n-p$, we have

$$\begin{aligned}
\sum_{1\leqslant i\neq j\leqslant n}\left(\overline{\omega}_{i,j}^2+\overline{\omega}_{i,j}\overline{\omega}_{j,i}\right) &= \sum_{i,j\in T,\, i\neq j}\left(\overline{\omega}_{i,j}^2+\overline{\omega}_{i,j}\overline{\omega}_{j,i}\right)+\sum_{1\leqslant i\leqslant n, j\in T^c, i\neq j}\underbrace{\left(\overline{\omega}_{i,j}^2+\overline{\omega}_{i,j}\overline{\omega}_{j,i}\right)}_{=0} \\
&\quad +\sum_{i\in T^c,\, j\in T}\left(\overline{\omega}_{i,j}^2+\overline{\omega}_{i,j}\underbrace{\overline{\omega}_{j,i}}_{=0}\right) \\
&= \frac{2(n-p-1)}{(n-p)^3}+\frac{4}{p(n-p)}
\end{aligned}$$

and

$$\sum_{i=1}^n\overline{\omega}_{i,i}^2=\frac{1}{(n-p)^3} \qquad \sum_{i=1}^n\overline{\omega}_{i,i}\overline{\sigma}_i=0 \qquad \sum_{i=1}^n\overline{\sigma}_i^2=\frac{4}{p} \ .$$

Therefore, by Lemma 16,

$$\begin{aligned}
&\operatorname{Var}\left(\mathrm{crit}_{\mathrm{HO}}(m_1,T)-\mathrm{crit}_{\mathrm{HO}}(m_2,T)\right) \\
&=\left(\frac{2(n-p-1)}{(n-p)^3}+\frac{4}{p(n-p)}\right)\operatorname{Var}\left(U_{m_1}(\xi_1,\xi_2)-U_{m_2}(\xi_1,\xi_2)\right) \\
&\quad+\frac{1}{(n-p)^3}\operatorname{Var}\left(U_{m_1}(\xi_1,\xi_1)-U_{m_2}(\xi_1,\xi_1)\right)+\frac{4}{p}\operatorname{Var}\left(s_{m_1}(\xi_1)-s_{m_2}(\xi_1)\right) \\
&=\left(\frac{2(n-p-1)}{(n-p)^3}+\frac{4}{p(n-p)}\right)\mathbf{B}(m_1,m_2) \\
&\quad+\frac{1}{(n-p)^3}\operatorname{Var}\left((\Psi_{m_1}-\Psi_{m_2})(\xi_1)-2(s_{m_1}-s_{m_2})(\xi_1)\right)+\frac{4}{p}\operatorname{Var}\left(s_{m_1}(\xi_1)-s_{m_2}(\xi_1)\right) \\
&=\left(\frac{1}{\tau_n^2}+\frac{2}{\tau_n(1-\tau_n)}-\frac{1}{n\tau_n^3}\right)\frac{2}{n^2}\mathbf{B}(m_1,m_2) \qquad (79) \\
&\quad+\frac{4}{n}\frac{1}{n^2\tau_n^3}\operatorname{Var}\left((s_{m_1}-s_{m_2})(\xi_1)-\frac{1}{2}(\Psi_{m_1}-\Psi_{m_2})(\xi_1)\right) \\
&\quad+\frac{4}{n}\frac{1}{1-\tau_n}\operatorname{Var}\left(s_{m_1}(\xi_1)-s_{m_2}(\xi_1)\right)
\end{aligned}$$

where we used that $\operatorname{Var}(U_{m_1}(\xi_1,\xi_2)-U_{m_2}(\xi_1,\xi_2))=\mathbf{B}(m_1,m_2)$ as proved at the beginning of Section A.4, and that $U_m(\xi_1,\xi_1)=\Psi_m(\xi_1)-2s_m(\xi_1)+\|s_m\|^2$.

Combining Eq. (79) and (78) with Lemma 25, we get Eq. (72). ■

Appendix D. Results on Hold-Out Penalization

This section gathers the proof of Theorem 12 (oracle inequality for hold-out penalization) and the variance computations we can make for hold-penalization.

D.1 Proof of Theorem 12

The hold-out penalty is equal to

$$\begin{aligned}\mathrm{pen}_{\mathrm{HO}}(m,T,x) &= 2x(1-\tau_n)^2\Big(P_n^{(T)} - P_n^{(T^c)}\Big)\Big(\widehat{s}_m^{(T)} - \widehat{s}_m^{(T^c)}\Big)\\ &= 2x(1-\tau_n)^2 \sum_{\lambda\in\Lambda_m}\Big[\big(P_n^{(T)} - P_n^{(T^c)}\big)(\psi_\lambda)\Big]^2 \enspace ,\end{aligned}$$

where we recall that $\tau_n = |T|/n$. As for Theorem 5, the oracle inequality is based on a concentration result for $\mathrm{pen}_{\mathrm{HO}}(m,T,x)$. Let us start with an exact formula for the hold-out penalty (Lemma 26, analogous to Lemma 13).

Lemma 26 *For all $m\in\mathcal{M}_n$, we have*

$$\mathrm{pen}_{\mathrm{HO}}(m,T,x) = 2x(1-\tau_n)^2\Big[\big\|\widehat{s}_m^{(T)} - s_m\big\|^2 + \big\|\widehat{s}_m^{(T^c)} - s_m\big\|^2 - 2\Big(P_n^{(T)} - P\Big)\Big(\widehat{s}_m^{(T^c)} - s_m\Big)\Big].$$

In particular, we have

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

Proof By definition

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

■

Lemma 27 *For all $m\in\mathcal{M}_n$ and $x>0$, with probability larger than $1-2\mathrm{e}^{-x}$, for all $\eta>0$, we have*

$$\Big|\big(P_n^{(T)} - P\big)\big(\widehat{s}_m^{(T^c)} - s_m\big)\Big| \leqslant \frac{\eta}{2}\big\|\widehat{s}_m^{(T^c)} - s_m\big\|^2 + \frac{2\|s\|_\infty x}{\eta\tau_n n} + \frac{b_m^2 x^2}{9\eta(\tau_n n)^2} \enspace .$$

Proof Let us apply Bernstein's inequality to the function $(\widehat{s}_m^{(T^c)} - s_m)$, conditionally to $(\xi_i)_{i\notin T}$. Recall that $v_m^2 \leqslant \|s\|_\infty$, hence

$$\begin{aligned}&\big\|\widehat{s}_m^{(T^c)} - s_m\big\|_\infty \leqslant \big\|\widehat{s}_m^{(T^c)} - s_m\big\| b_m\\ \text{and}\qquad &\mathrm{Var}\Big(\widehat{s}_m^{(T^c)}(\xi) - s_m(\xi)\,\big|\,(\xi_i)_{i\notin T}\Big) \leqslant \big\|\widehat{s}_m^{(T^c)} - s_m\big\|^2 v_m^2 \leqslant \big\|\widehat{s}_m^{(T^c)} - s_m\big\|^2\|s\|_\infty \enspace .\end{aligned}$$

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

$$\left| \left(P_n^{(T)} - P \right) \left(\widehat{s}_m^{(T^c)} - s_m \right) \right| \leqslant \left\| \widehat{s}_m^{(T^c)} - s_m \right\| \left(\sqrt{\frac{2\|s\|_\infty x}{\tau_n n}} + \frac{b_m x}{3\tau_n n} \right)$$
$$\leqslant \frac{\eta}{2} \left\| \widehat{s}_m^{(T^c)} - s_m \right\|^2 + \frac{1}{\eta} \left(\frac{2\|s\|_\infty x}{\tau_n n} + \frac{b_m^2 x^2}{9(\tau_n n)^2} \right) \enspace .$$

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

Proof of Theorem 12 From Lerasle (2011, Theorem 4.1)—a result recalled with Proposition 29 in Section F—, Lemma 26 and Lemma 27, an absolute constant κ exists such that, for all $x > 0$, with probability larger than $1 - 8\mathrm{e}^{-x}$, for all $\epsilon \in (0, 1]$, we have

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

We can then conclude the proof as in Theorem 5. ■

D.2 Variance

Proposition 28 *Let $(\psi_\lambda)_{\lambda \in \Lambda_{m_1}}$ and $(\psi_\lambda)_{\lambda \in \Lambda_{m_2}}$ denote two orthonormal families in $L^4(\mu)$. Assume that $|T| \in [\![n-1]\!]$ and denote for any $m \in \{m_1, m_2\}$,*

$$\mathcal{C}_{(C,T)}^{\mathrm{HO}}(m) = P_n \gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{HO}} \big(m, T, C\tau_n / (1 - \tau_n) \big) \enspace .$$

Then, with the notations introduced in Theorem 6, we have

$$\mathrm{Var}\Big(\mathcal{C}_{(C,T)}^{\mathrm{HO}}(m_1) \Big) = \frac{4}{n} \mathrm{Var}\left(\left(1 + \frac{2C - 1}{n} \right) s_{m_1}(\xi) - \frac{2C - 1}{2n} \Psi_{m_1}(\xi) \right)$$
$$+ \frac{2}{n^2} \left[1 + 4C^2 - \frac{(2C - 1)^2}{n} \right] \beta(\Lambda_{m_1}, \Lambda_{m_1})$$
$$+ \frac{4C^2}{n^3} \frac{(1 - 2\tau_n)^2}{\tau_n (1 - \tau_n)} \Big(\mathrm{Var}\big(\Psi_{m_1}(\xi) - 2 s_{m_1}(\xi) \big) - 2\beta(m, m) \Big)$$

and

$$\mathrm{Var}\Big(\mathcal{C}_{(C,T)}^{\mathrm{HO}}(m_1) - \mathcal{C}_{(C,T)}^{\mathrm{HO}}(m_2) \Big)$$
$$= \frac{4}{n} \mathrm{Var}\left(\left(1 + \frac{2C - 1}{n} \right) \big(s_{m_1}(\xi) - s_{m_2}(\xi) \big) - \frac{2C - 1}{2n} \big(\Psi_{m_1}(\xi) - \Psi_{m_2}(\xi) \big) \right)$$
$$+ \frac{2}{n^2} \left(1 + 4C^2 - \frac{(2C - 1)^2}{n} \right) \mathbf{B}(m_1, m_2) \qquad (80)$$

$$+ \frac{4C^2}{n^3} \frac{(1-2\tau_n)^2}{\tau_n(1-\tau_n)} \left(\operatorname{Var}\left(\left(\Psi_{m_1}(\xi) - \Psi_{m_2}(\xi) \right) - 2\left(s_{m_1}(\xi) - s_{m_2}(\xi) \right) \right) - 2\mathbf{B}(m_1, m_2) \right) \enspace .$$

Proof By definition

$$\begin{aligned} \operatorname{pen}_{\mathrm{HO}}(m_1, T, x) &= 2x \sum_{\lambda \in \Lambda_{m_1}} \left[\left(P_n^{(T)} - P_n \right) \psi_\lambda \right]^2 \\ &= \frac{2x}{n^2} \sum_{\lambda \in \Lambda_{m_1}} \left(\sum_{i=1}^n \left(\frac{1}{\tau_n} \mathbb{1}_{i \in T} - 1 \right) \psi_\lambda(\xi_i) \right)^2 \\ &= \frac{2x}{n^2} \sum_{i,j=1}^n E_{i,j}^{(\mathrm{HO})} U_{m_1}(\xi_i, \xi_j) \enspace , \end{aligned} \tag{81}$$

where, for all $i, j \in \{1 \ldots, n\}$, we recall that

$$\begin{aligned} U_{m_1}(\xi_i, \xi_j) &= \sum_{\lambda \in \Lambda_{m_1}} \left(\psi_\lambda(\xi_i) - P\psi_\lambda \right) \left(\psi_\lambda(\xi_j) - P\psi_\lambda \right) \\ \text{and} \qquad E_{i,j}^{(\mathrm{HO})} &= \left(\frac{1}{\tau_n} \mathbb{1}_{i \in T} - 1 \right) \left(\frac{1}{\tau_n} \mathbb{1}_{j \in T} - 1 \right) \enspace . \end{aligned}$$

