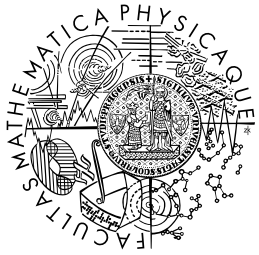

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Aibat Kossumov

Cross-validation and its use in statistics

Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D.

Study programme: Probability, Mathematical Statistics and Econometrics

Prague 2024

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to express my gratitude to doc. Ing. Marek Omelka, Ph.D., for his guidance throughout this work, assistance, valuable insights, and the time he dedicated to me. I would also like to thank my mother for everything she has given me and for the opportunity to continue my master's degree. Additionally, I thank all my friends for their support throughout my studies at Charles University.

Title: Cross-validation and its use in statistics

Author: Aibat Kossumov

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D.

Abstract: In this thesis, the use of cross-validation methods in different areas of statistics is studied. Firstly, the application of leave-one-out cross-validation, CV(1), for bandwidth selection in kernel density estimation and kernel regression tasks is considered. Theoretical findings are demonstrated on simulated data. Then, the selection of a linear model with the best predictive ability is explored. It is illustrated that, in the context of linear models, the use of CV(n_v) instead of the leave-one-out approach is advisable, where $n_v/n \to 1$ as $n \to \infty$. The studied methods are applied on real data from parliamentary and presidential elections in the Czech Republic in 2021 and 2023.

Keywords: cross-validation, kernel density estimation, nonparametric kernel regression, linear model

Contents

Introduction

Cross-validation is a set of methods that divide the observed data into two parts, commonly referred to as the validation and construction parts. These methods find their application in nonparametric approaches such as kernel density estimation and kernel regression. Additionally, the cross-validation idea of dividing data into two parts works well for assessing prediction error. Therefore, these methods play a crucial role in model selection based on the model's predictive ability, making cross-validation techniques very popular in fields such as machine learning and data science.

In the first two chapters of this thesis, the use of cross-validation methods is studied for bandwidth selection in kernel density estimation and kernel regression. The popularity of cross-validation in these areas increased after the paper Stone [1984], where it is shown that the bandwidth selected by cross-validation is optimal in some sense. In these chapters, theoretical justifications of cross-validation techniques are provided, and several other methods for bandwidth selection are introduced based on stronger assumptions on the distribution of observed data (see Section 1.2.2) or asymptotic approximations (see Sections 1.2.1 and 2.2.1). Both chapters are concluded by smaller simulation studies, where the advantages of cross-validation are demonstrated. All results are provided only for leave-one-out cross-validation CV(1), which means that for validation purposes only one observation is considered.

In the last chapter the focus is on the use of cross-validation techniques in the context of linear model selection. Suppose that there are n observations available for selecting a model from a class of linear models. For validation purposes, n_v observations are used, and for model construction, n_c are used, such that $n_v+n_c = n$. Obviously, there are $\binom{n}{n_v}$ different ways to split the dataset. Computational complexity of cross-validation increases as n_v increases. Therefore, leave-one-out cross-validation with $n_v = 1$ is the simplest one. However, unlike in the first two chapters, it may be shown that CV(1) is no longer a useful method for linear model selection. In linear model selection from all possible models, there exists an optimal one, denoted as M_*, in some sense (for details, see Section 3.3). The problem with the CV(1) method is that it is not asymptotically consistent in the following sense:

$$\mathrm{P}\left[\text{the model selected by CV(1) is not } M_*\right] \underset{n\to\infty}{\not\to} 0.$$

Chapter 3 demonstrates that instead of leave-one-out cross-validation, leave- n_v-out cross-validation should be used. Additionally, it should be ensured that $n_v/n \to 1$ as $n \to \infty$, which is completely opposite to the leave-one-out cross-validation principle. However, when n_v is large, the amount of computation required to use the cross-validation may be impractical. Therefore, a so-called "balanced incomplete" cross-validation, BICV(n_v) ,should be used. The idea of the method is that only a smaller part of $\binom{n}{n_v}$ splits are made according to a systematic manner. Additionally, there exists a Monte Carlo alternative to the last method, which is quite useful in practical applications. Finally, in Chapter 3, the aforementioned methods will be applied to real data from parliamentary and presidential elections in the Czech Republic in 2021 and 2023

1. Kernel density estimation

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables sampled from a distribution with a probability density function f. Depending on the task, one can construct appropriate statistical procedures based on these observations. Very often, such procedures are derived by assuming specific parametric models. However, these assumptions can sometimes be too restrictive. Therefore, there may be instances where one is interested in understanding the data structure without relying on any parametric model. In such cases, considering a nonparametric approach can be beneficial because it allows the data to speak for themselves.

If the goal is to nonparametrically estimate the distribution of these observations, one option is to use the empirical cumulative distribution function $\widehat{F}_n$. However, one can get deeper insights about the nature of the data by estimating the probability density function f. The most famous and oldest nonparametric density estimator is the histogram. The histogram is created by partitioning the real line into equally-sized intervals (bins). Subsequently, it takes the form of a step function, with the height of each step corresponding to the proportion of the sample within that bin divided by the width of the bin (binwidth). Simplicity of histogram ensures its popularity. However, one of the main disadvantages of the histogram is its high sensitivity to the placement of the bin edges. Moreover, most densities are not represented as step functions. The histogram, unfortunately, approximates all densities using a step function, which can be seen as another disadvantage. A well-known alternative for nonparametrically estimating the probability density function f is the *kernel density estimator*.

This chapter is based on Chapters 2 and 3 of Wand and Jones [1995] and the course notes Nagy and Omelka [2024].

1.1 Basic properties

From this point onwards, it is assumed that $X_1, \ldots, X_n$ are independent and identically distributed (iid) random variables. The following definition provides the recipe for constructing a family of estimates of $f(x)$, which are consistent and asymptotically normal (see below Theorem 1).

Definition 1. *Let $K : \mathbb{R} \to \mathbb{R}$ and $\{h_n\}_{n=1}^{\infty}$ be a sequence of positive real numbers. Then, the* ***kernel density estimator*** *is defined as*

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right), \text{ for } x \in \mathbb{R},$$

where the function K is called a ***kernel function*** *and h_n is usually called* ***bandwidth****.*

First, note that compared to a histogram, the estimator $\widehat{f}_n$ is defined everywhere on $\mathbb{R}$ and does not require specifying the placement of edge points. It is evident from Definition 1 that the kernel density estimator depends on the kernel function K and bandwidth h_n. As demonstrated in Wand and Jones [1995]

(Section 2.7), selecting the shape of the kernel function is not particularly crucial. Typically, the function K is chosen to be a unimodal probability density function that is symmetric about zero. For instance, one of the possible choices is the standard normal kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Remark. Note that if K is a density symmetric around zero, then

$$x \longmapsto \frac{1}{h_n} K\left(\frac{X_i - x}{h_n}\right)$$

is also a density symmetric around the point X_i. Thus, $\widehat{f_n}$ can be interpreted as the average of n densities of the form $\frac{1}{h_n} K\left(\frac{X_i - x}{h_n}\right)$.

The important aspect lies in choosing the bandwidth h_n, a topic which will be discussed in the next section. In most cases, it is assumed that $\{h_n\}_{n=1}^{\infty}$ is a non-increasing sequence of positive numbers converging to zero such that $nh_n \to \infty$, i.e., h_n must not converge to zero too quickly.

Suppose that K is a density symmetric around zero. Then, the expected value of $\widehat{f}_n(x)$ takes the following form

$$\mathbb{E}\widehat{f}_n(x) = \frac{1}{h_n} \mathbb{E} K\left(\frac{X_1 - x}{h_n}\right) = \frac{1}{h_n} \int_{\mathbb{R}} f(z) K\left(\frac{x - z}{h_n}\right) \mathrm{d}z,$$

which means that $\mathbb{E}\widehat{f}_n(x)$ is a convolution of densities $\frac{1}{h_n} K(\frac{\cdot}{h_n})$ and $f(\cdot)$. Such convolutions naturally appear in various situations. For example, consider a random variable Z with density K, implying that $h_n Z$ has density $\frac{1}{h_n} K(\frac{\cdot}{h_n})$. Let X be a random variable independent of Z, then $\mathbb{E}\widehat{f}_n(x)$ is the density value at x of a random variable $X + h_n Z$.

The kernel density estimator possesses the important properties of consistency and asymptotic normality, as summarized by the following theorem.

Theorem 1. *Let $X_1, \ldots, X_n$ be a random sample with common probability density function $f(x)$, and let $\widehat{f}_n(x)$ be a kernel density estimator. Suppose that the function K satisfies the following conditions:*

(A1) $\int_{\mathbb{R}} |K(y)| \mathrm{d}y < \infty$ *and* $\int_{\mathbb{R}} K(y) \mathrm{d}y = 1$,

(A2) $\lim_{|y| \to \infty} |yK(y)| = 0$,

(A3) $h_n \searrow 0$ *as* $n \to \infty$ *and* $(nh_n) \to \infty$ *as* $n \to \infty$.

Then, at each continuity point x of the density f, the following holds:

1. $\lim_{n \to \infty} nh_n \operatorname{var}(\widehat{f}_n(x)) = f(x) \int_{\mathbb{R}} K^2(y) \mathrm{d}y$,
2. $\widehat{f}_n(x) \xrightarrow[n \to \infty]{P} f(x)$,
3. *if $x \in \mathbb{R}$ such that $f(x) > 0$, then*

$$\frac{\widehat{f}_n(x) - \mathbb{E}\widehat{f}_n(x)}{\sqrt{\operatorname{var}(\widehat{f}_n(x))}} \xrightarrow[n \to \infty]{d} \mathsf{N}(0, 1).$$

The proof is provided in Parzen [1962], Chapter 2.

Remark.

- If K is a density, then $\int_{\mathbb{R}} |K(y)|\mathrm{d}y = \int_{\mathbb{R}} K(y)\mathrm{d}y = 1 < \infty$, i.e., the first assumption *(A1)* holds.
- The second assumption *(A2)* means that $|K(y)|$ converges to zero more rapidly than $\frac{1}{|y|}$ as $|y| \to \infty$. This condition is automatically met if K has a bounded support.

In Parzen [1962], the consistency of $\widehat{f}_n(x)$ is proven through L_2 convergence. One might wonder if it is possible to establish consistency of $\widehat{f}_n(x)$ through the law of large numbers. It is possible, but one must employ the law of large numbers for triangular arrays, as discussed in Dewan and Rao [1997] (Theorem 3.3). Additionally, it is important to note that Theorem 1 only asserts point-wise consistency. Demonstrating a similar result uniformly across $x \in \mathbb{R}$ is considerably more challenging, see, e.g., Wied and Weißbach [2012] (Theorem 2).

1.2 Bandwidth selection

1.2.1 Asymptotic approximation

In Theorem 1, it is shown that under certain assumptions on the function K, the kernel density estimator $\widehat{f}_n$ is a consistent estimator of the true density f. The catch is that the kernel density estimator $\widehat{f}_n$ implicitly depends on the sequence $\{h_n\}_{n=1}^{\infty}$, which should also be chosen in a reasonable manner.

A well-known measure that summarizes the quality of estimating the density f by the estimator $\widehat{f}_n$ at the point x is the *mean squared error*, defined as follows

$$\mathrm{MSE}(\widehat{f}_n(x)) = \mathbb{E}(\widehat{f}_n(x) - f(x))^2 = \mathrm{var}(\widehat{f}_n(x)) + [\mathrm{bias}(\widehat{f}_n(x))]^2.$$

The error in estimating the density at a single point is quantified by MSE. It's important to note that MSE is a local characteristic of the density estimator $\widehat{f}_n$. To obtain a measure that globally summarizes the quality of estimation (i.e., one that does not depend on x), it is useful to define the *mean integrated squared error*

$$\mathrm{MISE}(\widehat{f}_n) = \int_{\mathbb{R}} \mathrm{MSE}(\widehat{f}_n(x))\mathrm{d}x.$$

The choice of bandwidth h_n is often based on the approximation of MISE.

Let's begin by approximating MSE. According to the first part of Theorem 1, it follows that

$$\mathrm{var}(\widehat{f}_n(x)) = \frac{f(x)R(K)}{nh_n} + o\left(\frac{1}{nh_n}\right), \tag{1.1}$$

where $R(K) = \int_{\mathbb{R}} K^2(y)\mathrm{d}y$ is one possible measure of the roughness of K. The next step is to approximate the bias. Suppose that x is an interior point of f, and f is twice differentiable at x. Let the kernel K satisfy the following conditions: $\int_{\mathbb{R}} K(t)\mathrm{d}t = 1$, $\int_{\mathbb{R}} tK(t)\mathrm{d}t = 0$, and $\int_{\mathbb{R}} |t^2K(t)|\mathrm{d}t < \infty$. One can get

$$\mathbb{E}\widehat{f}_n(x) = \frac{1}{h_n}\mathbb{E}K\left(\frac{X_1 - x}{h_n}\right) = \int_{\mathbb{R}} \frac{1}{h_n}K\left(\frac{y-x}{h_n}\right) f(y)\mathrm{d}y = \int_{\mathbb{R}} K(t)f(x + th_n)\mathrm{d}t.$$

Using the second-order Taylor expansion of the density f around x, for sufficiently large n, one obtains

$$f(x+th_n) = f(x) + th_n f'(x) + \frac{1}{2}t^2 h_n^2 f''(x) + o(t^2 h_n^2).$$

By combining this Taylor approximation with the assumptions on the kernel function K, one has

$$\text{bias}(\widehat{f}_n(x)) = \mathbb{E}\widehat{f}_n(x) - f(x) = \frac{1}{2}h_n^2 f''(x)\mu_2 + o(h_n^2), \tag{1.2}$$

where $\mu_2 = \int_{\mathbb{R}} y^2 K(y)dy$. Thus, from equations (1.1) and (1.2), the following asymptotic approximation of the mean squared error is derived

$$\text{MSE}(\widehat{f}_n(x)) = \frac{1}{nh_n}f(x)R(K) + \frac{1}{4}h_n^4[f''(x)]^2\mu_2^2 + o\left(\frac{1}{nh_n}\right) + o(h_n^4). \tag{1.3}$$

This approximation of the mean squared error serves as motivation to define the *asymptotic mean integrated squared error* of $\widehat{f}_n$ by integrating the main terms and disregarding the remainder $o(\cdot)$ terms in (1.3)

$$\text{AMISE}(\widehat{f}_n) = \frac{R(K)}{nh_n} + h_n^4\frac{R(f'')\mu_2^2}{4}, \tag{1.4}$$

where $R(f'') = \int_{\mathbb{R}}[f''(x)]^2 \mathrm{d}x$. AMISE can be understood as an approximation of MISE.

It is known that MISE is a quantity that summarizes the quality of estimating the density f by $\widehat{f}_n$. Therefore it is natural to choose the bandwidth h_n in a way that minimizes $\text{MISE}(\widehat{f}_n)$. However, since $\text{MISE}(\widehat{f}_n)$ cannot be expressed simply as a function of h_n, $\text{AMISE}(\widehat{f}_n)$ will be minimized instead. By minimizing (1.4) with respect to h_n, the so-called *asymptotically optimal global bandwidth* is obtained

$$h_n^{(opt)} = n^{-\frac{1}{5}}\left[\frac{R(K)}{R(f'')\mu_2^2}\right]^{\frac{1}{5}}. \tag{1.5}$$

The asymptotically optimal global bandwidth $h_n^{(opt)}$ is obtained by applying the second-order Taylor approximation to the probability density function f. A similar procedure can generally be employed to derive the optimal bandwidth from any p-order approximation of f (under the assumption that f is p-times differentiable). To do this, the kernel function must satisfy the following conditions $\int_{\mathbb{R}} K(t)\mathrm{d}t = 1$ and

$$\int_{\mathbb{R}} t^j K(t)\mathrm{d}t = 0, \quad j = 1, ..., p-1 \quad \text{and} \quad \int_{\mathbb{R}} t^p K(t)\mathrm{d}t \neq 0.$$

The disadvantage of this method is that for $p > 2$, it must hold that

$$\int_{\mathbb{R}} t^2 K(t)\mathrm{d}t = 0.$$

Therefore K cannot be non-negative. As a consequence, it might happen that the estimator $\widehat{f}_n$ is negative. Detailed information about high-order kernels could be found in Wand and Jones [1995] (Section 2.8).

