Project: Cross-validation for model selection (notes on papers)

Diego Urdapilleta de la Parra

March 30, 2025

Contents

1 Introduction **2**

2 CV for Linear Model Selection **2**

2.1 Setup and preliminary results 2

2.2 A Result on the Leave-one-out: Shao, 1993 5

2.3 A General Perspective: Shao, 1997 8

2.3.1 The case of $\lambda_n \equiv 2$ 8

2.3.2 The case of $\lambda_n \to \infty$ 9

2.3.3 Cross-validation 10

3 CV for Nonparametric Model Selection **10**

3.1 Comparison of Distinct Procedures: Yang, 2007 10

3.1.1 Single-split cross-validation (the Hold-out) 10

3.1.2 Voting cross-validation with multiple splits 11

4 Aggregation **12**

4.1 Bunea et al., 2007 12

4.1.1 Four types of aggregation 12

4.1.2 Evaluating the aggregate 13

4.1.3 A BIC-type penalty 14

1 Introduction

The goal of this project is to study the asymptotic properties of cross-validation (CV) methods for model selection in a variety of scenarios. [(...) motivation, brief description of CV methods, outline of the project.]

Throughout the paper, we consider the usual regression setup: Let n, p_n be positive integers and $\mathcal{D}_n := \{(y_i, \boldsymbol{x}_i) : i \in [n]\}$ be a set of independent data points drawn from a distribution $\mathbb{P}_{y,\boldsymbol{x}}$ for $(y, \boldsymbol{x}) \in \mathbb{R} \times \mathcal{X}$. We treat the $\boldsymbol{x}_i$'s as predictors of the outcome y_i, and we assume a model

$$y_i = f(\boldsymbol{x}_i) + e_i, \qquad i \in [n], \tag{1}$$

where f is an unknown Borel-measurable function $f : \mathcal{X} \to \mathbb{R}$ with $f(x_i) \overset{\text{a.s.}}{=} \mathbb{E}[y_i \mid \boldsymbol{x}_i]$ and the e_i's are zero-mean random variables.

2 CV for Linear Model Selection

2.1 Setup and preliminary results

In this section, the regression function f in (1) is assumed to be linear, so that the data is generated from a linear model of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times p_n}$ is the design matrix, $\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_n]^\top$, and $\boldsymbol{e}$ is a mean-zero random vector with $\mathrm{Cov}(\boldsymbol{e}) = \sigma_2 \boldsymbol{I}_n$.

In the context of linear models, the model selection procesure reduces to selecting a subset of covariates from a set of candidate covariates of size p_n. This is also known as *variable selection*. We remark that the number p_n may depend on n, and some assumptions on the growth of p_n will be established later.

We let $\mathcal{A}_n \subset 2^{[p_n]}$ be a family of index sets representing candidate models. For $\alpha \in \mathcal{A}_n$, we denote by $p_n(\alpha)$ the cardinality of α and consider the model given by

$$f_\alpha(\boldsymbol{X}) = \boldsymbol{X}_\alpha \boldsymbol{\beta}_\alpha,$$

where $\boldsymbol{X}_\alpha$ is the sub-matrix of $\boldsymbol{X}$ containing only the columns indexed by α, and $\boldsymbol{\beta}_\alpha$ is the coefficient vector containing only the entries indexed by α in $\boldsymbol{\beta}$.

1. We say $\alpha \in \mathcal{A}_n$ is *correct* if $\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}] \overset{\text{a.s.}}{=} f_\alpha(\boldsymbol{X})$, and we denote by $\mathcal{T}_n$ the set of correct models in $\mathcal{A}_n$

2. We say $\alpha \in \mathcal{A}_n$ is *wrong* if it is not correct, and we denote by $\mathcal{T}_n^c$ the set of wrong models in $\mathcal{A}_n$

3. We say $\mathcal{A}_n$ is *embedded* if there exists an enumeration $\alpha_1, \alpha_2, \ldots, \alpha_k$ of all elements in $\mathcal{A}_n$ such that

$$\alpha_1 \subset \alpha_2 \subset \cdots \subset \alpha_k.$$

Definition 2.1. *For $\alpha \in \mathcal{A}_n$, let $\hat{\boldsymbol{\beta}}_\alpha$ be the OLS estimator of $\boldsymbol{\beta}_\alpha$ and $\hat{f}_\alpha(\boldsymbol{X}) := \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha$. We denote the average squared error of $\hat{f}_\alpha$ by*

$$L_n(\alpha) := \frac{1}{n}\|f(\boldsymbol{X}) - \hat{f}_\alpha(\boldsymbol{X})\|^2.$$

Additionally, we write

$$R_n(\alpha) := \mathbb{E}\left[L_n(\alpha) \mid \boldsymbol{X}\right].$$

The following conditions will be used throughout out treatment of linear models:

$$\begin{aligned}
&\textbf{H1}: && \liminf_{n\to\infty} \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 > 0 \text{ for all } \alpha \in \mathcal{A}_n. \\
&\textbf{H2}: && \boldsymbol{X}^\top \boldsymbol{X} = O(n) \quad \text{and} \quad \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = O\left(n^{-1}\right). \\
&\textbf{H3}: && \lim_{n\to\infty} \max_{i\le n} h_{ii,\alpha} = 0 \text{ for all } \alpha \in \mathcal{A}_n. \\
&\textbf{H4}: && \sum_{\alpha \in \mathcal{T}_n^c} \frac{1}{(nR_n(\alpha))^m} \to_{\mathbb{P}} 0 \quad \text{for some } m \ge 1.
\end{aligned}$$

Proposition 2.1. *Assumimg a linear model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}$,*

$$L_n(\alpha) = \frac{1}{n}\|H_\alpha \boldsymbol{e}\|^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \textit{and} \quad R_n(\alpha) = \frac{1}{n}\sigma^2 p_n(\alpha) + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2$$

almost surely, where $H_\alpha = \boldsymbol{X}_\alpha\left(\boldsymbol{X}_\alpha^\top \boldsymbol{X}_\alpha\right)^{-1}\boldsymbol{X}_\alpha^\top$ and $M_\alpha = I_n - H_\alpha$.