Therefore, from Eq. (57), if $x = C\tau_n / (1-\tau_n)$, we have

$$\begin{aligned} \mathcal{C}_{(C,T)}^{\mathrm{HO}}(m_1) &:= P_n \gamma(\widehat{s}_{m_1}) + \operatorname{pen}_{\mathrm{HO}}(m_1, T, x) \\ &= \sum_{1 \leqslant i,j \leqslant n} \frac{2x E_{i,j}^{(\mathrm{HO})} - 1}{n^2} U_{m_1}(\xi_i, \xi_j) - \sum_{i=1}^n \frac{2}{n} s_{m_1}(\xi_i) + \left\| s_{m_1} \right\|^2 \enspace . \end{aligned}$$

By definition

$$E_{i,j}^{(\mathrm{HO})} = \left(\frac{1-\tau_n}{\tau_n} \right)^2 \mathbb{1}_{i,j \in T} - \frac{1-\tau_n}{\tau_n} \mathbb{1}_{i \in T, j \notin T} - \frac{1-\tau_n}{\tau_n} \mathbb{1}_{i \notin T, j \in T} + \mathbb{1}_{i,j \notin T} \enspace .$$

Therefore, we can compute

$$\sum_{i=1}^n E_{i,i}^{(\mathrm{HO})} = n \left[\tau_n \left(\frac{1-\tau_n}{\tau_n} \right)^2 + 1 - \tau_n \right] = n \frac{1-\tau_n}{\tau_n} \enspace , \tag{82}$$

$$\sum_{i=1}^n \left(E_{i,i}^{(\mathrm{HO})} \right)^2 = n \left[\tau_n \left(\frac{1-\tau_n}{\tau_n} \right)^4 + 1 - \tau_n \right] = n(1-\tau_n) \frac{(1-\tau_n)^3 + \tau_n^3}{\tau_n^3} \enspace . \tag{83}$$

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

$$\sum_{1 \leqslant i,j \leqslant n} E_{i,j}^{(\mathrm{HO})} = \mathbb{E}\left[\left(\sum_{i=1}^n \left(\frac{1}{\tau_n} \mathbb{1}_{i \in T} - 1 \right) \right)^2 \right] = 0 \enspace ,$$

so Eq. (82) implies that

$$\sum_{1 \leqslant i \neq j \leqslant n} E_{i,j}^{(\mathrm{HO})} = - \sum_{i=1}^{n} E_{i,i}^{(\mathrm{HO})} = -n \frac{1-\tau_n}{\tau_n} \enspace . \tag{84}$$

In addition, we compute

$$\begin{aligned}
\sum_{1 \leqslant i \neq j \leqslant n} \left(E_{i,j}^{(\mathrm{HO})} \right)^2 &= 2n^2 \tau_n (1-\tau_n) \left(\frac{1-\tau_n}{\tau_n} \right)^2 + n\tau_n (n\tau_n - 1) \left(\frac{1-\tau_n}{\tau_n} \right)^4 \\
&\qquad + n(1-\tau_n) \bigl[n(1-\tau_n) - 1 \bigr] \\
&= n^2 (1-\tau_n)^2 \left[2\frac{1-\tau_n}{\tau_n} + \left(\frac{1-\tau_n}{\tau_n} \right)^2 + 1 \right] \\
&\qquad - n(1-\tau_n) \left[\left(\frac{1-\tau_n}{\tau_n} \right)^3 + 1 \right] \\
&= n^2 \left(\frac{1-\tau_n}{\tau_n} \right)^2 - n(1-\tau_n) \frac{(1-\tau_n)^3 + \tau_n^3}{\tau_n^3} \enspace .
\end{aligned} \tag{85}$$

According to Eq. (81) and (57), $\operatorname{Var}(\mathcal{C}_{(C,T)}^{\mathrm{HO}}(m_1))$ can be computed using Lemma 16 with

$$\forall i,j \in \{1, \ldots, n\}, \quad \omega_{i,j} = \frac{1}{n^2} \left(2x E_{i,j}^{(\mathrm{HO})} - 1 \right) \qquad f_{m_1} = \frac{-2 s_{m_1}}{n} \qquad \text{and} \qquad \sigma_i = 1 \enspace .$$

So, using Eq. (82), (83), (84) and (85), we have

$$\begin{aligned}
\sum_{i=1}^{n} \omega_{i,i}^2 &= \frac{1}{n^4} \left[4x^2 \sum_{i=1}^{n} \left(E_{i,i}^{(\mathrm{HO})} \right)^2 - 4x \sum_{i=1}^{n} E_{i,i}^{(\mathrm{HO})} + n \right] \\
&= \frac{1}{n^3} \left[4x^2 (1-\tau_n) \frac{(1-\tau_n)^3 + \tau_n^3}{\tau_n^3} - 4x \frac{1-\tau_n}{\tau_n} + 1 \right] \\
\sum_{\substack{1 \leqslant i,j \leqslant n \\ i \neq j}} \omega_{i,j}^2 &= \frac{1}{n^4} \left[4x^2 \sum_{1 \leqslant i \neq j \leqslant n} \left(E_{i,j}^{(\mathrm{HO})} \right)^2 - 4x \sum_{1 \leqslant i \neq j \leqslant n} E_{i,j}^{(\mathrm{HO})} + n(n-1) \right] \\
&= \frac{1}{n^4} \left[4x^2 \left(n^2 \left(\frac{1-\tau_n}{\tau_n} \right)^2 - n(1-\tau_n) \frac{(1-\tau_n)^3 + \tau_n^3}{\tau_n^3} \right) + 4xn \frac{1-\tau_n}{\tau_n} + n(n-1) \right] \\
\sum_{i=1}^{n} \omega_{i,i} \sigma_i &= \frac{1}{n} \left(2x \frac{1-\tau_n}{\tau_n} - 1 \right) \enspace .
\end{aligned}$$

Therefore, by Lemma 16 with $m = m' = m_1$, we deduce

$$\begin{aligned}
\operatorname{Var}\Bigl(\mathcal{C}_{(C,T)}^{\mathrm{HO}}(m_1) \Bigr) &= \frac{1}{n^3} \left(4C^2 \frac{(1-2\tau_n)^2}{\tau_n(1-\tau_n)} + (2C-1)^2 \right) \operatorname{Var}\bigl(\Psi_{m_1}(\xi) - 2 s_{m_1}(\xi) \bigr) \\
&\qquad + \frac{2}{n^2} \left[1 + 4C^2 - \frac{1}{n} \left(4C^2 \frac{(1-2\tau_n)^2}{\tau_n(1-\tau_n)} + (2C-1)^2 \right) \right] \beta(m_1, m_1)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{4}{n^2}(2C-1)\operatorname{Cov}\bigl(\Psi_{m_1}(\xi) - 2s_{m_1}(\xi), s_{m_1}(\xi)\bigr) + \frac{4}{n}\operatorname{Var}\bigl(s_{m_1}(\xi)\bigr) \\
&= \frac{4}{n}\operatorname{Var}\left(\left(1 + \frac{2C-1}{n}\right)s_{m_1}(\xi) - \frac{2C-1}{2n}\Psi_{m_1}(\xi)\right) \\
&\quad + \frac{2}{n^2}\left(1 + 4C^2 - \frac{(2C-1)^2}{n}\right)\beta(\Lambda_{m_1}, \Lambda_{m_1}) \\
&\quad + \frac{4C^2}{n^3}\frac{(1-2\tau_n)^2}{\tau_n(1-\tau_n)}\Bigl[\operatorname{Var}\bigl(\Psi_{m_1}(\xi) - 2s_{m_1}(\xi)\bigr) - 2\beta(m_1, m_1)\Bigr] \ .
\end{aligned}
$$

Eq (80) follows from similar computations. ■

Appendix E. Additional Comments on Computational Issues

This section is an appendix to Section 7. We first detail a naive algorithm for computing V-fold criteria, Algorithm 2. Then, we prove Proposition 8 which shows that Algorithm 1 also computes correctly the V-fold criteria, much faster than Algorithm 2.

E.1 Naive Implementation

Algorithm 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 $\operatorname{Card}(m) = d_m$.*

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. *Compute the empirical risk:*
   (a) *train $\widehat{s}_m(\cdot)$ with the data set $(\xi_i)_{1 \leqslant i \leqslant n}$, 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 that the computational cost of evaluating ψ_λ at some point $\xi \in \Xi$ is of order 1, the computational cost of this naive algorithm 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.

E.2 Proof of Proposition 8

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

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

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

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

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

Therefore,

$$\lVert \widehat{s}_m \rVert^2 = -P_n \gamma(\widehat{s}_m) = P_n(\widehat{s}_m) = \frac{1}{V^2} \sum_{1 \leqslant i,j \leqslant V} P_n^{(\mathcal{B}_i)} \left(\widehat{s}_m^{(\mathcal{B}_j)} \right) = \frac{1}{V^2} \mathcal{S}$$

and

$$\begin{aligned} &\mathrm{crit}_{\mathrm{VFCV}}(m) \\ &= \frac{1}{V} \sum_{j=1}^V P_n^{(\mathcal{B}_j)} \gamma\left(\widehat{s}_m^{(-\mathcal{B}_j)} \right) \\ &= \frac{1}{V} \sum_{j=1}^V \left[\left\lVert \widehat{s}_m^{(-\mathcal{B}_j)} \right\rVert^2 - 2 P_n^{(\mathcal{B}_j)} \left(\widehat{s}_m^{(-\mathcal{B}_j)} \right) \right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{V} \sum_{j=1}^{V} \left(\frac{1}{(V-1)^2} \sum_{\substack{1 \leqslant i, \ell \leqslant V \\ i, \ell \neq j}} P_n^{(\mathcal{B}_i)} \left(\widehat{s}_m^{(\mathcal{B}_\ell)} \right) - \frac{2}{V-1} \sum_{i \neq j} P_n^{(\mathcal{B}_j)} \left(\widehat{s}_m^{(\mathcal{B}_i)} \right) \right) \\
&= \frac{1}{V(V-1)^2} \sum_{1 \leqslant i, \ell \leqslant V} \left[P_n^{(\mathcal{B}_i)} \big(\widehat{s}_m^{(\mathcal{B}_\ell)} \big) \sum_{j=1}^{V} \mathbb{1}_{i \neq j, \, \ell \neq j} \right] - \frac{2}{V(V-1)} \sum_{1 \leqslant i \neq j \leqslant V} P_n^{(\mathcal{B}_j)} \left(\widehat{s}_m^{(\mathcal{B}_i)} \right) \\
&= \frac{1}{V(V-1)^2} \sum_{1 \leqslant i, \ell \leqslant V} \Big[P_n^{(\mathcal{B}_i)} \big(\widehat{s}_m^{(\mathcal{B}_\ell)} \big) (V - 1 - \mathbb{1}_{i \neq \ell}) \Big] - \frac{2}{V(V-1)} (\mathcal{S} - \mathcal{T}) \\
&= \frac{1}{V(V-1)} \sum_{1 \leqslant i \leqslant V} \Big[P_n^{(\mathcal{B}_i)} \big(\widehat{s}_m^{(\mathcal{B}_i)} \big) \Big] + \frac{V-2}{V(V-1)^2} \sum_{1 \leqslant i \neq \ell \leqslant V} \Big[P_n^{(\mathcal{B}_i)} \big(\widehat{s}_m^{(\mathcal{B}_\ell)} \big) \Big] \\
&\qquad - \frac{2}{V(V-1)} (\mathcal{S} - \mathcal{T}) \\
&= \frac{1}{V(V-1)} \mathcal{T} + \frac{V-2}{V(V-1)^2} (\mathcal{S} - \mathcal{T}) - \frac{2}{V(V-1)} (\mathcal{S} - \mathcal{T}) \\
&= \frac{1}{V(V-1)} \mathcal{T} - \frac{1}{(V-1)^2} (\mathcal{S} - \mathcal{T}) \ ,
\end{aligned}
$$

