MATH 410 Project: The Asymptotics of Cross-Validation for Model Selection in the Regression Setting

Supervised by: Prof. Mehdi Dagdoug

Diego Urdapilleta de la Parra

McGill University

2025

Table of Contents

The problem of model selection

Setting: Multiple competing candidate models for a regression task.
Some considerations:

- Striking a balance: Simpler models offer efficiency and stability, but may underfit; complex models capture structure, but risk overfitting.
- Theoretical assumptions behind models are often unverifiable, motivating comparisons among alternatives.

This work explores the asymptotic behavior of cross-validation as a model selection critetion.

Outline

Setup and Notation

For positive integers n and p_n, let $(y, \boldsymbol{x}) : \Omega \to \mathbb{R} \times [0,1]^{p_n}$ be a real-valued random vector with distribution $\mu_{y,\boldsymbol{x}}$ such that

- $\mathbb{E}|y|^2 < \infty$
- $\mathbb{E}\|\boldsymbol{x}\|^2 < \infty$
- $\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] \succ 0$

A Borel-measurable function $f : [0,1]^{p_n} \to \mathbb{R}$ that satisfies

$$f(\boldsymbol{x}) \overset{\text{a.s.}}{=} \mathbb{E}\left[y \mid \boldsymbol{x}\right]$$

is called the *regression* function of y on $\boldsymbol{x}$.

Let $\mathcal{D}_n := \{(y_i, \boldsymbol{x}_i) : i \in [n]\}$ be a sample of independent data points drawn from $\mu_{y,\boldsymbol{x}}$. Define the residual $\epsilon_i := y_i - f(\boldsymbol{x}_i)$, which yields the decomposition

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

Cross-validation

Let J be an index set and let $\{E_j\}_{j \in J}$ be a family of subsets $E_j \subset [n]$ such that $|E_j| = n_1$ for all $j \in J$, and write $V_j := E_j^c$.

For each subset E_j, we consider the estimation sample $\mathcal{D}_n^{E_j}$ of n_1 data ponts given by

$$\mathcal{D}_n^{E_j} = \{(y_i, \boldsymbol{x}_i) \in \mathcal{D}_n : i \in E_j\}.$$

For each $j \in J$, we fit the model $\hat{f}$ on the estimation sample $\mathcal{D}_n^{E_j}$ and compute the *hold-out* loss against the remaining $n - n_1 =: n_2$ data points in $\mathcal{D}_n^{V_j}$:

$$\hat{R}_n^{E_j} := \frac{1}{n_2} \sum_{i \in V_j} \left(y_i - \hat{f}\left(\boldsymbol{x}_i; \mathcal{D}_n^{E_j} \right) \right)^2$$

Cross-validation

The cross-validation loss estimator is

$$\hat{R}_n^{CV} := \frac{1}{|J|} \sum_{j \in J} \hat{R}_n^{E_j}.$$

Different choices of J and estimation size n_1 yield different variants of cross-validation.
Examples:

- $n_1 = n - 1$ for the *leave-one-out* estimator,
- $|J| = k$ and $n_1 = n(k-1)/k$ for the *k-fold* estimator.

Table of Contents

Linear models: Setup

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 write by $p_n(\alpha) := |\alpha|$ and consider

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

1. We say $\alpha \in \mathcal{A}_n$ is *correct* if $\boldsymbol{X}\beta \overset{\text{a.s.}}{=} f_\alpha(\boldsymbol{X})$, 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

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

Linear models: Setup

Example

Suppose that we have the underlying model $\mathbb{E}[y \mid \boldsymbol{x}] \overset{\text{a.s.}}{=} \boldsymbol{X}\boldsymbol{\beta}$ given by

$$\boldsymbol{X} = \begin{bmatrix} | & | & | & | \\ \boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} & \boldsymbol{x}^{(3)} & \boldsymbol{x}^{(4)} \\ | & | & | & | \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} 2 \\ 5 \\ 3 \\ 0 \end{bmatrix}$$

