Project: Cross-validation for model selection (rough draft)

Diego Urdapilleta de la Parra

March 5, 2025

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

Assumptions

We assume that the following are generally satisfied:

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

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

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

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

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

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

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

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

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 $\mathbb{E}\left[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}\right] = \sigma^2 p_n(\alpha)$ 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. ☐

2 Leave-One-Out CV [1]

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

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

Lemma 2.1 (Shao, 1993)

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

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,

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

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

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

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

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

so that $1/n\left\|H_\alpha \boldsymbol{e}\right\|^2 \to_{\mathbb{P}} 0$ 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,

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

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

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

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

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

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

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 2.3

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

3 Shao, 1997 [2]

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

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

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

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

Proposition 3.1 (Shao, 1997)

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

References

[1] J. Shao, "Linear Model Selection by Cross-validation," *Journal of the American Statistical Association*, vol. 88, pp. 486–494, June 1993.

[2] J. Shao, "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, vol. 7, pp. 221–264, Apr. 1997.