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

$$V \times \mathrm{Card}(\Lambda_m) \times \frac{n}{V} = n \, \mathrm{Card}(\Lambda_m) \ .$$

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 $j = 1, \ldots, n$, find the unique $i(j) \in \{1, \ldots, V\}$ 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 $V^2 \, \mathrm{Card}(\Lambda_m) + n$. ■

Appendix F. Probabilistic Tool

Proposition 29 (Lerasle, 2011, Theorem 4.1 of the supplementary material) *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 us define*

$$P_N t = \frac{1}{N} \sum_{i=1}^{N} t(\xi_i) \qquad v^2 = \sup_{t \in S} P\big[(t - Pt)^2 \big]$$

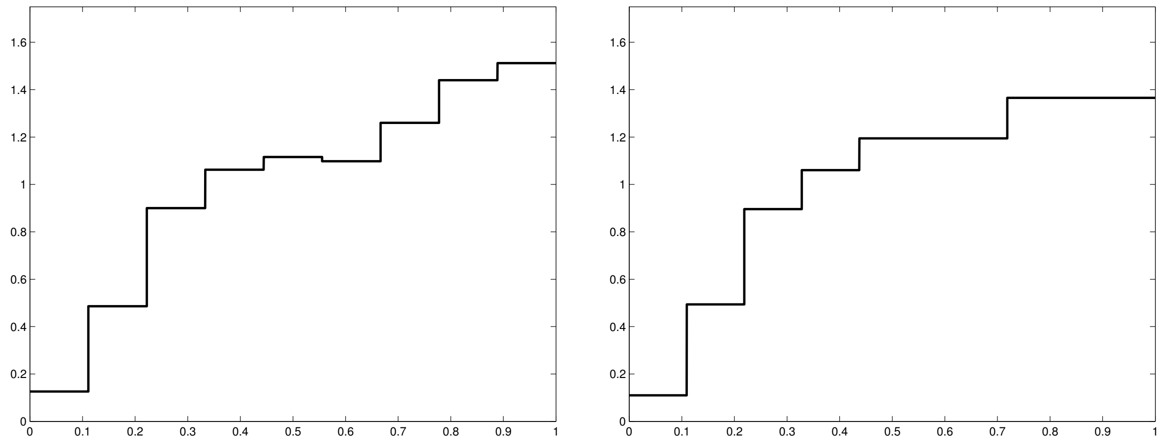


Figure 7: Oracle model for some sample of size $n = 500$, in setting L. Left: Regu. Right: Dya2.

$$Z = \sup_{t \in S} \{(P_N - P)(t)\} \qquad \textit{and} \qquad D = N \mathbb{E}\left[Z^2\right] \enspace .$$

There exists an absolute constant κ such that, for all $x > 0$, with probability larger than $1 - 2\mathrm{e}^{-x}$, for all $\epsilon \in (0, 1]$,

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

For instance, taking $S = \mathbb{B}_m$, by Eq. (3), this result applies to

$$Z = \sup_{t \in \mathbb{B}_m} \left\{ (P_n - P)(t) \right\} = \| \widehat{s}_m - s_m \|^2 \enspace .$$

Appendix G. Additional Simulation Results

This section provides simulation results in addition to the ones of Section 6.

Figure 7 is an analogous of Figure 2 in setting L, that illustrates the difference between the model collections Regu and Dya2.

Table 3 is an extended version of Table 2, with more procedures compared and two additional settings (L-Regu and S-Regu). Table 4 provides a similar comparison of model selection performances with a reduced sample size $n = 100$, again from $N = 10\,000$ independent samples.

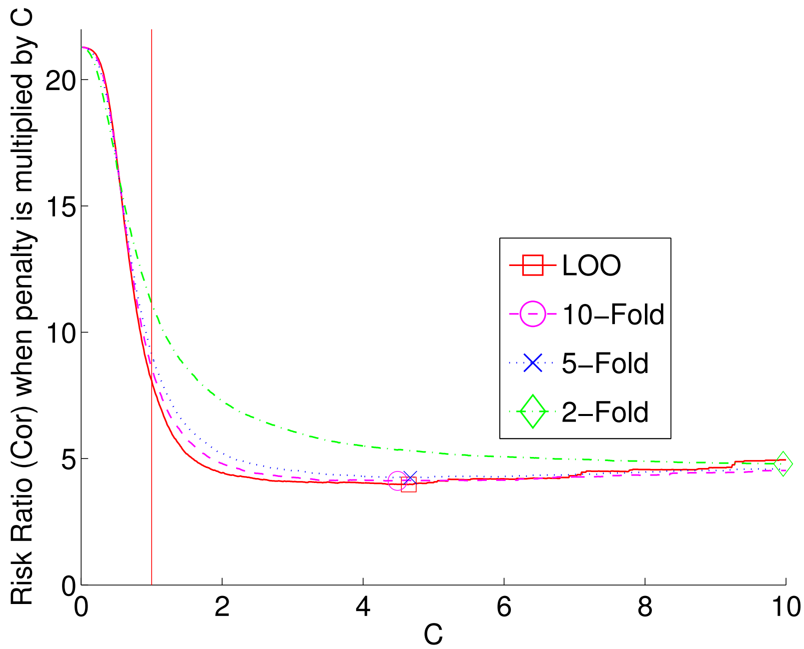
Experiment	L–Dya2	L–Regu	S–Dya2	S–Regu
$\mathbb{E}[\text{pen}_{\text{id}}]$	6.52 ± 0.05	2.33 ± 0.01	2.07 ± 0.01	1.75 ± 0.01
$1.25 \times \mathbb{E}[\text{pen}_{\text{id}}]$	4.81 ± 0.04	2.01 ± 0.01	1.94 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \mathbb{E}[\text{pen}_{\text{id}}]$	4.12 ± 0.03	$\mathbf{1.93 \pm 0.01}$	$\mathbf{1.92 \pm 0.01}$	1.65 ± 0.003
$2 \times \mathbb{E}[\text{pen}_{\text{id}}]$	$\mathbf{3.61 \pm 0.02}$	1.96 ± 0.01	2.01 ± 0.01	1.84 ± 0.004
pen_{dim}	8.27 ± 0.07	2.33 ± 0.01	3.21 ± 0.01	1.75 ± 0.01
$1.25 \times \text{pen}_{\text{dim}}$	5.95 ± 0.05	2.01 ± 0.01	3.01 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \text{pen}_{\text{dim}}$	4.99 ± 0.04	$\mathbf{1.94 \pm 0.01}$	3.03 ± 0.01	1.66 ± 0.003
$2 \times \text{pen}_{\text{dim}}$	4.38 ± 0.03	1.97 ± 0.01	3.24 ± 0.01	1.85 ± 0.004
pen_{LOO}	6.35 ± 0.05	2.33 ± 0.01	2.06 ± 0.01	1.75 ± 0.01
$1.25 \times \text{pen}_{\text{LOO}}$	4.62 ± 0.04	2.01 ± 0.01	1.92 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \text{pen}_{\text{LOO}}$	3.97 ± 0.03	$\mathbf{1.94 \pm 0.01}$	$\mathbf{1.90 \pm 0.005}$	1.66 ± 0.003
$2 \times \text{pen}_{\text{LOO}}$	$\mathbf{3.55 \pm 0.02}$	1.97 ± 0.01	1.98 ± 0.01	1.85 ± 0.004
pen_{VF} $(V{=}10)$	6.89 ± 0.06	2.42 ± 0.02	2.11 ± 0.01	1.77 ± 0.01
$1.25 \times \text{pen}_{\text{VF}}$ $(V{=}10)$	5.01 ± 0.04	2.04 ± 0.01	1.95 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$1.5 \times \text{pen}_{\text{VF}}$ $(V{=}10)$	4.27 ± 0.03	1.94 ± 0.01	1.92 ± 0.01	1.63 ± 0.004
$2 \times \text{pen}_{\text{VF}}$ $(V{=}10)$	3.68 ± 0.02	1.94 ± 0.01	1.98 ± 0.01	1.78 ± 0.004
pen_{VF} $(V{=}5)$	7.47 ± 0.06	2.55 ± 0.02	2.16 ± 0.01	1.80 ± 0.01
$1.25 \times \text{pen}_{\text{VF}}$ $(V{=}5)$	5.50 ± 0.04	2.10 ± 0.01	1.98 ± 0.01	1.63 ± 0.004
$1.5 \times \text{pen}_{\text{VF}}$ $(V{=}5)$	4.58 ± 0.03	1.96 ± 0.01	1.93 ± 0.01	$\mathbf{1.62 \pm 0.004}$
$2 \times \text{pen}_{\text{VF}}$ $(V{=}5)$	3.86 ± 0.02	$\mathbf{1.93 \pm 0.01}$	1.98 ± 0.01	1.73 ± 0.004
pen_{VF} $(V{=}2)$	10.21 ± 0.08	3.37 ± 0.03	2.39 ± 0.01	2.01 ± 0.01
$1.25 \times \text{pen}_{\text{VF}}$ $(V{=}2)$	7.69 ± 0.06	2.49 ± 0.02	2.15 ± 0.01	1.71 ± 0.01
$1.5 \times \text{pen}_{\text{VF}}$ $(V{=}2)$	6.41 ± 0.05	2.18 ± 0.01	2.05 ± 0.01	1.63 ± 0.004
$2 \times \text{pen}_{\text{VF}}$ $(V{=}2)$	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 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 $\mathbb{E}[\text{pen}_{\text{id}}]$.

Experiment	L–Dya2	L–Regu	S–Dya2	S–Regu
$\mathbb{E}[\text{pen}_{\text{id}}]$	8.38 ± 0.08	3.29 ± 0.03	1.97 ± 0.01	2.09 ± 0.01
$1.25 \times \mathbb{E}[\text{pen}_{\text{id}}]$	6.53 ± 0.07	2.61 ± 0.02	$\mathbf{1.93 \pm 0.01}$	1.72 ± 0.01
$1.5 \times \mathbb{E}[\text{pen}_{\text{id}}]$	5.59 ± 0.06	$\mathbf{2.46 \pm 0.02}$	$\mathbf{1.92 \pm 0.01}$	1.61 ± 0.01
$2 \times \mathbb{E}[\text{pen}_{\text{id}}]$	$\mathbf{4.72 \pm 0.05}$	2.57 ± 0.01	1.94 ± 0.005	$\mathbf{1.60 \pm 0.004}$
pen_{dim}	9.67 ± 0.09	3.28 ± 0.03	2.17 ± 0.01	2.09 ± 0.01
$1.25 \times \text{pen}_{\text{dim}}$	7.85 ± 0.08	2.62 ± 0.02	2.10 ± 0.01	1.72 ± 0.01
$1.5 \times \text{pen}_{\text{dim}}$	6.74 ± 0.07	$\mathbf{2.48 \pm 0.02}$	2.05 ± 0.01	1.62 ± 0.01
$2 \times \text{pen}_{\text{dim}}$	5.70 ± 0.06	2.60 ± 0.01	2.00 ± 0.01	1.61 ± 0.004
pen_{LOO}	8.10 ± 0.08	3.29 ± 0.03	1.97 ± 0.01	2.09 ± 0.01
$1.25 \times \text{pen}_{\text{LOO}}$	6.20 ± 0.06	2.62 ± 0.02	1.92 ± 0.01	1.72 ± 0.01
$1.5 \times \text{pen}_{\text{LOO}}$	5.18 ± 0.05	$\mathbf{2.49 \pm 0.02}$	1.91 ± 0.01	1.62 ± 0.01
$2 \times \text{pen}_{\text{LOO}}$	$\mathbf{4.44 \pm 0.04}$	2.59 ± 0.01	1.94 ± 0.005	1.61 ± 0.004
pen_{VF} (V=10)	8.61 ± 0.08	3.54 ± 0.04	1.97 ± 0.01	2.21 ± 0.01
$1.25 \times \text{pen}_{\text{VF}}$ (V=10)	6.76 ± 0.07	2.76 ± 0.02	1.92 ± 0.01	1.78 ± 0.01
$1.5 \times \text{pen}_{\text{VF}}$ (V=10)	5.77 ± 0.06	2.52 ± 0.02	$\mathbf{1.90 \pm 0.01}$	1.64 ± 0.01
$2 \times \text{pen}_{\text{VF}}$ (V=10)	4.81 ± 0.05	2.57 ± 0.01	1.91 ± 0.01	$\mathbf{1.60 \pm 0.004}$
pen_{VF} (V=5)	9.14 ± 0.08	3.92 ± 0.04	1.98 ± 0.01	2.34 ± 0.02
$1.25 \times \text{pen}_{\text{VF}}$ (V=5)	7.38 ± 0.07	2.90 ± 0.03	1.93 ± 0.01	1.85 ± 0.01
$1.5 \times \text{pen}_{\text{VF}}$ (V=5)	6.31 ± 0.06	2.60 ± 0.02	$\mathbf{1.91 \pm 0.01}$	1.68 ± 0.01
$2 \times \text{pen}_{\text{VF}}$ (V=5)	5.21 ± 0.05	2.56 ± 0.02	$\mathbf{1.90 \pm 0.01}$	$\mathbf{1.60 \pm 0.005}$
pen_{VF} (V=2)	11.15 ± 0.09	6.14 ± 0.08	2.01 ± 0.01	2.92 ± 0.02
$1.25 \times \text{pen}_{\text{VF}}$ (V=2)	9.61 ± 0.08	4.05 ± 0.05	1.97 ± 0.01	2.24 ± 0.01
$1.5 \times \text{pen}_{\text{VF}}$ (V=2)	8.60 ± 0.07	3.30 ± 0.03	1.94 ± 0.01	1.94 ± 0.01
$2 \times \text{pen}_{\text{VF}}$ (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