We consider $\mathcal{A}_n = \{\alpha_1, \alpha_2, \alpha_3\}$ with

$$\alpha_1 = \{1, 2\}, \quad \alpha_2 = \{1, 2, 3\}, \quad \text{and} \quad \alpha_3 = \{1, 2, 3, 4\}$$

The model α_2 is correct:

$$\boldsymbol{X}_{\alpha_2}\boldsymbol{\beta}_{\alpha_2} = \begin{bmatrix} | & | & | \\ \boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} & \boldsymbol{x}^{(3)} \\ | & | & | \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta}$$

Linear models: Consistency and efficiency

Define the losses

$$L_n(\alpha) := \frac{1}{n}\|f(\boldsymbol{X}) - \hat{f}_\alpha(\boldsymbol{X})\|^2 \quad \text{and} \quad R_n(\alpha) := \mathbb{E}\left[L_n(\alpha) \mid \boldsymbol{X}\right]$$

for $\alpha \in \mathcal{A}_n$, with $\hat{f}_\alpha(\boldsymbol{X}) := \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha$.
Let $\hat{R}_n$ be a model selection criterion and let $\hat{\alpha}_n$ be the model selected by minimizing $\hat{R}_n$ over $\mathcal{A}_n$. Let α_n^* denote the model minimizing R_n over $\mathcal{A}_n$. We say that $\hat{R}_n$ 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 assymptotically loss efficient if

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

The leave-one-out

We define the leave-one-out loss estimator for a model $\alpha \in \mathcal{A}_n$ to be

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

Proposition

For $\alpha \in \mathcal{A}_n$, the leave-one-out estimator $\hat{R}_n^{(1)}(\alpha)$ satisfies the following equality:

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

where $h_{ii,\alpha} = \boldsymbol{x}_{i\alpha}^{\top}(\boldsymbol{X}_{\alpha}^{\top}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{x}_{i\alpha}$ denotes the ith leverage and $\hat{\boldsymbol{\beta}}_{\alpha}$ is the OLS estimator for model α fitted on the whole data set.

The leave-one-out

Proposition (Shao 1993)

Suppose that $\mathcal{T}_n$ is non-empty and let $\hat{\alpha}^{(1)}$ be the model minimizing $\hat{R}_n^{(1)}(\alpha)$.

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

The leave-one-out

Interpretation: The leave-one-out places too much weight on the estimation and too little on the evaluation.

- $R_{n_1}(\alpha) = \sigma^2 p(\alpha)/n_1$ for $\alpha \in \mathcal{T}_n$.
- The larger n_1, the closer $R_{n_1}(\alpha)$ is to a flat line.
- The leave-one-out, adopts the largest possible $n_1 = n - 1$, making it difficult for the estimator to distinguish between correct models.

Conjecture: A smaller estimation set and a larger validation set might improve the performance of cross-validation procedures for model selection.

A general perspective: the GIC

A large portion of selection criteria in the literature can be reduced to a general penalized criterion with a penalty of the type

$$\mathrm{pen}_{\lambda_n}(\alpha) = \frac{1}{n}\lambda_n \hat{\sigma}_n^2 p_n(\alpha),$$

for some some estimator $\hat{\sigma}_n^2$ of σ^2 and a sequence of real numbers $\{\lambda_n\}_{n\geq 1}$ satisfying $\lambda_n \geq 2$ and $\lambda_n/n \to 0$. This penalty yields the *Generalized Information Criterion* (Shao 1997):

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

We consider two cases: $\lambda_n \equiv 2$ and $\lambda_n \to \infty$

GIC: The case of $\lambda_n \equiv 2$

Theorem

Suppose that $\hat{\sigma}_n^2$ is consistent for σ^2. Under some regularity assumptions,

1. If $|\mathcal{T}_n| \leq 1$ for all but finitely many n, then $\hat{R}_{n,2}$ is 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

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

then $\hat{R}_{n,2}$ is asymptotically loss efficient.