Proof. First, we have that

$$\begin{aligned}
\|f(\boldsymbol{X}) - \hat{f}_\alpha(\boldsymbol{X})\|^2 &= \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha\|^2 \\
&= \|\boldsymbol{X}\boldsymbol{\beta} - H_\alpha(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e})\|^2 \\
&= \|M_\alpha \boldsymbol{X}\boldsymbol{\beta} - H_\alpha \boldsymbol{e}\|^2.
\end{aligned}$$

Notice that $M_\alpha \boldsymbol{X}\boldsymbol{\beta}$ and $H_\alpha \boldsymbol{e}$ are orthogonal:

$$\boldsymbol{e}^\top H_\alpha M_\alpha \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{e}^\top H_\alpha (I_n - H_\alpha)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{e}^\top H_\alpha \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{e}^\top H_\alpha \boldsymbol{X}\boldsymbol{\beta} = 0.$$

Hence, the first part follows from the Pythagorean theorem.

For the second part, we note that $\mathbb{E}\left[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}\right] \overset{\text{a.s.}}{=} \sigma^2 p_n(\alpha)$ by the "trace trick", where $p_n(\alpha)$ denotes the size of model α. □

Proposition 2.2. *Suppose that that $\mathcal{T}_n$ is non-empty, and let α_n^* be the smallest correct model in $\mathcal{T}_n$. Then, α_n^* minimizes $R_n(\alpha)$ (with probability 1?) over $\alpha \in \mathcal{A}_n$.*

Proof. Let $\alpha \in \mathcal{A}_n$ be arbitrary and suppose that $\alpha \in \mathcal{T}_n$. Then, $\boldsymbol{X}_\alpha \boldsymbol{\beta}_\alpha = \boldsymbol{X}\boldsymbol{\beta}$ and $p_n(\alpha_n^*) \le p_n(\alpha)$. Thus,

$$\begin{aligned}
R_n(\alpha) &= \frac{1}{n}\sigma^2 p_n(\alpha) + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 \\
&= \frac{1}{n}\sigma^2 p_n(\alpha) + \frac{1}{n}\underbrace{\|M_\alpha \boldsymbol{X}_\alpha \boldsymbol{\beta}_\alpha\|^2}_{0} \\
&= \frac{1}{n}\sigma^2 p_n(\alpha) \ \ge \ \frac{1}{n}\sigma^2 p_n(\alpha_n^*) = R_n(\alpha_n^*).
\end{aligned}$$

Now suppose that $\alpha \in \mathcal{T}_n^c$. If $p_n(\alpha) \geq p_n(\alpha_n^*)$, the result follows immediately by assumption H1. On the other hand, if $p_n(\alpha) \leq p_n(\alpha_n^*)$, we must verify that

$$\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 \geq \sigma^2 \left(p_n(\alpha_n^*) - p_n(\alpha)\right). \tag{2}$$

To this end, we note that $\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|M_\alpha \boldsymbol{X}_{\alpha_n^*}\boldsymbol{\beta}_{\alpha_n^*}\|^2$ and that

$$\boldsymbol{X}_{\alpha_n^*}\boldsymbol{\beta}_{\alpha_n^*} = \boldsymbol{X}_\alpha \boldsymbol{\beta}_\alpha + \boldsymbol{X}_{\alpha_n^* \setminus \alpha}\boldsymbol{\beta}_{\alpha_n^* \setminus \alpha}.$$

Thus, if we let λ denote the smallest eigenvalue of $\boldsymbol{X}_{\alpha_n^*}^\top M_\alpha \boldsymbol{X}_{\alpha_n^*}$, we have that

$$\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|M_\alpha \boldsymbol{X}_{\alpha_n^* \setminus \alpha}\boldsymbol{\beta}_{\alpha_n^* \setminus \alpha}\|^2 \geq \lambda \|\boldsymbol{\beta}_{\alpha * \setminus \alpha}\|^2$$

(for a proof of the latter inequality, see [2]). This is as far as I got. I don't know how to show that $\lambda \|\boldsymbol{\beta}_{\alpha_n^* \setminus \alpha}\|^2 \geq \sigma^2 \left(p_n(\alpha_n^*) - p_n(\alpha)\right)$, but it seems reasonable if the coefficients in $\boldsymbol{\beta}$ are not too small. ∎

From Proposition 2.2, we see that R_n is a good choice of selection criterion. Unfortunately, R_n depends on the unkown regression function, and therefore cannot be used in practice. Instead, we may try to approximate it through other empirically feasible criteria.

Definition 2.2. *Let $\hat{\alpha}_n$ be the model selected by minimizing some criterion $\hat{R}_n$ over $\mathcal{A}_n$, and let α_n^* denote the model minimizing R_n over $\mathcal{A}_n$. We say $\hat{R}_n$ is consistent if*

$$\mathbb{P}\left\{\hat{\alpha}_n = \alpha_n^*\right\} \to 1$$

as $n \to \infty$. We say that $\hat{R}_n$ is assomptotically loss efficient if

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_n^*)} \xrightarrow{\mathbb{P}} 1.$$

Lemma 2.3. *If $\hat{R}_n$ is consistent, then it is asymptotically loss efficient.*

Proof. Suppose that $\hat{R}_n$ is consistent. Clearly, if $\hat{\alpha}_n = \alpha_n^*$, then $L_n(\hat{\alpha}) = L_n(\alpha_n^*)$. Therefore,

$$\mathbb{P}\left\{\hat{\alpha}_n = \alpha_n^*\right\} \leq \mathbb{P}\left\{L_n(\hat{\alpha}) = L_n(\alpha_n^*)\right\}.$$

By consistency, the left-hand side converges to 1, so that the right-hand side must also converge to 1. ∎

Proposition 2.4 (Shao, 1997 [3])**.** *Suppose H1, $p_n/n \to 0$, and that $\mathcal{T}_n$ is non-empty for all but finitely many n.*

1. *If $|\mathcal{T}_n| = 1$ for all but finitely many n, then consistency is equivalent to efficiency in the sense of Definition 2.2.*

2. *If $p_n(\alpha_n^*) \overset{\mathbb{P}}{\not\to} \infty$, then consistency is equivalent to efficiency in the sense of Definition 2.2.*

Proof. From Lemma 2.3, it remains to show that, under the given conditions, assymptotic loss efficiency implies consistency. We show the contrapositive:

1. Suppose that $\hat{R}_n$ is not consistent. By Proposition 2.2, α_n^* must be the correct model in $\mathcal{T}_n$ minimizing R_n. Therefore, $L_n(\alpha_n^*) = (1/n)\|H_\alpha \boldsymbol{e}\|^2$, and

$$\mathbb{E}\left[L_n(\alpha_n^*)\right] = \frac{1}{n}\sigma^2 p_n(\alpha_n^*) \leq \frac{1}{n}\sigma^2 p_n \to 0 \quad \text{as } n \to \infty$$

by assumption. We have shown that $L_n(\alpha_n^*) \xrightarrow{L_1} 0$, which implies $L_n(\alpha_n^*) \xrightarrow{\mathbb{P}} 0$.