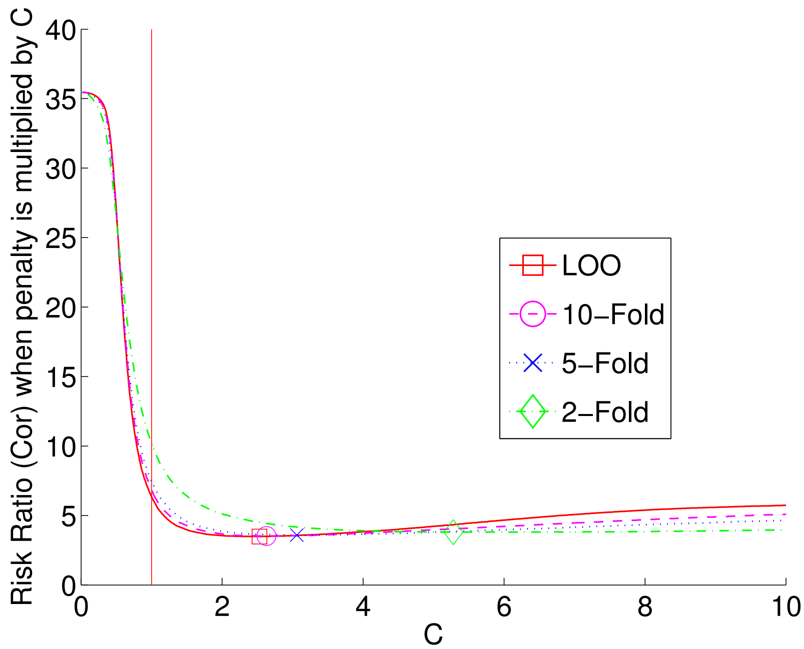
Table 4: Simulation results: settings L and S, $n = 100$. The best procedures (up to standard-deviations) are bolded, where the data-driven procedures are considered separately from the procedures using the knowledge of $\mathbb{E}[\text{pen}_{\text{id}}]$.

The influence of overpenalization is considered in Figures 8–15. As on Figure 3, the top graph represents the estimated model selection performance $C_{\mathrm{or}}(\mathcal{C}_{(C,\mathcal{B})})$ as a function of C, for various values of $V = |\mathcal{B}|$. Error bars are not shown on these graphs for clarity; all visible differences on the graph correspond to significant differences, as can be seen in Tables 3–4 for instance. The bottom tables in Figures 8–15 show the estimated model selection performance for three key values of C: the optimal one $C_n^\star$, the unbiased case ($C = 1$, which corresponds to an AIC-type penalty) and the value $C = \log(n)/2$ (which corresponds to a BIC-type penalty). The estimated value of the optimal overpenalizing constant $C_n^\star$ was obtained by minimizing over $C \in [0, 10]$ the estimated value of $C_{\mathrm{or}}(\mathcal{C}_{(C,\mathcal{B})})$. Error bars on $C_n^\star$ show the maximum of $|C_n^\star - C|$ over the set of values of C that are "not significantly worse than $C_n^\star$", where we define by convention "significantly worse" as having a $|C_{\mathrm{or}}(\mathcal{C}_{(C_n^\star,\mathcal{B})}) - C_{\mathrm{or}}(\mathcal{C}_{(C,\mathcal{B})})|$ larger than the sum of the corresponding error bars.



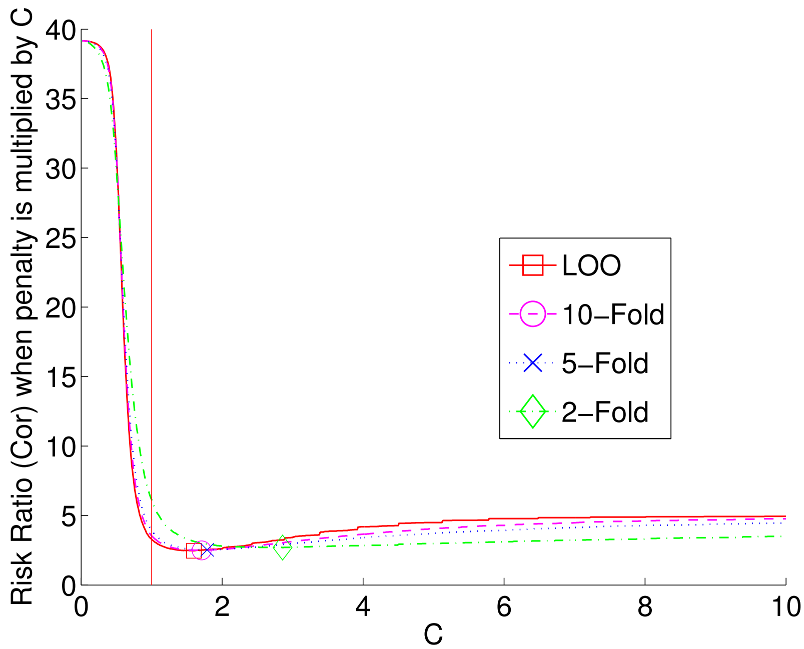
Penalty	$C_n^\star$	$C_{\mathrm{or}}(C = C_n^\star)$	$C_{\mathrm{or}}(C = 1)$	$C_{\mathrm{or}}(C = 2.30)$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	6.58 ± 1.33	3.89 ± 0.02	8.38 ± 0.08	$\mathbf{4.46 \pm 0.04}$
$\mathrm{pen}_{\mathrm{LOO}}$	4.65 ± 0.54	3.98 ± 0.02	8.10 ± 0.08	$\mathbf{4.29 \pm 0.04}$
pen 10F	4.49 ± 1.71	4.12 ± 0.03	8.61 ± 0.08	$\mathbf{4.53 \pm 0.04}$
pen 5F	4.67 ± 1.88	4.25 ± 0.03	9.14 ± 0.08	$\mathbf{4.90 \pm 0.05}$
pen 2F	9.96 ± 1.61	4.79 ± 0.04	11.15 ± 0.09	$\mathbf{6.80 \pm 0.06}$

Figure 8: Overpenalization in setting L-Dya2, $n = 100$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.



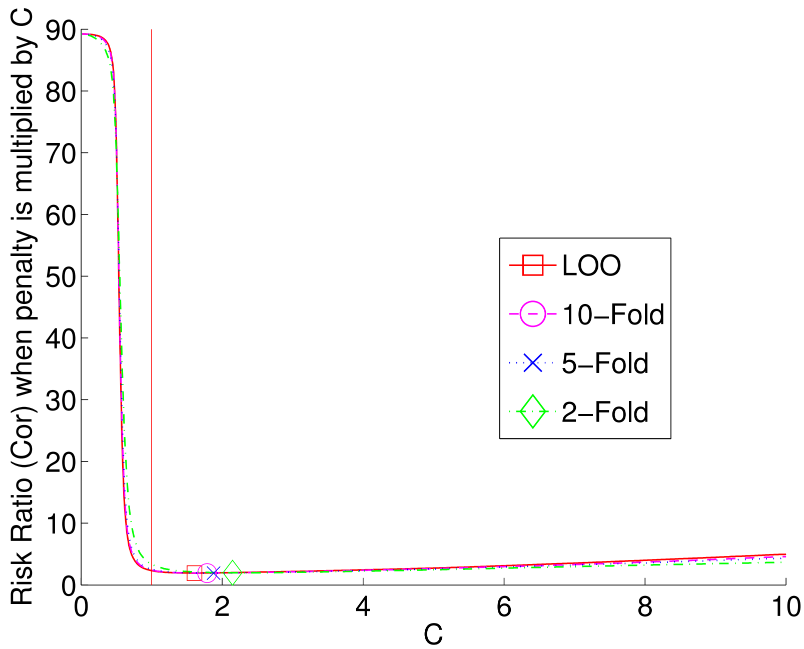
Penalty	$C_n^\star$	$C_{\mathrm{or}}(C = C_n^\star)$	$C_{\mathrm{or}}(C = 1)$	$C_{\mathrm{or}}(C = 3.11)$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	2.56 ± 0.40	3.53 ± 0.02	6.52 ± 0.05	$\mathbf{3.59 \pm 0.02}$
$\mathrm{pen}_{\mathrm{LOO}}$	2.53 ± 0.43	3.49 ± 0.02	6.35 ± 0.05	$\mathbf{3.58 \pm 0.02}$
pen 10F	2.63 ± 0.49	3.52 ± 0.02	6.89 ± 0.06	$\mathbf{3.55 \pm 0.02}$
pen 5F	3.06 ± 0.67	3.59 ± 0.02	7.47 ± 0.06	$\mathbf{3.59 \pm 0.02}$
pen 2F	5.28 ± 1.90	3.80 ± 0.02	10.21 ± 0.08	$\mathbf{4.14 \pm 0.03}$

Figure 9: Overpenalization in setting L-Dya2, $n = 500$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.