1.2.2 Normal reference rule of thumb

Note that (1.5) cannot be computed because the true density f is not known, which also means that $R(f'')$ is not known. Since $f''(x)$ measures the curvature of the function f at the point x, one could interpret $R(f'')$ as the overall curvature of the function f. Therefore, for functions with high curvature, $h_n^{(opt)}$ will be small. Conversely, for functions with low curvature, $h_n^{(opt)}$ will be large.

In practice, so-called 'plug-in methods' are popular, where it is assumed that f belongs to a parametric family. This allows one to compute $R(f'')$ and then simply plug this quantity into (1.5). For example, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, it can be shown that $R(f'') = \frac{3}{8\sigma^5\sqrt{\pi}}$. Thus the asymptotically optimal global bandwidth would be

$$h_n^{(opt)} = \sigma \; n^{-\frac{1}{5}} \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2^2}\right]^{\frac{1}{5}}.$$

By using a standard normal kernel $K(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$, the asymptotically optimal global bandwidth simplifies to

$$h_n^{(opt)} = \left[\frac{4}{3}\right]^{\frac{1}{5}} \sigma \; n^{-\frac{1}{5}} \approx 1.06 \; \sigma \; n^{-\frac{1}{5}}.$$

The standard normal reference rule is given by

$$h_n^{(NR)} = 1.06 \; n^{-\frac{1}{5}} \; \min\{S_n, \widetilde{IQR}_n\}, \tag{1.6}$$

where

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \;\text{ and }\; \widetilde{IQR}_n = \frac{\widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)}.$$

In this context, Φ^{-1} denotes the quantile function of the standard normal distribution, while $\widehat{F}_n^{-1}$ represents the quantile function of the empirical distribution. The normalizing factor in the denominator of $\widetilde{IQR}_n$ is the population interquartile range of the standard normal density and is approximately equal to 1.349. Further discussion about why the minimum of the quantities S_n and $\widetilde{IQR}_n$ is taken can be found in Silverman [1986](Section 3.4.2).

If the underlying distribution is close to a normal distribution, then the normal reference rule provide good results. However, if the true density is, for example, multimodal, then this bandwidth selector tends to over-smooth and mask important features in the data. That is why some authors prefer to use the following modification of the standard normal reference rule

$$h_n^{(MNR)} = 0.9 \; n^{-\frac{1}{5}} \; \min\{S_n, \widetilde{IQR}_n\}. \tag{1.7}$$

1.2.3 Least-squares cross-validation

In the previous subsections, the bandwidth was selected based on the asymptotic approximation of MISE. Now, instead of approximating MISE, an attempt will be made to decompose MISE and find an unbiased estimator for the part of the decomposition that depends on the bandwidth. Then, the bandwidth that

minimizes this unbiased estimator will be selected. By using Fubini's theorem, the following holds

$$\begin{aligned}\mathrm{MISE}(\widehat{f}_n) &= \int_{\mathbb{R}} \mathbb{E}(\widehat{f}_n(x) - f(x))^2 \mathrm{d}x \overset{\text{Fubini}}{=} \mathbb{E}\int_{\mathbb{R}} (\widehat{f}_n(x) - f(x))^2 \mathrm{d}x \\ &= \mathbb{E}\int_{\mathbb{R}} \left(\left[\widehat{f}_n(x)\right]^2 - 2\widehat{f}_n(x) f(x) + \left[f(x)\right]^2 \right) \mathrm{d}x \\ &= \mathbb{E}\int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x - 2\mathbb{E}\int_{\mathbb{R}} \widehat{f}_n(x) f(x) \mathrm{d}x + \int_{\mathbb{R}} \left[f(x)\right]^2 \mathrm{d}x. \end{aligned} \tag{1.8}$$

Note that only the first two terms of the equation (1.8) depend on the bandwidth.

Obviously, the unbiased estimator of the first term is $\int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2$. At first glance, it may seem that the estimator defined in this way is challenging to compute. However, it can actually be computed directly from the random sample $X_1, \ldots, X_n$, as

$$\begin{aligned}\int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x &= \int_{\mathbb{R}} \left[\frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)\right]^2 \mathrm{d}x \\ &= \frac{1}{(nh_n)^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \int_{\mathbb{R}} K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) \mathrm{d}x \\ &= \frac{1}{n^2 h_n} \sum_{i=1}^{n}\sum_{j=1}^{n} \int_{\mathbb{R}} K(u) K\left(u + \frac{X_j - X_i}{h_n}\right) \mathrm{d}u \\ &= \frac{1}{n^2 h_n} \sum_{i=1}^{n}\sum_{j=1}^{n} \int_{\mathbb{R}} K(u) K\left(u - \frac{X_i - X_j}{h_n}\right) \mathrm{d}u. \end{aligned}$$

One can see that the integral from the last expression reminds one of convolution. Assuming that K is symmetric, which is satisfied in many cases, one indeed gets convolution. Denote $\widetilde{K}(t) = \int_{\mathbb{R}} K(u)K(t-u)\mathrm{d}u$. Then, for a symmetric kernel function K, it holds that

$$\int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x = \frac{1}{n^2 h_n} \sum_{i=1}^{n}\sum_{j=1}^{n} \widetilde{K}\left(\frac{X_i - X_j}{h_n}\right).$$

If $K(u) = \exp(-u^2/2)/\sqrt{2\pi}$, which is a standard normal kernel, then the convolution

$$\widetilde{K}(t) = \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{t^2}{4}\right),$$

is a normal density with mean zero and variance two. This holds true because the sum of two independent $\mathsf{N}(0,1)$ random variables results in a random variable with $\mathsf{N}(0,2)$ distribution. This illustrates that $\int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x$ could be calculated explicitly.

It remains to find an unbiased estimator for $A_n = \mathbb{E}\int_{\mathbb{R}} \widehat{f}_n(x) f(x) \mathrm{d}x$. Let X be a random variable with the same distribution as X_1 and independent from the random sample $X_1, \ldots, X_n$. Then, it holds

$$A_n = \mathbb{E}\widehat{f}_n(X).$$

Define the following estimator

$$\widehat{A}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i),$$

where

$$\widehat{f}_{-i}(x) = \frac{1}{(n-1)h_n}\sum_{j=1, j\neq i}^{n} K\left(\frac{X_j - x}{h_n}\right)$$

is the estimate of $f(x)$ that is based on the sample without the i-th observation X_i, where $i \in \{1, ..., n\}$. From the definition of the function $\widehat{f}_{-i}$, one can observe a general principle of cross-validation. In this particular situation, the estimator was 'trained' on all observations except for one. Such cross-validation technique is usually referred to as leave-one-out cross-validation.

Now, compute the expected value of individual terms in the estimator $\widehat{A}_n$

$$\begin{aligned}
\mathbb{E}\widehat{f}_{-i}(X_i) &= \mathbb{E}\left[\frac{1}{(n-1)h_n}\sum_{j=1, j\neq i}^{n} K\left(\frac{X_j - X_i}{h_n}\right)\right] = \frac{1}{h_n}\mathbb{E}K\left(\frac{X_1 - X_2}{h_n}\right) \\
&= \frac{1}{h_n}\int_{\mathbb{R}}\int_{\mathbb{R}} K\left(\frac{y-x}{h_n}\right) f(x)f(y)\mathrm{d}x\ \mathrm{d}y \\
&= \int_{\mathbb{R}}\left[\int_{\mathbb{R}} \frac{1}{h_n} K\left(\frac{y-x}{h_n}\right) f(y)\mathrm{d}y\right] f(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} \mathbb{E}\widehat{f}_n(x) f(x)\mathrm{d}x \overset{\text{Fubini}}{=} \mathbb{E}\int_{\mathbb{R}} \widehat{f}_n(x) f(x)\mathrm{d}x.
\end{aligned}$$

Since $X_1, ..., X_n$ are identically distributed, $\widehat{A}_n$ is an unbiased estimator of A_n.

Unbiased estimators were found for the first two terms of the equation (1.8). Therefore, define the following function

$$\mathcal{L}(h_n) = \int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i).$$

This function can be understood as an estimate of the part of MISE that depends on h_n. In other words, $\mathcal{L}(h_n)$ is an unbiased estimate of the shifted MISE$(\widehat{f}_n)$. Generally, one aims to minimize MISE, thus the bandwidth is selected as

$$h_n^{(LCSV)} = \underset{h_n > 0}{\arg\min}\ \mathcal{L}(h_n). \tag{1.9}$$

Further, the following notation $\mathrm{ISE}(h_n) = \int_{\mathbb{R}} \left(\widehat{f}_n(x) - f(x)\right)^2 \mathrm{d}x$ will be used. Note that $\mathrm{ISE}(h_n)$ represents the integrated squared error of the kernel density estimator $\widehat{f}_n$ and typically is denoted as $\mathrm{ISE}(\widehat{f}_n)$. However, in this context, ISE is considered as a function of the bandwidth h_n.

The next theorem justifies that $h_n^{(LSCV)}$ is asymptotically optimal in terms of minimizing ISE.

Theorem 2. *Let $X_1, \ldots, X_n$ be a random sample with a common probability density function $f(x)$. Assume that the density f is a bounded function. Further, let the kernel function K satisfy the following conditions:*

(B1) $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$,

(B2) *K is symmetric about the origin and has compact support,*

(B3) $\widetilde{K}(0) < 2K(0)$*, where* $\widetilde{K}$ *denotes the convolution of* K *with itself,*

(B4) *K is Hölder continuous, i.e., there exist* $\beta > 0$ *and* $\alpha > 0$ *such that*

$$|K(y) - K(x)| < \alpha|y - x|^{\beta} \ \forall y, x \in \mathbb{R}.$$

Then

$$\frac{\mathrm{ISE}(h_n^{(LSCV)})}{\min_{h_n} \mathrm{ISE}(h_n)} \xrightarrow[n\to\infty]{a.s.} 1.$$

The general proof in multivariate settings is presented in Stone [1984].

Remark.

- The second assumption *(B2)* requires that K has a compact support. In the above text, the standard normal kernel has always been considered for illustrative purposes, which does not have bounded support. In Wand and Jones [1995] (Section 2.7), it is shown that among non-negative kernels, the optimal one is the so-called Epanechnikov kernel, which is defined as follows:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{for } |u| \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1.10}$$

  Note that the Epanechnikov kernel has a compact support $[-1, 1]$. Also, this kernel is Hölder continuous because it is continuously differentiable on its compact support and is otherwise defined as zero.

  In general, restricting the kernel to compact support allows for simpler proofs with more powerful asymptotic results. However, in practical applications, kernels with unbounded support, such as the standard normal kernel, are very popular.

- The assumption *(B3)* is satisfied if the kernel K is nonnegative and $K(0) = \max_{u\in\mathbb{R}} K(u)$, as

$$\widetilde{K}(0) = \int_{\mathbb{R}} K^2(u)\mathrm{d}u \leq \max_{u\in\mathbb{R}} K(u) \int_{\mathbb{R}} K(u)\mathrm{d}u = K(0) < 2K(0).$$

Hence, according to Theorem 2, cross-validated smoothing parameter selection is asymptotically optimal in terms of minimizing the integrated squared error. Despite the strong theoretical attractiveness of least square cross-validation, it was demonstrated in Hall and Marron [1987] that the variance of $h_n^{(LSCV)}$ (for not too big sample sizes) is rather large.

1.2.4 Biased cross-validation

Due to the limitations of *least square cross-validation*, another bandwidth selector known as *biased cross-validation* was proposed. In general, biased cross-validation can be seen as a hybrid approach. It combines elements of both cross-validation and 'plug-in' bandwidth selection, as will be shown later.

Recall the defintion of AMISE given by (1.4), which is

$$\mathrm{AMISE}(\widehat{f}_n) = \frac{R(K)}{nh_n} + h_n^4 \frac{R(f'')\mu_2^2}{4},$$

where $R(f'') = \int_{\mathbb{R}} [f''(x)]^2 dx$. Note that in the expression of AMISE, the only quantity that is not known is $R(f'')$. One can consider $R(\widehat{f}_n'')$ as a natural estimator of $R(f'')$.

Lemma 3. *Consider a random sample $X_1, \ldots, X_n$ with a common probability density function f. Assume that the following conditions are satisfied.*

- *K is a symmetric, non-negative, two times continuously differentiable kernel function with a bounded support $[-1, 1]$, and $\int_{-1}^{1} K(u)\mathrm{d}u = 1$. Additionally, suppose that $K(\pm 1) = 0$ and $K'(\pm 1) = 0$.*
- *f is four times continuously differentiable and it holds that*

$$\sup_{x \in \mathbb{R}} |f''(x)| < \infty. \tag{1.11}$$

  Also, suppose that there exists $\delta_0 > 0$ such that

$$\int_{\mathbb{R}} \sup_{\eta \in [-\delta_0, \delta_0]} \left[f^{(4)}(x + \eta) \right]^2 \mathrm{d}x < \infty. \tag{1.12}$$

- *$h_n \searrow 0$ as $n \to \infty$ and $(nh_n^2)^{-1} = O(1)$.*

Then the following holds

$$\mathbb{E}R\left(\widehat{f}_n''\right) = R(f'') + \frac{R(K'')}{nh_n^5} + O(h_n^2).$$

Proof. Since K is symmetric, one has

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right).$$

By differentiating twice with respect to the variable x, one obtains

$$\widehat{f}_n''(x) = \frac{1}{nh_n^3} \sum_{i=1}^{n} K''\left(\frac{x - X_i}{h_n}\right).$$

Hence, one gets that

$$\begin{aligned}
\mathbb{E}R\left(\widehat{f}_n''\right) &= \mathbb{E} \int_{\mathbb{R}} \left[\widehat{f}_n''(x)\right]^2 \mathrm{d}x \\
&= \mathbb{E} \int_{\mathbb{R}} \left(\frac{1}{n^2 h_n^6} \sum_{i=1}^{n} \sum_{j=1}^{n} K''\left(\frac{x - X_i}{h_n}\right) K''\left(\frac{x - X_j}{h_n}\right) \right) \mathrm{d}x
\end{aligned}$$

$$= \underbrace{\frac{1}{n^2 h_n^6} \sum_{i=1}^{n} \mathbb{E} \int_{\mathbb{R}} \left[K'' \left(\frac{x - X_i}{h_n} \right) \right]^2 \mathrm{d}x}_{Q_1^n =}$$

$$+ \underbrace{\frac{1}{n^2 h_n^6} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{E} \int_{\mathbb{R}} K'' \left(\frac{x - X_i}{h_n} \right) K'' \left(\frac{x - X_j}{h_n} \right) \mathrm{d}x}_{Q_2^n =} . \tag{1.13}$$

By using the substitution $u = \frac{x - X_i}{h_n}$, where $i = 1, \ldots, n$, it holds that

$$\int_{\mathbb{R}} \left[K'' \left(\frac{x - X_i}{h_n} \right) \right]^2 \mathrm{d}x = h_n \int_{\mathbb{R}} [K''(u)]^2 \mathrm{d}u = h_n R(K''). \tag{1.14}$$

Therefore, one obtains the following equation

$$Q_1^n = \frac{1}{n^2 h_n^6} \sum_{i=1}^{n} \mathbb{E} \int_{\mathbb{R}} \left[K'' \left(\frac{x - X_i}{h_n} \right) \right]^2 \mathrm{d}x \overset{(1.14)}{=} \frac{R(K'')}{n h_n^5}. \tag{1.15}$$

Since $X_1, \ldots, X_n$ are independent and identically distributed random variables, the second term Q_2^n in (1.13) can be expressed as follows

$$Q_2^n = \frac{(n-1)}{n h_n^6} \int_{\mathbb{R}} \left[\mathbb{E} K'' \left(\frac{x - X_i}{h_n} \right) \right]^2 \mathrm{d}x$$

$$= \frac{(n-1)}{n h_n^4} \int_{\mathbb{R}} \left[\underbrace{\int_{\mathbb{R}} K''(u) f(x - h_n u) \, \mathrm{d}u}_{Q_3^n(x) =} \right]^2 \mathrm{d}x. \tag{1.16}$$

