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

Diego Urdapilleta de la Parra

March 25, 2025

Note: I removed the boxes to keep the proposition enumeration consistent, and to make the text easier to read.

Contents

1 CV for Linear Model Selection

1.1 Setup and preliminary results

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 $(y, \boldsymbol{x}) \in \mathbb{R}^{1+p_n}$. We treat the $\boldsymbol{x}_i$ as predictors of the outcome y_i, and we assume a linear model

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

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 $\mathrm{Cov}\,(\boldsymbol{e}) = \sigma_2 \boldsymbol{I}_n$.

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

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

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 $\alpha \in \mathcal{A}$ is *correct* if $\mathbb{E}\,[\boldsymbol{y} \mid \boldsymbol{X}] \overset{\text{a.s.}}{=} m_\alpha(\boldsymbol{X})$, and we denote by $\mathcal{A}_c$ the set of correct models in $\mathcal{A}$
2. We say $\alpha \in \mathcal{A}$ is *wrong* if it is not correct, and we denote by $\mathcal{A}_w$ the set of wrong models in $\mathcal{A}$
3. We say $\mathcal{A}$ is *embedded* if there exists an enumeration $\alpha_1, \alpha_2, \ldots, \alpha_k$ of all elements in $\mathcal{A}$ such that
$$\alpha_1 \subset \alpha_2 \subset \cdots \subset \alpha_k.$$

Definition 1.1. *For $\alpha \in \mathcal{A}$, let $\hat{\boldsymbol{\beta}}_\alpha$ be the OLS estimator of $\boldsymbol{\beta}_\alpha$ and $\hat{m}_\alpha(\boldsymbol{X}) := \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha$. We denote the average squared error of $\hat{m}_\alpha$ by*

$$L_n(\alpha) := \frac{1}{n} \|\mathbb{E}\,[\boldsymbol{y} \mid \boldsymbol{X}] - \hat{m}_\alpha(\boldsymbol{X})\|^2.$$

Additionally, we write $R_n(\alpha) := \mathbb{E}\,[L_n(\alpha) \mid \boldsymbol{X}]$.

The following conditions will be used throughout this section:

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

Proposition 1.1. *If we assume a linear model* $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}$*, then*

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

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

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}] - \hat{m}_\alpha(\boldsymbol{X})\|^2 &= \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha\|^2 \\ &= \|\boldsymbol{X}\boldsymbol{\beta} - H_\alpha(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e})\|^2 \\ &= \|M_\alpha \boldsymbol{X}\boldsymbol{\beta} - H_\alpha \boldsymbol{e}\|^2.\end{aligned}$$

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 $\mathbb{E}[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}] = \sigma^2 p_n(\alpha)$ by the "trace trick", where $p_n(\alpha)$ denotes the size of model α. □

Proposition 1.2. *Suppose that the set of correct candidate models* $\mathcal{A}_c \subset \mathcal{A}$ *is non-empty, and let* α^* *be the smallest correct model in* $\mathcal{A}_c$*. Then,* α^* *minimizes* $R_n(\alpha)$ *over* $\alpha \in \mathcal{A}$*.*

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

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

Now suppose that $\alpha \in \mathcal{A}_w$. If $p_n(\alpha) \geq p_n(\alpha^*)$, the result follows by assumption H1. If $p_n(\alpha) \geq p_n(\alpha^*)$, then ... MISSING. □

1.2 Shao, 1993: Notes on Leave-One-Out CV

In this section, we assume that $p(\alpha) := p_n(\alpha)$ is constant for each $\alpha \in \mathcal{A}$.

Definition 1.2. *The LOOCV estimator of* $R_n(\alpha)$ *is definded 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 1.3 (Shao, 1993 [3])**.**

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

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,

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

Let $\xi_{\alpha,n}$ and $\zeta_{\alpha,n}$ denote the first and second terms in (1), respectively. Note that

$$\begin{aligned} \xi_{\alpha,n} &= \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta} + M_\alpha \boldsymbol{e}\|^2 \\ &= \frac{1}{n}\left(\|M_\alpha \boldsymbol{e}\|^2 + \|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta}\right) \qquad (2) \\ &= \frac{1}{n}\|\boldsymbol{e}\|^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{1}{n}\|H_\alpha \boldsymbol{e}\|^2 + \frac{2}{n}\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta} \\ &= \frac{1}{n}\|\boldsymbol{e}\|^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 + o_{\mathbb{P}}(1). \qquad (3) \end{aligned}$$

The equality at (3) follows from the fact that $\mathbb{E}\left[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}\right] = \sigma^2 p(\alpha)$ and

$$\mathbb{E}\left[\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{X}\right]^2 = \sigma^2 \|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 = O_{\mathbb{P}}(n), \qquad (?)$$

so that $1/n\|H_\alpha \boldsymbol{e}\|^2 \to_{\mathbb{P}} 0$ and $2/n\left(\boldsymbol{e}^\top M_\alpha \boldsymbol{X}\boldsymbol{\beta}\right) = O_{\mathbb{P}}(1)$. (?)

