MATH 410 Report: The Asymptotics of Cross-Validation and Aggregation in the Regression Setting

Diego Urdapilleta de la Parra
Supervised by Prof. Mehdi Dagdoug

April 24, 2025

Contents

1 Introduction: The problem of model selection

Model selection lies at the heart of statistical learning and predictive modeling. Given a set of candidate models, each representing different assumptions and levels of complexity, the central challenge is to identify the one that balances interpretability, computational feasibility, and predictive accuracy. This task is rendered especially delicate by the fundamental tension between model simplicity and flexibility.

From a computational standpoint, simpler models are often preferable: they require less processing time, are easier to implement, and tend to be more stable numerically. In statistical terms, parsimony can translate into greater efficiency, as models with fewer parameters typically have a lower variance in their estimates. However, an overly simplistic model may fail to capture essential patterns in the data, resulting in high bias and poor predictive performance. Conversely, highly flexible models, while potentially better at capturing complex relationships, may obfuscate the true signal and are more prone to overfitting noise.

Each model or regression procedure typically rests on a set of theoretical assumptions that are rarely, if ever, fully verifiable in practice. This uncertainty makes it prudent to consider and compare multiple competing models rather than rely solely on a priori reasoning.

Cross-validation methods provide a data-driven framework for navigating these trade-offs. By estimating a model's predictive performance on unseen data, cross-validation enables an empirical basis for comparison, reducing the reliance on unverifiable assumptions and helping to guard against both under- and overfitting. As such, it plays a crucial role in modern approaches to model selection, where the goal is not just to fit the data well, but to generalize effectively to new observations.

Thus, as is generally the case in statistics, the need for model selection stems from an uncertainty or lack of knowledge about whatever system is being studied, and while selection procedures may give us improved confidence on our estimation endeavours, it certainly is not a perfect cure for our ignorance. Rather than committing to a single estimator, another way of coping with our lack of knowledge is to combine or *aggregate* our candidate models. This way, we may obtain a more robust and potentially better-performing estimator.

In this paper, we study two frameworks for model selection in the regression setting, with a focus on the asymptotic properties of cross-validation methods. We then investigate aggregation as an alternative approach, examining the asymptotic guarantees that it can provide us with. The goal of this project is to offer insight into how different model selection and aggregation strategies perform in asymptotic scenarios. We aim to provide a nuanced understanding of when and why certain methods may be preferable, thereby guiding more informed decisions in real-world predictive modeling tasks.

1.1 Setup and notation

Throughout this report, we consider the following regression setup. 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 $\mathbb{E}|y|^2 < \infty$, $\mathbb{E}\|\boldsymbol{x}\|^2 < \infty$, and $\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] \succ 0$. A Borel-measurable function $f : [0,1]^{p_n} \to \mathbb{R}$ that satisfies

$$f(\boldsymbol{x}) \overset{\text{a.s.}}{=} \mathbb{E}\left[y \mid \boldsymbol{x}\right] \tag{1}$$

is called the *regression* function of y on $\boldsymbol{x}$.[1]

[1] Technically, it would be more accurate to refer to *a version of* the regression function f, since it is only unique in the almost-sure sense. In this report, however, we adopt the common convension of treating f as a single, well-defined function.

We are treat y as a response and $\boldsymbol{x}$ as a covariate vector, and we would like to estimate the regression function f from sampled data. To this end, suppose that we have a sample $\mathcal{D}_n := \{(y_i, \boldsymbol{x}_i) : i \in [n]\}$ of independent data points drawn from the distribution $\mu_{y,\boldsymbol{x}}$. We define the residual $\epsilon_i := y_i - f(\boldsymbol{x}_i)$, which yields the decomposition

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

(here and throughout this paper, the notation $[n]$ denotes the set $\{1, \ldots, n\}$). In general, we will assume that the second moment $\mathbb{E}\left[\epsilon_i^2 \mid \boldsymbol{x}\right]$ is bounded almost surely. We may also abuse notation by writing

$$\boldsymbol{y}_n = f(\boldsymbol{X}_n) + \boldsymbol{\epsilon}_n,$$

where $\boldsymbol{y}_n$ is the vector of responses, $\boldsymbol{\epsilon}_n$ is the vector of residuals, $\boldsymbol{X}_n = [\boldsymbol{x}_1^\top \ \cdots \ \boldsymbol{x}_n^\top]^\top$ denotes the design matrix, and $f(\boldsymbol{X}_n) = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))$.

We will use the notation $\|\cdot\|$ to denote the Euclidean norm on $\mathbb{R}^n$ (and, in some contexts, other finite-dimensional vector spaces on $\mathbb{R}$). At the same time, for $q \geq 1$ and $g \in L_q([0,1], \mathcal{B}[0,1], \mu_{\boldsymbol{x}})$, we write

$$\|g\|_q := \begin{cases} \left(\int |g|^q \, d\mu_{\boldsymbol{x}}\right)^{1/q} & \text{if } 1 \leq q < \infty, \\ \sup\left\{c \in \mathbb{R} : \mu_{\boldsymbol{x}}\left\{|f| > c\right\} > 0\right\} & \text{if } q = \infty, \end{cases}$$

where $\mathcal{B}[0,1]$ denotes the Borel σ-field over $[0,1]$ and $\mu_{\boldsymbol{x}}$ denotes the marginal distribution of $\boldsymbol{x}$.

1.2 Cross-validation

Cross-validation refers to a family of methods for loss/risk estimation. In its most general form, cross-validation proceeds as follows.

Suppose that we wish to estimate the loss L_n of a model $\hat{f}$ against f. 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 $V_j := E_j^c$. The integer n_1, chosen beforehand, is called the *estimation size*, and it denotes the number of data points used for fitting the model $\hat{f}$ on the training phase. For each subset E_j, we consider the estimation sample $\mathcal{D}_n^{E_j}$ of n_1 data ponts given by

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

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 $\mathcal{D}_n^{V_j}$:

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

Finally, we combine the hold-out estimates into an overall cross-validation loss estimate

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

Many popular variants of this procedure exist, and these are often characterized by the choices of J (e.g., all possible subsets versus a fixed number) and estimation size n_1 (e.g., $n_1 = n - 1$ for the *leave-one-out* method, $n_1 = n - d$ for the *delete-d* method, or $n_1 = n(k-1)/k$ for the *k-fold* method). Sometimes, as we will see in Section 3, alternative approaches for combining the hold-outs are carried out.

We will begin our exploration of the properties of some of these methods with a treatment of linear models. Then, we will move on to a more general framework where no such restrictions are imposed on f, and we will treat problems of selection on a much broader class of estimators.