GIC: The case of $\lambda_n \equiv 2$

Theorem (continued)

- Suppose that $|\mathcal{T}_n| > 1$ for all but finitely many n. If $|\mathcal{T}_n|$ is bounded, then the condition that

$$p_n(\alpha_n^*) \to \infty \quad \text{or} \quad \min_{\substack{\alpha \in \mathcal{T}_n, \\ \alpha \neq \alpha^*}} \left(p_n(\alpha) - p_n(\alpha^*)\right) \to \infty \tag{2}$$

is necessary and sufficient for the asymptotic loss efficiency of $\hat{R}_{n,2}$.

Takeaway: The GIC estimator with $\lambda_n \equiv 2$ is asymptotically loss efficient whenever there is at most one correct model with fixed dimension.

GIC: The case of $\lambda_n \to \infty$

Theorem (Shao 1997)

Suppose that

$$\limsup_{n\to\infty} \sum_{\alpha\in\mathcal{T}_n} \frac{1}{p_n(\alpha)^m} < \infty \tag{3}$$

for some $m \geq 1$ with $\mathbb{E}\left[e_i^{4m}\right] < \infty$. Under some regularity conditions,

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

Takeaway: The GIC with $\lambda_{n\to\infty}$ performs well when there exist fixed-dimension correct models.

Cross-validation and the GIC

Theorem (Shao 1997)

Under regularity conditions, the following hold.

1. The assertions in about the GIC with $\lambda_n \equiv 2$ apply for the leave-one-out cross-validation estimator $\hat{R}_n^{(1)}$.
2. If $d_n \leq n$ is chosen so that $d_n/n \to 1$ as $n \to \infty$, then the delete-d_n cross-validation estimator $\hat{R}_n^{(d_n)}$ has the same asymptotic behavior as the GIC with $\lambda \to \infty$. Specifically, if

$$\frac{p_n}{n - d_n} \to 0$$

and the splits are well "balanced," then $\hat{R}_n^{(d_n)}$ is consistent in selection whenever $\mathcal{A}_n$ contains at least one correct model with fixed dimension.

Cross-validation and the GIC

Letting $n_1 := n - d_n$ and $n_2 := d_n$, the conditions in 2. of the latter Theorem can be written as

$$\frac{n_2}{n} \to 1 \quad \text{and} \quad \frac{p_n}{n_1} \to 0.$$

If p_n is fixed for large enough n, we can equivalently write

$$\frac{n_2}{n_1} \to \infty \quad \text{and} \quad n_1 \to \infty. \tag{4}$$

This confirms our conjecture from before: a dominating validation size is necessary for cross-validation methods to be able to discriminate among correct models.

Table of Contents

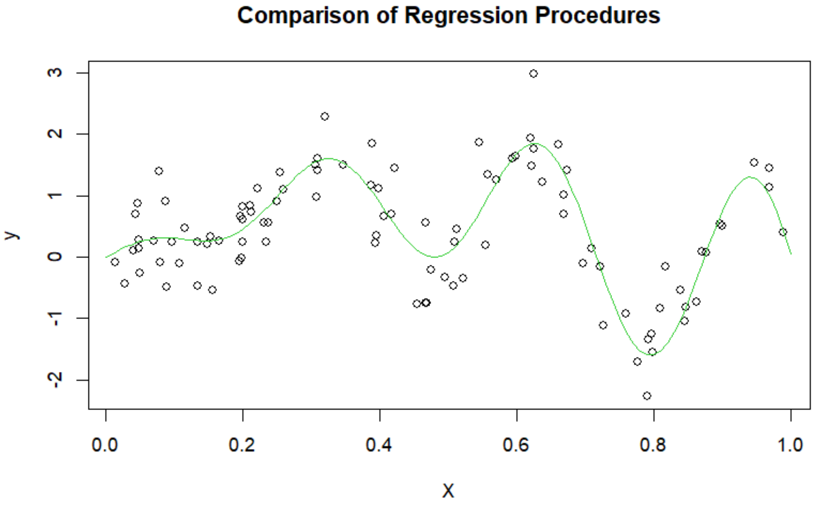
The nonparametric setting

- In many applications, the goal is accurate prediction, not a precise model of the data-generating process.
- The idea of a single "true" or "correct" model is less relevant.
- For many estimators, we can prove risk bounds of the form