On the other hand, since $\hat{R}_n$ is not consistent, there must exist $\tilde{\alpha}_n \neq \alpha^*$ for infinitely many n such that $\mathbb{P}\{\hat{\alpha}_n = \tilde{\alpha}_n\} \neq 0$. Notice that, since $\mathcal{T}_n = \{\alpha^*\}$, it must be the case that $\tilde{\alpha}_n \in \mathcal{T}_n^c$. We have the following:

$$L_n(\hat{\alpha}_n) \geq \mathbb{1}_{[\hat{\alpha}_n = \tilde{\alpha}_n]} L_n(\tilde{\alpha}_n) = \mathbb{1}_{[\hat{\alpha}_n = \tilde{\alpha}_n]} \left(\frac{1}{n}\|H_{\tilde{\alpha}_n}\boldsymbol{e} + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2\|^2\right).$$

By assumption **H1**, the latter expression cannot not converge to 0. We conclude that the ratio $L_n(\hat{\alpha}_n)/L_n(\alpha^*) \stackrel{\mathbb{P}}{\not\to} 1$.

2. Suppose again that $\hat{R}_n$ is not consistent. Since $\mathcal{T}_n$ contains at least two models, there must exist $\tilde{\alpha}_n \in \mathcal{T}_n$ such that $\tilde{\alpha}_n \neq \alpha^*$ and $\mathbb{P}\{\hat{\alpha}_n = \tilde{\alpha}_n\} \not\to 0$. Hence,

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_n^*)} - 1 \geq \left(\frac{L_n(\tilde{\alpha}_n)}{L_n(\alpha_n^*)} - 1\right)\mathbb{1}_{[\hat{\alpha}_n = \tilde{\alpha}_n]} = \left(\frac{\|H_{\tilde{\alpha}_n}\|^2}{\|H_{\alpha_n^*}\|^2} - 1\right)\mathbb{1}_{[\hat{\alpha}_n = \tilde{\alpha}_n]} \stackrel{\mathbb{P}}{\not\to} 0.$$

□

2.2 A Result on the Leave-one-out: Shao, 1993

[Brief introduction to LOOCV and motivation]

For this section, we will consider the case where the set $\mathcal{A}_n =: \mathcal{A}$ and all its elemets are constant across all $n \geq 1$. That is, the candidate models are not changed by the number of observations.

Definition 2.3. *The Leave-one-out estimator of $R_n(\alpha)$ is defined as*

$$\hat{R}_n^{(1)}(\alpha) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{x}_{i\alpha}^\top \hat{\boldsymbol{\beta}}_\alpha}{1 - h_{ii,\alpha}}\right)^2$$

Lemma 2.5 (Shao, 1993 [4])**.**

$$\hat{R}_n^{(1)}(\alpha) = \begin{cases} R_n(\alpha) + \sigma^2 + o_{\mathbb{P}}(1) & \text{if } \alpha \in \mathcal{T}^c \\ \frac{1}{n}\|M_\alpha \boldsymbol{e}\|^2 + \frac{2}{n}\sigma^2 p(\alpha) + o_{\mathbb{P}}(n^{-1}) & \text{if } \alpha \in \mathcal{T} \end{cases} \tag{3}$$

Proof. Using the Taylor expansion of $1/(1-x)^2 = 1 + 2x + O(x^2)$, we have

$$\frac{1}{(1-h_{ii,\alpha})^2} = 1 + 2h_{ii,\alpha} + O_{\mathbb{P}}\left(h_{ii,\alpha}^2\right).$$

Thus,

$$\hat{R}_n^{(1)}(\alpha) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_{i\alpha}^\top \hat{\boldsymbol{\beta}}_\alpha\right)^2}_{\xi_{\alpha,n}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(2h_{ii,\alpha} + O_{\mathbb{P}}\left(h_{ii,\alpha}^2\right)\right)\left(y_i - \boldsymbol{x}_{i\alpha}^\top \hat{\boldsymbol{\beta}}_\alpha\right)^2}_{\zeta_{\alpha,n}} \tag{4}$$

Let $\xi_{\alpha,n}$ and $\zeta_{\alpha,n}$ denote the first and second terms in (4), respectively. Note that

$$\begin{aligned}
\xi_{\alpha,n} &= \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta} + M_\alpha \boldsymbol{e}\|^2 \\
&= \frac{1}{n}\left(\|M_\alpha \boldsymbol{e}\|^2 + \|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta}\right) && (5) \\
&= \frac{1}{n}\|\boldsymbol{e}\|^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{1}{n}\|H_\alpha \boldsymbol{e}\|^2 + \frac{2}{n}\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta} && (6)
\end{aligned}$$

From here, we emphasize four intermediate steps:

i. Using Markov's inequality, for $\varepsilon > 0$,

$$\mathbb{P}\left\{\|H_\alpha \boldsymbol{e}\|^2 \geq n\varepsilon\right\} \leq \frac{\sigma^2 p_n(\alpha)}{n\varepsilon} \to 0$$

$$\implies \frac{1}{n}\|H_\alpha \boldsymbol{e}\|^2 = o_{\mathbb{P}}(1).$$

ii. Since M_α is a projection matrix, $\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 \leq \|\boldsymbol{X}\boldsymbol{\beta}\|^2 = O_{\mathbb{P}}(n)$, so that

$$\mathbb{E}\left[\left(\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta}\right)^2 \mid \boldsymbol{X}\right] = \frac{4}{n^2}\sigma^2\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 = o_{\mathbb{P}}(1).$$