2 Variable Selection for Linear Models

2.1 Setup

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 $\boldsymbol{X} = [\boldsymbol{x}_1^\top \ \boldsymbol{x}_2^\top \ \cdots \ \boldsymbol{x}_n^\top]^\top \in \mathbb{R}^{n \times p_n}$ 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 procedure 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 $\boldsymbol{\beta}$. To perform the regression task on these models, we utilize the *Ordinary Least Squares* (OLS) estimator $\hat{\boldsymbol{\beta}}_\alpha$ of $\boldsymbol{\beta}_\alpha$, which is given by

$$\hat{\boldsymbol{\beta}}_\alpha = \left(\boldsymbol{X}_\alpha^\top \boldsymbol{X}_\alpha\right) \boldsymbol{X}_\alpha^\top \boldsymbol{y}.$$

We will use the variable selection framework outlined by the following definitions:

1. We say $\alpha \in \mathcal{A}_n$ is *correct* if $\mathbb{E}\left[\boldsymbol{y} \mid \boldsymbol{X}\right] \overset{\text{a.s.}}{=} f_\alpha(\boldsymbol{X})$, 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
$$\alpha_1 \subset \alpha_2 \subset \cdots \subset \alpha_k.$$

Throughout this section, we will also use the notations $H_\alpha = \boldsymbol{X}_\alpha \left(\boldsymbol{X}_\alpha^\top \boldsymbol{X}_\alpha\right) \boldsymbol{X}_\alpha^\top$ for the hat matrix, $M_\alpha := I_n - H_\alpha$ for the anihilator matrix, and $h_{ii,\alpha} := \boldsymbol{x}_{i\alpha}^\top (\boldsymbol{X}_\alpha^\top \boldsymbol{X}_\alpha)^{-1} \boldsymbol{x}_{i\alpha}$ for the ith leverage corresponding to $\alpha \in \mathcal{A}_n$.

In order to select one model over another, we need some way to measure and compare the quality of their predictive ability. A natural approach is to consider the distance between their estimated mean outcomes and the true mean structure given by f. We define th

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 useful throughout our treatment of linear models:

$$\textbf{H1}: \quad \liminf_{n\to\infty} \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 > 0 \text{ almost surely for all } \alpha \in \mathcal{T}_n^c.$$

$$\textbf{H2}: \quad \mathbb{E}\left[\|\boldsymbol{x}\|^2\right] < \infty.$$

$$\textbf{H3}: \quad \lim_{n\to\infty} \max_{i\leq n} h_{ii,\alpha} \overset{\text{a.s.}}{=} 0 \text{ for all } \alpha \in \mathcal{A}_n.$$

$$\textbf{H4}: \quad \sum_{\alpha\in\mathcal{T}_n^c} \frac{1}{\left(nR_n(\alpha)\right)^m} \to_{\mathbb{P}} 0 \text{ and } \mathbb{E}\left[e_i^{4m} \mid \boldsymbol{x}_i\right] < \infty \quad \text{for some } m \geq 1.$$

Condition **H1** establishes a minimal difference in predictive ability between correct and wrong models. It ensures that the signal be strong enough that the distinction between correct and wrong matters. Indeed, if there is no meaningful difference in performance between the two sets, we need not to worry about selection. **H2** and **H4** are moment conditions that will allow us to derive convergence and apply the law of large numbers to derive asymptotic results. Finally, condition **H3** ensures that there are no overly-influential data points in the models. It will be utilized, for instance, to prove some asymptotic properties of the Leave-one-out loss estimator.

We now state two preliminary results the justify are choice of L_n and R_n for selection purposes.

Proposition 2.1. *Assuming the setup of Definition 2.1, the following equalities hold almost surely:*

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

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

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

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

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

Hence, the first part follows from the Pythagorean theorem.
For the second part, we note that

$$\mathbb{E}\left[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}\right] \overset{\text{a.s.}}{=} \operatorname{tr}\left(\mathbb{E}\left[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}\right]\right) \overset{\text{a.s.}}{=} \mathbb{E}\left[\operatorname{tr}\left(\|H_\alpha \boldsymbol{e}\|^2\right) \mid \boldsymbol{X}\right],$$

so that

$$\begin{aligned}
\mathbb{E}\left[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}\right] &= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{e}^\top H_\alpha \boldsymbol{e}\right) \mid \boldsymbol{X}\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{e}\boldsymbol{e}^\top H_\alpha\right) \mid \boldsymbol{X}\right] \\
&= \operatorname{tr}\left(\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^\top \mid \boldsymbol{X}\right] H_\alpha\right) \\
&= \sigma^2 \operatorname{tr}\left(H_\alpha\right) \\
&= \sigma^2 p_n(\alpha),
\end{aligned}$$

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 $R_n(\alpha)$ over $\alpha \in \mathcal{A}_n$ with probability 1.*

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 $p_n(\alpha_n^*) \leq p_n(\alpha)$. Thus,

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

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 Hansen (2022)). 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 an effective selection criterion. Unfortunately, R_n is an unkown expectation that depends on the regression function f, and therefore cannot be used in practice. Instead, we may try to approximate it through some other empirically feasible criterion, $\hat{R}_n$, which we treat as an "estimator" of the random quantity R_n.

Definition 2.2. *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

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

Consistency as defined above is a naturally desirable property for any selection criterion $\hat{R}_n$: it ensures that, asymptotically, $\hat{R}_n$ will select the same model as the optimal risk R_n. Asymptotic loss efficiency is a weaker property that captures a certain degree of "closeness" between $\hat{\alpha}_n$ and α_n^*. All consistent criteria are asymptotically loss efficient:

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,

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

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

The converse of Lemma 2.3 is not necessarily true. However, since asymptotic loss efficiency is easier to prove than consistency, it is in our interest to find conditions for their equivalency. Proposition 2.4 illustrates two such cases.

Proposition 2.4 (Shao 1997)**.** *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 $|\mathcal{T}_n| \geq 2$ and $p_n(\alpha_n^*) \overset{\mathbb{P}}{\not\to} \infty$, 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, asymptotic 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

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

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:

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

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

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

□

2.2 A result on the leave-one-out

Having established some essential results in the previous section, we now turn our attention to a particular variant of cross-validation, namely the leave-one-out. For this section, we will only

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.

The leave-one-out cross-validation method (otherwise known as delete-1 CV) consists in estimating a model using the data with one point removed, and then evaluating our estimate on the single removed point. This procedure is repeated for each data point and the results are averaged out onto a single loss estimate $\hat{R}_n^{(1)}(\alpha)$ of $R_n(\alpha)$. Formally, we define the leave-one-out loss estimator for a model $\alpha \in \mathcal{A}$ to be

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

The leave-one-out is a very popular tool for linear models due to its low computational cost. The following proposition shows that it is not necessary to fit all n coefficient vectors $\hat{\boldsymbol{\beta}}_\alpha^{(i)}$: it suffices to fit the model on the complete dataset just once. The proof uses the Sherman-Morrison inversion formula and will not be presented here.

Proposition 2.5. *For alpha $\in \mathcal{A}$, 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.

Our goal in this section is to show some asymptotic properties of the leave-one-out as a criterion for model selection. The main result, 2.7 below, uses the following decomposition of $\hat{R}_n^{(1)}(\alpha)$ in terms of R_n and other familiar quanitities.

Lemma 2.6 (Shao 1993)**.**

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

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

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

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,

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

proving the second case. □

Proposition 2.7 (Shao 1993)**.** *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 $p(\alpha^*) < p$,*
$$\lim_{n\to\infty} \mathbb{P}\left\{\hat{\alpha}^{(1)} = \alpha^*\right\} \neq 1.$$

Before continuing with the proof, a few comments are in order. The first part of the proposition shows that, under the given conditions, the leave-one-out will select a correct model with probability approaching 1 as n approaches infinity. This is a relatively strong result, as it implies that the leave-one-out is reliable at excluding wrong models. However, parts 2 and 3 tell us that the leave-one-out is not consistent for selection in the sense of Definition 2.2: it selects overly complex correct models with non-vanishing probability. In other words, the leave-one-out is prone to overfitting.