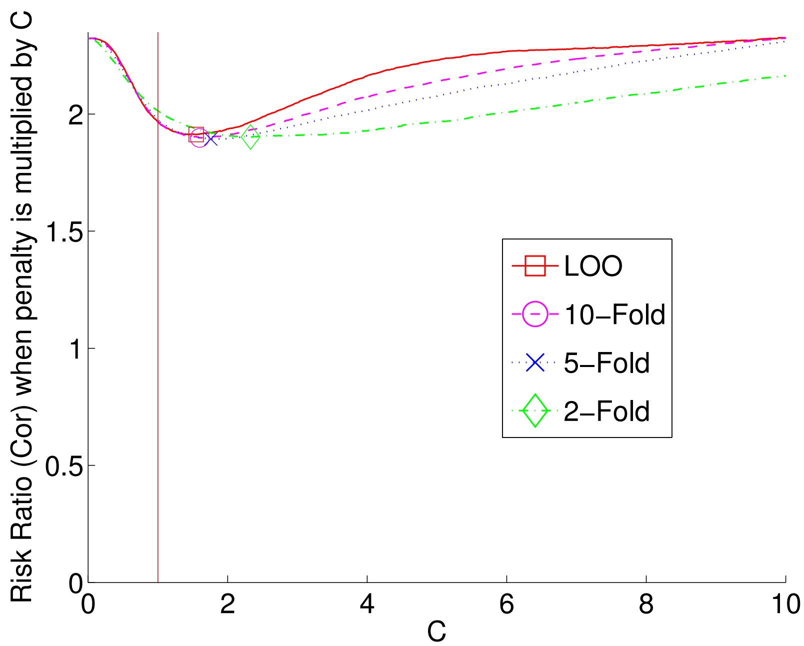
Penalty	$C_n^\star$	$C_{\mathrm{or}}(C = C_n^\star)$	$C_{\mathrm{or}}(C = 1)$	$C_{\mathrm{or}}(C = 2.30)$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	1.66 ± 0.21	2.44 ± 0.01	3.29 ± 0.03	$\mathbf{2.77 \pm 0.01}$
$\mathrm{pen}_{\mathrm{LOO}}$	1.60 ± 0.18	2.47 ± 0.01	3.29 ± 0.03	$\mathbf{2.81 \pm 0.01}$
pen 10F	1.71 ± 0.22	2.49 ± 0.02	3.54 ± 0.04	$\mathbf{2.71 \pm 0.02}$
pen 5F	1.79 ± 0.31	2.53 ± 0.02	3.92 ± 0.04	$\mathbf{2.67 \pm 0.02}$
pen 2F	2.86 ± 0.58	2.70 ± 0.02	6.14 ± 0.08	$\mathbf{2.74 \pm 0.02}$

Figure 10: Overpenalization in setting L-Regu, $n = 100$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.



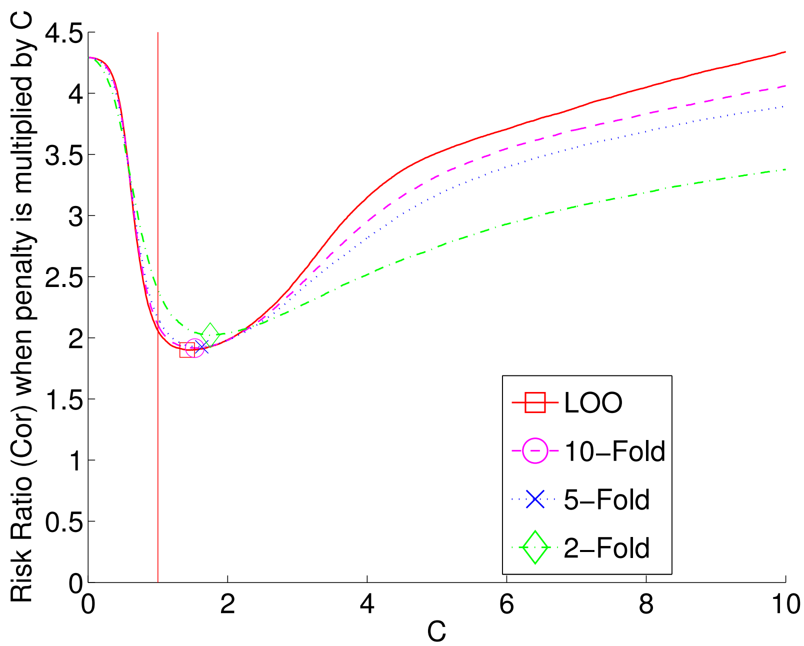
Penalty	$C_n^\star$	$C_{\text{or}}(C = C_n^\star)$	$C_{\text{or}}(C = 1)$	$C_{\text{or}}(C = 3.11)$
$\mathbb{E}[\text{pen}_{\text{id}}]$	1.63 ± 0.25	1.93 ± 0.01	2.33 ± 0.01	$\mathbf{2.20 \pm 0.01}$
pen_{LOO}	1.61 ± 0.23	1.93 ± 0.01	2.33 ± 0.01	$\mathbf{2.21 \pm 0.01}$
pen 10F	1.79 ± 0.26	1.92 ± 0.01	2.42 ± 0.02	$\mathbf{2.16 \pm 0.01}$
pen 5F	1.88 ± 0.29	1.92 ± 0.01	2.55 ± 0.02	$\mathbf{2.11 \pm 0.01}$
pen 2F	2.15 ± 0.34	1.97 ± 0.01	3.37 ± 0.03	$\mathbf{2.07 \pm 0.01}$

Figure 11: Overpenalization in setting L-Regu, $n = 500$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.



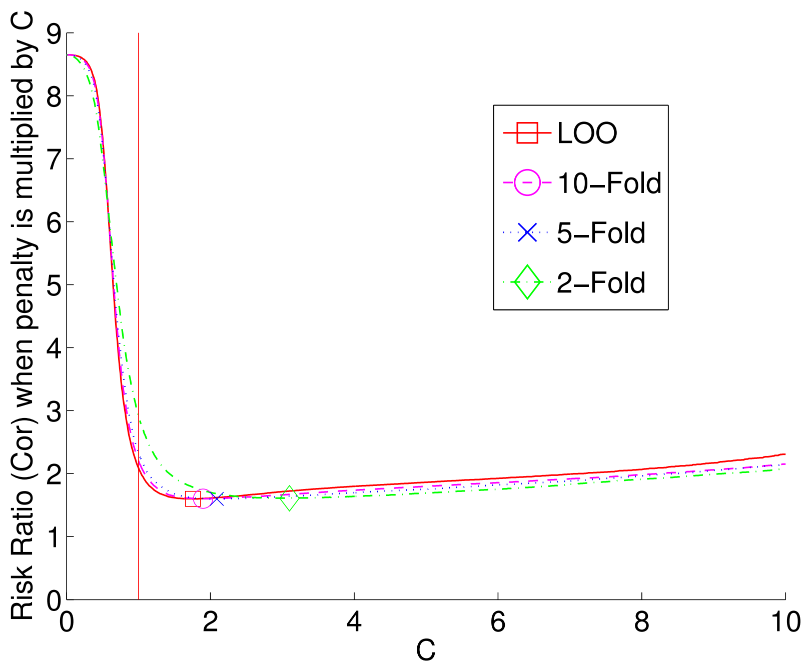
Penalty	$C_n^\star$	$C_{\mathrm{or}}(C = C_n^\star)$	$C_{\mathrm{or}}(C = 1)$	$C_{\mathrm{or}}(C = 2.30)$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	1.54 ± 0.28	1.91 ± 0.01	$\mathbf{1.97 \pm 0.01}$	$\mathbf{1.97 \pm 0.00}$
$\mathrm{pen}_{\mathrm{LOO}}$	1.55 ± 0.32	1.91 ± 0.01	$\mathbf{1.97 \pm 0.01}$	$\mathbf{1.96 \pm 0.00}$
pen 10F	1.60 ± 0.31	1.90 ± 0.01	1.97 ± 0.01	$\mathbf{1.93 \pm 0.01}$
pen 5F	1.76 ± 0.39	1.89 ± 0.01	1.98 ± 0.01	$\mathbf{1.91 \pm 0.01}$
pen 2F	2.33 ± 1.09	1.90 ± 0.01	2.01 ± 0.01	$\mathbf{1.90 \pm 0.01}$

Figure 12: Overpenalization in setting S-Dya2, $n = 100$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.



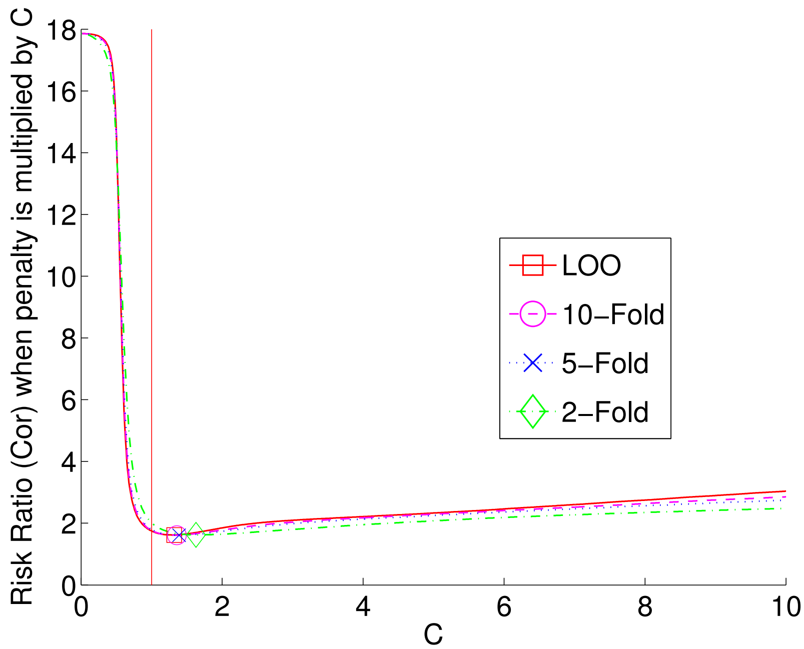
Penalty	$C_n^\star$	$C_{\mathrm{or}}(C = C_n^\star)$	$C_{\mathrm{or}}(C = 1)$	$C_{\mathrm{or}}(C = 3.11)$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	1.44 ± 0.15	1.92 ± 0.01	$\mathbf{2.07 \pm 0.01}$	2.58 ± 0.01
$\mathrm{pen}_{\mathrm{LOO}}$	1.42 ± 0.19	1.90 ± 0.01	$\mathbf{2.06 \pm 0.01}$	2.56 ± 0.01
pen 10F	1.53 ± 0.19	1.92 ± 0.01	$\mathbf{2.11 \pm 0.01}$	2.47 ± 0.01
pen 5F	1.63 ± 0.21	1.93 ± 0.01	$\mathbf{2.16 \pm 0.01}$	2.42 ± 0.01
pen 2F	1.75 ± 0.23	2.02 ± 0.01	2.39 ± 0.01	$\mathbf{2.28 \pm 0.01}$

Figure 13: Overpenalization in setting S-Dya2, $n = 500$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.



Penalty	$C_n^\star$	$C_{\mathrm{or}}(C = C_n^\star)$	$C_{\mathrm{or}}(C = 1)$	$C_{\mathrm{or}}(C = 2.30)$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	1.77 ± 0.21	1.59 ± 0.00	2.09 ± 0.01	$\mathbf{1.63 \pm 0.00}$
$\mathrm{pen}_{\mathrm{LOO}}$	1.76 ± 0.27	1.60 ± 0.00	2.09 ± 0.01	$\mathbf{1.64 \pm 0.00}$
pen 10F	1.90 ± 0.26	1.60 ± 0.00	2.21 ± 0.01	$\mathbf{1.62 \pm 0.00}$
pen 5F	2.09 ± 0.34	1.60 ± 0.00	2.34 ± 0.02	$\mathbf{1.60 \pm 0.00}$
pen 2F	3.10 ± 0.50	1.61 ± 0.01	2.92 ± 0.02	$\mathbf{1.65 \pm 0.01}$

Figure 14: Overpenalization in setting S-Regu, $n = 100$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.



Penalty	$C_n^\star$	$C_{\mathrm{or}}(C = C_n^\star)$	$C_{\mathrm{or}}(C = 1)$	$C_{\mathrm{or}}(C = 3.11)$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	1.33 ± 0.11	1.62 ± 0.00	$\mathbf{1.75 \pm 0.01}$	2.11 ± 0.00
$\mathrm{pen}_{\mathrm{LOO}}$	1.32 ± 0.10	1.62 ± 0.00	$\mathbf{1.75 \pm 0.01}$	2.11 ± 0.00
pen 10F	1.36 ± 0.09	1.61 ± 0.00	$\mathbf{1.77 \pm 0.01}$	2.05 ± 0.00
pen 5F	1.39 ± 0.08	1.61 ± 0.00	$\mathbf{1.80 \pm 0.01}$	2.01 ± 0.00
pen 2F	1.63 ± 0.23	1.62 ± 0.00	2.01 ± 0.01	$\mathbf{1.81 \pm 0.01}$

Figure 15: Overpenalization in setting S-Regu, $n = 500$.
Top: same as Figure 3 (estimated loss ratio as a function of the overpenalization constant C).
Bottom: Table showing the estimated optimal overpenalization constant $C_n^\star$ as well as the estimated loss ratio for several values of C: $C = C_n^\star$ (optimal value), $C = 1$ (AIC-type penalty) and $C = \log(n)/2$ (BIC-type penalty). See text for details.