Combining the latter with $\mathbb{E}\left[\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{X}\right] = 0$, we obtain that

$$\frac{2}{n}\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta} = o_{\mathbb{P}}(1).$$

iii. Combining i. and ii. with (6) yields

$$\xi_{\alpha,n} = \frac{1}{n}\|\boldsymbol{e}\|^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + o_{\mathbb{P}}(1).$$

Furthermore, since $\|\boldsymbol{e}\|^2 = O_{\mathbb{P}}(n)$, we have that $\xi_{\alpha,n} = O_{\mathbb{P}}(1)$.

iv. Finally, since $0 < h_{ii,\alpha} < 1$, $2h_{ii,\alpha} + O_{\mathbb{P}}\left(h_{ii,\alpha}^2\right) \leq O_{\mathbb{P}}\left(\max_i h_{ii,\alpha}\right)$. Thus,

$$\zeta_{\alpha,n} \leq O_{\mathbb{P}}\left(\max_i h_{ii,\alpha}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_{i\alpha}^\top \hat{\boldsymbol{\beta}}_\alpha\right)^2\right) = O_{\mathbb{P}}\left(\max_i h_{ii,\alpha}\right)\xi_{\alpha,n}.$$

From assumption **H3**, $\zeta_{\alpha,n} = o_{\mathbb{P}}(1)\xi_{\alpha,n} = o_{\mathbb{P}}(1)$.

It follows that

$$\hat{R}_n^{(1)}(\alpha) = \frac{1}{n}\|e\|^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + o_{\mathbb{P}}(1) \overset{\text{(LLN)}}{=} \sigma^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + o_{\mathbb{P}}(1).$$

Noting that $R_n(\alpha) = \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + o_{\mathbb{P}}(1)$ yields the first case in (3).

If $\alpha \in \mathcal{T}$, it is easy to see from (5) that $\xi_{\alpha,n} = 1/n\|M_\alpha \boldsymbol{e}\|^2$, Furthermore,

$$\zeta_{\alpha,n} = \frac{2}{n}\sigma^2 p(\alpha) + o_{\mathbb{P}}(1), \qquad \text{(?)}$$

proving the second case. □

Proposition 2.6 (Shao, 1993 [4]). *Suppose that $\mathcal{T}$ is non-empty and let $\hat{\alpha}^{(1)}$ be the model minimizing $\hat{R}_n^{(1)}(\alpha)$.*

1. *Under H1, H2, and H3,*
$$\lim_{n\to\infty} \mathbb{P}\left\{\hat{\alpha}^{(1)} \in \mathcal{T}^c\right\} = 0.$$

2. *For $\alpha \in \mathcal{T}$ with $\alpha \neq \alpha^*$,*
$$\mathbb{P}\left\{\hat{R}_n^{(1)}(\alpha) \leq \hat{R}_n^{(1)}(\alpha^*)\right\} = \mathbb{P}\left\{2\left(p(\alpha) - p(\alpha^*)\right)\sigma^2 < \boldsymbol{e}^\top (H_\alpha - H_{\alpha^*})\boldsymbol{e}\right\} + o_{\mathbb{P}}(1).$$
In particular, if $\boldsymbol{e} \sim \mathcal{N}(0_n, \sigma^2 I_n)$,
$$\mathbb{P}\left\{\hat{R}_n^{(1)}(\alpha) \leq \hat{R}_n^{(1)}(\alpha^*)\right\} = \mathbb{P}\left\{2k < \chi^2(k)\right\} + o_{\mathbb{P}}(1) > 0$$
for $k = p(\alpha) - p(\alpha^*)$.

3. *If $p(\alpha^*) < p$,*
$$\lim_{n\to\infty} \mathbb{P}\left\{\hat{\alpha}^{(1)} = \alpha^*\right\} \neq 1.$$

Proof.
1. Let $\bar{\alpha} \in \mathcal{T}$ and $\tilde{\alpha} \in \mathcal{T}^c$. By, Lemma 2.5, we have that

$$\begin{aligned}
\mathbb{P}\left\{\hat{R}_n^{(1)}(\tilde{\alpha}) \leq \hat{R}_n^{(1)}(\bar{\alpha})\right\} &= \mathbb{P}\Big\{\frac{1}{n}\sigma^2 p(\tilde{\alpha}) + \frac{1}{n}\|M_{\tilde{\alpha}}\boldsymbol{X}\boldsymbol{\beta}\|^2 + \sigma^2 + o_{\mathbb{P}}(1) \\
&\qquad\qquad \leq \frac{1}{n}\|M_{\bar{\alpha}}\boldsymbol{e}\|^2 + \frac{1}{n}\sigma^2 p(\bar{\alpha}) + o_{\mathbb{P}}(n^{-1})\Big\} \\
&= \mathbb{P}\left\{\frac{1}{n}\sigma^2\left(p(\tilde{\alpha}) - p(\bar{\alpha})\right) + \sigma^2 + \frac{1}{n}\|M_{\tilde{\alpha}}\boldsymbol{X}\boldsymbol{\beta}\|^2 - \frac{1}{n}\|M_{\bar{\alpha}}\boldsymbol{e}\|^2 \leq o_{\mathbb{P}}(1)\right\}
\end{aligned}$$

From **H1**, the latter probability goes to zero as $n \to \infty$. Therefore, $\mathbb{1}_{\left[\hat{R}_n^{(1)}(\tilde{\alpha}) \leq \hat{R}_n^{(1)}(\bar{\alpha})\right]} = o_p(1)$. We now observe that

$$\mathbb{P}\left\{\hat{\alpha}^{(1)} \in \mathcal{T}^c\right\} = \mathbb{E}\left[\mathbb{1}_{\left[[\hat{\alpha}^{(1)} \in \mathcal{T}^c]\right]}\right] = \mathbb{E}\left[\sum_{\tilde{\alpha}\in\mathcal{T}^c}\prod_{\alpha\in\mathcal{A}} \mathbb{1}_{\left[\hat{R}_n^{(1)}(\tilde{\alpha}) \leq \hat{R}_n^{(1)}(\alpha)\right]}\right] \to 0.$$

