

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

Diego Urdapilleta de la Parra

March 18, 2025

1 Setup and preliminary results

Let n, p_n be positive integers and $\mathcal{D}_n := \{(y_i, \mathbf{x}_i) : i \in [n]\}$ be a set of independent data points drawn from a distribution $\mathbb{P}_{y, \mathbf{x}}$ for $(y, \mathbf{x}) \in \mathbb{R}^{1+p_n}$. We treat the \mathbf{x}_i as predictors of the outcome y_i , and we assume a linear model

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

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

We consider the following setup for model selection. 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 model

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

where \mathbf{X}_α is the sub-matrix of \mathbf{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 $\mathbb{E}[\mathbf{y} \mid \mathbf{X}] \stackrel{\text{a.s.}}{=} m_\alpha(\mathbf{X})$, 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 $\alpha_1, \alpha_2, \dots, \alpha_k$ 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 $\hat{m}_\alpha(\mathbf{X}) := \mathbf{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha$. We denote the average squared error of \hat{m}_α by

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

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

Assumptions

The following conditions will be used throughout the paper:

$$\begin{aligned}
\mathbf{H1} : \quad & \liminf_{n \rightarrow \infty} \frac{1}{n} \|M_\alpha \mathbf{X} \boldsymbol{\beta}\|^2 > 0 \text{ for all } \alpha \in \mathcal{A} \\
\mathbf{H2} : \quad & \mathbf{X}^\top \mathbf{X} = O(n) \quad \text{and} \quad (\mathbf{X}^\top \mathbf{X})^{-1} = O(n^{-1}) \\
\mathbf{H3} : \quad & \lim_{n \rightarrow \infty} \max_{i \leq n} h_{ii, \alpha} = 0 \text{ for all } \alpha \in \mathcal{A} \\
\mathbf{H4} : \quad & \sum_{\alpha \in \mathcal{A}_w} \frac{1}{(n R_n(\alpha))^m} \xrightarrow{\mathbb{P}} 0 \quad \text{for some } m \geq 1
\end{aligned}$$

Proposition 1.1

If we assume a linear model $\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{e}$, then

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

where $H_\alpha = \mathbf{X}_\alpha (\mathbf{X}_\alpha^\top \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha^\top$ and $M_\alpha = I_n - H_\alpha$.

Proof. First, we have that

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

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

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

Hence, the first part follows from the Pythagorean theorem.

For the second part, we note that $\mathbb{E}[\|H_\alpha \mathbf{e}\|^2 \mid \mathbf{X}] = \sigma^2 p_n(\alpha)$ by the “trace trick”, where $p_n(\alpha)$ denotes the size of model α . \square

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, $\mathbf{X}_\alpha \boldsymbol{\beta}_\alpha = \mathbf{X} \boldsymbol{\beta}$ and $p_n(\alpha^*) \leq$

$p_n(\alpha)$. Thus,

$$\begin{aligned}
R_n(\alpha) &= \frac{1}{n} \sigma^2 p_n(\alpha) + \frac{1}{n} \|M_\alpha \mathbf{X} \boldsymbol{\beta}\|^2 \\
&= \frac{1}{n} \sigma^2 p_n(\alpha) + \frac{1}{n} \underbrace{\|M_\alpha \mathbf{X}_\alpha \boldsymbol{\beta}_\alpha\|^2}_0 \\
&= \frac{1}{n} \sigma^2 p_n(\alpha) \geq \frac{1}{n} \sigma^2 p_n(\alpha^*) = R_n(\alpha^*).
\end{aligned}$$

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

2 Leave-One-Out CV

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

Definition 2.1

The LOOCV estimator of $R_n(\alpha)$ is

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

Lemma 2.1 (Shao, 1993)

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

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

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

Thus,

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

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

$$\begin{aligned}\xi_{\alpha,n} &= \frac{1}{n} \|M_\alpha \mathbf{X} \boldsymbol{\beta} + M_\alpha \mathbf{e}\|^2 \\ &= \frac{1}{n} (\|M_\alpha \mathbf{e}\|^2 + \|M_\alpha \mathbf{X} \boldsymbol{\beta}\|^2 + 2\mathbf{e}^\top M_\alpha \mathbf{X} \boldsymbol{\beta})\end{aligned}\tag{2}$$

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

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

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

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

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

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

(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 $\xi_{\alpha,n} = 1/n \|M_\alpha \mathbf{e}\|^2$, Furthermore,

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

proving the second case. □

Proposition 2.2 (Shao, 1993)

Let $\hat{\alpha}^{(1)}$ be the model minimizing $\hat{R}_n^{(1)}(\alpha)$.