The study of variance of Section 6.4 (setting S with $n = 100$) is completed with Figure 16, which tests the validity of the heuristic of Section 4, Figure 17, which is the equivalent of Figure 5 without zooming on the smallest dimensions, and Figure 18, which shows that

$$\forall m \neq m^\star, \quad \mathrm{SNR}(m) \approx \frac{\mathbb{E}\bigl[\Delta(m, m^\star)\bigr]}{\sqrt{\mathrm{Var}\bigl(\Delta(m, m^\star)\bigr)}} \enspace .$$

The next figures present the same results as the ones of Section 6.4 about the variance, for other experimental settings.

Figures 18–23 show the results for setting L with $n = 100$, based upon $N = 10\,000$ independent samples.

Figures 24–34 show the results for settings S and L with $n = 500$, based upon $N = 1\,000$ independent samples.

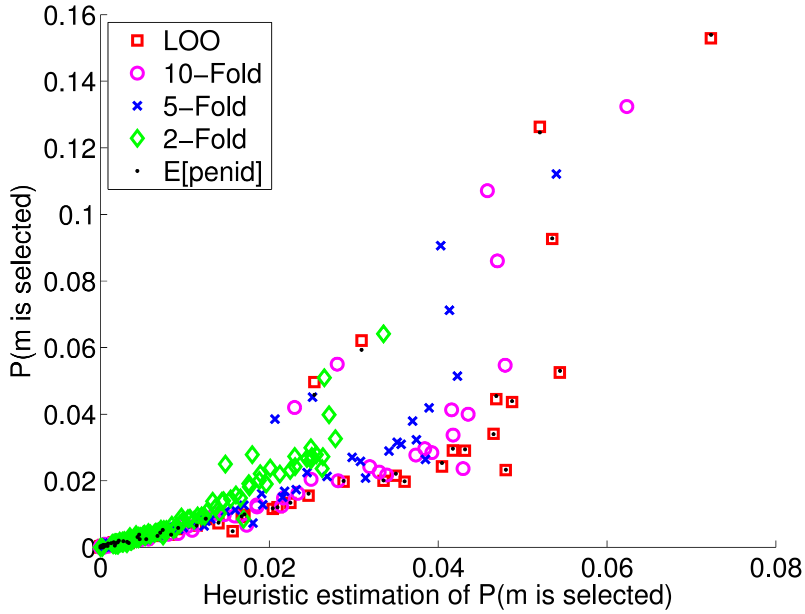


Figure 16: Illustration of the variance heuristic: $\mathbb{P}(\widehat{m} = m)$ as a function of $\overline{\Phi}(\mathrm{SNR}(m))$ (renormalized to have a sum equal to one). Setting S-Regu, $n = 100$.

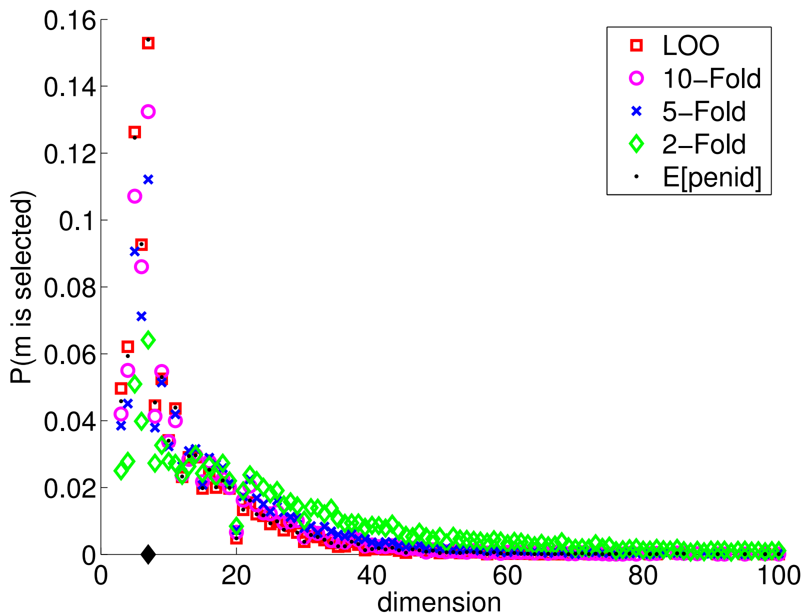


Figure 17: Setting S-Regu, $n = 100$. $\mathbb{P}(\widehat{m} = m)$ as a function of m. The black diamond shows $m^{\star} = 7$.

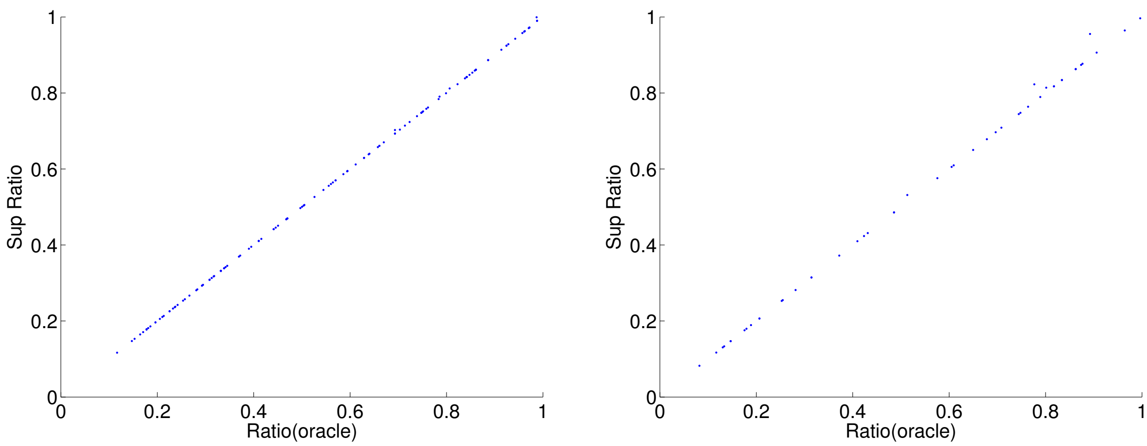


Figure 18: SNR(m) as a function of the ratio at $m' = m^\star$. $n = 100$. Left: S-Regu. Right: L-Regu.

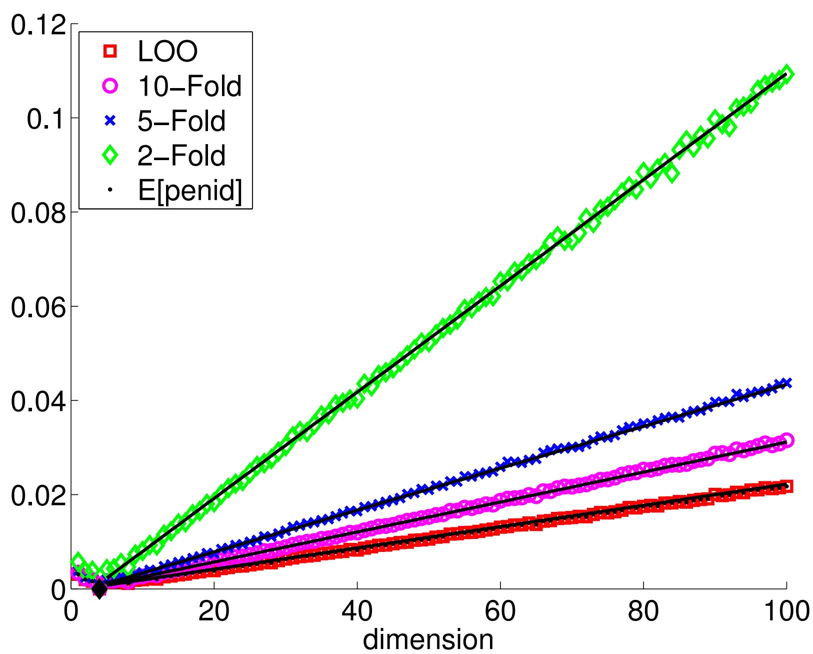


Figure 19: L-Regu, $n = 100$. $\mathrm{Var}(\Delta_{\mathcal{C}}(m, m^\star))$ as a function of m. The black lines show the linear approximation $n^{-2}[5.6(1+\frac{1.1}{V-1})+2.2(1+\frac{4.2}{V-1})(m-m^\star)]$ for $m > m^\star = 4$.

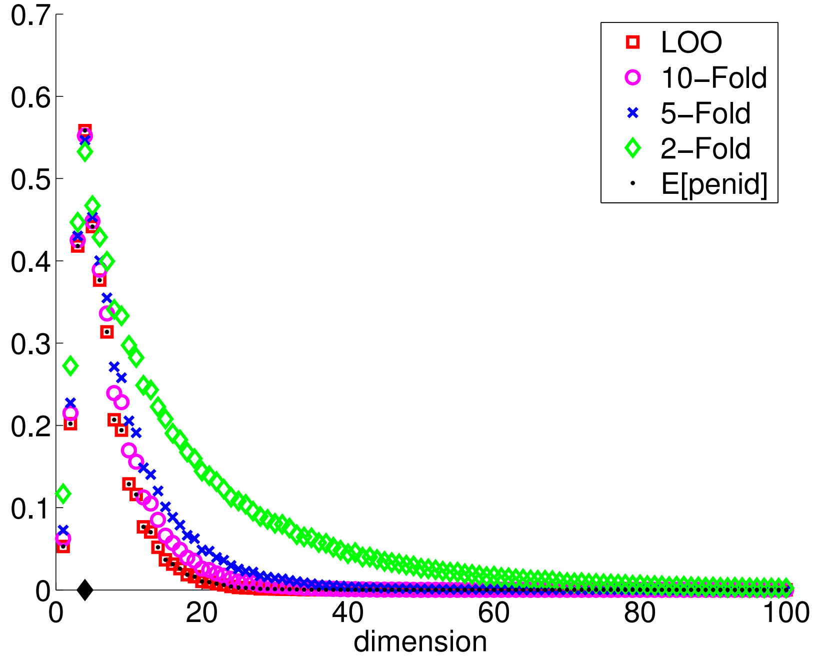


Figure 20: L-Regu, $n = 100$. $\overline{\Phi}(\mathrm{SNR}_{\mathcal{C}}(m))$ as a function of m.

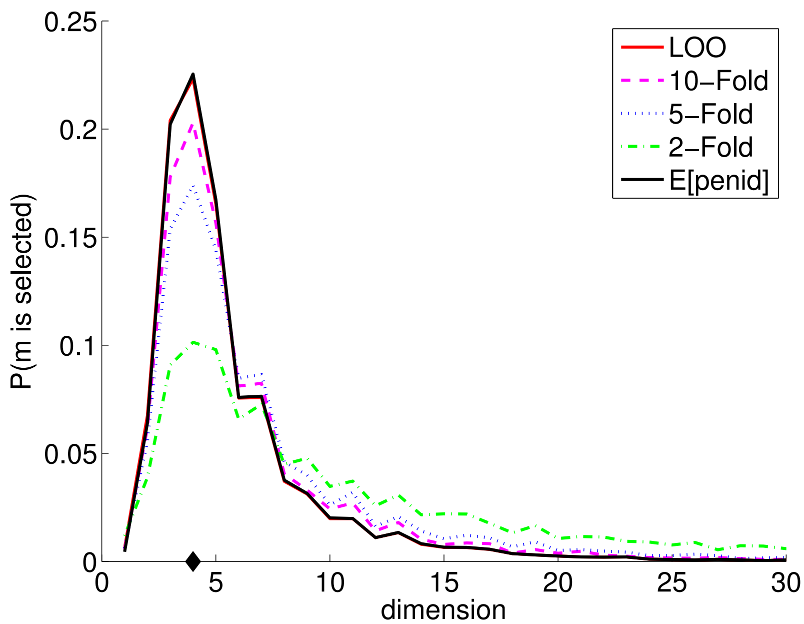


Figure 21: L-Regu, $n = 100$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ as a function of m.