The lemma assumes that f is four times continuously differentiable. Therefore, by employing the Lagrange form of the remainder (see Bartle and Sherbert [2011], Theorem 6.4.1), one obtains that for sufficiently large n.

$$Q_3^n(x) = f(x) \int_{\mathbb{R}} K''(u) \mathrm{d}u - h_n f'(x) \int_{\mathbb{R}} u K''(u) \mathrm{d}u + \frac{h_n^2 f''(x)}{2!} \int_{\mathbb{R}} u^2 K''(u) \mathrm{d}u$$

$$- \frac{h_n^3 f^{(3)}(x)}{3!} \int_{\mathbb{R}} u^3 K''(u) \mathrm{d}u + \frac{h_n^4}{4!} \int_{\mathbb{R}} u^4 K''(u) f^{(4)}(\xi_{x,n}^u) \mathrm{d}u, \tag{1.17}$$

where $\xi_{x,n}^u$ is between x and $x - h_n u$. Since K is an even function, then K' is an odd function and K'' is an even function. Hence $\int_{\mathbb{R}} u K''(u) \mathrm{d}u = \int_{\mathbb{R}} u^3 K''(u) \mathrm{d}u = 0$. Furthermore, with the help of conditions on K, note that

$$\int_{\mathbb{R}} K''(u) \mathrm{d}u = \int_{-1}^{1} K''(u) \mathrm{d}u = K'(1) - K'(-1) = 0,$$

$$\int_{\mathbb{R}} u^2 K''(u) \mathrm{d}u = \int_{-1}^{1} u^2 K''(u) \mathrm{d}u = [u^2 K'(u)]_{-1}^{1} - 2 \int_{-1}^{1} u K'(u) \mathrm{d}u$$

$$= [u^2 K'(u)]_{-1}^{1} - 2[u K(u)]_{-1}^{1} + 2 \int_{-1}^{1} K(u) \mathrm{d}u = 2.$$

Overall with the help of (1.17), one obtains the following

$$Q_3^n(x) = h_n^2 f''(x) + \frac{h_n^4}{4!}\int_{-1}^{1} u^4 K''(u) f^{(4)}(\xi_{x,n}^u)\mathrm{d}u.$$

Note that it is possible to rewrite

$$\int_{\mathbb{R}} \Big[Q_3^n(x)\Big]^2 \mathrm{d}x = h_n^4 R(f'') + \int_{\mathbb{R}} a_n(x)\mathrm{d}x + \int_{\mathbb{R}} b_n(x)\mathrm{d}x, \tag{1.18}$$

where the functions $a_n(x)$ and $b_n(x)$ are defined as

$$a_n(x) = \frac{h_n^6 f''(x)}{12}\int_{-1}^{1} u^4 K''(u) f^{(4)}(\xi_{x,n}^u)\mathrm{d}u,$$

$$b_n(x) = \frac{h_n^8}{(4!)^2}\left[\int_{-1}^{1} u^4 K''(u) f^{(4)}(\xi_{x,n}^u)\mathrm{d}u\right]^2.$$

First, by employing Hölder's inequality, one can bound $a_n(x)$ as

$$\begin{aligned}
|a_n(x)| &\leq \frac{h_n^6|f''(x)|}{12}\int_{-1}^{1} u^4 \Big|K''(u) f^{(4)}(\xi_{x,n}^u)\Big|\mathrm{d}u \\
&\overset{\text{Hölder}}{\leq} \frac{h_n^6|f''(x)|}{12}\left(\int_{-1}^{1} u^8 \Big[K''(u)\Big]^2 \mathrm{d}u\right)^{\frac{1}{2}}\left(\int_{-1}^{1}\Big[f^{(4)}(\xi_{x,n}^u)\Big]^2\mathrm{d}u\right)^{\frac{1}{2}} \\
&\leq \frac{h_n^6|f''(x)|}{12}\left(\int_{-1}^{1} u^8 \Big[K''(u)\Big]^2 \mathrm{d}u\right)\left(\int_{-1}^{1}\Big[f^{(4)}(\xi_{x,n}^u)\Big]^2\mathrm{d}u\right) \\
&\leq \frac{h_n^6|f''(x)|}{12}\underbrace{\sup_{u\in[-1,1]}\Big[K''(u)\Big]^2}_{<\infty \text{ by assump. on } K}\left(\int_{-1}^{1} u^8\mathrm{d}u\right)\left(\int_{-1}^{1}\Big[f^{(4)}(\xi_{x,n}^u)\Big]^2\mathrm{d}u\right) \\
&\leq \left(\int_{-1}^{1}\Big[f^{(4)}(\xi_{x,n}^u)\Big]^2\mathrm{d}u\right)|f''(x)|O(h_n^6).
\end{aligned} \tag{1.19}$$

Note that for sufficiently large n, it holds for all real x and u

$$\xi_{x,n}^u \in \Big(\min\{x-\delta_0 u, x\}, \max\{x-\delta_0 u, x\}\Big).$$

Therefore, with the help of (1.19), one obtains the following

$$\begin{aligned}
|a_n(x)| &\leq \left(\int_{-1}^{1}\sup_{\delta\in[-\delta_0,\delta_0]}\Big[f^{(4)}(x+u\delta)\Big]^2\mathrm{d}u\right)|f''(x)|O(h_n^6) \\
&\leq \left(2\sup_{\eta\in[-\delta_0,\delta_0]}\Big[f^{(4)}(x+\eta)\Big]^2\right)|f''(x)|O(h_n^6) \\
&\leq \left(\sup_{\eta\in[-\delta_0,\delta_0]}\Big[f^{(4)}(x+\eta)\Big]^2\right)\underbrace{\left(\sup_{x\in\mathbb{R}}|f''(x)|\right)}_{<\infty \text{ by assump. (1.11)}}O(h_n^6) \\
&\leq \left(\sup_{\eta\in[-\delta_0,\delta_0]}\Big[f^{(4)}(x+\eta)\Big]^2\right)O(h_n^6).
\end{aligned} \tag{1.20}$$

Now, by using Jensen's inequality, bound $b_n(x)$

$$
\begin{aligned}
|b_n(x)| &\overset{\text{Jensen}}{\leq} \frac{h_n^8}{(4!)^2}\int_{-1}^{1} u^8\Big[K''(u)\Big]^2\Big[f^{(4)}(\xi_{x,n}^u)\Big]^2\mathrm{d}u \\
&\leq \frac{h_n^8}{(4!)^2}\underbrace{\sup_{u\in[-1,1]}\Big[K''(u)\Big]^2}_{<\infty \text{ by assump. on } K}\int_{-1}^{1}\Big[f^{(4)}(\xi_{x,n}^u)\Big]^2\mathrm{d}u \\
&\leq \left(\int_{-1}^{1}\sup_{\delta\in[-\delta_0,\delta_0]}\Big[f^{(4)}(x+u\delta)\Big]^2\mathrm{d}u\right)O(h_n^8) \\
&\leq \left(\sup_{\eta\in[-\delta_0,\delta_0]}\Big[f^{(4)}(x+\eta)\Big]^2\right)O(h_n^8). \qquad (1.21)
\end{aligned}
$$

From (1.12), (1.20) and (1.21), one can get that

$$
\begin{aligned}
&\left|\int_{\mathbb{R}} a_n(x)\mathrm{d}x + \int_{\mathbb{R}} b_n(x)\mathrm{d}x\right| \leq \int_{\mathbb{R}}|a_n(x)|\mathrm{d}x + \int_{\mathbb{R}}|b_n(x)|\mathrm{d}x \\
&\leq O(h_n^6) + O(h_n^8) = O(h_n^6). \qquad (1.22)
\end{aligned}
$$

Thus, by employing (1.18) and (1.22), one obtains that

$$
\int_{\mathbb{R}}\Big[Q_3^n(x)\Big]^2\mathrm{d}x = h_n^4 R(f'') + O(h_n^6). \qquad (1.23)
$$

Therefore, from (1.16) and (1.23), it follows that

$$
\begin{aligned}
Q_2^n &= \frac{n-1}{nh_n^4}\int_{\mathbb{R}}\Big[Q_3^n(x)\Big]^2\mathrm{d}x = \frac{n-1}{nh_n^4}\Big(h_n^4 R(f'') + O(h_n^6)\Big) \\
&= \frac{n-1}{n}R(f'') + \frac{n-1}{n}O(h_n^2) = R(f'') + O(h_n^2), \qquad (1.24)
\end{aligned}
$$

where the last equation holds because of the assumption $(nh_n^2)^{-1} = O(1)$. Overall, by combining (1.13), (1.15) and (1.24), one can derive the statement of the Lemma

$$
\mathbb{E}R\left(\widehat{f}_n''\right) = R(f'') + \frac{R(K'')}{nh_n^5} + O(h_n^2).
$$

□

Remark.

- According to the previous lemma, it holds that the bias of $R(\widehat{f}_n'')$ with respect to $R(f'')$ is

  $$
  \mathbb{E}R(\widehat{f}_n'') - R(f'') = \frac{R(K'')}{nh_n^5} + O(h_n^2).
  $$

  If h_n decreases towards zero sufficiently quickly, more precisely, it must hold that $h_n = o(n^{-\frac{1}{7}})$. Then, for sufficiently large n, the term $\frac{R(K'')}{nh_n^5}$ would dominate over the term $O(h_n^2)$.

- Note that for the asymptotically optimal global bandwidth, defined in (1.5), it holds that $h_n^{(opt)} = O(n^{-\frac{1}{5}})$. Hence, $h_n^{(opt)}$ satisfies the conditions of the lemma.

The analogy of the lemma can be proved for higher-order derivatives, see Scott and Terrell [1987](Lemma 3.2). However, the authors of that paper do not discuss the rate of convergence of the residual term. Therefore, they do not consider conditions (1.11) and (1.12). The natural question is which densities satisfy those conditions. It is worth noting that densities that are four times continuously differentiable and have bounded support satisfy conditions (1.11) and (1.12).

Assume that f is the density of $\mathsf{N}(0,1)$. It is evident that f is 4 times continuously differentiable and satisfies condition (1.11). Now, it will be shown that f also satisfies the second condition (1.12). It could be calculated that

$$f^{(4)}(x) = e^{-\frac{x^2}{2}} p_4(x),$$

where $p_4(x)$ is a polynomial of the fourth degree. Note that there exist $C > 0$ and $K > 0$ such that for all $x \geq K$, it holds that

$$\begin{aligned} \sup_{\eta\in[-\delta_0,\delta_0]} \left[f^{(4)}(x+\eta)\right]^2 &= \sup_{\eta\in[-\delta_0,\delta_0]} \left|e^{-(x+\eta)^2} p_4^2(x+\eta)\right| \\ \leq C \sup_{\eta\in[-\delta_0,\delta_0]} \left|e^{-(x+\eta)^2}(x+\eta)^8\right| &\leq C e^{-(x-\delta_0)^2}(x+\delta_0)^8 \end{aligned} \tag{1.25}$$

and for all $x \leq -K$, it holds that

$$\sup_{\eta\in[-\delta_0,\delta_0]} \left[f^{(4)}(x+\eta)\right]^2 \leq C e^{-(x+\delta_0)^2}(x-\delta_0)^8. \tag{1.26}$$

Therefore, with the help of (1.25) and (1.26), one obtains the following

$$\begin{aligned} \int_{\mathbb{R}} \sup_{\eta\in[-\delta_0,\delta_0]} \left[f^{(4)}(x+\eta)\right]^2 \mathrm{d}x &\leq 2K \sup_{|x|\leq K} \sup_{\eta\in[-\delta_0,\delta_0]} \left[f^{(4)}(x+\eta)\right]^2 \\ + C\int_K^{\infty} e^{-(x-\delta_0)^2}(x+\delta_0)^8 \mathrm{d}x + C\int_{-\infty}^{-K} e^{-(x+\delta_0)^2}(x-\delta_0)^8 \mathrm{d}x &< \infty. \end{aligned}$$

The result of Lemma 3 is that $R(\widehat{f}_n'')$ is an asymptotically positively biased estimator of $R(f'')$. That is why, as an approximation of $R(f'')$, one will use $R(\widehat{f}_n'') - \frac{R(K'')}{nh_n^5}$ instead. Therefore, in this method, the bandwidth is selected as

$$h_n^{(BCV)} = \arg\min_{h_n>0} \mathcal{B}(h_n), \tag{1.27}$$

where

$$\mathcal{B}(h_n) = \frac{R(K)}{nh_n} + \frac{1}{4}h_n^4\mu_2^2\left[R(\widehat{f}_n'') - \frac{R(K'')}{nh_n^5}\right] \tag{1.28}$$

is the estimated counterpart of AMISE. At the beginning of the chapter, it is mentioned that biased cross-validation combines elements of both cross-validation and 'plug-in' bandwidth selection. It is apparent how the 'plug-in' method was

employed in formula (1.28). However, it is not entirely clear from (1.28) at which step the cross-validation approach was applied. From the proof of Lemma 3 one can deduce that

$$R(\widehat{f}_n'') - \frac{R(K'')}{nh_n^5} = \frac{1}{n^2h_n^5}\sum_{i=1}^{n}\sum_{j\neq i}\int_{\mathbb{R}} K''(u)K''\left(u + \frac{X_i - X_j}{h_n}\right)\mathrm{d}u,$$

which could be interpreted as leave-out-diagonals cross-validation.

The next theorem demonstrates the asymptotic equivalence between biased cross-validation and the asymptotically optimal global bandwidth. For the purpose of the next theorem, it will be useful to define

$$h_{n,b}^{(BCV)} = \underset{0<h_n<bh_n^{(opt)}}{\arg\min}\ \mathcal{B}(h_n), \tag{1.29}$$

where $b > 1$ and $h_n^{(opt)}$ is the asymptotically optimal global bandwidth defined in (1.5).

Theorem 4. *Let $X_1, \ldots, X_n$ be a random sample with common probability density function $f(x)$. Suppose that the following conditions hold:*

(C1) $f^{(3)}$ is absolutely continuous, $f^{(4)}$ is integrable, and also assume that

$$R\Big(f^{(4)}(f)^{\frac{1}{2}}\Big) < \infty \quad \textit{and} \quad R\Big((f^{(4)})^{\frac{1}{2}}f\Big) < \infty.$$

(C2) $K \geq 0$ symmetric on $[-1, 1]$, K' is Hölder continuous and $\mu_2 < \infty$.

(C3) K'' is absolutely continuous, $K^{(3)}$ is continuous and $R(K^{(3)}) < \infty$.

Then, for each $b > 1$, it holds that

$$\frac{h_{n,b}^{(BCV)}}{h_n^{(opt)}} \xrightarrow[n\to\infty]{P} 1.$$

The proof is given in Scott and Terrell [1987] (Corollary 3.2).

Remark.

- Note that Theorem 4 presents a significantly different result compared to Theorem 2. The latter asserts that the bandwidth selected by the least squares cross-validation $h_n^{(LSCV)}$ is asymptotically optimal in the sense of minimizing integrated square error. In contrast, Theorem 4 states that the bandwidth chosen by the biased cross-validation $h_{n,b}^{(BCV)}$ is asymptotically equal to $h_n^{(opt)}$. Another thing to note is that $h_{n,b}^{(BCV)}$ minimizes $\mathcal{B}(h_n)$ over the finite interval $(0, bh_n^{(opt)})$, where $b > 1$. On the other hand, $h_n^{(LSCV)}$ minimizes $\mathcal{L}(h_n)$ over $(0, \infty)$.

- Note that the conditions on the kernel function from Theorem 4 are stronger than those from Theorem 2. For example, the Epanechnikov kernel satisfies all the conditions *(B1)*$-$*(B4)* from Theorem 2. However, the simplest kernel that satisfies conditions *(C2)* and *(C3)* is the triweight kernel, defined as follows

$$K(t) = \frac{35}{32}(1 - t^2)^3\mathbb{I}_{[-1,1]}(t).$$

Additionally, both the triweight kernel and the Epanechnikov kernel are functions that belong to the so-called 'symmetric Beta family', which is the following set

$$\left\{ \frac{1}{\text{Beta}(1/2, \gamma+1)} \left[(1-t^2)_+ \right]^\gamma : \gamma \in \mathbb{R}_+ \right\}. \tag{1.30}$$

The subscript + denotes the positive part, and $\text{Beta}(\cdot,\cdot)$ represents a Beta function. As noted in Marron and Nolan [1988], Section 4, the standard normal kernel does not belong to (1.30), but it is obtained by taking the limit $\gamma \to \infty$.