1. Under H1, H2, and H3,

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

2. For $\alpha \in \mathcal{A}_c$ with $\alpha \neq \alpha^*$,

$$\mathbb{P}\left(\hat{R}_n^{(1)}(\alpha) \leq \hat{R}_n^{(1)}(\alpha^*)\right) = \mathbb{P}\left(2(p(\alpha) - p(\alpha^*))\sigma^2 < \mathbf{e}^\top (H_\alpha - H_{\alpha^*})\mathbf{e}\right) + o_{\mathbb{P}}(1).$$

In particular, if $\mathbf{e} \sim \mathcal{N}(0_n, \sigma^2 I_n)$,

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

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

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

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

Proof.

1. **MISSING**

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

$$\frac{\mathbf{e}^\top}{\sigma} (H_\alpha - H_{\alpha^*}) \frac{\mathbf{e}}{\sigma} \sim \chi^2(\text{tr}(H_\alpha - H_{\alpha^*})).$$

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

Corollary 2.3

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

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

3 Shao, 1997

Definition 3.1

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

$$\mathbb{P}\{\hat{\alpha} = \alpha^*\} \rightarrow 1$$

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

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

Proposition 3.1 (Shao, 1997)

Suppose H1, $p_n/n \rightarrow 0$, 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 $p_n(\alpha_n^*) \not\rightarrow_{\mathbb{P}} \infty$, then consistency is equivalent to efficiency in the sense of Definition 3.1

Proof. **MISSING** \square

Definition 3.2

We define the GIC loss estimator to be

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

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 $\lambda_n/n \rightarrow 0$.

3.1 The case of $\lambda_n \equiv 2$

Proposition 3.2 (Shao, 1997)

Suppose that $\lambda_n = 2$ for all $n \geq 1$ and that $\hat{\sigma}_n^2$ is a consistent estimator of σ^2 . Then,

$$\hat{R}_{n,2}(\alpha) = \begin{cases} \text{ TBD } \end{cases}$$

Theorem 3.3 (Shao, 1997)

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 $\mathbb{E}[y_1 - \mathbf{x}_1^\top \boldsymbol{\beta}]^{4m} < \infty$ and

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

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

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

$$\liminf_{n \rightarrow \infty} \inf_{Q_n \in \mathcal{Q}_{n,q}} \mathbb{P}\{\mathbf{e}_n^\top Q_n \mathbf{e}_n > c\sigma^2 q\} > 0, \quad (6)$$

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

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

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

Corollary 3.4 (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 ?? and Proposition ??. □

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

4 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

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

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

Definition 4.1

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

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

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

4.0.1 Single-split cross-validation (Hold-out CV)

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,

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

The results rely on the following conditions:

- **C0:** $\mathbb{E}[\epsilon_i^2 \mid \mathbf{x}_i]$ is bounded a.s. for $i \in [n]$.
- **C1:** There exists A_n such that $\left\| f - \hat{f}_{n,j} \right\|_\infty = O_{\mathbb{P}}(A_n)$ for $j = 1, 2$.
- **C2:** One procedure is asymptotically better than the other.

- **C3:** There exists M_n such that $\left\|f - \hat{f}_{n,j}\right\|_4 / \left\|f - \hat{f}_{n,j}\right\|_4 = o_{\mathbb{P}}(M_n)$ for $j = 1, 2$.

Theorem 4.1 (Yang, 2007)

Suppose that **C0–C3** hold. Suppose, furthermore, that

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

Then, the hold-out CV procedure is consistent.

4.0.2 Voting cross-validation with multiple splits

Theorem 4.2 (Yang, 2007)

Under the same conditions as in Theorem 4.1 and that the data is iid, the voting CV procedure with multiple splits is consistent.