$$\sup_{f \in \mathcal{F}} \mathbb{E}\|f - \hat{f}_n\|^2 \leq C\psi_n^2$$

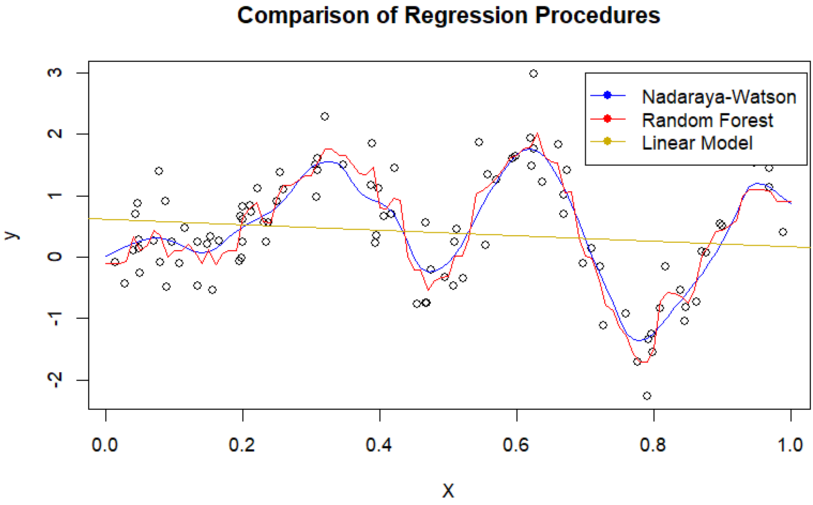
  for certain constants C, positive sequences $\psi_n \to 0$, and classes of functions $\mathcal{F}$.
- Inclusion/exclusion of relevant covariates remains important.

The nonparametric setting



True model

The nonparametric setting



Estimators

The nonparametric setting

We will consider the simplified scenario of selecting between 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 f.

Definition: Let L_n be a loss function. We say δ_1 is *asymptotically better* than δ_2 under L_n if, for $0 < \varepsilon < 1$, there exists $c_\varepsilon > 0$ such that

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

for all but finitely many n.

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 $n \to \infty$.

Nonparametric selection: The hold-out

Recall the hold-out loss estimator, defined by

$$\hat{R}_{\mathrm{ho}}(\hat{f}_{n,j}) = \sum_{i=n_1+1}^{n} \Big(y_i - \hat{f}_{n_1,j}(\boldsymbol{x}_i)\Big)^2 \quad \text{for } j = 1, 2. \tag{5}$$

We write $\hat{f}_n^{(\mathrm{ho})}$ to denote the estimator selected by $\hat{R}_{\mathrm{ho}}$. To show the consistency of $\hat{R}_{\mathrm{ho}}$, we establish two assumptions:

- We assume the existance of two positive sequences $\{A_n\}_{n\geq 1}$ and $\{M_n\}_{n\geq 1}$ such that

$$\|f - \hat{f}_{n,j}\|_\infty = O_{\mathbb{P}}\left(A_n\right) \quad \text{and} \quad \frac{\|f - \hat{f}_{n,j}\|_4}{\|f - \hat{f}_{n,j}\|_2} = o_{\mathbb{P}}\left(M_n\right) \tag{6}$$

- We will assume that one of δ_1 and δ_2 is asymptotically better than the other.

Nonparametric selection: The hold-out

Theorem (Yang 2007)

Suppose that the conditions established above hold. Suppose, furthermore, that

1. $n_1 \to \infty$ as $n \to \infty$
2. $n_2 \to \infty$ as $n \to \infty$
3. $n_2 M_{n_1}^{-4} \to \infty$ as $n \to \infty$
4. $\sqrt{n_2} \max(p_{n_1}, q_{n_1}) / \left(1 + A_{n_1}\right) \to \infty$ as $n \to \infty$

Then, the hold-out CV procedure is consistent.