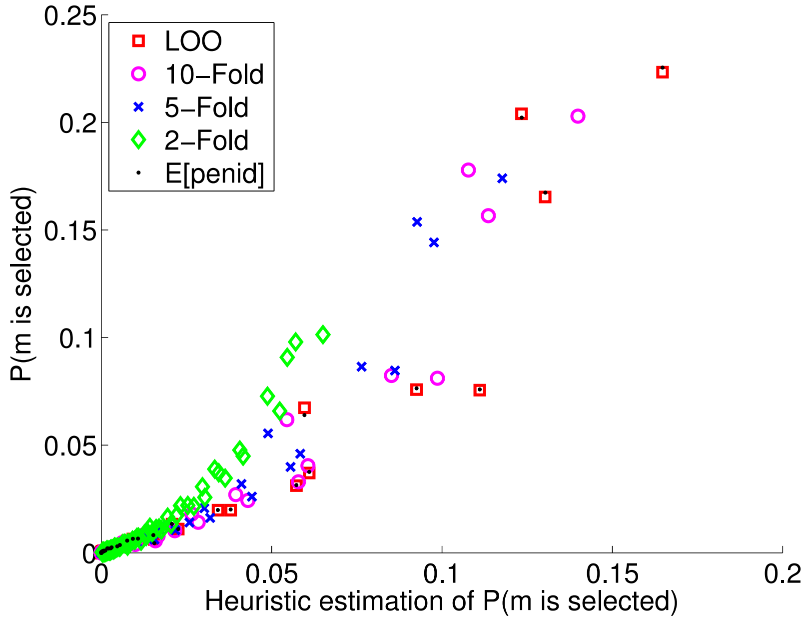


Figure 22: L-Regu, $n = 100$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ as a function of $\overline{\Phi}(\mathrm{SNR}_{\mathcal{C}}(m))$.

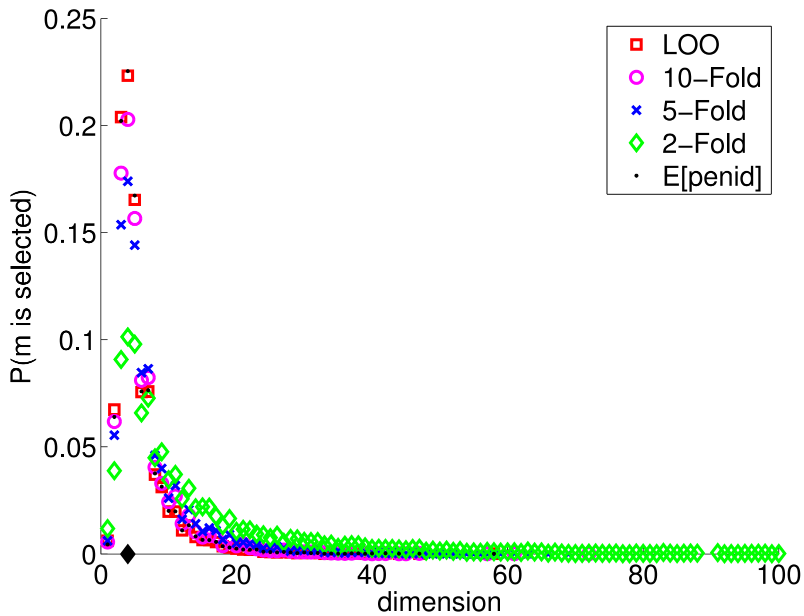


Figure 23: L-Regu, $n = 100$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ as a function of m.

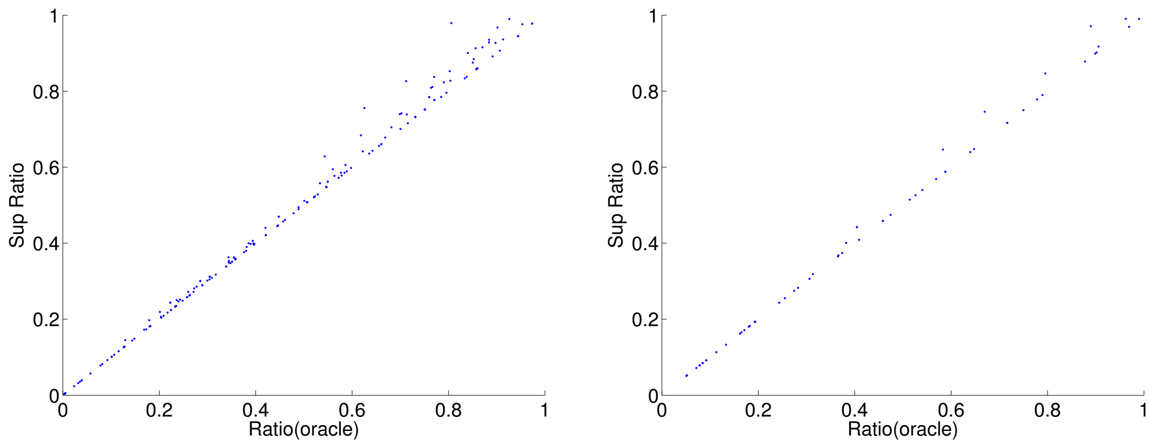


Figure 24: SNR(m) as a function of the ratio at $m' = m^\star$. $n = 500$. Left: S-Regu. Right: L-Regu.

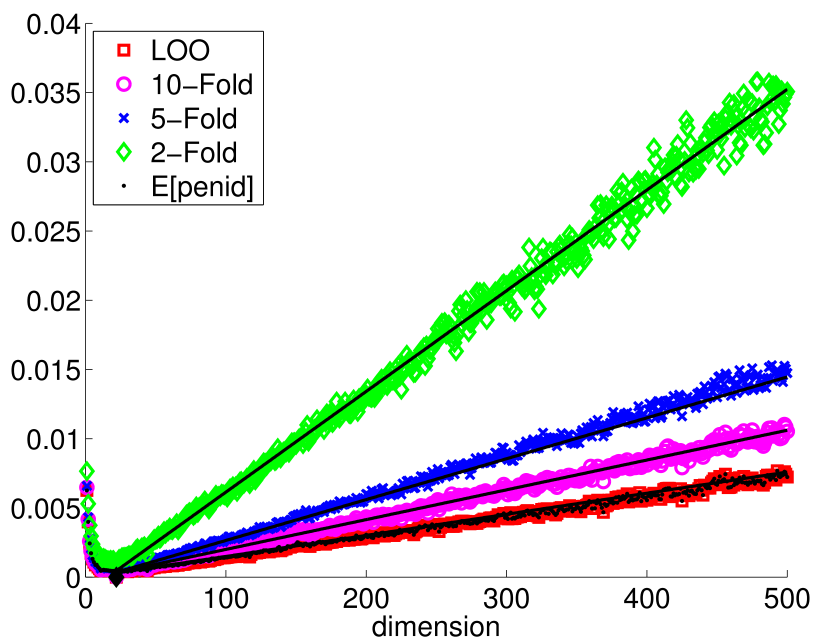


Figure 25: S-Regu, $n = 500$. $\mathrm{Var}(\Delta_{\mathcal{C}_V}(m, m^\star))$ as a function of m. The black lines show the linear approximation $n^{-2}[75(1+\frac{0.52}{V-1})+3.8(1+\frac{3.8}{V-1})(m-m^\star)]$ for $m > m^\star = 22$.

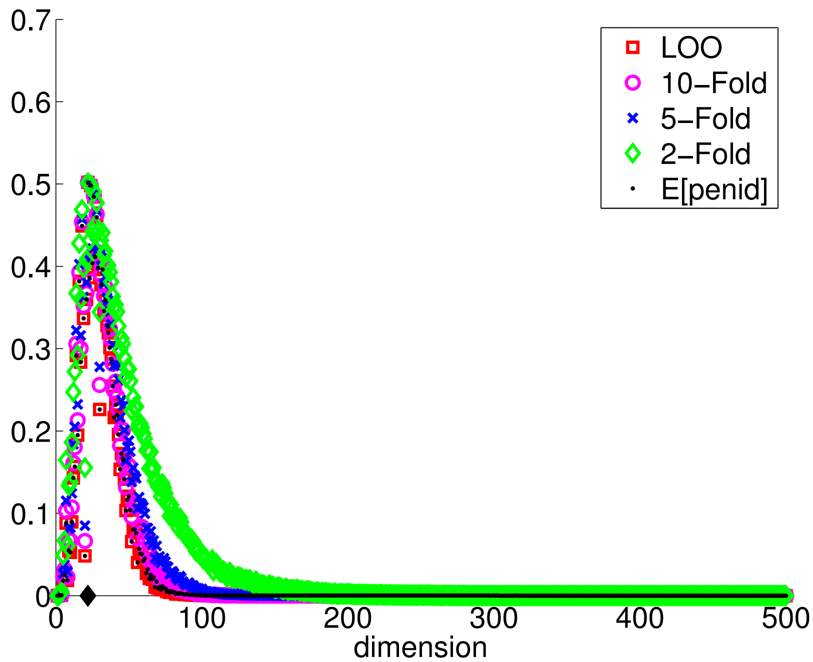


Figure 26: S-Regu, $n = 500$. $\overline{\Phi}(\mathrm{SNR}_{\mathcal{C}_V}(m))$ as a function of m.

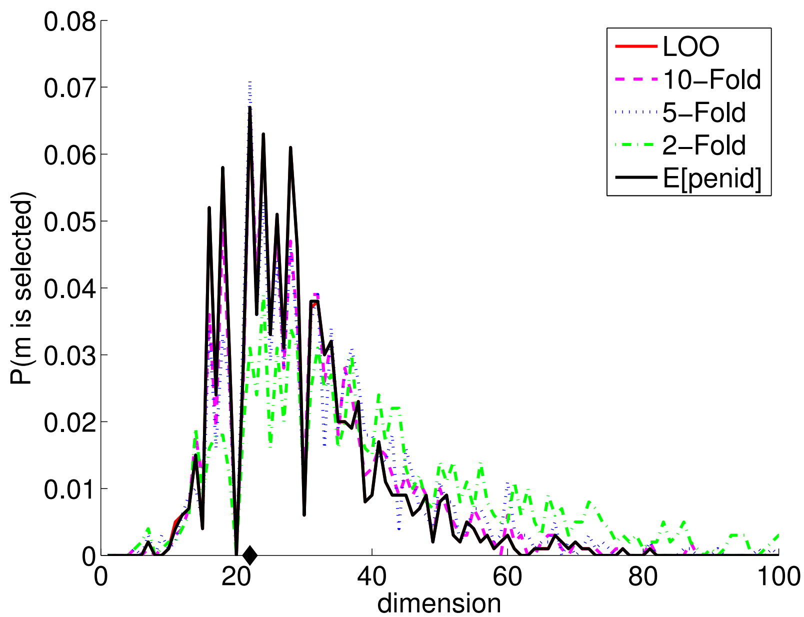


Figure 27: S-Regu, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ as a function of m.