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

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

(3) and (4) imply the first case in the Lemma.DOES IT?

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

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

proving the second case. $\square$

Proposition 1.4 (Shao, 1993 [3])**.** *Let $\hat{\alpha}^{(1)}$ be the model minimizing $\hat{R}_n^{(1)}(\alpha)$.*

1. Under H1, H2, and H3,

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\alpha}^{(1)} \in \mathcal{A}_w\right) = 0.$$

2. For $\alpha \in \mathcal{A}_c$ 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 $\boldsymbol{e} \sim \mathcal{N}(0_n, \sigma^2 I_n)$,

$$\mathbb{P}\left(\hat{R}_n^{(1)}(\alpha) \leq \hat{R}_n^{(1)}(\alpha^*)\right) = \mathbb{P}\left(2k < \chi^2(k)\right) + o_{\mathbb{P}}(1)$$

for $k = p(\alpha) - p(\alpha^*)$.

3. If $p(\alpha^*) \neq p$,

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

Proof.
1. MISSING
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

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

3. If $p(\alpha^*) = p$, then $\mathcal{A}_c = \{\alpha^*\}$. It follows from 1. that $\lim_{n\to\infty} \mathbb{P}\left(\hat{\alpha}^{(1)} = \alpha^*\right) = 1$. Conversely, if $\lim_{n\to\infty} \mathbb{P}\left(\hat{\alpha}^{(1)} = \alpha^*\right) = 1$ MISSING ☐

Corollary 1.5. *LOOCV is not consistent. In particular, overfits with non-vanishing probability.*

A subsequent result states that cross-validation is consistent if $n_v/n \to 1$ as $n \to \infty$, where n_v is the number of validation samples.

1.3 Shao, 1997

Definition 1.3. *Let $\hat{\alpha}_n$ be the model selected by minimizing some criterion $\hat{R}_n$ over $\mathcal{A}$, and let α_n^* denote the model minimizing R_n over $\mathcal{A}$. We say $\hat{R}_n$ is* consistent *if*

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

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

$$\frac{R_n(\hat{\alpha})}{R_n(\alpha_n^*)} \to 1 \quad a.s.$$

Proposition 1.6 (Shao, 1997 [2]). *Suppose H1, $p_n/n \to 0$, and that $\mathcal{A}_c$ is non-empty for all but finitely many n.*

1. *If $|\mathcal{A}_c| = 1$ for all but finitely many n, then consistency is equivalent to efficiency in the sense of Definition 3.1*

2. *If $p_n(\alpha_n^*) \not\to_{\mathbb{P}} \infty$, then consistency is equivalent to efficiency in the sense of Definition 3.1*

Proof. MISSING □

Definition 1.4. *We define the GIC loss estimator to be*

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

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 $\lambda_n/n \to 0$.

1.3.1 The case of $\lambda_n \equiv 2$

Proposition 1.7 (Shao, 1997 [2])**.** *Suppose that $\lambda_n = 2$ for all $n \geq 1$ and that $\hat{\sigma}_n^2$ is a consistent estimator of σ^2. Then,*

$$\hat{R}_{n,2}(\alpha) = \Big\{ TBD$$

Theorem 1.8 (Shao, 1997 [2])**.** *Suppose that H4 holds and that $\hat{\sigma}_n^2$ is consistent for σ^2. Then, $\hat{\alpha}_n^2$ is consistent and asymptotically loss efficient.*

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

2. *Suppose that $|\mathcal{A}_c| > 1$ 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*

$$\sum_{\alpha \in \mathcal{A}_c} \frac{1}{(p_n(\alpha))^m} \to 0 \quad \text{or} \quad \sum_{\substack{\alpha \in \mathcal{A}_c, \\ \alpha \neq \alpha^*}} \frac{1}{(p_n(\alpha) - p_n(\alpha^*))^m}, \tag{5}$$