2. The first part follows from Lemma 2.1 by algebraic manipulation. The second part follows by noting that, if $\boldsymbol{e} \sim \mathcal{N}(0_n, \sigma^2 I_n)$, then

$$\frac{\boldsymbol{e}^\top}{\sigma}\left(H_\alpha - H_{\alpha^*}\right)\frac{\boldsymbol{e}}{\sigma} \sim \chi^2\left(\operatorname{tr}\left(H_\alpha - H_{\alpha^*}\right)\right).$$

3. It is easy to see that $p(\alpha^*) = p$ if and only if $\mathcal{T} = \{\alpha^*\}$. Thus, if $p(\alpha^*) < p$, there exists $\alpha \in \mathcal{T}^c$ with $\alpha \neq \alpha^*$. The result then follows by part 2 above. □

Corollary 2.7. *Leave-one-out cross-validaton is not consistent for selection. In particular, it overfits with non-vanishing probability.*

2.3 A General Perspective: Shao, 1997

[Brief intro to section]

For this section, we allow the number of candidates in $\mathcal{A}_n$, as well as the candidates $\alpha \in \mathcal{A}_n$ themselves, to vary with n (though we assume that both remain finite). To illustrate why this might be useful, we consider two examples from [3]:

If we wish to approximate a univariate regression function $x \mapsto f(x)$ by a polynomial of degree at most $p_n < n$, we may consider the models indexed by $\mathcal{A}_n := \{\alpha_d : d \in [p_n]\}$, with $\alpha_d = \{1, \ldots, d\}$ and $f_{\alpha_d}(x) = \beta_0 + \beta_1 x + \cdots + \beta_d x^d$. Clearly, the number of candidate models increases as more observations become available.

The *Generalized Information Criterion* (GIC), defined below, is a generalization of multiple empirical criteria for model selection. Various cross-validation methods can be seen as special cases of the GIC.

Definition 2.4. *We define the GIC loss estimator to be*

$$\hat{R}_{n,\lambda_n}(\alpha) := \frac{\|\boldsymbol{y} - \hat{m}(\boldsymbol{X})\|^2}{n} + \frac{1}{n}\lambda_n \hat{\sigma}_n^2 p_n(\alpha) \quad \text{for } \alpha \in \mathcal{A}_n,$$

where $\hat{\sigma}_n^2$ is an estimator of σ^2 and λ_n is a sequence of positive real numbers satisfying $\lambda_n \geq 2$ and $\lambda_n/n \to 0$.

2.3.1 The case of $\lambda_n \equiv 2$

Proposition 2.8 (Shao, 1997 [3]). *Suppose that $\lambda_n = 2$ for all $n \geq 1$ and that $\hat{\sigma}_n^2$ is a consistent estimator of σ^2. Then,*

$$\hat{R}_{n,2}(\alpha) = \begin{cases} \frac{1}{n}\|e\|^2 + \frac{2}{n}\hat{\sigma}_n^2 p_n(\alpha) - \frac{1}{n}\|H_\alpha \boldsymbol{e}\|^2 & \text{if } \alpha \in \mathcal{T}_n \\ \frac{1}{n}\|e\|^2 + L_n(\alpha) + o_{\mathbb{P}}(L_n(\alpha)) & \text{if } \alpha \in \mathcal{T}_n^c \end{cases}$$

Theorem 2.9 (Shao, 1997 [3]). *Suppose that H4 holds and that $\hat{\sigma}_n^2$ is consistent for σ^2.*

1. *If $|\mathcal{T}_n| \leq 1$ for all but finitely many n, then $\hat{\alpha}_n^2$ asymptotically loss efficient.*

2. *Suppose that* $|\mathcal{T}_n| > 1$ *for all but finitely many* n. *If there exists a positive integer* m *such that* $\mathbb{E}\left[y_1 - \boldsymbol{x}_1^\top \boldsymbol{\beta}\right]^{4m} < \infty$ *and*

$$\sum_{\alpha \in \mathcal{T}_n} \frac{1}{(p_n(\alpha))^m} \to 0 \quad \text{or} \quad \sum_{\substack{\alpha \in \mathcal{T}_n,\\ \alpha \neq \alpha^*}} \frac{1}{(p_n(\alpha) - p_n(\alpha^*))^m}, \tag{7}$$

then $\hat{\alpha}_n^2$ *is assymptotically loss efficient.*

3. *Suppose that* $|\mathcal{T}_n| > 1$ *for all but finitely many* n. *Suppose, furthermore, that for some constant* $c > 2$,

$$\liminf_{n \to \infty} \inf_{Q_n \in \mathcal{Q}_{n,q}} \mathbb{P}\left\{\|Q_n \boldsymbol{e}\|^2 > c\sigma^2 q\right\} > 0, \tag{8}$$

where $\mathcal{Q}_{n,q}$ *is the set of all projection matrices of rank* q. *Then, if* $|\mathcal{T}_n|$ *is bounded or* $\mathcal{A}_n$ *is embedded, the condition that*

$$p_n(\alpha_n^*) \to \infty \quad \text{or} \quad \min_{\substack{\alpha \in \mathcal{T}_n,\\ \alpha \neq \alpha^*}} (p_n(\alpha) - p_n(\alpha^*)) \to \infty \tag{9}$$

is necessary and sufficient for the asymptotic loss efficiency of $\hat{\alpha}_n^2$.

Proof. The proof for 1. is given in the last paragraph of page 226. I don't understand it. □

Note that condition (8) is satisfied if $\boldsymbol{e} \sim \mathcal{N}(0_n, \sigma^2 I_n)$. Condition (9) is satisfied if $\mathcal{A}_n$ does not contain two correct models with fixed dimension for all but finitely many n.

Corollary 2.10 (Shao, 1997)**.** *If* $\mathcal{T}_n$ *contains exactly one model with fixed dimension for all but finitely many* n, *then* $\hat{\alpha}_n^2$ *is consistent.*

Proof. This follows immediately from Theorem 2.9 and Proposition 2.8. □

2.3.2 The case of $\lambda_n \to \infty$

The proofs are missing here.