An interpretation on what causes this behavior, as noted by Shao (1993), is that the leave-one-out places too much weight on the estimation and too little on the evaluation. If we let n_1 denote the size of the fitting sets and $n_2 := 1 - n_1$ the size of the validation sets, from 2.6 we have that $R_{n_1}(\alpha) = \sigma^2 p(\alpha)/n_1$ for any correct model $\alpha \in \mathcal{T}$. Clearly, optimizing $R_{n_1}(\alpha)$ over $\mathcal{T}$ becomes difficult for large n_1: the larger n_1, the closer $R_{n_1}(\alpha)$ is to a flat line. With the leave-one-out, we are choosing the largest possible $n_1 = n - 1$, making it difficult for the estimator to distinguish between correct models.

Proof of Proposition 2.7:

1. Let $\bar{\alpha} \in \mathcal{T}$ and $\tilde{\alpha} \in \mathcal{T}^c$. By, Lemma 2.6, we have that
$$\begin{aligned}
&\mathbb{P}\left\{\hat{R}_n^{(1)}(\tilde{\alpha}) \leq \hat{R}_n^{(1)}(\bar{\alpha})\right\} \\
&\quad = \mathbb{P}\Big\{\frac{1}{n}\sigma^2 p(\tilde{\alpha}) + \frac{1}{n}\|M_{\tilde{\alpha}}\boldsymbol{X}\boldsymbol{\beta}\|^2 + \sigma^2 + o_{\mathbb{P}}(1) \leq \frac{1}{n}\|M_{\bar{\alpha}}\boldsymbol{e}\|^2 + \frac{1}{n}\sigma^2 p(\bar{\alpha}) + o_{\mathbb{P}}(n^{-1})\Big\} \\
&\quad = \mathbb{P}\left\{\frac{1}{n}\sigma^2\left(p(\tilde{\alpha}) - p(\bar{\alpha})\right) + \sigma^2 + \frac{1}{n}\|M_{\tilde{\alpha}}\boldsymbol{X}\boldsymbol{\beta}\|^2 - \frac{1}{n}\|M_{\bar{\alpha}}\boldsymbol{e}\|^2 \leq o_{\mathbb{P}}(1)\right\}.
\end{aligned}$$
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
$$\mathbb{P}\left\{\hat{\alpha}^{(1)} \in \mathcal{T}^c\right\} = \mathbb{E}\left[\mathbb{1}_{\left[\left[\hat{\alpha}^{(1)} \in \mathcal{T}^c\right]\right]}\right] = \mathbb{E}\left[\sum_{\tilde{\alpha}\in\mathcal{T}^c}\prod_{\alpha\in\mathcal{A}} \mathbb{1}_{\left[\hat{R}_n^{(1)}(\tilde{\alpha}) \leq \hat{R}_n^{(1)}(\alpha)\right]}\right] \to 0.$$
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. □

2.3 A General Perspective [2]

Our previous observations suggest that a smaller estimation set and a larger validation set might improve the performance of cross-validation procedures for model selection. In this section, we explore this idea from a different perspective, understanding cross-validation as a special case of a more general approach to loss estimation.

In the context of selection, overfitting occurs because our choice of empirical loss may underestimate the true loss for large models. The inclusion of irrelevant information can make the estimator $\hat{R}_n$ "hallucinate" a signal that is not there. To avoid this problem, we may use *penalization*: we modify our estimator $\hat{R}_n$ to favour the choice of less complex models. Given some loss estimator $\hat{R}_n$, we may define the corresponding *penalized selection criterion* as

$$\hat{R}'_n(\alpha) = \hat{R}_n(\alpha) + \operatorname{pen}(\alpha)$$

for some penalty function $\operatorname{pen} : \mathcal{A}_n \to \mathbb{R}$.

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

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

for some some estimator $\hat{\sigma}_n^2$ of σ^2 and a sequence of real numbers $\{\lambda_n\}_{n\geq 1}$ satisfying certain conditions. This type of penalty yields the *Generalized Information Criterion* (Shao 1997), defined below.

Definition 2.3. *We define the Generalized Informaton Criterion (GIC) loss estimator to be*

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

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 $\lambda_n/n \to 0$.

In what follows, we study some asymptotic properties of the GIC in two cases of interest: a constant $\lambda_n \equiv 2$ and a growing $\lambda_n \to \infty$.

[2] For this section, we allow the set of candidates $\mathcal{A}$ to vary with n, though we assume that it remains finite. To illustrate why this might be useful, we consider an example drawn from Shao (1997):

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 $\mathcal{A} := \{\alpha_d : d \in [p_n]\}$, 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. Furthermore, the dimension of the optimal model α_n^*, for instance, may also increase with n.

2.3.1 The case of $\lambda_n \equiv 2$

Theorem 2.8 (Shao 1997). *Suppose that **H4** holds and that $\hat{\sigma}_n^2$ is consistent for σ^2. Let*

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*

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

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

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

$$p_n(\alpha_n^*) \to \infty \quad or \quad \min_{\substack{\alpha \in \mathcal{T}_n, \\ \alpha \neq \alpha^*}} (p_n(\alpha) - p_n(\alpha^*)) \to \infty \tag{8}$$

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

In short, Theorem 2.8 guarantees the asymptotic loss efficiency of $\hat{R}_{n,2}$ in some cases where there exists at most one correct model with fixed dimension. In particular, the first part shows that $\hat{R}_{n,2}$ asymptotically chooses the best-performing wrong model in the case that no correct model exists. If only one correct model exists in $\mathcal{T}_n$ and it has fixed dimension for all but finitely many n, an application of Proposition 2.4 in section 2.1 yields consistency in selection. In other words, $\hat{R}_{n,2}$ is able to identify the correct model among a set of incorrect models with probability tending to 1.

Parts 2. and 3. specify conditions for asymptotic loss efficiency in when multiple correct model exist. In these cases, $\hat{R}_{n,2}$ performs well if no two correct models have fixed dimension. This is what conditions (7) and (8) imply: they require that the dimensions of the models in $\mathcal{T}_n$ diverge, either absolutely or relative to the smallest correct model α_n^*. The following example illustrates this phenomenon:

Example 2.1: Suppose that $\mathcal{A}_n = \mathcal{T}_n = \{\alpha_{1n}, \alpha_{2n}\}$ with $\alpha_{1n} \subset \alpha_{2n}$. Note that

$$\|\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha\|^2 = \|\boldsymbol{e}\|^2 - \|H_\alpha \boldsymbol{e}\|^2.$$

Then, form Definition 2.3, $\hat{\alpha}_n = \alpha_{1n}$ if and only if

$$\|\left(H_{\alpha_{2n}} - H_{\alpha_{1n}}\right)\boldsymbol{e}\|^2 < 2\hat{\sigma}_n^2 \left(p_{2n} - p_{1n}\right), \tag{9}$$

where p_{1n} and p_{2n} denote the dimensions of α_{1n} and α_{2n}, respectively. If $p_{1n} \to \infty$ and $p_{2n} - p_{1n} \to \infty$, then

$$\frac{\|\left(H_{\alpha_{2n}} - H_{\alpha_{1n}}\right)\boldsymbol{e}\|^2}{p_{2n} - p_{1n}} \xrightarrow{\mathbb{P}} \sigma^2,$$

so that (9) is satisfied with probability approaching 1 (by consistency of $\hat{\sigma}_n^2$), and $\hat{R}_{n,2}(\alpha)$ is consistent. If $p_{1n} \to \infty$ but $p_{2n} - p_{1n} \leq c$ for some $c > 0$, then $p_{2n}/p_{1n} \to 1$, which yields $L_n(\alpha_{2n})/L_n(\alpha_{1n}) \to 1$. On the other hand, if α_{1n} and $p_{2n} - p_{1n} \not\to \infty$, then