1.3 Simulation Study

In this section, different bandwidth selectors will be compared on simulated data. The underlying probability density functions chosen for data simulations include:

- The standard normal density $\mathsf{N}(0,1)$.
- The mixture of normal densities with different means

$$0.5\ \mathsf{N}(-10,1) + 0.5\ \mathsf{N}(10,1).$$

The sample sizes considered here are $n = 100$ and $n = 400$. In the simulation study, the standard normal kernel is used. Each setting involves the use of 500 Monte Carlo repetitions. Mean integrated squared errors (MISE) are compared to evaluate bandwidth selectors. All results of the simulation study are presented in Table 1.1. Each row in the first column of the table corresponds to a specific method of bandwidth selection, with labels summarized as follows

- `Optimal`: Asymptotically optimal global bandwidth, $h_n^{(opt)}$, which is defined in (1.5).
- `NR`: The normal reference rule, $h_n^{(NR)}$, as defined in (1.6).
- `MNR`: Modification of the normal reference rule, $h_n^{(MNR)}$, defined in (1.7).
- `LSCV`: Least-squares cross-validation, $h_n^{(LSCV)}$, which is defined in (1.9).
- `BCV`: Biased cross-validation, $h_n^{(BCV)}$, defined in (1.27).

The simulations were conducted using statistical software R Core Team [2023a]. For kernel density estimation, the density function from the package R Core Team [2023b] was used. One of the arguments of this function allows for choosing an appropriate method for bandwidth selection. Note that `Optimal`, `NR`, and `MNR` can be computed straightforwardly. However, to use `LSCV` and `BCV`, one should specify a finite interval over which cross-validation is conducted. All cross-validation results from Table 1.1 are based on the following interval $(0.1\ \widehat{h}_{\max}, \widehat{h}_{\max})$, where $\widehat{h}_{\max}$ is given by

$$\widehat{h}_{\max} = 1.144\ n^{-\frac{1}{5}}\ S_n,$$

and $S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2}$ represents the sample standard deviation. This interval $(0.1\ \widehat{h}_{\max}, \widehat{h}_{\max})$ is recommended by default settings within the density function.

1.3.1 Monte Carlo

By using Fubini's theorem, it holds that

$$\mathrm{MISE}(\widehat{f}_n) = \int_{\mathbb{R}} \mathrm{MSE}(\widehat{f}_n(x))\mathrm{d}x \overset{\text{Fubini}}{=} \mathbb{E}\int_{\mathbb{R}} (\widehat{f}_n(x) - f(x))^2 \mathrm{d}x = \mathbb{E}\Big[\mathrm{ISE}(\widehat{f}_n)\Big].$$

One may observe that $\mathrm{MISE}(\widehat{f}_n)$ can be approximated using the law of large numbers because it can be represented as an expectation.

To estimate $\mathrm{MISE}(\widehat{f}_n)$, where n is fixed, one will require a Monte Carlo approach. Consider random samples $X_1^b, \ldots, X_n^b$ from a standard normal distribution or a mean mixture distribution, which are independent for $b = 1, \ldots, B$. In the simulation study, it is considered that $n = 100$ (or $n = 400$) and $B = 500$. As the kernel function, the standard normal kernel will be considered. The kernel density estimator from the b-th sample will be denoted as $\widehat{f}_n^b$. The integrated squared error of $\widehat{f}_n^b$ has the following form

$$\mathrm{ISE}(\widehat{f}_n^b) = \int_{\mathbb{R}} \Big[\widehat{f}_n^b(x) - f(x)\Big]^2 \mathrm{d}x. \tag{1.31}$$

It is obvious that $\mathrm{ISE}(\widehat{f}_n^b)$ is random, i.e., it depends on the realization of the b-th random sample. To approximate (1.31), the `integrate` function from the package R Core Team [2023b] was used.

After applying the strong law of large numbers one can get

$$\overline{\mathrm{ISE}}_B = \frac{1}{B}\sum_{b=1}^{B} \mathrm{ISE}(\widehat{f}_n^b) \xrightarrow[B\to\infty]{\text{a.s.}} \mathrm{MISE}(\widehat{f}_n).$$

Analogously, by the law of large numbers, it holds that

$$\widehat{\sigma}_B^2 = \frac{1}{B-1}\sum_{b=1}^{B} \Big(\mathrm{ISE}(\widehat{f}_n^b) - \overline{\mathrm{ISE}}_B\Big)^2 \xrightarrow[B\to\infty]{\text{a.s.}} \mathrm{var}\Big[\mathrm{ISE}(\widehat{f}_n)\Big].$$

Then, by combining the Central Limit Theorem and Cramér-Slucky Theorem, one can obtain

$$\frac{\sqrt{B}\left[\overline{\mathrm{ISE}}_B - \mathrm{MISE}(\widehat{f}_n)\right]}{\widehat{\sigma}_B} \xrightarrow[B\to\infty]{D} \mathsf{N}(0,1).$$

This asymptotic result justifies a $100(1-\alpha)\%$ confidence interval for $\mathrm{MISE}(\widehat{f}_n)$

$$\left(\overline{\mathrm{ISE}}_B \pm \frac{u(1-\frac{\alpha}{2})\widehat{\sigma}_B}{\sqrt{B}}\right), \tag{1.32}$$

where $u(1-\frac{\alpha}{2})$ represents the $(1-\frac{\alpha}{2})$-quantile of $\mathsf{N}(0,1)$. The third column in Table 1.1, $100(1-\alpha)\%$ *Monte Carlo Confidence Interval of MISE*, corresponds to the confidence interval (1.32).

1.3.2 Discussion of results

From Table 1.1, one observes that in all cases, the lowest Monte Carlo approximation of MISE is provided by the asymptotically optimal global bandwidth $h_n^{(opt)}$. In practical applications, it is impossible to calculate $h_n^{(opt)}$ as one does not

know the true density $f(x)$. The next observation is that if the true distribution is standard normal, biased cross-validation performs better than least-squares cross-validation. This is evident as the confidence intervals for MISE when the `BCV` method is employed are disjointly shifted downward from those of the `LSCV` method. However, if the true distribution is a mixture of normal distributions, then conversely, `LSCV` performs better than `BCV`. Also, if the underlying distribution is a mixture of normal distributions, then it can be observed that least-squares cross-validation is the best choice among all the provided data-driven methods. Note that its confidence intervals are more similar to the optimal choice for a larger sample size (see the results in Table 1.1 for n=400). In the case when the underlying distribution is standard normal, one observes that the normal reference rule performs as well as biased cross-validation because their confidence intervals for MISE are similar.

If the underlying distribution is a mixture of normal distributions, it results in a bimodal distribution. This bimodal distribution deviates significantly from normality. As a consequence, the performance of the normal reference rule is very poor, as expected. Additionally, $h_n^{(\mathrm{MNR})}$ exhibits better performance on MISE than $h_n^{(\mathrm{NR})}$ because their confidence intervals for MISE do not overlap. However, if the underlying distribution is standard normal, the `MNR` method does not provide such significant an improvement. In general, one can see that the results of the simulation study correspond to the previous discussions.

In the fourth column of Table 1.1, median values of bandwidths calculated from simulations are presented. It can be seen that if data were generated from a standard normal distribution, then median values of bandwidths for the `Optimal`, `NR`, `LSCV`, and `BCV` methods are quite similar. However, the median values of bandwidths for the `MNR` method are shifted downward compared to the `Optimal` method. This indicates that density estimates with bandwidths selected by the `MNR` method generally under-smooth.

Additionally, if data were generated from a mixture of normal distributions, a big difference can be observed in the median values of bandwidths between the `Optimal` method and the normal reference rule methods, such as the `NR` and `MNR` methods. In other words, density estimates computed from normal reference rule methods tend to over-smooth.

Density function	Monte Carlo Approx. of MISE	95% Monte Carlo Conf. Interval of MISE	Median of bandwidths
STANDARD NORMAL			
$n = 100$			
Optimal	**5.439**	**(5.055, 5.823)**	0.422
NR	5.914	(5.518, 6.310)	0.402
MNR	6.543	(6.129, 6.957)	0.342
LSCV	7.731	(7.053, 8.409)	0.429
BCV	5.720	(5.324, 6.116)	0.447
$n = 400$			
Optimal	**1.948**	**(1.836, 2.060)**	0.320
NR	2.016	(1.902, 2.130)	0.314
MNR	2.189	(2.075, 2.303)	0.267
LSCV	2.596	(2.408, 2.784)	0.331
BCV	2.016	(1.900, 2.132)	0.338
MIXTURE OF NORMAL DISTRIBUTIONS			
$n = 100$			
Optimal	**5.608**	**(5.341, 5.875)**	0.484
NR	84.687	(84.638, 84.736)	4.245
MNR	76.098	(76.039, 76.157)	3.604
LSCV	*5.949*	*(5.675, 6.223)*	*0.558*
BCV	88.304	(88.257, 88.351)	4.563
$n = 400$			
Optimal	**1.908**	**(1.814, 2.002)**	0.367
NR	69.504	(69.473, 69.535)	3.214
MNR	59.972	(59.935, 60.009)	2.729
LSCV	*1.992*	*(1.892, 2.092)*	*0.394*
BCV	11.524	(9.405, 13.643)	0.427

Table 1.1: Results are based on 500 Monte Carlo replications. The underlying generative mechanism includes the STANDARD NORMAL distribution $\mathsf{N}(0, 1)$ and the MIXTURE OF NORMAL DISTRIBUTIONS $0.5\ \mathsf{N}(-10, 1) + 0.5\ \mathsf{N}(10, 1)$. In each case, kernel density estimators were fitted using a standard normal kernel. For better presentation, all results were multiplied by 1000.

2. Kernel regression

In the previous chapter, the task of estimating the probability density function using the kernel density estimator was studied. Another common scenario where kernel smoothing concepts find application is in the context of regression problems. In this chapter, it is assumed that $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ are independent and identically distributed random vectors. Also, it is assumed that the observations satisfy the following model for all $i = 1, \ldots, n$

$$Y_i = m(X_i) + \varepsilon_i, \tag{2.1}$$

where $m(x) = \mathbb{E}[Y_1 | X_1 = x]$ for $x \in \mathbb{R}$ and conditionally on $\mathbb{X} = (X_1, \ldots, X_n)$, error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed. The first two conditional moments satisfy $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\mathrm{var}[\varepsilon_i | X_i] < \infty$.

The natural idea would be to estimate m using linear regression, which implies that m belongs to a specific parametric family. Here, as in Chapter 1, the focus will be on nonparametric approaches, i.e., no parametric form for m will be assumed. This method leads to what is commonly known as *nonparametric regression*. Nonparametric approaches for estimating the conditional mean function, m, are valuable when the relationship between the covariate and the response variable is complex. Currently, there are various methods available for addressing the nonparametric regression problem. Among the commonly favored approaches are those centered around kernel functions, spline functions, and wavelets. In this chapter, the focus will be on those methods based on kernel functions, specifically *kernel regression*, due to their close relationship to kernel density estimation.

This chapter is based on Chapters 3 and 4 of Fan and Gijbels [1996] and the course notes Nagy and Omelka [2024].

2.1 Framework for local polynomial regression

In the context of kernel regression, the traditional approach is *local polynomial regression*. This approach locally fits a polynomial at a particular point $x \in \mathbb{R}$. Suppose that K is some prespecified kernel function. Define vector $\widehat{\boldsymbol{\beta}}(x) = \left(\widehat{\beta}_0(x), \ldots, \widehat{\beta}_p(x)\right)^T$ as the minimizer of the following quantity

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j \right]^2 K\left(\frac{X_i - x}{h_n}\right), \tag{2.2}$$

where h_n is a bandwidth controlling the size of the local neighborhood, and the terms $K\left(\frac{X_i - x}{h_n}\right)$ represent the weights assigned to each data point. In Fan and Gijbels [1996], it is shown on page 58 that the optimization task (2.2) could be reformulated as the weighted least squares problem.

The form of the optimization problem (2.2) arises from the following motivation. Assume that the conditional mean function m is sufficiently smooth. Then, for a point y in the neighborhood of x, the following approximation holds

$$m(y) \approx \sum_{j=0}^{p} b_j (y - x)^j. \tag{2.3}$$

Therefore, the main interest is in the estimate of the intercept $\widehat{\beta}_0(x)$ since it estimates the conditional mean $m(x)$. This interest arises from (2.3), where it is observed that $m(x) \approx b_0$. Instead of $\widehat{\beta}_0(x)$, the notation $\widehat{m}_p(x)$ will be used, indicating the order of local polynomial regression.

From the expression (2.2), it is evident that the local polynomial estimator is influenced by the following factors:

1. the kernel function K,
2. the degree of polynomial approximation p,
3. the choice of bandwidth h_n.

In terms of the selection of the kernel function K, Theorem 3.4 of Fan and Gijbels [1996] illustrates that the Epanechnikov kernel, as defined in (1.10), is optimal among symmetric and nonnegative kernels. Regarding the determination of the appropriate order p, Section 3.3.3 of Fan and Gijbels [1996] presents several data-driven procedures. Therefore, the main focus here will be primarily on the choice of the bandwidth h_n.

2.1.1 Asymptotic representation of bias and variance

In the next section, the bandwidth selection problem will be discussed. Here, as in Chapter 1, criteria for bandwidth selection will be based on MSE. Therefore, gaining insights about the bias and variance of the defined estimators is necessary. Here, one would need the conditional variance function denoted as

$$\sigma^2(x) = \operatorname{var}[Y_1|X_1 = x] \quad \text{for } x \in \mathbb{R}.$$

In Theorem 3.1 of Fan and Gijbels [1996], it is demonstrated that given certain regularity assumptions on $m(\cdot)$, $\sigma^2(\cdot)$, the marginal density $f_X(\cdot)$, and the kernel K, for every $p \in \mathbb{N}_0$, there exist constants V_p such that the following relation holds

$$\operatorname{var}\Big(\widehat{m}_p(x)|\mathbb{X}\Big) = \frac{V_p \sigma^2(x)}{f_X(x) n h_n} + o_P\left(\frac{1}{n h_n}\right), \tag{2.4}$$

where $V_0 = V_1 < V_2 = V_3 < V_4 = V_5 < \ldots$, and so forth. Moreover, for odd values of p, there are constants B_p such that the asymptotic conditional bias is expressed as

$$\operatorname{bias}\Big(\widehat{m}_p(x)|\mathbb{X}\Big) = B_p h_n^{p+1} m^{(p+1)}(x) + o_P(h_n^{p+1}). \tag{2.5}$$

Conversely, for even values of p, there exist constants $\widetilde{B}_p$ such that the conditional bias takes the form

$$\begin{aligned} &\operatorname{bias}\Big(\widehat{m}_p(x)|\mathbb{X}\Big) \\ &= \widetilde{B}_p h_n^{p+2} \left[m^{(p+2)}(x) + (p+2) m^{(p+1)}(x) \frac{f_X'(x)}{f_X(x)} \right] + o_P(h_n^{p+2}). \end{aligned} \tag{2.6}$$

One can observe that if p is even, then it holds that the conditional biases of $\widehat{m}_p(x)$ and $\widehat{m}_{p+1}(x)$ are both of the same order $O_P(h_n^{p+2})$. However, the conditional bias of $\widehat{m}_{p+1}(x)$ has a simpler structure compared to that of $\widehat{m}_p(x)$. Regarding the

conditional variance from (2.4), it can be observed that the main term remains the same for p and $p+1$ if p is even. Therefore, when p is even, it is reasonable to increase the order from p to $p+1$. This adjustment preserves the conditional variance without increasing it, while also potentially reducing bias.