We now consider the case of the GIC $\hat{R}_{n,n}(\lambda_n)$ with $\lambda \to \infty$ as $n \to \infty$. Unlike in the previous case, the following results do not require that $\hat{\sigma}_n^2$ be consistent for σ^2.

Proposition 2.11 (Shao, 1997 [3])**.** *Suppose that* $\lambda_n = 2$ *for all* $n \geq 1$ *and that* $\hat{\sigma}_n^2$ *is a consistent estimator of* σ^2. *Then,*

$$\hat{R}_{n,2}(\alpha) = \begin{cases} \frac{1}{n}\|e\|^2 + \frac{2}{n}\hat{\sigma}_n^2 p_n(\alpha) - \frac{1}{n}\|H_\alpha \boldsymbol{e}\|^2 & \text{if } \alpha \in \mathcal{T}_n \\ \frac{1}{n}\|e\|^2 + L_n(\alpha) + \frac{1}{n}p_n(\lambda_n \hat{\sigma}_n^2 - 2\sigma^2) + o_{\mathbb{P}}(L_n(\alpha)) & \text{if } \alpha \in \mathcal{T}_n^c \end{cases}$$

Theorem 2.12 (Shao, 1997 [3])**.** *Suppose that* ***H4*** *holds and that*

$$\limsup_{n \to \infty} \sum_{\alpha \in \mathcal{T}_n} \frac{1}{p_n(\alpha)^m} \tag{10}$$

for some m *with* $\mathbb{E}\left[e_i^{4m}\right] < \infty$.

1. *If **H1**, $\lambda_n \to \infty$, and $\lambda_n p_n / n \to 0$ are satisfied, then $\hat{R}_{n,n}(\lambda_n)$ is asymptotically loss efficient.*

2. *If there exists $\alpha_0 \in \mathcal{T}_n$ with $p_n(\alpha_0)$ constant for all but finitely many n, $\lambda_n \to \infty$, and $\lambda_n / n \to 0$, then $\hat{R}_{n,n}(\lambda_n)$ is consistent.*

Remark: Condition (10) is satisfied whenever $|\mathcal{T}_n|$ is bounded or $\mathcal{A}_n$ is embedded. It implies that

$$\max_{\alpha \in \mathcal{T}_n} \frac{\|H_\alpha \boldsymbol{e}\|^2}{\lambda_n \hat{\sigma}_n^2 p_n(\alpha)} \xrightarrow{\mathbb{P}} 0.$$

2.3.3 Cross-validation

Theorem 2.13 (Shao, 1997 [3]). 1. *If **H3** holds, then Theorem 2.9 applies for the leave-one-out estimator.*

2. *Suppose that **H1**, **H4**, and (10) hold. If the splits are "balanced", and d is chosen so that $d/n \to 1$ and $p_n/(n-d) \to 0$, then the delete-d cross-validation estimator is asymptotically loss efficient*

MISSING: Discussion.

3 CV for Nonparametric Model Selection

3.1 Comparison of Distinct Procedures: Yang, 2007

Here we consider two regression procedures, denoted δ_1 and δ_2, that yield estimators $\hat{f}_{n,1}$ and $\hat{f}_{n,2}$ of the regression function stisfying

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i \quad i \in [n], \tag{11}$$

for $\boldsymbol{x}_i$ iid, $\mathbb{E}[\epsilon_i \mid \boldsymbol{X}] \overset{\text{a.s.}}{=} 0$ and $\mathbb{E}[\epsilon_i^2 \mid \boldsymbol{X}] \overset{\text{a.s.}}{<} \infty$.

Definition 3.1. *We say δ_1 is* assymptotically better *than δ_2 under the loss function L if, for $0 < \varepsilon < 1$, there exists $c_\varepsilon > 0$ such that*

$$\mathbb{P}\left\{L_n\left(\hat{f}_{n,2}\right) \geq (1 + c_\epsilon) L_n\left(\hat{f}_{n,2}\right)\right\} \geq 1 - \varepsilon.$$

Given that δ_1 is assymptotically better than δ_2, we say that a selection procedure is consistent if it selects δ_1 with probability tending to 1 as $n \to \infty$.

3.1.1 Single-split cross-validation (the Hold-out)

For this section, we assume that the first n_1 elements in $\mathcal{D}_n$ are used as a training/estimation sample and the remaining n_2 elements make up the validation sample. We write p_n and q_n for the rates of convergence of the estimators $\hat{f}_{n,1}$ and $\hat{f}_{n,2}$, respectively. That is,

$$O_{\mathbb{P}}(p_n) = \|f - \hat{f}_{n,1}\|_2 \quad \text{and} \quad O_{\mathbb{P}}(q_n) = \|f - \hat{f}_{n,2}\|_2.$$

The hold-out cross-validation method consists in selecting the estimator that minimizes the hold-out loss

$$L_{\text{ho}}(\hat{f}_{n,j}) = \sum_{i=n_1+1}^{n} \left(y_i - \hat{f}_{n,j}(\boldsymbol{x}_i) \right)^2 \quad \text{for } j = 1, 2.$$

The propositions in this section rely on the following conditions:

- **A3.1:** $\mathbb{E}\left[\epsilon_i^2 \mid \boldsymbol{x}_i\right]$ is bounded a.s. for $i \in [n]$.
- **A3.2:** There exists A_n such that $\|f - \hat{f}_{n,j}\|_\infty = O_{\mathbb{P}}(A_n)$ for $j = 1, 2$.
- **A3.3:** One procedure is asymptotically better than the other.
- **A3.4:** There exists M_n such that $\|f - \hat{f}_{n,j}\|_4 / \|f - \hat{f}_{n,j}\|_4 = o_{\mathbb{P}}(M_n)$ for $j = 1, 2$.

Theorem 3.1 (Yang, 2007 [6])**.** *Suppose that **A3.1**–**A3.4** hold. Suppose, furthermore, that*

1. $n_1 \to \infty$
2. $n_2 \to \infty$
3. $n_2 M_n^{-4} \to \infty$
4. $\sqrt{n_2} \max(p_{n_1}, q_{n_1})$

Then, the hold-out CV procedure is consistent.