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_{1n})} = \mathbb{1}_{[\hat{\alpha}_n = \alpha_{1n}]} + \frac{\|H_{\alpha_{2n}}\boldsymbol{e}\|^2}{\|H_{\alpha_{1n}}\boldsymbol{e}\|^2} \mathbb{1}_{[\hat{\alpha}_n = \alpha_{1n}]}$$

(see Prop. 2.1). By Proposition 2.2, $\|H_{\alpha_{2n}}\boldsymbol{e}\|^2/\|H_{\alpha_{1n}}\boldsymbol{e}\|^2 \geq 1$ almost surely. Hence,

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_{1n})} = 1 + \left(\frac{\|H_{\alpha_{2n}}\boldsymbol{e}\|^2}{\|H_{\alpha_{1n}}\boldsymbol{e}\|^2} - 1\right)\mathbb{1}_{[\hat{\alpha}_n=\alpha_{1n}]} = 1 + \frac{\|(H_{\alpha_{2n}} - H_{\alpha_{1n}})\boldsymbol{e}\|^2}{\|H_{\alpha_{1n}}\boldsymbol{e}\|^2} \overset{\text{a.s.}}{>} 1.$$

Thus, $\hat{R}_{n,2}(\alpha)$ is not asymptotically loss efficient.

Notice that these results resemble those presented in section 2.2 about the leave-one-out cross-validation estimator. Indeed, the correspondance between the leave-one-out and the GIC with $\lambda_n \equiv 2$ will be discussed in section 2.3.3.

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

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

Theorem 2.9 (Shao 1997)**.** *Suppose that **H1**-**H4** hold and that*

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

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

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.*

As in the case of Theorem 2.8, condition (10) restricts the asymptotic behavior of the size of correct models in $\mathcal{T}_n$. In contrast to the setup of of Theorem 2.8, however, here we want correct models that eventually stop growing. In other words, Theorem 2.9 concerns itself with the case were we have correct models with fixed dimension.

It can be shown that condition (10) is satisfied whenever $|\mathcal{T}_n|$ is bounded or $\mathcal{A}_n$ is embedded. It also implies that

$$\max_{\alpha\in\mathcal{T}_n} \frac{\|H_\alpha\boldsymbol{e}\|^2}{\lambda_n\hat{\sigma}_n^2 p_n(\alpha)} \xrightarrow{\mathbb{P}} 0. \tag{11}$$

The proof of Theorem 2.9 relies on the following decomposition, which we state without proof.

Lemma 2.10 (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,*

$$\hat{R}_{n,\lambda_n}(\alpha) = \begin{cases} \frac{1}{n}\|\boldsymbol{e}\|^2 + \frac{1}{n}\lambda_n p_n(\alpha)\hat{\sigma}_n^2 - \frac{1}{n}\|H_\alpha\boldsymbol{e}\|^2 & \text{if } \alpha \in \mathcal{T}_n \\ \frac{1}{n}\|\boldsymbol{e}\|^2 + L_n(\alpha) + \frac{1}{n}p_n(\lambda_n\hat{\sigma}_n^2 - 2\sigma^2) + o_{\mathbb{P}}(L_n(\alpha)) & \text{if } \alpha \in \mathcal{T}_n^c \end{cases} \tag{12}$$

Proof of Theorem 2.9: From (11), we have that

$$\frac{\|H_\alpha\boldsymbol{e}\|^2}{\lambda_n p_n(\alpha)} = O_{\mathbb{P}}\left(\lambda_n^{-1}\right). \tag{13}$$

By an argument similar to the one in Example 2.1, it follows that

$$\mathbb{P}\left\{\hat{\alpha}^n \in \mathcal{T}_n \text{ but } \hat{\alpha}^n \neq \alpha_n^*\right\} \to 0$$

as $n \to \infty$.
For part 1., since $\lambda_n p_n / n \to 0$ and from (12),

$$\hat{R}_{n,\lambda_n}(\alpha) = \frac{1}{n}\|e\|^2 + L_n(\alpha) + o_{\mathbb{P}}\left(L_n(\alpha)\right)$$

for $\alpha \in \mathcal{T}_n^c$. Additionally, if $\mathcal{T}_n$ is nonempty,

$$\hat{R}_{n,\lambda_n}(\alpha) = o_{\mathbb{P}}\left(L_n(\alpha)\right)$$

for $\alpha \in \mathcal{T}_n$. We then have that the minimizer $\hat{\alpha}^n$ or $\hat{R}_{n,\lambda_n}(\alpha)$ satisfies $L_n(\hat{\alpha}^n)/L_n(\alpha_n^*) \to 1$, with α_n^* being the minimizer of the loss L_n.
For the second part, consistency follows immediately from (13), **H1**, and Proposition 2.2 (that is, that the smallest correct model is the model minimizing L_n). □

2.3.3 Cross-validation and the GIC

So far in this section, we have studied the properties of two general approaches to loss estimation with distinct properties, namely the GIC estimators with $\lambda_n \equiv 2$ and $\lambda_n \to \infty$. We will now connect these criteria with cross-validation methods, with the goal of "inheriting" some of former's properties to the latter.

Theorem 2.11 (Shao 1997)**.** *Suppose that **H1** to **H4** hold.*

1. *The assertions in Theorem 2.8 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.

The conditions on d_n are important. Returning to the notation from the introduction, we can write $n_1 := n - d_n$ for the size of estimation samples and $n_2 := d_n$ for the size of validation samples. The conditions in Theorem 2.11 2. 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

$$\frac{n_2}{n_1} \to \infty \quad \text{and} \quad n_1 \to \infty. \tag{14}$$

This confirms our conjecture from section 2.2, where we argued that a larger validation size was necessary for cross-validation methods to be able to discriminate among correct models. Indeed, (14) shows that the validation size must be the dominating.

This is a major result in our investigation that can guide our intuition and inform choice of estimators. From a practical viewpoint, however, it might seem discouraging. One of the great appeals of the leave-one-out is its accessibility and exceptionally low computational cost. The delete-d approach, on the other hand, is less compelling: there is no general shortcut as there is for the leave-one-out, and, with a dominating validation size, the number of necessary model fittings can grow with n at an absurdly large rate. Fortunately, a result in the following section suggests that, as far as consistency goes, this might not always be the case.

3 Selection of Nonparametric Procedures

3.1 Setup

In statistical learning, we are oftentimes more interested in finding a practical means for prediction rather than a precise description of the data-generating mechanism. In these cases, where the goal is optimal predictive performance, the notion of a "true" or "correct" model becomes less meaningful. Indeed, for many kinds of nonparametric estimators, it is possible to establish inequalities 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 $\mathcal{F}$ (see Tsybakov 2009). These inequalities imply that the risk is guaranteed to approach 0 as n increases, and therefore, unlike in linear models, model specification is less of a concern (although the inclusion/exclusion of relevant covariates remains important).

The problem of selecting a nonparametric estimator of the regression function, then, focuses more on the rates of convergence ψ_n of each procedure. In this section, 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 3.1. *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 $n \to \infty$.

In what follows, we will study the asymptotic performance of two variations on the cross-validation method for selecting an estimator. First, we will show that cross-validation with a single data split, also known as the *hold-out*, is consistent in selection under some conditions on how the split is performed. Afterwards, we will show that this result extends to the multiple-split case with majority vote.