If one wants to explicitly express constants V_p, B_p, and $\widetilde{B}_p$, it will be useful to introduce the following notation

$$\mu_j = \int_{\mathbb{R}} y^j K(y)\mathrm{d}y \quad \text{and} \quad \nu_j = \int_{\mathbb{R}} y^j K^2(y)\mathrm{d}y, \tag{2.7}$$

where $j \in \{0, 1, \ldots, p\}$. Additionally, one would need the following matrices and vectors

$$\mathbb{S}_p = (\mu_{j+l})_{0\le j,l\le p} \qquad \boldsymbol{c}_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$$
$$\mathbb{S}_p^* = (\nu_{j+l})_{0\le j,l\le p} \qquad \widetilde{\boldsymbol{c}}_p = (\mu_{p+2}, \ldots, \mu_{2p+2})^T.$$

Then the following holds (also see Fan and Gijbels [1996], Theorem 3.1)

$$V_p = \boldsymbol{e}_1^T \mathbb{S}_p^{-1} \mathbb{S}_p^* \mathbb{S}_p^{-1} \boldsymbol{e}_1, \quad \text{where} \quad \boldsymbol{e}_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{p+1}, \tag{2.8}$$

$$B_p = \frac{\boldsymbol{e}_1^T \mathbb{S}_p^{-1} \boldsymbol{c}_p}{(p+1)!} \quad \text{for } p \text{ odd}, \tag{2.9}$$

$$\widetilde{B_p} = \frac{\boldsymbol{e}_1^T \mathbb{S}_p^{-1} \widetilde{\boldsymbol{c}}_p}{(p+2)!} \quad \text{for } p \text{ even}. \tag{2.10}$$

2.2 Bandwidth selection

2.2.1 Asymptotic approximation

Bandwidth selection is a crucial aspect of how accurately kernel regression would perform. A bandwidth that is too large may over-smooth the regression function, while a bandwidth that is too small may under-smooth the regression function. In this section, similarly to Chapter 1, the mean squared error will be used to assess the accuracy of a kernel regression. In practical applications, it is popular to choose p small due to the simpler expression of conditional bias and variance. Therefore, the local linear estimator $\widehat{m}_1(x)$ will be employed for estimating $m(x)$. Also, the choice of $\widehat{m}_1(x)$ is justified due to its simpler bias structure, compared to $\widehat{m}_0(x)$, as shown in (2.5) and (2.6).

First, note that expressions for the conditional variance in (2.4) and conditional biases (2.5) and (2.6) are derived under some regularity assumptions on K. Specifically, it is assumed that the kernel K is bounded, symmetric around zero, positive on its bounded support $(-1, 1)$, and such that $\int_{-1}^{1} K(t)\mathrm{d}t = 1$. Therefore, it holds that $\mu_0 = 1$, and for odd p, it holds that $\mu_p = 0$, where μ_p is defined in (2.7).

To obtain explicit expressions for the conditional bias and variance of $\widehat{m}_1$, it is needed to calculate V_1 from (2.8) and B_1 from (2.9). Note that (2.4) reveals that $V_0 = V_1$, hence V_0 can be calculated instead. Therefore

$$V_0 = \mathbb{S}_0^{-1} \mathbb{S}_0^* \mathbb{S}_0^{-1} = \frac{\nu_0}{\mu_0^2} = \nu_0, \text{ where } \mathbb{S}_0 = \mu_0 \text{ and } \mathbb{S}_0^* = \nu_0.$$

Thus, from (2.4), one obtains that

$$\operatorname{var}\Big(\widehat{m}_1(x)|\mathbb{X}\Big) = \frac{\sigma^2(x)\nu_0}{f_X(x)nh_n} + o_P\left(\frac{1}{nh_n}\right). \tag{2.11}$$

On the other hand, in order to determine the constant B_1, knowledge of $\mathbb{S}_1^{-1}$ and $\boldsymbol{c}_1$ is required, where

$$\mathbb{S}_1 = \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{c}_1 = \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} \mu_2 \\ 0 \end{pmatrix}.$$

Thus, it follows that

$$\mathbb{S}_1^{-1} = \frac{1}{\mu_2}\begin{pmatrix} \mu_2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Consequently, the expression for B_1 can be derived as

$$B_1 = \frac{\boldsymbol{e}_1^T\mathbb{S}_1^{-1}\boldsymbol{c}_1}{2!} = \frac{1}{2\mu_2}\begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} \mu_2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \mu_2 \\ 0 \end{pmatrix} = \frac{\mu_2}{2}.$$

Therefore, according to (2.5), the conditional bias of $\widehat{m}_1$ can be expressed as

$$\operatorname{bias}\Big(\widehat{m}_1(x)|\mathbb{X}\Big) = h_n^2\mu_2\frac{m''(x)}{2} + o_P(h_n^2). \tag{2.12}$$

It is known that the conditional mean squared error can be decomposed as follows

$$\operatorname{MSE}\Big(\widehat{m}_1(x)|\mathbb{X}\Big) = \Big[\operatorname{bias}\Big(\widehat{m}_1(x)|\mathbb{X}\Big)\Big]^2 + \operatorname{var}\Big(\widehat{m}_1(x)|\mathbb{X}\Big).$$

By using (2.11) and (2.12), one gets

$$\begin{aligned}\operatorname{MSE}\Big(\widehat{m}_1(x)|\mathbb{X}\Big) &= \left[h_n^2\mu_2\frac{m''(x)}{2} + o_P(h_n^2)\right]^2 + \frac{\sigma^2(x)\nu_0}{f_X(x)nh_n} + o_P\left(\frac{1}{nh_n}\right) \\ &= \frac{1}{nh_n}\frac{\sigma^2(x)\nu_0}{f_X(x)} + \frac{1}{4}h_n^4\left[m''(x)\right]^2\mu_2^2 + o_P\left(\frac{1}{nh_n}\right) + o_P(h_n^4).\end{aligned}$$

Ignoring $o_P(\cdot)$ terms, one could get that the *asymptotic mean squared error* of $\widehat{m}_1(x)$ is given by

$$\operatorname{AMSE}\Big(\widehat{m}_1(x)|\mathbb{X}\Big) = \frac{1}{nh_n}\frac{\sigma^2(x)\nu_0}{f_X(x)} + \frac{1}{4}h_n^4\left[m''(x)\right]^2\mu_2^2. \tag{2.13}$$

The asymptotic mean squared error $\operatorname{AMSE}\Big(\widehat{m}_1(x)|\mathbb{X}\Big)$ can be considered as an approximation of the mean squared error $\operatorname{MSE}\Big(\widehat{m}_1(x)|\mathbb{X}\Big)$. To obtain the optimal bandwidth, one could attempt to minimize (2.13) with respect to h_n. However, the problem is that such a bandwidth selector would depend on x, which would not be very useful in practical applications. Therefore, define the *asymptotic mean integrated squared error* as follows

$$\operatorname{AMISE}\Big(\widehat{m}_1|\mathbb{X}\Big) = \int_{\mathbb{R}} \operatorname{AMSE}\Big(\widehat{m}_1(x)|\mathbb{X}\Big)w(x)f_X(x)\mathrm{d}x$$

$$= \frac{\nu_0}{nh_n} \int_{\mathbb{R}} \sigma^2(x) w(x) \mathrm{d}x + \frac{1}{4} h_n^4 \mu_2^2 \int_{\mathbb{R}} [m''(x)]^2 w(x) f_X(x) \mathrm{d}x, \quad (2.14)$$

where $w(x)$ is a given weight function introduced to ensure that the integral is finite. Minimising (2.14) with respect to h_n yields the *asymptotically optimal global bandwidth*

$$h_n^{(opt)} = n^{-\frac{1}{5}} \left[\frac{\nu_0 \int_{\mathbb{R}} \sigma^2(x) w(x) \mathrm{d}x}{\mu_2^2 \int_{\mathbb{R}} [m''(x)]^2 w(x) f_X(x) \mathrm{d}x} \right]^{\frac{1}{5}}. \quad (2.15)$$

Note that $h_n^{(opt)}$ does not depend on a specific value of x. However, to calculate $h_n^{(opt)}$, one needs to integrate characteristics of the true distribution, such as $\sigma^2(x)$, $m''(x)$, and $f_X(x)$, which are typically unknown in practice. Therefore some methods are needed to circumvent this problem.

2.2.2 Rule of thumb and homoscedasticity

Suppose homoscedastic settings, which means that $\sigma^2(x) = \sigma^2 > 0$ is constant. Then the asymptotically optimal global bandwidth $h_n^{(opt)}$, defined in (2.15), can be rewritten as follows

$$h_n^{(opt)} = n^{-\frac{1}{5}} \left[\frac{\sigma^2 \nu_0 \int_{\mathbb{R}} w(x) \mathrm{d}x}{\mu_2^2 \int_{\mathbb{R}} [m''(x)]^2 w(x) f_X(x) \mathrm{d}x} \right]^{\frac{1}{5}}. \quad (2.16)$$

To estimate $m(x)$, the function $\widetilde{m}(x)$ will be used, which is obtained by fitting a standard polynomial regression of order 4.

Remark. In general, the recommended order of the fitted standard polynomial regression is $p+3$. Because $h_n^{(opt)}$ was derived from the local linear estimator, i.e., $p = 1$, therefore polynomial regression of order 4 is used.

The unknown variance σ^2 from (2.16) can be estimated by

$$\widetilde{\sigma}^2 = \frac{1}{n-5} \sum_{i=1}^{n} [Y_i - \widetilde{m}(X_i)]^2.$$

Note the following

$$\int_{\mathbb{R}} [m''(x)]^2 w(x) f_X(x) \mathrm{d}x = \mathbb{E}_X [m''(X)]^2 w(X), \quad (2.17)$$

where $\mathbb{E}_X$ denotes the expectation taken with respect to the probability distribution of the random variable X. As an estimation of the expected value in (2.17), one could simply use

$$\frac{1}{n} \sum_{i=1}^{n} [\widetilde{m}''(X_i)]^2 w(X_i).$$

This results to the following bandwidth selector

$$h_n^{(ROT)} = n^{-\frac{1}{5}} \left[\frac{\widetilde{\sigma}^2 \nu_0 \int_{\mathbb{R}} w(x) \mathrm{d}x}{\mu_2^2 \frac{1}{n} \sum_{i=1}^{n} [\widetilde{m}''(X_i)]^2 w(X_i)} \right]^{\frac{1}{5}}. \quad (2.18)$$

Note that $h_n^{(ROT)}$ does not depend on any unknown characteristics and serves as a global bandwidth selector, as it does not rely on any specific value of x.

2.2.3 Cross-validation

Recall that $\widehat{m}_p$ denotes the estimation of the conditional mean using a local polynomial regression of order p, where $p \in \mathbb{N}_0$. Let X' be a random variable with the same distribution as X_1 and independent of observations $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$. The integrated squared error of $\widehat{m}_p$ can be rewritten as

$$\begin{aligned} \mathrm{ISE}(\widehat{m}_p) &= \int_{\mathbb{R}} \Big(\widehat{m}_p(x) - m(x)\Big)^2 f_X(x) w(x) \mathrm{d}x \\ &= \mathbb{E}_{X'}\Big(\widehat{m}_p(X') - m(X')\Big)^2 w(X'). \end{aligned} \tag{2.19}$$

Note that $\mathrm{ISE}(\widehat{m}_p)$ is random since it implicitly depends on the random sample $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ through the estimator $\widehat{m}_p$. The form of the integrated squared error (2.19) motivates the following estimator

$$\mathcal{M}(h_n) = \frac{1}{n} \sum_{i=1}^{n} \Big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w(X_i),$$

where $\widehat{m}_p^{(-i)}$ is based on a sample that leaves out the i-th observation. The issue with this estimator is its dependency on the unknown conditional mean m. Therefore, instead of $\mathcal{M}(h_n)$, the following estimator will be used

$$\mathcal{CV}(h_n) = \frac{1}{n} \sum_{i=1}^{n} \Big[Y_i - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w(X_i). \tag{2.20}$$

The choice of $\mathcal{CV}(h_n)$ is justified because it is shown below that the conditional expectations $\mathbb{E}[\mathcal{CV}(h_n)|\mathbb{X}]$ and $\mathbb{E}[\mathcal{M}(h_n)|\mathbb{X}]$ are equal up to a term that does not depend on h_n.

$$\begin{aligned} \mathbb{E}[\mathcal{CV}(h_n)|\mathbb{X}] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ \Big[\varepsilon_i + m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w(X_i) \Big| \mathbb{X} \right\} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big\{ \varepsilon_i^2 w(X_i) \Big| \mathbb{X} \Big\}}_{\mathrm{I}_n =} + \underbrace{\frac{2}{n} \sum_{i=1}^{n} \mathbb{E}\Big\{ \varepsilon_i \Big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big] w(X_i) \Big| \mathbb{X} \Big\}}_{\mathrm{II}_n =} \\ &\quad + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big\{ \Big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w(X_i) \Big| \mathbb{X} \Big\}}_{\mathrm{III}_n =}. \end{aligned}$$

Obviously, I_n does not depend on h_n and $\mathrm{III}_n = \mathbb{E}[\mathcal{M}(h_n)|\mathbb{X}]$. Thus it remains to show that the second term II_n is equal to 0. Since the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are conditionally independent given $\mathbb{X}$ and ε_i was not used in the computation of $\widehat{m}_p^{(-i)}$ for all $i = 1, \ldots, n$, then one can conclude that

$$\begin{aligned} \mathrm{II}_n &= \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}\Big\{\varepsilon_i | \mathbb{X}\Big\} \mathbb{E}\Big\{ \Big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w(X_i) \Big| \mathbb{X} \Big\} \\ &= \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}\Big\{\varepsilon_i | X_i\Big\} \mathbb{E}\Big\{ \Big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w(X_i) | \mathbb{X} \Big\} = 0. \end{aligned}$$

Therefore, define the bandwidth selector $h_n^{(CV)}$ as follows

$$h_n^{(CV)} = \underset{h_n > 0}{\arg\min}\, \mathcal{CV}(h_n). \tag{2.21}$$

Remark. Suppose that $\binom{X'}{Y'}$ is a random vector with the same distribution as $\binom{X_1}{Y_1}$ and independent of observations $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$. Another important characteristic of the estimator $\widehat{m}_p$ is its ability to make predictions. Let measure the *prediction error* in the following way

$$\mathcal{R}(h_n) = \mathbb{E}_{X',Y'}\Big(Y' - \widehat{m}_p(X')\Big)^2 w(X'). \tag{2.22}$$

Note that $\mathcal{CV}(h_n)$ was derived from the integrated squared error of $\widehat{m}_p$. Alternatively, $\mathcal{CV}(h_n)$ could be seen as a natural estimator of the prediction error (2.22). Also note that in the form of the estimator $\mathcal{CV}(h_n)$, the general principle of leave-one-out cross-validation is apparent. This is because for all $i = 1, \ldots, n$, the estimator of the conditional mean $\widehat{m}_p^{(-i)}$ was 'trained' on all observations except for the i-th observation.

One could be interested in whether the bandwidth choice $h_n^{(CV)}$ is optimal in some sense analogously to $h_n^{(LSCV)}$, as seen in Theorem 2. To formulate such a result, the focus will be on the case when $p = 0$. In this special case, it can easily be shown that the following holds

$$\widehat{m}_0(x) = \sum_{i=1}^{n} w_{n,i}(x) Y_i,$$

where

$$w_{n,i}(x) = \frac{K\left(\frac{X_i - x}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h_n}\right)}, \quad i = 1, \ldots, n.$$

Asymptotic optimality could be measured with respect to different distances, for example:

- Average Squared Error:

$$\mathrm{ASE}_0(h_n) = \frac{1}{n} \sum_{i=1}^{n} \Big[\widehat{m}_0(X_i) - m(X_i)\Big]^2 w(X_i);$$

- Integrated Squared Error:

$$\mathrm{ISE}_0(h_n) = \int_{\mathbb{R}} \Big(\widehat{m}_0(x) - m(x)\Big)^2 f_X(x) w(x) \mathrm{d}x;$$

- Conditional Mean Integrated Squared Error:

$$\mathrm{MISE}_0(h_n|\mathbb{X}) = \mathbb{E}\Big[\mathrm{ISE}_0(h_n)|\mathbb{X}\Big].$$

Theorem 5. *Suppose the weight function $w \geq 0$ is bounded and has a compact support. Assume that the following conditions are satisfied:*

(D1) The kernel function K is β-Hölder continuous, i.e., there exist $\beta > 0$ and $\alpha > 0$ such that

$$|K(y) - K(x)| < \alpha |y - x|^\beta \ \forall y, x \in \mathbb{R},$$

also assume that $\int_{\mathbb{R}} K(y)\mathrm{d}y = 1$ and $\int_{\mathbb{R}} |y|^\beta |K(y)|\mathrm{d}y < \infty$.