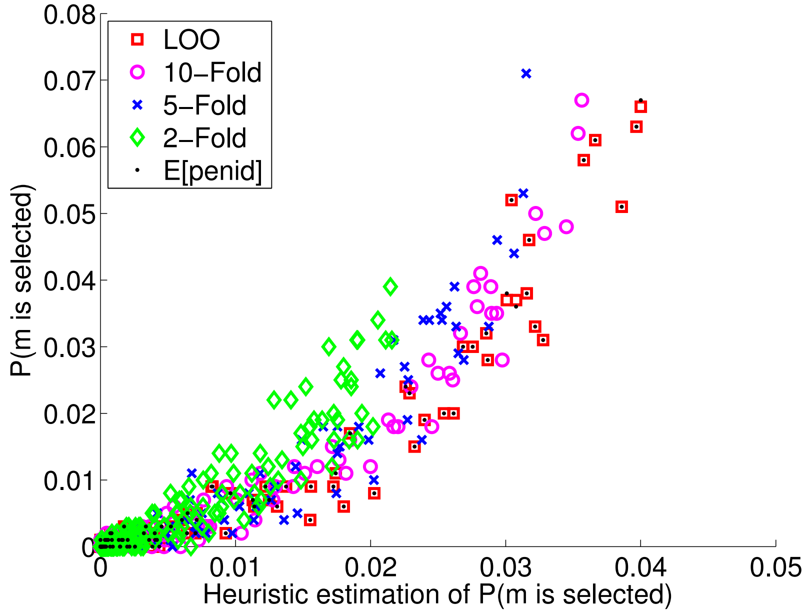


Figure 28: S-Regu, $n = 500$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ as a function of $\overline{\Phi}(\mathrm{SNR}(m))$.

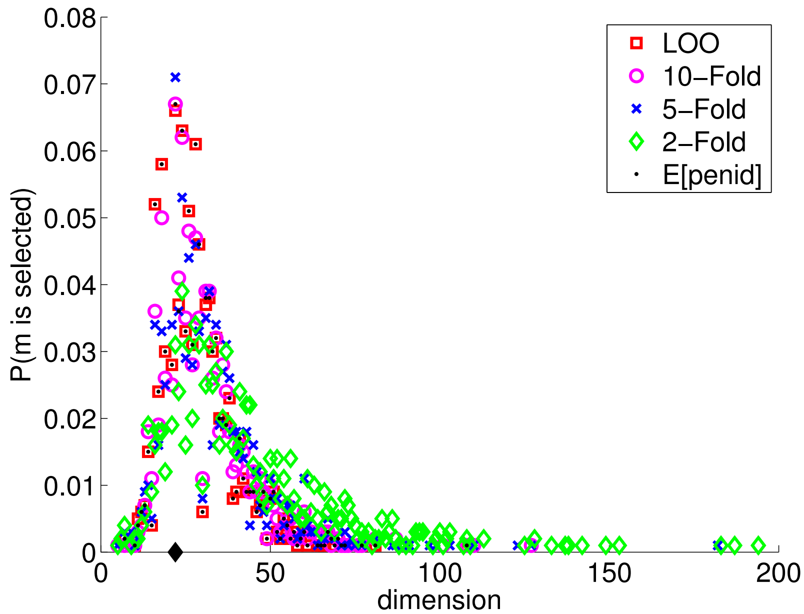


Figure 29: S-Regu, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ as a function of m.

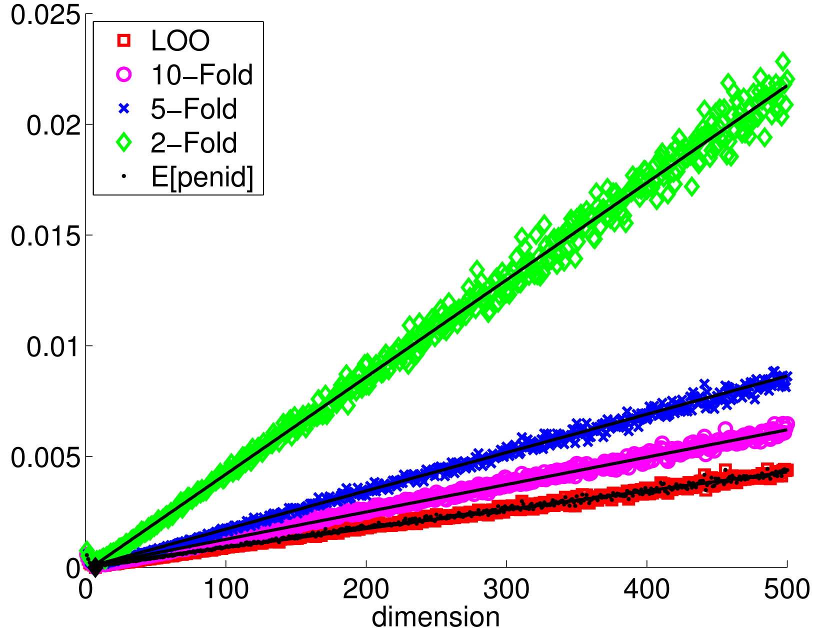


Figure 30: L-Regu, $n = 500$. $\mathrm{Var}(\Delta_{\mathcal{C}_V}(m, m^\star))$ as a function of m. The black lines show the linear approximation $n^{-2}[28(1 + \frac{0.06}{V-1}) + 2.1(1 + \frac{4.2}{V-1})(m - m^\star)]$ for $m > m^\star = 7$.

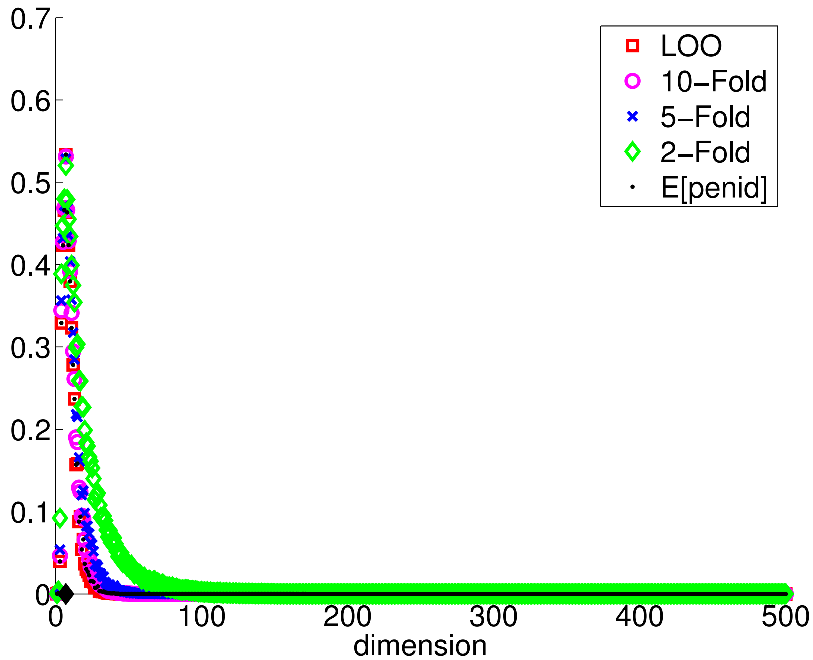


Figure 31: L-Regu, $n = 500$. $\overline{\Phi}(\mathrm{SNR}_{\mathcal{C}_V}(m))$ as a function of m.

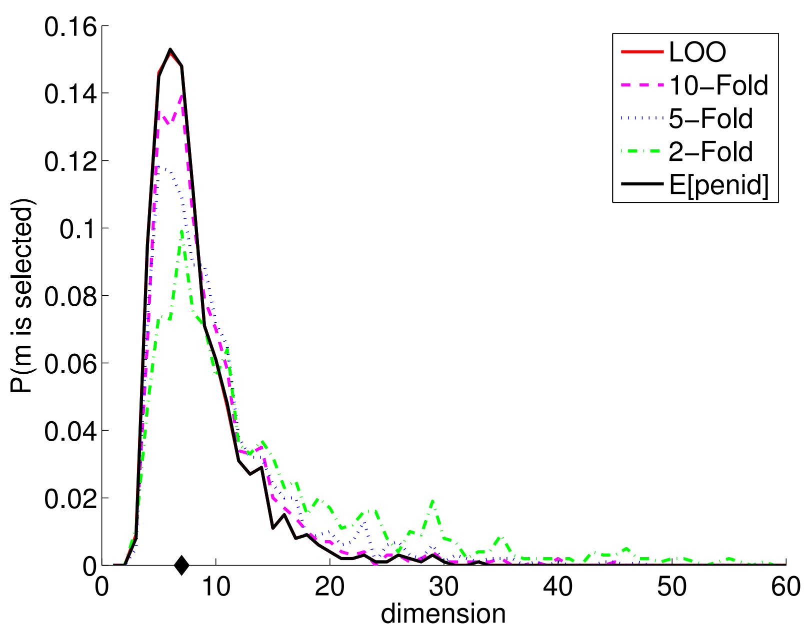


Figure 32: L-Regu, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ as a function of m.

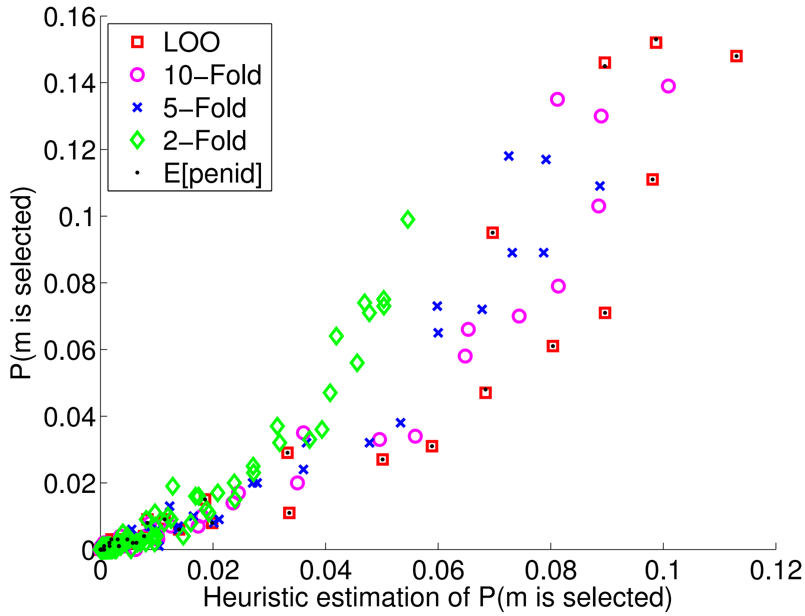


Figure 33: L-Regu, $n = 500$. $\mathbb{P}(\widehat{m}(\mathcal{C}) = m)$ as a function of $\overline{\Phi}(\mathrm{SNR}(m))$.

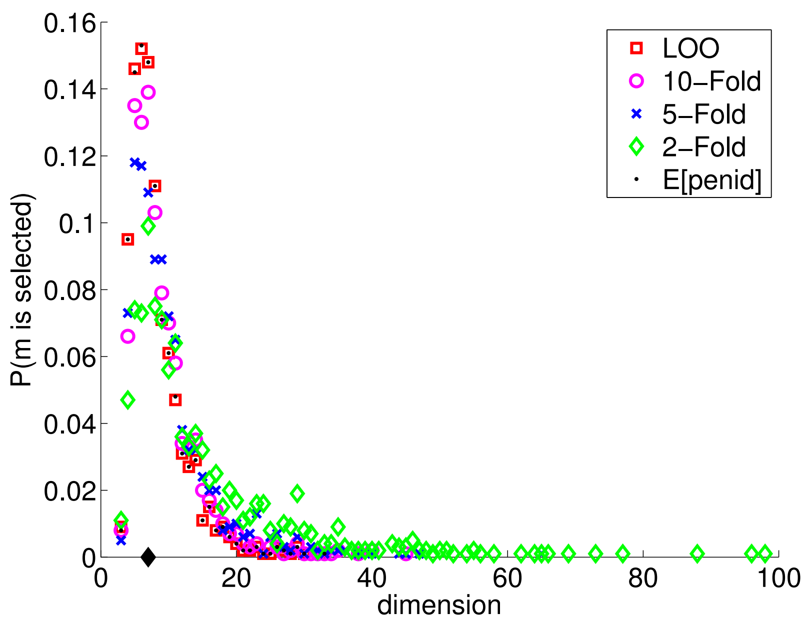


Figure 34: L-Regu, $n = 500$. $\mathbb{P}(\widehat{m} = m)$ as a function of m.

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

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

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