3.2 Single-split cross-validation (the *hold-out*)

For this section, we analyze a single-split approach where 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

assume that the estimators $\hat{f}_{n,1}$ and $\hat{f}_{n,2}$ converge exactly at rates $\{p_n\}_{n\geq 1}$ and $\{q_n\}_{n\geq 1}$ under the $L_2[0,1]$ loss, respectively. In other words, we assume that

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

for all $\varepsilon \in (0,1)$ and for some $c_\varepsilon > 0$, and similarly for $\hat{f}_{n,2}$ and q_n.

The hold-out loss estimator is defined by

$$\hat{L}_{\text{ho}}(\hat{f}_{n,j}) = \sum_{i=n_1+1}^{n} \left(y_i - \hat{f}_{n_1,j}(\boldsymbol{x}_i)\right)^2 \quad \text{for } j = 1,2. \tag{15}$$

The selection procedure consists in selecting the learning method δ_j whose estimator $\hat{f}_{n,j}$ minimizes $\hat{L}_{\text{ho}}$ for $j \in \{1,2\}$. We write $\hat{f}_n^{(\text{ho})}$ to denote the estimator selected by $\hat{L}_{\text{ho}}$.

To show the consistency of $\hat{L}_{\text{ho}}$, we will establish a few assumptions. First, we will assume the existance of two positive sequences $\{A_n\}_{n\geq 1}$ and $\{M_n\}_{n\geq 1}$ such that

$$\|f - \hat{f}_{n,j}\|_\infty = O_{\mathbb{P}}(A_n) \quad \text{and} \quad \frac{\|f - \hat{f}_{n,j}\|_4}{\|f - \hat{f}_{n,j}\|_2} = o_{\mathbb{P}}(M_n) \tag{16}$$

for $j = 1,2$. These rates of convergence will be used in the proof of Theorem 3.1 below. Note that the supnorm contition in (16) is satisfied if the regression function f and the estimators $\hat{f}_{n,j}$ are bounded.

In addition to the above, we will assume that one of δ_1 and δ_2 is asymptotically better than the other. Indeed, if the latter is not satisfied, choosing one over the other may be irrelevant.

Theorem 3.1 (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. $\sqrt{n_2}\max(p_{n_1}, q_{n_1})/(1 + A_{n_1}) \to \infty$ *as* $n \to \infty$

Then, the hold-out CV procedure is consistent.

A very detailed proof of this result is given in Yang (2007), so it will be skipped here. We illustrate the conditions of Theorem 3.1 via an ideal-scenario example.

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 (16) 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.
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.

3.3 Voting cross-validation with multiple splits

The majority-vote cross-validation method proceeds as follows:[3] 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

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

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

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

We then define our selection criterion as follows:

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

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

Theorem 3.2 (Yang 2007)**.** *Under the conditions of Theorem 3.1 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

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

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

[3]The procedure desctibed here is theoretical in nature. In practice, we would not compute the hold-outs for all permutations of the data.

the splitting conditions are met. Moreover, Yang (2007) 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.

3.4 Some remarks

The theoretical results presented in this chapter highlight important distinctions between cross-validation for parametric and nonparametric procedures. Unlike Shao's results for parametric model selection, where the validation set must asymptotically dominate the training set to achieve consistency, our analysis shows that for nonparametric procedures, the training set can be dominant provided certain convergence conditions are satisfied. This contrast emphasizes that the optimal splitting strategy depends critically on the nature of the procedures being compared.

Cross-validation proves particularly useful for selection when comparing estimators with different convergence rates, as demonstrated in our theorems. The consistency of both single-split and majority-vote methods suggests that cross-validation can reliably identify asymptotically superior procedures under appropriate conditions on the L_2, L_4, and L_∞ norms of the estimation error. Not unlike the results in Shao (1993), however, the results in this section also indicate that leave-one-out cross-validation, where $n_1 = n - 1$ and $n_2 = 1$, is generally inadequate for consistent selection since the validation size condition $n_2 \to \infty$ is violated.

While this section focused the voting approach, it is worth noting that the more common averaging approach (described, for example, in 1.2) should, in theory, retain more information from the data. From a consistency perspective, Yang (2007) hypothesizes that both approaches are likely to be equivalent: if a majority of the permutations favour $\hat{f}_{n,1}$, say, with high probability, then it is also likely that the average $(1/n!)\sum_{\pi\in\Pi}\hat{L}_\pi(\hat{f}_{n,1})$ will be smaller than that of $\hat{f}_{n,2}$.

Several practical considerations emerge from this analysis, including the optimal choice of splitting ratios and the relative merits of k-fold procedures versus single splits. While our theoretical framework provides guidance on asymptotic properties, the finite-sample performance of these methods under various conditions remains an important question.

4 Aggregation

As is generally the case in statistics, the need for model selection stems from an uncertainty or lack of knowledge about the system being studied. And while selection procedures may give us improved confidence on our estimation endeavours, it certainly will not cure our ignorance. For instance, without any means of verifying underlying hypotheses (say, linearity of f for linear models, or its membership to a Soblovev class for projection estimators), we have no guarantee that a single selected procedure will reliably yield good predictions. Moreover, selection may not even be helpful if the models being considered are hard to distinguish or if no clear winner exists among them.

As stated in the Introduction, a good alternative to selection is *aggregation*: the combining of multiple candidate models into a single estimator. Aggregation of multiple estimators consists in combining them via a weighted average. In fact, as will be shown below, model selection can be seen a special case of this broader framework, where the computed "weighted average" has weights equal to 0 for all but one candidate model.

This section explores some perspectives through which aggregation schemes can be analyzed and presents asymptotic results on a particular aggregate estimator drawn from Bunea et al. (2007).

4.1 Setup

As before, we consider independent pairs in $\mathcal{D}_n := \{(y_i, \boldsymbol{x}_i) : i \in [n]\}$ satisfying the conditions established in Section 1.1. Suppose that we have M candidate estimators of the regression function, denoted $\hat{f}_{n,1}, \hat{f}_{n,2}, \ldots, \hat{f}_{n,M}$. Given a set $\Lambda \subset \mathbb{R}^M$ of admissible weights, we combine the estimators 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 $\hat{\lambda} := \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_M\right) \in \Lambda$ chosen to satisfy some optimality criteria.

We establish the following assumptions for this section.

H4.1: The residuals ϵ_i have a normal distribution $\mathcal{N}(0, \sigma^2)$ for $\sigma^2 < \infty$.

H4.2: The regression funcion f and the base estimators $\hat{f}_{n,1}, \hat{f}_{n,2}, \ldots, \hat{f}_{n,M}$ belong to the space $L_\infty([0,1], \mu_{\boldsymbol{x}})$.

Both assumptions grant well-behaved estimators and allow for the application of certain minimax results relating, in particular, to the Kullback-Liebler divergence between the distributions of estimators in the class $\mathcal{F}_0$.

4.1.1 Four types of aggregation

There are four main aggregation schemes that are treated in the literature. Each scheme is characterized by a different set Λ of admissible weights, and they are all suitable for distinct use cases. The schemes are:

- Model Selection Aggregation (MS): Only one estimator is selected among the candidates. We take Λ_{MS} to be the standard basis on $\mathbb{R}^M$.
- Linear Aggregation (L): The aggregate $\tilde{f}_{\hat{\lambda}}$ is chosen among all linear combinations of the estimators. We let $\Lambda_{\mathrm{L}} = \mathbb{R}^M$.
- Convex Aggregation (C): The aggregate $\tilde{f}_{\hat{\lambda}}$ 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): Given a positive integer $D \leq M$, at most D estimators from the candidates are selected and combined. That is,

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

  To denote the number of nonzero entries in a vector λ, we may use the notations $M(\lambda)$ and $\|\lambda\|_0$ interchangeably.