(D2) For $n \in \mathbb{N}$, define the interval $H_n = [\underline{h}_n, \overline{h}_n]$ where

$$\underline{h}_n = b^{-1} n^{\delta - 1} \quad \textit{and} \quad \overline{h}_n = b n^{-\delta},$$

for some constants $b > 1$ and $0 < \delta < \frac{1}{2}$.

*Under appropriate regularity assumptions * on $m(\cdot)$, the marginal density $f_X(\cdot)$, and the boundedness of conditional moments, the following holds*

$$\frac{\mathrm{SE}_0(h_n^{(CV)})}{\inf_{h_n \in H_n} \mathrm{SE}_0(h_n)} \xrightarrow[n \to \infty]{a.s.} 1, \tag{2.23}$$

where $\mathrm{SE}_0(\cdot)$ can be one of the previously defined distances $\mathrm{ASE}_0(\cdot)$, $\mathrm{ISE}_0(\cdot)$, or $\mathrm{MISE}_0(\cdot|\mathbb{X})$.

The general proof for multivariate predictors is presented in Hardle and Marron [1985].

Remark. The width of the interval $|H_n| = \overline{h}_n - \underline{h}_n$ is converging to zero as $n \to \infty$. If $\delta = \frac{1}{5}$, then under some regularity conditions on the kernel K and the marginal density f_X, it could be shown that $\overline{h}_n = b h_n^{(opt)}$, where $h_n^{(opt)}$ is defined in (1.5). In other words, in the special case where $\delta = \frac{1}{5}$, the width of the interval H_n is converging to zero at the same rate as the width of the biased cross-validation interval $(0, b h_n^{(opt)})$ from the expression (1.29).

Note that in Theorem 5 the conditions on the kernel are weaker than those in Theorem 2 and Theorem 4. In the asymptotic results from Chapter 1, it was assumed that K has a bounded support. However, such an assumption is not made here, as integrability issues are now controlled by the weight function w, which has a bounded support. Also, one might think that the results of Theorem 5 are quite restrictive, as they hold only for the special case when the degree of the local polynomial estimator is $p = 0$. For general p, asymptotic optimality with respect to the mean integrated squared error can be found in Xia and Li [2002], Theorem 2.1.

2.3 Simulation Study

Let us illustrate the performance of different bandwidth selectors in the setting of kernel regression. Data were simulated from the following model

$$Y_i = X_i + 2\exp(-16 X_i^2) + \varepsilon_i, \text{ where } i \in \{1, \ldots, n\}. \tag{2.24}$$

Also, assume that the model (2.24) satisfies the following conditions:

- $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathsf{N}\left(0, \left(\frac{2}{5}\right)^2\right)$,
- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$,
- The random sample $(\varepsilon_1, \ldots, \varepsilon_n)$ is independent of $(X_1, \ldots, X_n)$.

*See Hardle and Marron [1985], Section 2.

The model (2.24) was considered in Fan and Gijbels [1992], Section 5. The authors showed that a modification of the local linear estimator reasonably estimates the shape of the conditional mean of the model (2.24). In this section, it will be demonstrated how previously defined bandwidth selection methods perform under the model (2.24).

For the calculation, the standard normal kernel was employed along with the following weight function $w(x) = \mathbf{1}_{[-2,2]}(x)$, which is an indicator of the compact interval $[-2, 2]$. The justification for using such a weight function is that approximately only 5% of observations of X do not belong to the interval $[-2, 2]$. Each setting involves the use of 1200 Monte Carlo repetitions. Results of the simulation study can be found in Table 2.1. As in Chapter 1, each row of the first column corresponds to a specific method, with labels summarized as follows

- `Optimal`: Asymptotically optimal global bandwidth, $h_n^{(opt)}$, which is defined in (2.15).
- `ROT`: Rule of thumb derived under homoscedasticity, $h_n^{(ROT)}$, as defined in (2.18).
- `CV`: Cross-validation, $h_n^{(CV)}$, defined in (2.21).

For local constant estimation and local linear estimation, the functions locCteSmootherC and locLinSmootherC from the package Cabrera and Quast [2022] were used. To compute `ROT` and `CV`, the functions thumbBw and regCVBwSelC from the package Cabrera and Quast [2022] were used. Also, it is important to note that the function regCVBwSelC conducts cross-validation over the following finite interval $(5 * 10^{-4}, 1.5)$ by default settings.

2.3.1 Monte Carlo

Generally, the *conditional mean integrated squared error* of $\widehat{m}_p$ is defined as

$$\mathrm{MISE}\Big(\widehat{m}_p|\mathbb{X}\Big) = \mathbb{E}[\mathrm{ISE}(\widehat{m}_p)|\mathbb{X}], \text{ where } p \in \mathbb{N}_0. \tag{2.25}$$

Note that the conditional mean integrated squared error, as indicated in (2.25), also implicitly depends on the sample size of $\mathbb{X}$. Here, it is considered that $n = 200$ and $n = 500$. Due to the fixed value of n, a Monte Carlo approach is necessary. One needs to estimate the conditional MISE of $\widehat{m}_p$. Hence, before the Monte Carlo repetitions, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$ were generated. Overall, $B = 1200$ Monte Carlo repetitions were conducted. In each Monte Carlo repetition $b = 1, \ldots, B$, the following random variables $Y_1^b, \ldots, Y_n^b$ were generated from the model (2.24), i.e.,

$$Y_i^b = X_i + 2\exp(-16X_i^2) + \varepsilon_i^b, \text{ for } i = 1, \ldots, n,$$

where $\varepsilon_1^b, \ldots, \varepsilon_n^b \overset{\text{iid}}{\sim} \mathsf{N}\left(0, \left(\frac{2}{5}\right)^2\right)$ and the random sample $(\varepsilon_1^b, \ldots, \varepsilon_n^b)$ is independent of $(X_1, \ldots, X_n)$ for all $b = 1, \ldots, B$.

Therefore, for each $b = 1, \ldots, B$, one would obtain $\mathrm{ISE}(\widehat{m}_p^b)$. By averaging those $\mathrm{ISE}(\widehat{m}_p^b)$, one gets an estimation of the conditional mean integrated squared error (2.25).

Recall that the integrated squared error of $\widehat{m}_p^b$ is defined as

$$\mathrm{ISE}(\widehat{m}_p^b) = \int_{\mathbb{R}} (\widehat{m}_p^b(x) - m(x))^2 w(x) f_X(x) \mathrm{d}x, \tag{2.26}$$

where f_X represents the marginal density of X. Here, f_X corresponds to the probability density function of the standard normal distribution $\mathsf{N}(0,1)$. To estimate (2.26), the random sample $Z_1^b, \ldots, Z_m^b \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ with $m = 1000$ was generated independently of $Y_1^b, \ldots, Y_n^b$ and $X_1, \ldots, X_n$. Then, by the law of large numbers, $\mathrm{ISE}(\widehat{m}_p^b)$ could be approximated by

$$\frac{1}{m} \sum_{i=1}^{m} (\widehat{m}_p^b(Z_i^b) - m(Z_i^b))^2 w(Z_i^b).$$

2.3.2 Discussion of results

Table 2.1 reveals that for a larger sample size (see the results for $n = 500$), cross-validation `CV` performs better than the rule of thumb `ROT`. This is evident as the confidence intervals for MISE when the `CV` method is employed are disjointly shifted downward from those of the `ROT` method. However, for $n = 500$, cross-validation performs worse than the `Optimal` method. Recall that the `Optimal` method is impossible to calculate in real applications as one needs to know characteristics of the true distribution, such as $\sigma^2(x)$, $m''(x)$, and $f_X(x)$.

In the case of local constant estimation, for a smaller sample size (see the results for $n = 200$), one could see that the rule of thumb has better performance than cross-validation. However, in the case of local linear estimation, for a smaller sample size $n = 200$ the superiority of the `ROT` method under the `CV` method is not so evident as their confidence intervals for MISE overlap.

Also, it can be seen that almost in all cases, the confidence intervals for MISE in the case of the local linear estimator are shifted downwards compared to the corresponding confidence intervals of the local constant estimator. This indicates that generally under the model (2.24), the local linear estimator performs better than the local constant estimator. The only exception is the `ROT` method for a larger sample size $n = 500$.

In the fourth column of Table 2.1, median values of bandwidths calculated from simulations are presented. It can be observed that for local constant estimation, the median values of data-driven selected bandwidths, such as `CV` and `ROT`, are smaller than the median value of the asymptotically optimal bandwidths. Conversely, for local linear estimation, the median values of data-driven selected bandwidths are larger. This indicates that regression estimates computed from the local constant estimation with bandwidths selected by data-driven methods generally under-smooth. In the case of local constant estimator regression estimates with bandwidths selected by data-driven methods tend to over-smooth.

Local polynomial regression	Monte Carlo Approx. of MISE	95% Monte Carlo Conf. Interval of MISE	Median of bandwidths
LOCAL CONSTANT ESTIMATOR			
$n = 200$			
`Optimal`	**19.765**	**(19.451, 20.079)**	0.081
`ROT`	20.064	(19.739, 20.389)	0.063
`CV`	*21.034*	*(20.679, 21.389)*	0.078
$n = 500$			
`Optimal`	**8.178**	**(8.053, 8.303)**	0.068
`ROT`	9.384	(9.255, 9.513)	0.044
`CV`	*8.472*	*(8.339, 8.605)*	0.064
LOCAL LINEAR ESTIMATOR			
$n = 200$			
`Optimal`	**18.186**	**(17.835, 18.537)**	0.081
`ROT`	18.968	(18.654, 19.282)	0.107
`CV`	*19.173*	*(18.775, 19.571)*	0.088
$n = 500$			
`Optimal`	**7.656**	**(7.531, 7.781)**	0.068
`ROT`	9.926	(9.777, 10.075)	0.096
`CV`	*7.996*	*(7.865, 8.127)*	0.071

Table 2.1: Results are based on 1200 Monte Carlo replications. In each case, a standard normal kernel and the indicator weight function over the interval $[-2, 2]$ were used. For better presentation, all results were multiplied by 1000.

3. Linear Model Selection

Throughout the previous two chapters, the application of cross-validation in kernel density estimation and kernel regression was discussed. In both of these tasks, leave-one-out cross-validation was considered. In the second chapter, it was pointed out that cross-validation emerges as a natural estimator of the prediction error, see (2.22). This principle can be generalized for linear model selection. In other words, if one aims to select the linear model with the best predictive ability, cross-validation techniques can be applied. However, this section demonstrates that instead of the classical leave-one-out cross-validation, leave-n_v-out cross-validation should be used, where n_v represents the number of observations reserved for validation. Additionally, it should be ensured that $n_v/n \to 1$ as $n \to \infty$, which is completely opposite to the leave-one-out cross-validation principle, because $n_v = 1$ for the last case.

This chapter is based on Shao [1993].

3.1 Framework for linear model selection

Let $Y_1, \ldots, Y_n$ be independent random variables that satisfy a linear regression model with fixed covariates given by

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{3.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ ($p \in \mathbb{N}$), $\boldsymbol{x}_i \in \mathbb{R}^p$ for all $i = 1, \ldots, n$ and the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed, with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{var}[\varepsilon_i] = \sigma^2$ for some $\sigma^2 > 0$. Also, throughout the whole chapter, it will be assumed that $n > p$, which means that the number of observations is greater than the number of predictors.

As one does not know the true $\boldsymbol{\beta}$, it may happen that some of the components of $\boldsymbol{\beta}$ are actually equal to zero. Suppose that our goal is to make predictions based on the model (3.1). For this task, one needs to identify important predictors because it may potentially reduce the prediction error. Therefore, it may be useful to consider the compact form of the linear model (3.1)

$$Y_i = \boldsymbol{x}_{i,\alpha}^T \boldsymbol{\beta}_\alpha + \varepsilon_i, \tag{3.2}$$

where α is a subset of d_α positive integers from the set $\{1, \ldots, p\}$. In (3.2) $\boldsymbol{\beta}_\alpha$ (or $\boldsymbol{x}_{i,\alpha}$) is the d_α vector containing the components of $\boldsymbol{\beta}$ (or $\boldsymbol{x}_i$) that are indexed by the integers in α. Let $\mathcal{A}$ denote all nonempty subsets of $\{1, \ldots, p\}$. Then, obviously, there are $|\mathcal{A}| = 2^{p-1}$ different models in the compact form (3.2), each corresponding to some subset of predictors α. Further, a model of the form (3.2) will be denoted as M_α. In the special case when $\alpha = \{1, \ldots, p\}$, the notation M is used for the model. The number of predictors d_α will also be referred to as the dimension of M_α and denoted as $\dim(M_\alpha)$.

All models M_α for $\alpha \in \mathcal{A}$ can be classified into the following two categories:

- Category I: At least one nonzero component of $\boldsymbol{\beta}$ is not in $\boldsymbol{\beta}_\alpha$.
- Category II: $\boldsymbol{\beta}_\alpha$ contains all nonzero components of $\boldsymbol{\beta}$.

It is evident that all models from Category I miss at least one important predictor, which means that they are not of the main interest. On the other hand, models from Category II may consist of unrelated predictors, potentially resulting in an unnecessarily large dimension for such models. Therefore, the primary interest lies in the model within Category II with the smallest dimension, denoted as M_*, because this model consists of all important predictors and does not include any unrelated predictors.

From this point onwards, it is assumed that for all $\alpha \in \mathcal{A}$,

$$\mathbb{X}_\alpha = (\boldsymbol{x}_{1,\alpha}, \ldots, \boldsymbol{x}_{n,\alpha})^T$$

is an $n \times d_\alpha$ matrix of full rank. In other words, $\operatorname{rank}(\mathbb{X}_\alpha) = d_\alpha$ because it is assumed that $n > p$. Hence, under the model M_α, the least squares estimator of $\boldsymbol{\beta}_\alpha$ is given as

$$\widehat{\boldsymbol{\beta}}_\alpha = (\mathbb{X}_\alpha^T \mathbb{X}_\alpha)^{-1} \mathbb{X}_\alpha^T \boldsymbol{Y}, \tag{3.3}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is an $n \times 1$ response vector. In the special case when $\alpha = \{1, \ldots, p\}$, the matrix $\mathbb{X}_\alpha$ will be denoted simply as $\mathbb{X}$.

3.2 Prediction error

Suppose that the future observations $Z_1, \ldots, Z_n$ satisfy the following model

$$Z_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \widetilde{\varepsilon}_i, \tag{3.4}$$

where $i = 1, \ldots, n$ and the error terms $\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n, \varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed.

Let the least squares estimator $\widehat{\boldsymbol{\beta}}_\alpha$ be given by (3.3), i.e., it is fitted under the model M_α and based on the data $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$. Now, the estimator $\widehat{\boldsymbol{\beta}}_\alpha$ could be used to predict the future values Z_i from the model (3.4) as $\boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_\alpha$. Therefore, define the *average squared prediction error* as

$$\mathrm{ASPE}_\alpha^n = \frac{1}{n} \sum_{i=1}^{n} \left(Z_i - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_\alpha \right)^2. \tag{3.5}$$

It is important to note that ASPE_α^n is a random variable, where the randomness is hidden in the future observations $Z_1, \ldots, Z_n$ and in the estimator $\widehat{\boldsymbol{\beta}}_\alpha$, which is based on the 'construction' data $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$. The following lemma provides the expression for the expectation of ASPE_α^n.