Nonparametric selection: The hold-out

Example

Suppose that $\hat{f}_{n,1}$ and $\hat{f}_{n,2}$ are two nonparametric estimators with rates of convergence $p_n = O\left(n^{-4/9}\right)$ and $q_n = O\left(n^{-1/3}\right)$, respectively. Suppose that (6) is satisfied with $A_n = O(1)$ and $M_n = O(1)$. If we choose splits such that $n_1 \to \infty$ and $n_2 \to \infty$ as $n \to \infty$, then $n_2 M_{n_1}^{-4}$ is clearly satisfied and

$$\frac{\sqrt{n_2}\max(p_{n_1}, q_{n_1})}{1 + A_{n_1}} \geq \frac{n_2^{1/2}}{n_1^{1/3}} \to \infty$$

is satisfied if $n_1 = o\left(n_2^{3/2}\right)$. In other words, it is possible for the estimation size n_1 to be dominating.

Nonparametric selection: The hold-out

Example

On the other hand, if at least one of $\hat{f}_{n,1}$ and $\hat{f}_{n,2}$ has a parametric rate of convergence $O(n^{-1/2})$, then

$$\sqrt{n_2}\max(p_{n_1}, q_{n_1}) \geq \left(\frac{n_2}{n_1}\right)^{1/2} \to \infty$$

is satisfied whenever $n_2/n_1 \to \infty$. This agrees with the conclusion from Section 2, in which we showed that cross-validation is often consistent if the validation size dominates.

Nonparametric selection: Voting CV

We introduce the majority-vote cross-validation: 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,

$$\mathcal{D}_n^{E_1} = \left\{ \left(y_{\pi(1)}, \boldsymbol{x}_{\pi(1)}\right), \ldots, \left(y_{\pi(n_1)}, \boldsymbol{x}_{\pi(n_1)}\right) \right\},$$

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

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

Nonparametric selection: Voting CV

The chosen estimator is the one favored by the majority of the permutations. Let

$$\tau_\pi = \mathbb{1}_{\left[\hat{R}_\pi(\hat{f}_{n,1}) \leq \hat{R}_\pi(\hat{f}_{n,2})\right]}$$

The majority-vote estimator selection rule is 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 $[n]$.

Nonparametric selection: Voting CV

Theorem (Yang 2007)

Under the conditions of the previous Theorem 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

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

The equality at $(*)$ follows from the fact that the data are iid, hence exchangeable, and thus the τ_π are identically distributed. By the previous theorem, 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. □

Nonparametric selection: Voting CV

Remark 1: The proof does not require using the entire set Π of permutations for the majority vote.
Remark 2: These conditions are not merely sufficient but necessary $\implies$ The number of splits does not affect consistency. In other words, multiple splits in cross-validation cannot rescue an inconsistent single-split procedure.

Nonparametric selection: Key points

- Key distinction: In nonparametric settings, training set dominance is acceptable for consistency, unlike the parametric case.
- Cross-validation is effective for comparing estimators with different convergence rates.
- Both single-split and voting methods can yield consistent selection under suitable norm conditions.
- Leave-one-out CV is generally inadequate due to its small validation size.
- Averaging approaches may retain more data and are likely asymptotically equivalent to voting methods (Yang 2007).

Table of Contents

Conclusion: Cross-Validation for Model Selection

- **Linear Models:**
  - For consistent model selection, validation set size must dominate: $\frac{n_2}{n_1} \to \infty$ as $n \to \infty$
  - The leave-one-out CV is effective only when at most one correct model with fixed dimension exists.
  - The delete-d method performs well when fixed-dimension correct models exist.
- **Nonparametric Setting:**
  - Training set dominance can be acceptable for consistency.
  - Single-split and voting methods can both yield consistent selection under suitable norm conditions.
  - Leave-one-out CV remains inadequate due to minimal validation size
  - Multiple splits alone cannot rescue an inconsistent single-split procedure

Conclusion

Practical Implications:

- Though computationally costly, cross-validation may be worth using.
- The result on the majority-vote approach suggests that few splits can still yield consistency (i.e., k-fold remains useful).

Remain to be addressed:

- Split ratios.
- Number of splits.
- Voting versus averaging.

Questions?

- How do you use cross-validation?
- What behaviors have you observed in practice?