Notably, mentioned before, model selection can be viewed from this framework as a particular type of aggregation. Each of this schemes may perform differently on different tasks, and

4.1.2 Evaluating the aggregate

In an ideal scenario, we would like to select a weights vector λ^* that minimizes the largest possible expected error on a class of functions $\mathcal{F}$ containing f. That is, we would like to find λ^* satisfying

$$\sup_{f\in\Theta}\mathbb{E}\|f-\tilde{f}_{\lambda^*}\|_2^2=\inf_{\lambda\in\Lambda}\sup_{f\in\Theta}\mathbb{E}\|f-\tilde{f}_{\lambda}\|_2^2. \tag{17}$$

This is known as *minimax* extimation. However, there is no obvious way to compute the expectation $\mathbb{E}\|f-\tilde{f}_{\hat{\lambda}}\|_2^2$ for an arbitrary $f\in\Theta$, and so the minimax approach is not feasible in practice. The next best thing, then, would be to obtain weights $\hat{\lambda}$ that perform at least as good as λ^* in (17) plus some small, ideally vanishing remainder $\Delta_{n,M}$. That is, we seek $\hat{\lambda}$ such that

$$\mathbb{E}\|f-\tilde{f}_{\hat{\lambda}}\|_2\leq\inf_{\lambda\in\Lambda}\mathbb{E}\|f-\tilde{f}_{\lambda}\|_2+\Delta_{n,M}, \tag{18}$$

regardless of what f is. With this approach in mind, we now introduce the notion of an *oracle*.

Definition 4.1 (adapted from Tsybakov 2009)**.** *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 the L_2 norm.

The inequality at (18) is called an oracle inequality, and we say that the aggregate $\tilde{f}_{\hat{\lambda}}$ mimics (or is adaptive to) the oracle if (18) is satisfied for the smallest possible $\Delta_{n,M}>0$ independent of f.

As is suggested in the latter definition, the remainder $\Delta_{n,M}$, also known as the *rate of convergence* of $\tilde{f}_{\hat{\lambda}}$, is central to the evaluation of an aggregate's suitability. Both lower and upper bounds on these rates of convergence can be theoretically found for each of the above-mentioned schemes, and these bounds can guide us towards optimal choices of $\hat{\lambda}$. The lower bounds will be particularly useful since, from Definition 4.1, we are mainly interested in the smallest possible rate $\delta_{n,M}$

Definition 4.2 (Tsybakov 2009)**.** *Let $\Lambda\subset\mathbb{R}^M$ be a set of admissible weights and let $\mathcal{F}$ and $\mathcal{F}'$ be classes of Borel-measurable functions on $[0,1]$ with $\{\tilde{f}_\lambda\}_{\lambda\in\Lambda}\subset\mathcal{F}'$. A sequence $\{\psi_{n,M}\}_{n,M\geq1}$ of positive numbers is called an optimal rate of convergence if there exist constants $c,C>0$ such that, for any $n\geq1$,*

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}\|f-\tilde{f}\|_2^2-\inf_{\lambda\in\Lambda}\|f-\tilde{f}_\lambda\|_2^2\right)\leq C\psi_{n,M} \tag{19}$$

for some aggregate estimator $\tilde{f}$ of f, and

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}\|f-T_n\|^2-\inf_{\lambda\in\Lambda}\|f-\tilde{f}_\lambda\|^2\right)\geq c\psi_{n,M} \tag{20}$$

for any data-dependent estimator T_n of f.

We now state optimal rates of convergence for the four aggregation schemes presented above. Theorem 4.1 below relies on the following assumptions:

H4.3: There exists $S\subset[0,1]^{p_n}$ such that $\mu_{\boldsymbol{x}}$ has a bounded density $g_{\boldsymbol{x}}$ w.r.t the Lebesgue measure satisfying $g_{\boldsymbol{x}}(\boldsymbol{z})>0$ for $\boldsymbol{z}\in S$.

H4.4: $M\leq c_0 n$ and $\log(M)\leq c_0 n$ for some $c_0>0$.

In the following statement, we use the notation $a_n \asymp b_n$ for sequences a_n, b_n to denote that $cb_n \leq a_n \leq Cb_n$ for some constants $0 < c \leq C < \infty$.

Theorem 4.1 (Bunea et al. 2007)**.** *Suppose that **H4.1** to **H4.4** hold. For Λ_{MS}, Λ_{L}, Λ_{C}, and Λ_{S} as defined in Section 4.1.1, the inequality at (20) is satisfied for the rates given by*

$$\psi_{n,M} \asymp \begin{cases} \log(M)/n & \text{if } \Lambda = \Lambda_{\mathrm{MS}}, \\ M/n & \text{if } \Lambda = \Lambda_{\mathrm{L}}, \\ M/n & \text{if } \Lambda = \Lambda_{\mathrm{C}} \text{ and } M \leq \sqrt{n}, \\ \sqrt{\log\left(1 + M/\sqrt{n}\right)/n} & \text{if } \Lambda = \Lambda_{\mathrm{C}} \text{ and } M > \sqrt{n}, \\ (D \log(1 + M/D)/n) & \text{if } \Lambda = \Lambda_{\mathrm{S}}, \end{cases} \tag{21}$$

with $M \log(M/D + 1) \leq n$ and $M \geq D$ in the case of Subset aggregation (S).

The proof of Theorem 4.1 is beyond the scope of this report. A proof in the case of a random design (being considered here) can be found in Tsybakov (2003) and relies on a result on the -Liebler divergence between the distributions of two estimators vergence stated in page 99 of Tsybakov (2009).

The four general aggregation schemes were introduced here as common staples in the litterature. A natural question to ask is when and why should we use one scheme over the other. This problem remains open Bunea et al. 2007, and it is generally difficult to compare these procedures. In fact, results that achieve the oracle adaptability in the sense of Definition 4.1 and (18) are rather rare or restricted in scope. Oftentimes, results attain a relaxation of (18) is given in the form

$$\mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|_2 \leq (1 + \varepsilon) \inf_{\lambda \in \Lambda} \mathbb{E}\|f - \tilde{f}_{\lambda}\|_2 + \Delta_{n,M}, \tag{22}$$

for some small $\varepsilon > 0$ independent of f. This is the goal that we will adopt in this section.

4.2 The penalized least-squares approach

So far, we have studied properties that can inform our construction of an appropriate aggregate $\tilde{f}_{\hat{\lambda}}$ but we have yet to actually construct such an estimator. In this section, rather that focusing on a single one of the aggregation schemes given above, we examine a penalized least squares approach drawn from Bunea et al. (2007). We will consider aggregates $\tilde{f}_{\hat{\lambda}}$ with weights satisfying

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \left(\frac{1}{n} \|\boldsymbol{y}_n - \tilde{f}_{l\lambda}(\boldsymbol{X}_n)\|^2 + \mathrm{pen}(\lambda) \right) \tag{23}$$

for some penalty $\mathrm{pen}(\lambda)$. We will show that, for an appropriate choice of $\mathrm{pen}(\lambda)$, such an estimator gets close to all rates in 4.1 in the sense of (22).