Lemma 6. *The unconditional expectation of the average squared prediction error* $\Gamma_{\alpha,n} = \mathbb{E}\left[\mathrm{ASPE}_\alpha^n\right]$ *satisfies the following equation*

$$\Gamma_{\alpha,n} = \sigma^2 + n^{-1} d_\alpha \sigma^2 + \Delta_{\alpha,n},$$

where

$$\Delta_{\alpha,n} = \tfrac{1}{n}\boldsymbol{\beta}^T\mathbb{X}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta},$$

$$\mathbb{P}_\alpha = \mathbb{X}_\alpha(\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_\alpha^T \textit{ is the projection matrix under model } M_\alpha,$$

$$\textit{and } \mathbb{I}_n \textit{ is the identity matrix of order } n.$$

Proof. From the definition of ASPE_α^n in (3.5), the following holds

$$\begin{aligned}\Gamma_{\alpha,n} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(Z_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(Z_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2\right]\\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left(Z_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2}_{\mathrm{I}_n=} + \underbrace{\frac{2}{n}\sum_{i=1}^n\mathbb{E}\left[\left(Z_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)\right]}_{\mathrm{II}_n=}\\ &\quad + \underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2\right]}_{\mathrm{III}_n=}. \end{aligned} \tag{3.6}$$

Now each of the terms I_n, II_n, and III_n will be considered separately. By using the model (3.4), the term I_n can be expressed as

$$\mathrm{I}_n = \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left(Z_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 = \frac{1}{n}\sum_{i=1}^n\mathbb{E}\ (\widetilde{\varepsilon}_i)^2 = \sigma^2. \tag{3.7}$$

The second term II_n is actually equals to zero, because

$$\begin{aligned}\mathrm{II}_n &= \frac{2}{n}\sum_{i=1}^n\mathbb{E}\left\{\mathbb{E}\left[\left(Z_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)\Big|\boldsymbol{Y}\right]\right\}\\ &= \frac{2}{n}\sum_{i=1}^n\mathbb{E}\left\{\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)\mathbb{E}\left[\left(Z_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\Big|\boldsymbol{Y}\right]\right\}\\ &\overset{(3.4)}{=} \frac{2}{n}\sum_{i=1}^n\mathbb{E}\left\{\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)\mathbb{E}\left[\widetilde{\varepsilon}_i|\boldsymbol{Y}\right]\right\}\\ &\overset{\text{indp.}}{=} \frac{2}{n}\sum_{i=1}^n\mathbb{E}\left\{\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)\mathbb{E}(\widetilde{\varepsilon}_i)\right\} = 0. \end{aligned} \tag{3.8}$$

Finally, the third term III_n can be expressed as

$$\begin{aligned}\mathrm{III}_n &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2\right] = \tfrac{1}{n}\mathbb{E}\left[\left(\mathbb{X}\boldsymbol{\beta} - \mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right)^T\left(\mathbb{X}\boldsymbol{\beta} - \mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right)\right]\\ &= \tfrac{1}{n}\mathbb{E}\left[\left(\boldsymbol{\beta}^T\mathbb{X}^T - \widehat{\boldsymbol{\beta}}_\alpha^T\mathbb{X}_\alpha^T\right)\left(\mathbb{X}\boldsymbol{\beta} - \mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right)\right]\\ &= \tfrac{1}{n}\mathbb{E}\left[\boldsymbol{\beta}^T\mathbb{X}^T\mathbb{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbb{X}^T\mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha + \widehat{\boldsymbol{\beta}}_\alpha^T\mathbb{X}_\alpha^T\mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right]\\ &= \tfrac{1}{n}\boldsymbol{\beta}^T\mathbb{X}^T\mathbb{X}\boldsymbol{\beta} - \tfrac{2}{n}\boldsymbol{\beta}^T\mathbb{X}^T\mathbb{X}_\alpha\mathbb{E}\widehat{\boldsymbol{\beta}}_\alpha + \tfrac{1}{n}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_\alpha^T\mathbb{X}_\alpha^T\mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right]. \end{aligned} \tag{3.9}$$

Note that by using the vector form of the linear regression model (3.1) and formula (3.3) , one would obtain

$$\mathbb{E}\widehat{\boldsymbol{\beta}}_\alpha = (\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_\alpha^T\mathbb{E}\boldsymbol{Y} = (\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_\alpha^T\mathbb{E}\Big[\mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\Big] = (\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_\alpha^T\mathbb{X}\boldsymbol{\beta}, \tag{3.10}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and the last equation in (3.10) holds because $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}$. Also, the following holds

$$\begin{aligned}
&\tfrac{1}{n}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_\alpha^T \mathbb{X}_\alpha^T \mathbb{X}_\alpha \widehat{\boldsymbol{\beta}}_\alpha\right] = \tfrac{1}{n}\mathbb{E}\left[\boldsymbol{Y}^T \mathbb{X}_\alpha (\mathbb{X}_\alpha^T \mathbb{X}_\alpha)^{-1}\right] \mathbb{X}_\alpha^T \mathbb{X}_\alpha \left[(\mathbb{X}_\alpha^T \mathbb{X}_\alpha)^{-1} \mathbb{X}_\alpha^T \boldsymbol{Y}\right] \\
&= \tfrac{1}{n}\mathbb{E}\left[\boldsymbol{Y}^T \mathbb{P}_\alpha \boldsymbol{Y}\right] \overset{\text{Lem. A.1}}{=} \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} + \tfrac{1}{n}\mathrm{tr}(\sigma^2 \mathbb{P}_\alpha) \\
&= \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} + \tfrac{\sigma^2}{n}\mathrm{rank}(\mathbb{P}_\alpha) = \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} + \tfrac{\sigma^2}{n} d_\alpha. \qquad (3.11)
\end{aligned}$$

Then, by inserting (3.10) and (3.11) into (3.9), one would get that

$$\begin{aligned}
\mathrm{III}_n &= \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{X}\boldsymbol{\beta} - \tfrac{2}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} + \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} + \tfrac{\sigma^2}{n} d_\alpha \\
&= \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{X}\boldsymbol{\beta} - \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T \mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} + \tfrac{\sigma^2}{n} d_\alpha \\
&= \tfrac{1}{n}\boldsymbol{\beta}^T \mathbb{X}^T (\mathbb{I}_n - \mathbb{P}_\alpha) \mathbb{X}\boldsymbol{\beta} + \tfrac{\sigma^2}{n} d_\alpha = \Delta_{\alpha,n} + \tfrac{\sigma^2}{n} d_\alpha. \qquad (3.12)
\end{aligned}$$

Finally, by combining (3.6), (3.7), (3.8), and (3.12), one gets the statement of the lemma. □

The expected value of the squared prediction error $\Gamma_{\alpha,n}$ is a numerical characteristic that summarizes the predictive ability of a model M_α. According to the previous lemma, $\Gamma_{\alpha,n}$ can be decomposed into two parts:

- the variability of the future observations σ^2,
- the error in model selection and estimation, represented by $\frac{\sigma^2}{n} d_\alpha + \Delta_{\alpha,n}$.

Obviously, the term $\frac{\sigma^2}{n} d_\alpha$ vanishes as $n \to \infty$. Therefore, the main interest lies in the term $\Delta_{\alpha,n}$.

Note that $\mathbb{I}_n - \mathbb{P}_\alpha$ is an idempotent matrix because $\mathbb{P}_\alpha$ is a projection matrix. Therefore, $\mathbb{I}_n - \mathbb{P}_\alpha$ is also a positive-semidefinite matrix, which implies that $\Delta_{\alpha,n} \geq 0$ for all $n \in \mathbb{N}$ and for all $\alpha \in \mathcal{A}$. Additionally, $\Delta_{\alpha,n} = 0$ if and only if $\mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta}$. The last condition means that $\mathbb{X}\boldsymbol{\beta} \in \mathrm{Im}(\mathbb{X}_\alpha)$, which occurs only when M_α is a model in Category II. Overall, one would get that $\Delta_{\alpha,n} > 0$ if M_α is a model in Category I, and $\Delta_{\alpha,n} = 0$ if M_α is a model in Category II.

Many asymptotic results from this chapter will require the following condition to hold

$$\liminf_{n\to\infty} \Delta_{\alpha,n} > 0 \quad \forall \alpha \in \mathcal{A} \text{ such that } M_\alpha \text{ in Category I.} \qquad (3.13)$$

Now, it will be explained that the condition (3.13) is not strong and naturally appears as a type of asymptotic model identifiability. Suppose that M_α is a model in Category I and M_γ is a model in Category II such that $d_\alpha = d_\gamma$. Then, in order to decide which of the two models is better in terms of their predictive ability, one is interested in the asymptotic behavior of the following fraction

$$\frac{\Gamma_{\alpha,n}}{\Gamma_{\gamma,n}} = 1 + \frac{\Delta_{\alpha,n}}{\sigma^2 + \frac{\sigma^2}{n} d_\gamma}. \qquad (3.14)$$

From the expression (3.14), it can be easily seen that $\frac{\Gamma_{\alpha,n}}{\Gamma_{\gamma,n}} \geq 1$. If $\lim_{n\to\infty} \frac{\Gamma_{\alpha,n}}{\Gamma_{\gamma,n}}$ exists and equals 1, then it can be interpreted as models M_α and M_γ having asymptotically no difference in terms of their predictive ability. On the other

hand, to identify that the models M_α and M_γ are asymptotically different in terms of their predictive ability, it is sufficient to observe

$$\liminf_{n\to\infty} \frac{\Gamma_{\alpha,n}}{\Gamma_{\gamma,n}} > 1 \iff \liminf_{n\to\infty} \Delta_{\alpha,n} > 0.$$

3.3 Cross-validation

Let $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$ are data pairs that are sampled from the linear regression model (3.1). To introduce methods for model selection, one would first need to divide the dataset into two parts: one for fitting a model (model construction) and the other for assessing the predictive ability of the model (model validation). Let s be a subset of $\{1, \ldots, n\}$ with cardinality n_v and s^c be its complement with cardinality n_c, where $n_v + n_c = n$. Then, one would use the construction data $\{(Y_i, \boldsymbol{x}_i),\ i \in s^c\}$ to fit the model M_α and the validation data $\{(Y_i, \boldsymbol{x}_i),\ i \in s\}$ to assess the prediction error. Here, the validation data $\{(Y_i, \boldsymbol{x}_i),\ i \in s\}$ are treated as unobserved future values, therefore the average squared prediction error is given as

$$\mathrm{ASPE}_\alpha^s = \frac{1}{n_v} \sum_{i\in s} (Y_i - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_{\alpha,s^c})^2, \tag{3.15}$$

where $\widehat{\boldsymbol{\beta}}_{\alpha,s^c} = (\mathbb{X}_{\alpha,s^c}^T \mathbb{X}_{\alpha,s^c})^{-1} \mathbb{X}_{\alpha,s^c}^T \boldsymbol{Y}_{s^c}$, and $\mathbb{X}_{\alpha,s^c}$ is the $n_c \times d_\alpha$ matrix containing the rows of $\mathbb{X}_\alpha$ indexed by $i \in s^c$, and $\boldsymbol{Y}_{s^c} = (Y_i,\ i \in s^c)^T$. Analogously, one could introduce the notation of $\mathbb{X}_{\alpha,s}$ and $\boldsymbol{Y}_s$. From the equation (3.15), it can be seen that ASPE_α^s depends on $\widehat{\boldsymbol{\beta}}_{\alpha,s^c}$. It will be useful to express the dependency of ASPE_α^s on $\widehat{\boldsymbol{\beta}}_\alpha$, which is the subject of the following lemma.

Lemma 7. *Let $\mathbb{X}_\alpha$ be a matrix of full rank such that*

$$\mathit{rank}(\mathbb{X}_\alpha) = \mathit{rank}(\mathbb{X}_{\alpha,s}) = \mathit{rank}(\mathbb{X}_{\alpha,s^c}) = d_\alpha.$$

Then the average squared prediction error ASPE_α^s given by (3.15) satisfies the following equation

$$\mathrm{ASPE}_\alpha^s = \frac{1}{n_v} \left|\left| (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1} \left(\boldsymbol{Y}_s - \mathbb{X}_{\alpha,s} \widehat{\boldsymbol{\beta}}_\alpha \right) \right|\right|_2^2, \tag{3.16}$$

where $||\cdot||_2$ denotes the Euclidean norm and $\mathbb{Q}_{\alpha,s} = \mathbb{X}_{\alpha,s}(\mathbb{X}_\alpha^T \mathbb{X}_\alpha)^{-1} \mathbb{X}_{\alpha,s}^T$.

Proof. First, note that

$$\mathrm{ASPE}_\alpha^s = \frac{1}{n_v} \sum_{i\in s} (Y_i - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_{\alpha,s^c})^2 = \frac{1}{n_v} \left|\left| \boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_{\alpha,s^c} \right|\right|_2^2,$$

where $\widehat{\boldsymbol{Y}}_{\alpha,s^c} = \mathbb{X}_{\alpha,s} \widehat{\boldsymbol{\beta}}_{\alpha,s^c}$. Therefore, it will be sufficient to prove the following equation

$$\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_{\alpha,s^c} = (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1} \left(\boldsymbol{Y}_s - \mathbb{X}_{\alpha,s} \widehat{\boldsymbol{\beta}}_\alpha \right). \tag{3.17}$$

Rewrite the left hand side of the equation (3.17)

$$
\begin{aligned}
&\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_{\alpha,s^c} \\
&= \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}(\mathbb{X}_{\alpha,s^c}^T\mathbb{X}_{\alpha,s^c})^{-1}\mathbb{X}_{\alpha,s^c}^T\boldsymbol{Y}_{s^c} \\
&= \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}(\mathbb{X}_{\alpha,s^c}^T\mathbb{X}_{\alpha,s^c} + \mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s} - \mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s})^{-1}(\mathbb{X}_{\alpha}^T\boldsymbol{Y} - \mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s) \\
&= \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha} - \mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s})^{-1}(\mathbb{X}_{\alpha}^T\boldsymbol{Y} - \mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s) \\
&= \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}\left\{\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha}\left[\mathbb{I}_{d_\alpha} - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]\right\}^{-1}(\mathbb{X}_{\alpha}^T\boldsymbol{Y} - \mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s) \\
&= \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}\underbrace{\left[\mathbb{I}_{d_\alpha} - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^{-1}}_{\mathbb{H}_{\alpha,s}=}(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}(\mathbb{X}_{\alpha}^T\boldsymbol{Y} - \mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s) \\
&= \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}\mathbb{H}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right). \qquad (3.18)
\end{aligned}
$$

Since it is assumed that $\mathbb{X}_\alpha$, $\mathbb{X}_{\alpha,s}$, and $\mathbb{X}_{\alpha,s^c}$ are matrices of full rank, which also means that $\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha}$, $\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}$, and $\mathbb{X}_{\alpha,s^c}^T\mathbb{X}_{\alpha,s^c}$ are positive definite matrices. Then Lemma A.3 implies that all eigenvalues of the matrix $(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}$ are smaller than 1 in absolute value. Therefore, one can use Lemma 7.18 from Burden and Faires [2010], and get

$$
\mathbb{H}_{\alpha,s} = \left[\mathbb{I}_{d_\alpha} - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^{-1} = \sum_{j=0}^{\infty}\left[(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^j. \qquad (3.19)
$$

Hence, it holds that

$$
\begin{aligned}
\mathbb{X}_{\alpha,s}\mathbb{H}_{\alpha,s} &= \mathbb{X}_{\alpha,s}\left[\mathbb{I}_{d_\alpha} - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^{-1} \\
&\overset{(3.19)}{=} \mathbb{X}_{\alpha,s} + \sum_{j=1}^{\infty}\mathbb{X}_{\alpha,s}\left[(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^j \\
&= \mathbb{X}_{\alpha,s} + \sum_{j=1}^{\infty}\underbrace{\mathbb{X}_{\alpha,s}(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T}_{\mathbb{Q}_{\alpha,s}=}\mathbb{X}_{\alpha,s}\left[(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^{j-1} \\
&= \mathbb{X}_{\alpha,s} + \mathbb{Q}_{\alpha,s}\mathbb{X}_{\alpha,s}\sum_{j=1}^{\infty}\left[(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^{j-1} \\
&= \mathbb{X}_{\alpha,s} + \mathbb{Q}_{\alpha,s}\mathbb{X}_{\alpha,s}\sum_{j=0}^{\infty}\left[(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\mathbb{X}_{\alpha,s}\right]^{j} \overset{(3.19)}{=} \mathbb{X}_{\alpha,s} + \mathbb{Q}_{\alpha,s}\mathbb{X}_{\alpha,s}\mathbb{H}_{\alpha,s}. \qquad (3.20)
\end{aligned}
$$