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

3. *Suppose that $|\mathcal{A}_c > 1|$ for all but finitely many n and that for any integer q and constant $c > 2$,*

$$\liminf_{n\to\infty} \inf_{Q_n \in \mathcal{Q}_{n,q}} \mathbb{P}\left\{\boldsymbol{e}_n^\top Q_n \boldsymbol{e}_n > c\sigma^2 q\right\} > 0, \tag{6}$$

where $\mathcal{Q}_{n,q}$ is the set of all projection matrices of rank q. The condition that

$$p_n(\alpha_n^*) \to \infty \quad \text{or} \quad \min_{\substack{\alpha \in \mathcal{A}_c, \\ \alpha \neq \alpha^*}} (p_n(\alpha) - p_n(\alpha^*)) \to \infty \tag{7}$$

is necessary and sufficient for the asymptotic loss efficiency of $\hat{\alpha}_n^2$ whenever $|\mathcal{A}_c|$ is bounded or $\mathcal{A}$ is embedded.

Proof. MISSING □

Note that condition (6) is satisfied if $\boldsymbol{e} \sim \mathcal{N}(0_n, \sigma^2 I_n)$. Condition (7) is satisfied if $\mathcal{A}$ does not contain two correct models with fixed dimensions for all but finitely many n.

Corollary 1.9 (Shao, 1997)**.** *If $\mathcal{A}_c$ 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 1.8 and Proposition 1.7. □

INCOMPLETE: Missing $\lambda_n \to \infty$ and cross-validation discussion.

2 CV for Nonparametric Model Selection

2.1 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

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i \quad i \in [n], \tag{8}$$

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

Definition 2.1. *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 $n \to \infty$.

2.1.1 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,

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

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:

- **A2.1:** $\mathbb{E}\left[\epsilon_i^2 \mid \boldsymbol{x}_i\right]$ is bounded a.s. for $i \in [n]$.

- **A2.2:** There exists A_n such that $\|f - \hat{f}_{n,j}\|_\infty = O_{\mathbb{P}}(A_n)$ for $j = 1, 2$.
- **A2.3:** One procedure is asymptotically better than the other.
- **A2.4:** 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 2.1 (Yang, 2007 [5])**.** *Suppose that **A2.1**–**A2.4** 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 [5], so it will be skipped here.

2.1.2 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

$$\tau_\pi = \mathbb{1}_{L_\pi(\hat{f}_{n_1,1}) \leq L_\pi(\hat{f}_{n_1,2})}$$

We then define our selection criterion as follows:

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

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

Theorem 2.2 (Yang, 2007 [5])**.** *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

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

The equality at $(*)$ follows from the fact that the data are iid, hence exchangeable, and thus the τ_π are identically distributed. By Theorem 2.1, 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 2.2 does not require using the entire set Π of permutations for the majority vote. In fact, Theorem 2.1 establishes that even a single data split suffices for consistency, provided the splitting conditions are met. Moreover, Yang [5] 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 the simulation later on. (MAYBE?)

3 Aggregation

3.1 Bunea et al., 2007

As before, we consider independent pairs in $\mathcal{D}_n := \{(y_i, \boldsymbol{x}_i) : i \in [n]\}$ satisfying (8). 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

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

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

3.1.1 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,

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

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

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

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

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

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

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

3.1.2 Optimality criteria

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, this is not an easy task, as Θ may be too large or unknown. We also know that the expected risk goes to 0 as $n \to \infty$. not sure about this

Instead, we consider an alternative approach using *oracles.*

Definition 3.1 (adapted from Tsybakov, 2009 [4]). *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

$$\mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|_2 \leq \inf_{\lambda \in \Lambda} \mathbb{E}\|f - \tilde{f}_\lambda\|_2 + \Delta_{n,M}. \tag{9}$$

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

Definition 3.2 (Tsybakov, 2009 [4]). *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*

$$\limsup_{n \to \infty} \left(\psi_n^{-2} \inf_{\hat{f}} \sup_{f \in \Theta} \mathbb{E}\left[\|f - \hat{f}\|_2^2\right]\right) \leq C \tag{10}$$

$$\text{and} \qquad \liminf_{n \to \infty} \left(\psi_n^{-2} \inf_{\hat{f}} \sup_{f \in \Theta} \mathbb{E}\left[\|f - \hat{f}\|_2^2\right]\right) \geq c \tag{11}$$

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*

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

Theorem 3.1 (Bunea et al., 2007 [1]). *(Statement of lower bounds)*

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

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

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

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

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