Definition 4.3. *For $\lambda \in \Lambda$, let $M(\lambda) := \|\lambda\|_0$ (i.e., the number of non-zero coefficients in λ) and write*

$$L(\lambda) = 2 \log\left(\frac{eM}{\max(M(\lambda), 1)} \right).$$

For $a > 0$, we define the Bunea-Tsybakov-Wegkamp BIC-type penalty to be

$$\mathrm{pen}_{\mathrm{BIC}}(\lambda) := \frac{2\sigma^2}{n} M(\lambda) \left(1 + \frac{2 + a}{1 + a} \sqrt{L(\lambda)} + \frac{1 + a}{a} L(\lambda) \right).$$

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

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

Theorem 4.4 below, which get us closer to a result of the type in 22 for the $\tilde{f}_{\mathrm{BIC}}$, relies on the two following lemmas, which are stated without proof:

Lemma 4.2. *For positive integers* $m \leq M$,

$$\binom{M}{m} \leq \left(\frac{eM}{m} \right)^m.$$

Lemma 4.3 (Adapted from Birgé et al. (2001))**.** *Let* $\mathcal{A}_n$ *be an index set and let* pen_0 *be a penalty. Suppose that there exists a family of weight sets* $\{\Lambda_\alpha\}_{\alpha \in \mathcal{A}_n}$*, a collection of integers* $\{m_\alpha\}_{\alpha \in \mathcal{A}_n}$*, and a real sequence* $\{p_\alpha\}_{\alpha \in \mathcal{A}_n}$ *such that* $M(\lambda) = m_\alpha$ *and* $\mathrm{pen}_0(\lambda) = p_\alpha$ *for all* $\lambda \in \Lambda_\alpha$*. Suppose, furthermore, that there exists a collection of non-negative real numbers* $\{L_\alpha\}_{\alpha \in \mathcal{A}_n}$ *satisfying*

$$\Sigma := \sum_{\alpha \in \mathcal{A}_n} \exp(-m_\alpha L_\alpha) < \infty. \tag{24}$$

Let $\theta \in (0,1)$ *and* $K > 2 - \theta$*. If there exists a finite subset (possibly empty)* $\bar{\mathcal{A}}_n \subset \mathcal{A}_n$ *such that, for all* $\alpha \in \mathcal{A}_n \setminus \bar{\mathcal{A}}_n$,

$$p_\alpha \geq \frac{\sigma^2}{n} m_\alpha \left(K + 2(2-\theta)\sqrt{L_\alpha} + 2\theta^{-1} L_\alpha \right) =: Q_\alpha \qquad \textit{whenever } \lambda \in \Lambda_\alpha. \tag{25}$$

Then, the aggregate $\tilde{f}_{\hat{\lambda}}$ *corresponding to* pen_0 *as defined by (23) exists almost surely and satisfies*

$$\begin{aligned} (1-\theta)\,\mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|^2 \leq \inf_{\alpha \in \mathcal{A}_n} \left\{ \inf_{\lambda \in \Lambda_\alpha} \left(\|f - \tilde{f}_\lambda\|^2 \right) + p_\alpha - \frac{\sigma^2}{n} m_\alpha \right\} \\ + \sup_{\alpha \in \mathcal{A}_n} \{Q_\alpha - p_\alpha\} \\ + \frac{\sigma^2}{n} m_\alpha \Sigma \left((2-\theta)^2 (K + \theta - 2)^{-1} + 2\theta^{-1} \right) \end{aligned} \tag{26}$$

Lemma 4.3 may seem rather convoluted and abstract. However, it essentially tells us that the risk $\mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|^2$ can be upper-bounded if a suitable discretization of the admissible weights can be found. This utility will become apparent in the proof of the following theorem.

Theorem 4.4 (Bunea et al. 2007)**.** *Assume that* ***H4.1*** *and* ***H4.2*** *hold. Then, for all* $a > 0$, $M \geq 2$, *and* $n \geq 1$,

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

Proof: Define Λ_m for each integer $0 \leq m \leq M$ as

$$\Lambda_m := \left\{ \lambda \in \mathbb{R}^M : M(\lambda) = m \right\}$$

and let $\{J_{m,k}\}$ for $k = 1, \ldots, \binom{M}{m}$ be the collection of subsets all of $[M]$ such that $|J_{m,k}| = m$.

Define

$$\Lambda_{m,k} := \{\lambda \in \Lambda_m : \lambda_j \neq 0 \iff j \in J_{m,k}\}$$

and note that $\mathcal{P}_m := \left\{\Lambda_{m,k} : k = 1, \ldots, \binom{M}{m}\right\}$ forms a partition on Λ_m, while $\bigcup_m \mathcal{P}_m$ is a partition of $\mathbb{R}^M$. Notice also that

- the function $L(\lambda)$ from Definition 4.3 is constant on each Λ_m, $0 \leq m \leq M$, with

$$L(\lambda) = 2\ln\left(\frac{eM}{\max(m,1)}\right) =: L_m$$

- trivially, the zero-norm $M(\lambda)$ is a constant m on each Λ_m, $0 \leq m \leq M$;
- the penalty $\text{pen}_{\text{BIC}}(\lambda)$ is constant on each Λ_m, $0 \leq m \leq M$.

More formally, for $0 \leq m \leq M$ and $k = 1, \ldots, \binom{M}{m}$, we have that the images $\text{pen}_{\text{BIC}}(\Lambda_{m,k}) = \{p_m\}$ and $L(\Lambda_{m,k}) = \{L_m\}$ for some $\{p_m\}_{m=0}^M$ and $\{L_m\}_{m=0}^M$.
Positioning ourselves for an application of Lemma 4.3, we let $\mathcal{A}_n := \left\{(m,k) : m \leq M, k \leq \binom{M}{m}\right\}$, $K = 2$, and $\theta = a/(1+a)$ for some arbitrary $a > 0$. It is easy to see from the definition of $\text{pen}_{\text{BIC}}(\lambda)$ that condition (25) is satisfied with equality for all $(m,k) \in \mathcal{A}_n$. Furthermore, let $\Sigma := \sum_{\alpha \in \mathcal{A}_n} \exp(-m_\alpha L_\alpha)$ and note that

$$\begin{aligned}
\Sigma &= \sum_{m=1}^{M} \binom{M}{m} \left(e^{\ln(eM/m)}\right)^{-2m} = \sum_{m=1}^{M} \binom{M}{m} \left(\frac{eM}{m}\right)^{-2m} \\
&\leq \sum_{m=1}^{M} \left(\frac{eM}{m}\right)^{m} \left(\frac{eM}{m}\right)^{-2m} = \sum_{m=1}^{M} \left(\frac{m}{M}\right)^{m} e^{-m} \qquad \text{(Lemma 4.2)} \\
&\leq \sum_{m=1}^{M} e^{-m} \leq \frac{1}{e-1}. \qquad (27)
\end{aligned}$$

Hence, condition (24) is also satisfied. We apply Lemma 4.3 and obtain that

$$\begin{aligned}
\left(\frac{1}{1+a}\right)\mathbb{E}\|f - \tilde{f}_{\text{BIC}}\|^2 \leq & \inf_{(m,k)\in\mathcal{A}_n} \left\{\inf_{\lambda\in\Lambda_m}\left(\|f - \tilde{f}_\lambda\|^2\right) + p_m - \frac{\sigma^2}{n}m\right\} \\
& + \sup_{(m,k)\in\mathcal{A}_n}\{p_m - p_m\} \\
& + \frac{\sigma^2}{n}m\Sigma\left(\left(2 - \left(\frac{a}{1+a}\right)\right)^2\left(\frac{1+a}{a}\right) + \frac{2+2a}{a}\right) \\
= & \inf_{(m,k)\in\mathcal{A}_n} \left\{\inf_{\lambda\in\Lambda_m}\left(\|f - \tilde{f}_\lambda\|^2\right) + p_m - \frac{\sigma^2}{n}m\right\} \\
& \frac{(1+a)\sigma^2}{an}\left(\left(\frac{2+a}{1+a}\right)^2 + 2\right)\Sigma.
\end{aligned}$$