A very detailed proof of this result is provided in Yang [6], so it will be skipped here.

3.1.2 Voting cross-validation with multiple splits

The (theoretical) majority-vote cross-validation method proceeds as follows: for each permutation $i \mapsto \pi(i)$ of the data, we compute the estimators $\hat{f}_{n_1,1}$ and $\hat{f}_{n_1,2}$ using the first n_1 data points $(y_{\pi(1)}, \boldsymbol{x}_{\pi(1)}), \ldots, (y_{\pi(n_1)}, \boldsymbol{x}_{\pi(n_1)})$ as the training sample and the remaining $n_2 = n - n_1$ elements as the validation sample. We then find the estimator that minimizes the hold-out loss

$$L_\pi(\hat{f}_{n_1,j}) = \sum_{i=n_1+1}^{n} \left(y_{\pi(i)} - \hat{f}_{n_1,j}\left(\boldsymbol{x}_{\pi(i)}\right) \right)^2 \qquad \text{for } j = 1, 2.$$

The chosen estimator is the one favored by the majority of the permutations. More formally, we define

$$\tau_\pi = \mathbb{1}_{\left[L_\pi(\hat{f}_{n_1,1}) \le L_\pi(\hat{f}_{n_1,2})\right]}$$

We then define our selection criterion as follows:

$$\hat{f}_n = \begin{cases} \hat{f}_{n,1} & \text{if } \sum_{\pi \in \Pi} \tau_\pi \ge n!/2, \\ \hat{f}_{n,2} & \text{otherwise,} \end{cases}$$

where Π denotes the set of all permutations of $[n]$.

Theorem 3.2 (Yang, 2007 [6]). *Under the conditions of Theorem 4.1 and the condition that the data is iid, the majority-vote cross-validation method is consistent.*

Proof. Suppose that δ_1 is asymptotically better than δ_2. For $\pi \in \Pi$, we have that

$$\mathbb{P}\left\{L_\pi\left(\hat{f}_{n_1,1}\right) \leq L_\pi\left(\hat{f}_{n_1,2}\right)\right\} = \mathbb{E}\left[\tau_\pi\right] \stackrel{(*)}{=} \mathbb{E}\left[\frac{1}{n!}\sum_{\pi\in\Pi}\tau_\pi\right].$$

The equality at $(*)$ follows from the fact that the data are iid, hence exchangeable, and thus the τ_π are identically distributed. By Theorem 3.1, the right-hand side converges to 1 as $n \to \infty$. Since the average $1/n!\sum_\pi \tau_\pi$ is almost surely at most 1, it follows that $1/n!\sum_\pi \tau_\pi \to 1$ in probability, and the majority-vote cross-validation method is consistent. □

The proof of Theorem 3.2 does not require using the entire set Π of permutations for the majority vote. In fact, Theorem 3.1 establishes that even a single data split suffices for consistency, provided the splitting conditions are met. Moreover, Yang [6] presents a counterexample demonstrating that these conditions are not merely sufficient but necessary, hence showing that the number of splits does not affect consistency. In other words, multiple splits in cross-validation cannot rescue an inconsistent single-split procedure. A natural question, then, is: if multiple splits do not improve consistency, what is their benefit? This will be explored in a simulation later on.

4 Aggregation

4.1 Bunea et al., 2007

As before, we consider independent pairs in $\mathcal{D}_n := \{(y_i, \boldsymbol{x}_i) : i \in [n]\}$ satisfying (11). Suppose that we have M candidate estimators of the regression function, denoted $\hat{f}_{n,1}, \hat{f}_{n,2}, \ldots, \hat{f}_{n,M}$. Instead of selecting a single estimator, we combine them into an *aggregate* $\tilde{f}_{\hat{\lambda}}$ given by

$$\tilde{f}_{\hat{\lambda}} = \sum_{j=1}^{M} \hat{\lambda}_j \hat{f}_{n,j},$$

with $\hat{\lambda} := \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_M\right) \in \Lambda \subset \mathbb{R}^M$ chosen to satisfy

$$\hat{\lambda} = \underset{\lambda\in\Lambda}{\arg\min}\left\{\frac{1}{n}\|y - f_\lambda(\boldsymbol{x})\|^2 - \mathrm{pen}(\lambda)\right\} \tag{12}$$

for some penalty $\mathrm{pen}(\lambda)$ on the coefficients.

4.1.1 Four types of aggregation

There are four aggregation schemes considered in Bunea et al. [1], each of which is characterized by a different set Λ of admissible weights $\hat{\lambda}$:

- Model Selection Aggregation (MS): A single estimator is selected. That is,

$$\Lambda_{\mathrm{MS}} = \left\{\lambda \in \mathbb{R}^M : \lambda = \boldsymbol{e}_j \text{ for some } j \in [M]\right\}.$$

- Linear Aggregation (L): $\tilde{f}_{\hat{\lambda}}$ is chosen among all linear combinations of the estimators. That is,

$$\Lambda_{\mathrm{L}} = \mathbb{R}^M.$$

- Convex Aggregation (C): $\tilde{f}_{\hat{\lambda}}$ is chosen among all convex combinations of the estimators. That is,

$$\Lambda_{\mathrm{C}} = \left\{\lambda \in \mathbb{R}^M : \lambda \geq 0, \sum_{j=1}^{M} \lambda_j = 1\right\}.$$

- Subset Selection (S): We select and aggregate at most D estimators from the pool, for a given $D \leq M$. That is,

$$\Lambda_{\mathrm{S}} = \left\{\lambda \in \mathbb{R}^M : \lambda \text{ has at most } D \text{ non-zero entries}\right\}.$$

4.1.2 Evaluating the aggregate

In an ideal scenario, we would like to select weights λ^* satisying

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \mathbb{E}\left[d\left(f, \tilde{f}_\lambda\right)\right]$$

for some distance function d (e.g., the L_2 norm). However, since the true regression function f is unknown, this approach is clearly not feasible. Another way of constructing an estimator is to minimize its maximum risk on a class of functions Θ containing f. That is, we would like to find $\hat{\lambda}$ satisfying

$$\sup_{f \in \Theta} \mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|_2^2 = \inf_{\lambda \in \Lambda} \sup_{f \in \Theta} \mathbb{E}\|f - \tilde{f}_\lambda\|_2^2.$$

This is known as *minimax* extimation. However, once again, there is no obvious way to compute the expectation $\mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|_2^2$ for an arbitrary $f \in \Theta$.

For these reasons, we instead adopt the least-squares approach in (12). But how can we know if this approach is any good? We need a tool to evaluate the performance of our aggregate against *any* possible of regression function f. Oracles provide us with such a tool.

Definition 4.1 (adapted from Tsybakov, 2009 [5]). *Suppose that there exists $\lambda^* \in \Lambda$ such that*

$$\mathbb{E}\|f - \tilde{f}_{\lambda^*}\|_2^2 = \inf_{\lambda \in \Lambda} \mathbb{E}\|f - \tilde{f}_{\lambda^*}\|_2^2.$$

The function $f \mapsto \tilde{f}_{\lambda^*}$ is called the oracle of aggregation under L_2.

We say that the aggregate $\tilde{f}_{\hat{\lambda}}$ mimics the oracle if

$$\mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|_2 \leq \inf_{\lambda \in \Lambda} \mathbb{E}\|f - \tilde{f}_\lambda\|_2 + \Delta_{n,M}. \tag{13}$$

for the smalles possible $\Delta_{n,M} > 0$ independent of f.

In what follows, the goal is to find lower bounds on $\Delta_{n,M}$ for each of the aggregation schemes.

Definition 4.2 (Tsybakov, 2009 [5]). *For a class of functions Θ, a sequence $\{\psi_n\}_{n\geq 1}$ of positive numbers is called an* optimal rate of convergence *of estimators $\hat{f}$ on Θ under L_2 if there exist constants $c, C > 0$ such that*

$$\limsup_{n\to\infty} \left(\psi_n^{-2} \inf_{\hat{f}} \sup_{f\in\Theta} \mathbb{E}\left[\|f - \hat{f}\|_2^2 \right] \right) \leq C \tag{14}$$

$$\text{and} \qquad \liminf_{n\to\infty} \left(\psi_n^{-2} \inf_{\hat{f}} \sup_{f\in\Theta} \mathbb{E}\left[\|f - \hat{f}\|_2^2 \right] \right) \geq c \tag{15}$$

An estimator $\hat{f}_n$ is said to be rate-optimal *if*

$$\sup_{f\in\Theta} \mathbb{E}\left[\|f - \hat{f}_n\|_2^2 \right] \leq C' \psi_n^2$$

for some $C' > 0$. It is called asymptotically efficient *of Θ under L_2 if*

$$\lim_{n\to\infty} \frac{\sup_{f\in\Theta} \mathbb{E}\|f - \hat{f}_n\|^2}{\inf_{\hat{f}} \sup_{f\in\Theta} \mathbb{E}\|f - \hat{f}\|_2^2} = 1.$$

We adapt Theorem 5.1 in [1] to consider exclusively the L_2 norm:

Theorem 4.1 (Bunea et al., 2007 [1]). *(Statement of lower bounds)*

$$\sup_{f_1,\ldots,f_2\in\mathcal{F}_0} \inf_{T_n} \sup_{f\in\mathcal{F}_0} \left\{ \mathbb{E}\|f - T_n\|_2^2 - \min_{\lambda\in\Lambda} \|f - \tilde{f}_\lambda\|_2^2 \right\} \geq c\psi_n$$

INCOMPLETE SECTION

4.1.3 A BIC-type penalty

Definition 4.3. *Let $M(\lambda) := \|\lambda\|_0$ (i.e., the number of non-zero coefficients in λ). For $a > 0$, we define the Bunea-Tsybakov-Wegkamp (I don't know what to call it) penalty to be*

$$\text{pen}_{\text{BIC}}(\lambda) := \frac{2\sigma^2}{n} M(\lambda) \left(1 + \frac{2+a}{1+a} \sqrt{2\log\left(\frac{eM}{M(\lambda)\vee 1}\right)} + \frac{1+a}{a}\left[2\log\left(\frac{eM}{M(\lambda)\vee 1}\right) \right] \right).$$

This penalty yields the BIC-type least-squares aggregate $\tilde{f}_{\hat{\lambda}_{\text{BIC}}} =: \tilde{f}_{\text{BIC}}$ with

$$\hat{\lambda}_{BIC} = \arg\min_{\lambda\in\mathbb{R}^M} \left\{ \frac{1}{n}\|\boldsymbol{y} - f_\lambda(\boldsymbol{x})\|^2 - \text{pen}_{\text{BIC}}(\lambda) \right\}$$

Theorem 4.2 (Bunea et al., 2007 [1]). *Assume that the e_i are iid $\mathcal{N}(0,\sigma^2)$ and that the functions $f, \hat{f}_{n,1}, \ldots, \hat{f}_{n,M}$ are uniformly bounded. Then, for all $a > 0$, $M \geq 2$, and $n \geq 1$,*

$$\mathbb{E}\|\tilde{f}_{\text{BIC}} - f\|^2 \leq (1+a) \inf_{\lambda\in\mathbb{R}^M} \left\{ \|\tilde{f}_\lambda - f\|^2 + \frac{\sigma^2}{n}\left(5 + \frac{2+3a}{a}\left(2\log\left(\frac{eM}{M(\lambda)\vee 1}\right)\right)\right) M(\lambda) \right\} + \frac{6\sigma^2(1+a)^2}{an(e-1)}$$

References

[1] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. "Aggregation for Gaussian Regression". In: *The Annals of Statistics* 35.4 (Aug. 2007).

[2] Bruce E. Hansen. *Econometrics*. Princeton: Princeton University Press, 2022.

[3] Jun Shao. "An Asymptotic Theory for Linear Model Selection". In: *Statistica Sinica* 7.2 (Apr. 1997), pp. 221–264.

[4] Jun Shao. "Linear Model Selection by Cross-validation". In: *Journal of the American Statistical Association* 88.422 (June 1993), pp. 486–494.

[5] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. English Edition. Springer Series in Statistics. New York: Springer, 2009.

[6] Yuhong Yang. "Consistency of cross validation for comparing regression procedures". In: *The Annals of Statistics* 35.6 (Dec. 2007).