A simple argument can be made (using, for example, the quadratic formula on L_m) to show that

$$n\text{pen}_{\text{BIC}}(\lambda) - m\sigma^2 = \sigma^2 m\left(1 + 2\left(\frac{2+a}{1+a}\right)\sqrt{L_m} + 2\left(\frac{1+a}{a}\right)L_m\right) \leq \sigma^2 m\left(5 + \frac{2+3a}{a}L_m\right)$$

for any Λ_m, $m = 0, \ldots, M$, and $a > 0$. Combining this with (27),

$$\left(\frac{1}{1+a}\right)\mathbb{E}\|f - \tilde{f}_{\hat{\lambda}}\|^2 \leq \inf_{(m,k)\in\mathcal{A}_n}\left\{\inf_{\lambda\in\Lambda_m}\left(\|f - \tilde{f}_\lambda\|^2\right) + \frac{\sigma^2 m}{n}\left(5 + \frac{2+3a}{a}L_m\right)\right\} + \frac{6\sigma^2(1+a)}{an(e-1)}$$
$$= \inf_{\lambda\in\mathbb{R}^M}\left\{\|\tilde{f}_\lambda - f\|^2 + \frac{\sigma^2}{n}\left(5 + \frac{2+3a}{a}L(\lambda)\right)M(\lambda)\right\} + \frac{6\sigma^2(1+a)}{an(e-1)}.$$

□

Theorem 4.4 implies that the $\tilde{f}_{\text{BIC}}$ aggregate satisfies the bounds in (21).

Corollary 4.5 (Bunea et al. 2007)**.** *Under the conditions of Theorem 4.4, there exists a constant $C > 0$ such that, for all $a > 0$ and integers $n \geq 1$, $M \geq 2$, and $D \leq M$, the following holds:*

$$\mathbb{E}\|f - \tilde{f}_{\text{BIC}}\|^2 \leq (1+a)\inf_{\lambda\in\Lambda}\|f - \tilde{f}_\lambda\|^2 + C(1 + a + a^{-1})\frac{\sigma^2\psi_{n,M}}{n},$$

where Λ and $\psi_{n,M}$ are chosen accordingly as given in (21).

We prove the case for the Model Selection scheme only, taking

$$\Lambda = \Lambda_{\text{MS}} \quad \text{and} \quad \psi_{n,M} = \frac{\log(M)}{n}.$$

Note that, for $\lambda \in \Lambda_{\text{MS}}$, $M(\lambda) = 1$ and $L(\lambda) = 2\log(eM)$. Thus,

$$(1+a)\inf_{\lambda\in\Lambda_{\text{MS}}}\left\{\|\tilde{f}_\lambda - f\|^2 + \frac{\sigma^2}{n}\left(5 + \frac{2+3a}{a}L(\lambda)\right)M(\lambda)\right\}$$
$$= (1+a)\inf_{\lambda\in\mathbb{R}^M}\left\{\|\tilde{f}_\lambda - f\|^2\right\} + (1+a)\frac{\sigma^2}{n}\left(5 + \frac{2+3a}{a}\left(2\log(eM)\right)\right)$$
$$\leq (1+a)\inf_{\lambda\in\mathbb{R}^M}\left\{\|\tilde{f}_\lambda - f\|^2\right\} + C(1 + a + a^{-1})\frac{\sigma^2}{n}\log(M).$$

□

4.3 Some remarks

The BIC-type aggregation approach discussed in this section provides a general framework that adapts well to various model aggregation scenarios, with theoretical guarantees that approach the optimal rates established in Theorem 4.1. Several key observations are worth highlighting:

1. Flexibility of framework: The BIC-type penalty can be applied across multiple aggregation schemes while maintaining near-optimal convergence rates. This makes it particularly useful when the appropriate aggregation scheme is not obvious a priori.

2. Computational considerations: While the theoretical results are encouraging, the computation of $\hat{\lambda}_{\text{BIC}}$ as defined in (23) may be challenging in practice, especially for large values of M. For Linear and Convex aggregation schemes, the optimization problem is convex and

thus tractable. However, for Model Selection and Subset Selection, which involve discrete optimization over a combinatorial space, exact computation may be infeasible for large M.

3. Variance estimation: The practical implementation of the BIC-type penalty requires knowledge of σ^2, the variance of the residuals. In practice, this parameter is typically unknown and must be estimated from the data.

4. Extension to heteroscedastic settings: The framework presented assumes homoscedastic, Gaussian errors. Extending the results to settings with heteroscedastic errors would be valuable for many practical applications where the assumption of constant variance is unrealistic.

5 Conclusion

We have highlighted key theoretical insights into the use of cross-validation for model selection in both parametric and nonparametric settings. In the context of linear models, consistent model selection requires that the size of the validation set asymptotically dominates that of the training set. That is,

Under these conditions, leave-one-out cross-validation performs well only when there is at most one correct model of fixed dimension. In scenarios where such a model exists, the delete-d method also shows strong performance.

In contrast, the nonparametric setting presents a different picture. Here, consistency can be achieved even when the training set is larger than the validation set, provided that certain norm conditions are satisfied. Both single-split and majority-vote cross-validation procedures can yield consistent selection under these conditions. Nevertheless, leave-one-out cross-validation remains inadequate in this setting due to the validation set being too small. Additionally, employing multiple data splits does not remedy the inconsistency of a flawed single-split procedure.

From a practical perspective, although cross-validation can be computationally intensive, its use may be justified by its strong theoretical guarantees. Notably, the results on majority-vote methods indicate that even a small number of splits, as in k-fold cross-validation, can suffice for consistent model selection. However, several practical aspects remain unresolved and warrant further investigation, including the optimal choice of data split ratios, the number of folds or splits, and the comparative performance of voting versus averaging strategies.

Finally, the aggregation approach presented provides an alternative theoretically sound framework for dealing with multiple candidate models, with guarantees that approach optimal minimax rates. The flexibility of the framework makes it applicable across various model selection and aggregation scenarios, though practical implementation considerations remain to be addressed, particularly for large sets of candidate models.

References

Birgé, L. and P. Massart (2001). *A generalized C_p criterion for Gaussian model selection*. Prépublication 647. Paris: Laboratoire de Probabilités & Modèles Aléatoires, Universités de Paris VI & VII.

Bunea, Florentina, Alexandre B. Tsybakov, and Marten H. Wegkamp (2007). "Aggregation for Gaussian Regression". In: *The Annals of Statistics* 35.4, pp. 1674–1697.

Hansen, Bruce E. (2022). *Econometrics*. Princeton: Princeton University Press.

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

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

Tsybakov, Alexandre B. (2003). "Optimal Rates of Aggregation". In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf and Manfred K. Warmuth. Berlin, Heidelberg: Springer, pp. 303–313.

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

Yang, Yuhong (2007). "Consistency of cross validation for comparing regression procedures". In: *The Annals of Statistics* 35.6, pp. 2450–2473.