Combining (3.18) and (3.20) one obtains the following equation

$$
\begin{aligned}
&\boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}\mathbb{H}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right) \\
&= \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right) \\
&\quad - \mathbb{Q}_{\alpha,s}\mathbb{X}_{\alpha,s}\mathbb{H}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right). \qquad (3.21)
\end{aligned}
$$

The last equation (3.21) can be simplified to

$$\begin{aligned}
&\mathbb{X}_{\alpha,s}\mathbb{H}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right)\\
&= (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{X}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right) \qquad (3.22)
\end{aligned}$$

Finally, one obtains

$$\begin{aligned}
\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_{\alpha,s^c} &\overset{(3.18)}{=} \boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}\mathbb{H}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right)\\
&\overset{(3.22)}{=} \boldsymbol{Y}_s - (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{X}_{\alpha,s}\left(\widehat{\boldsymbol{\beta}}_\alpha - (\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\mathbb{X}_{\alpha,s}^T\boldsymbol{Y}_s\right)\\
&= \boldsymbol{Y}_s - (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{X}_{\alpha,s}\widehat{\boldsymbol{\beta}}_\alpha + (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{Q}_{\alpha,s}\boldsymbol{Y}_s\\
&= (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})\boldsymbol{Y}_s - (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{X}_{\alpha,s}\widehat{\boldsymbol{\beta}}_\alpha\\
&\quad + (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{Q}_{\alpha,s}\boldsymbol{Y}_s\\
&= (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\left(\boldsymbol{Y}_s - \mathbb{X}_{\alpha,s}\widehat{\boldsymbol{\beta}}_\alpha\right).
\end{aligned}$$

□

The cross-validation estimate for $\Gamma_{\alpha,n}$ is obtained by averaging the quantities in (3.16) over some (or all) subsets $s \subseteq \{1,\ldots,n\}$ of size n_v. More precisely, let $\mathcal{B}$ be a collection of subsets of the set $\{1,\ldots,n\}$ that have size n_v. Note that the maximal size of $\mathcal{B}$ is $\binom{n}{n_v}$. The cross-validation estimate of $\Gamma_{\alpha,n}$ is given as

$$\widehat{\Gamma}_{\alpha,n_v}^{\mathrm{CV}} = \frac{1}{|\mathcal{B}|}\sum_{s\in\mathcal{B}} \mathrm{ASPE}_\alpha^s. \qquad (3.23)$$

Then, the model selected by cross-validation is $M_{\widehat{\alpha}(n_v)}$, where

$$\widehat{\alpha}(n_v) = \arg\min_{\alpha\in\mathcal{A}} \widehat{\Gamma}_{\alpha,n_v}^{\mathrm{CV}}. \qquad (3.24)$$

Note that $\widehat{\alpha}(n_v)$ is random as it implicitly depends on ASPE_α^s. The method of model selection (3.94) will be referred as leave-n_v-out cross-validation, abbreviated as CV(n_v). The error rate of using the CV(n_v) for selecting the optimal model M_* is

$$\mathrm{P}\Big[\text{the selected model is not } M_*\Big] = \mathrm{P}\left[M_{\widehat{\alpha}(n_v)} \neq M_*\right]. \qquad (3.25)$$

Also, it is important to note that $\widehat{\alpha}(n_v)$ implicitly depends on n. It is said that the method CV(n_v) *works* if and only if the probability that this method does not select the optimal model M_* converges to 0 as $n \to \infty$, which means the following

$$\limsup_{n\to\infty} \mathrm{P}\left[M_{\widehat{\alpha}(n_v)} \neq M_*\right] = 0. \qquad (3.26)$$

In other words, condition (3.26) means that the error rate of using the CV(n_v), which is given by (3.25), vanishes as $n \to \infty$.

3.3.1 Leave-one-out cross-validation

In order to compute the cross-validation estimate $\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,n_v}$ given by (3.23), one would need to invert the $n_v \times n_v$ matrix $(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})$ from the expression (3.16). Therefore, from a computational point of view, the simplest CV(n_v) would be one with $n_v = 1$. This type of cross-validation is referred to as leave-one-out cross-validation, abbreviated as CV(1).

If $s = \{i\}$ for $i = 1, \ldots, n$, then

$$w^n_{i,\alpha} = \mathbb{Q}_{\alpha,\{i\}} = \mathbb{X}_{\alpha,\{i\}}(\mathbb{X}^T_\alpha \mathbb{X}_\alpha)^{-1}\mathbb{X}^T_{\alpha,\{i\}} = \boldsymbol{x}^T_{i,\alpha}(\mathbb{X}^T_\alpha \mathbb{X}_\alpha)^{-1}\boldsymbol{x}_{i,\alpha}. \tag{3.27}$$

Note that $w^n_{i,\alpha}$ is actually the $i-$th diagonal element of the projection matrix $\mathbb{P}_\alpha$. Therefore, by using (3.27) and the fact that $n_v = 1$ in the case when $s = \{i\}$, the average squared prediction, as given in (3.16), simplifies to

$$\mathrm{ASPE}^{\{i\}}_\alpha = \left[\frac{Y_i - \boldsymbol{x}^T_{i,\alpha}\widehat{\boldsymbol{\beta}}_\alpha}{1 - w^n_{i,\alpha}}\right]^2.$$

In the case of CV(1), it is natural to set $\mathcal{B} = \{\{i\} \text{ for } i = 1, \ldots, n\}$. Therefore, as a cross-validation estimate of $\Gamma_{\alpha,n}$, one would get the following

$$\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} = \frac{1}{n}\sum_{i=1}^n \mathrm{ASPE}^{\{i\}}_\alpha = \frac{1}{n}\sum_{i=1}^n \left[\frac{Y_i - \boldsymbol{x}^T_{i,\alpha}\widehat{\boldsymbol{\beta}}_\alpha}{1 - w^n_{i,\alpha}}\right]^2. \tag{3.28}$$

To state the main results of this chapter, it is necessary to introduce the big O and small o symbols for matrices.

Definition 2. *Let $\{A_n\}_{n=1}^\infty$ be a sequence of square matrices of the same order and $||\cdot||$ is a matrix norm, as defined in Definition* A.2*.*

- *This sequence of matrices is $O(1)$ if there exists $M > 0$ and $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, it holds that $||A_n|| \leq M$.*
- *The sequence of matrices is said to be $o(1)$ if $||A_n|| \underset{n\to\infty}{\longrightarrow} 0$.*

Remark. In this chapter, for simplicity, only natural matrix norms as defined in Definition A.3 will be considered. Due to the equivalence of matrix norms, as stated in Theorem A.2, it does not matter which norm one uses to determine if the matrix sequence $\{A_n\}_{n=1}^\infty$ is $O(1)$ or $o(1)$.

Lemma 8. *Assume that the following condition holds*

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}^T_i = O(1) \quad \textit{and} \quad \left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{i,\alpha}\boldsymbol{x}^T_{i,\alpha}\right]^{-1} = O(1). \tag{3.29}$$

Then

$$\tfrac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon} \underset{n\to\infty}{\xrightarrow{P}} 0 \quad \textit{and} \quad \tfrac{1}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta} \underset{n\to\infty}{\xrightarrow{P}} 0. \tag{3.30}$$

Proof. Recall that L_p convergence implies convergence in probability. Therefore, it will be sufficient to show that the terms $\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}$ and $\frac{1}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}$ converge to 0 in L_p for some $p \geq 1$.

Obviously, the projection matrix $\mathbb{P}_\alpha$ is positive-semidefinite, which means that $\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon} \geq 0$. Also, from Lemma A.1 it is known that $\mathbb{E}\left[\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}\right] = \sigma^2 d_\alpha$. Thus, one obtains $\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon} \xrightarrow[n\to\infty]{L_1} 0$. Now, the following will be shown

$$\tfrac{1}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta} \xrightarrow[n\to\infty]{L_2} 0. \tag{3.31}$$

To prove this, note that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}\right]^2 &= \mathbb{E}\left[\boldsymbol{\beta}^T\mathbb{X}^T(\mathbb{I}_n - \mathbb{P}_\alpha)^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}\right] \\ &= \boldsymbol{\beta}^T\mathbb{X}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right](\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta} \\ &= \sigma^2\boldsymbol{\beta}^T\mathbb{X}^T(\mathbb{I}_n - \mathbb{P}_\alpha)^2\mathbb{X}\boldsymbol{\beta} = \sigma^2\boldsymbol{\beta}^T\mathbb{X}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}. \end{aligned} \tag{3.32}$$

The condition (3.29) states that $\mathbb{X}^T\mathbb{X}$ is $O(n)$ under some natural matrix norm. The l_∞ norm is a natural matrix norm defined in (A.2). From Lemma A.5, it is obvious that if $\mathbb{X}^T\mathbb{X}$ is $O(n)$ under the l_∞ norm, then $\mathbb{X}^T\mathbb{X}_\alpha$ and $\mathbb{X}_\alpha^T\mathbb{X}$ are also $O(n)$ under the same norm. Thus, the following holds

$$\begin{aligned}&\left|\boldsymbol{\beta}^T\mathbb{X}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}\right| \leq p\cdot||\boldsymbol{\beta}||_\infty \cdot \left|\left|\mathbb{X}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}\right|\right|_\infty \\ &\overset{\text{Lem. A.4}}{\leq} p\cdot||\boldsymbol{\beta}||_\infty^2 \cdot \left|\left|\mathbb{X}^T\mathbb{X} - \mathbb{X}^T\mathbb{X}_\alpha\left[\mathbb{X}_\alpha^T\mathbb{X}_\alpha\right]^{-1}\mathbb{X}_\alpha^T\mathbb{X}\right|\right|_\infty \\ &\overset{\text{Def. A.2(4)}}{\leq} p\cdot||\boldsymbol{\beta}||_\infty^2 \cdot \left(\left|\left|\mathbb{X}^T\mathbb{X}\right|\right|_\infty + \left|\left|\mathbb{X}^T\mathbb{X}_\alpha\left[\mathbb{X}_\alpha^T\mathbb{X}_\alpha\right]^{-1}\mathbb{X}_\alpha^T\mathbb{X}\right|\right|_\infty\right) \\ &\overset{\text{Def. A.2(5)}}{\leq} p\cdot||\boldsymbol{\beta}||_\infty^2 \cdot \left(\left|\left|\mathbb{X}^T\mathbb{X}\right|\right|_\infty + \left|\left|\mathbb{X}^T\mathbb{X}_\alpha\right|\right|_\infty \left|\left|\left[\mathbb{X}_\alpha^T\mathbb{X}_\alpha\right]^{-1}\right|\right|_\infty \left|\left|\mathbb{X}_\alpha^T\mathbb{X}\right|\right|_\infty\right) \overset{(3.29)}{=} O(n). \end{aligned} \tag{3.33}$$

By combining (3.32) and (3.33) one would get that

$$\tfrac{1}{n^2}\mathbb{E}\left[\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}\right]^2 \leq O(\tfrac{1}{n}),$$

which implies (3.31).

□

Remark. In the next theorem, it will be important to note that by combining (3.32) and (3.33), one would get that

$$n\sigma^2\Delta_{\alpha,n} = O(n), \quad \text{which also means that} \quad \Delta_{\alpha,n} = O(1). \tag{3.34}$$

Consistency of leave-one-out cross validation

The next theorem demonstrates the asymptotic representation of $\widehat{\Gamma}_{\alpha,1}^{\text{CV}}$.

Theorem 9. *Assume the same conditions as in Lemma 8 and that*

$$\lim_{n\to\infty}\max_{1\leq i\leq n} w_{i,\alpha}^n = 0 \quad \textit{for any} \quad \alpha \in \mathcal{A}. \tag{3.35}$$

Then

$$\widehat{\Gamma}_{\alpha,1}^{\text{CV}} = \Gamma_{\alpha,n} + o_P(1).$$

Proof. First, note that the projection matrix $\mathbb{P}_\alpha$ at position (i,j) has the element $\boldsymbol{x}_{i,\alpha}^T(\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\boldsymbol{x}_{j,\alpha}$. Recall that the i-th diagonal element of the projection matrix $\mathbb{P}_\alpha$ is denoted as $w_{i,\alpha}^n$. It also holds that the matrix $\mathbb{P}_\alpha$ is idempotent, meaning $\mathbb{P}_\alpha = \mathbb{P}_\alpha\mathbb{P}_\alpha$. Therefore, from the definition of matrix multiplication, the following holds for all $i = 1,\ldots,n$

$$w_{i,\alpha}^n = (w_{i,\alpha}^n)^2 + \sum_{j\neq i}\left[\boldsymbol{x}_{i,\alpha}^T(\mathbb{X}_\alpha^T\mathbb{X}_\alpha)^{-1}\boldsymbol{x}_{j,\alpha}\right]^2.$$

Now, it is easy to observe that $w_{i,\alpha}^n \geq (w_{i,\alpha}^n)^2$, which also means that $w_{i,\alpha}^n \in [0,1]$ for all $i = 1,\ldots,n$. Consider the following function $\phi(w_{i,\alpha}^n) = (1-w_{i,\alpha}^n)^{-2}$ over the domain $[0,1]$. It can be easily shown that for all $w_{i,\alpha}^n \in [0,1)$, the following holds

$$\phi'(w_{i,\alpha}^n) = \frac{2}{(1-w_{i,\alpha}^n)^3} \quad \text{and} \quad \phi''(w_{i,\alpha}^n) = \frac{6}{(1-w_{i,\alpha}^n)^4}.$$

Therefore, by employing the Lagrange form of the remainder (see Bartle and Sherbert [2011], Theorem 6.4.1), there exists $\rho_{i,\alpha}^n \in (0, w_{i,\alpha}^n)$ such that

$$\begin{aligned}\frac{1}{(1-w_{i,\alpha}^n)^2} &= \phi(0) + \phi'(0)w_{i,\alpha}^n + \frac{\phi''(\xi_{i,\alpha})}{2}(w_{i,\alpha}^n)^2 \\ &= 1 + 2w_{i,\alpha}^n + \frac{3}{(1-\rho_{i,\alpha}^n)^4}(w_{i,\alpha}^n)^2. \end{aligned} \tag{3.36}$$

Obviously, there exists $\gamma > 0$ such that $\forall\delta \in (0,\gamma)$ it holds that $\frac{3}{(1-\delta)^4} < 1$. From the assumption (3.35), it holds that $w_{i,\alpha}^n < \gamma$ for all $i = 1,\ldots,n$, where n is sufficiently large. Therefore one obtains the following

$$\frac{3}{(1-\rho_{i,\alpha}^n)^4} < \frac{3}{(1-w_{i,\alpha}^n)^4} < 1, \tag{3.37}$$

which by combining with (3.36) will lead to

$$\frac{1}{(1-w_{i,\alpha}^n)^2} = 1 + 2w_{i,\alpha}^n + O\left((w_{i,\alpha}^n)^2\right). \tag{3.38}$$

By inserting (3.38) into the CV(1) estimate of $\Gamma_{\alpha,n}$ provided in (3.28), the following equation can be obtained

$$\widehat{\Gamma}_{\alpha,1}^{\mathrm{CV}} = \underbrace{\frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2}_{\xi_{\alpha,n}=} + \underbrace{\frac{1}{n}\sum_{i=1}^n \left[2w_{i,\alpha}^n + O((w_{i,\alpha}^n)^2)\right](Y_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2}_{\zeta_{\alpha,n}=}. \tag{3.39}$$

The first term $\xi_{\alpha,n}$ can be expressed as

$$\begin{aligned}\xi_{\alpha,n} &= \frac{1}{n}\sum_{i=1}^n \left[Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha) + \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2. \end{aligned} \tag{3.40}$$

The second sum in (3.40) could be rewritten as

$$\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)=\frac{2}{n}\Big[\underbrace{\sum_{i=1}^{n}\varepsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta}}_{=\boldsymbol{\varepsilon}^T\mathbb{X}\boldsymbol{\beta}}-\underbrace{\sum_{i=1}^{n}\varepsilon_i\boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha}_{=\boldsymbol{\varepsilon}^T\mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha}\Big]$$

$$\overset{(3.3)}{=}\tfrac{2}{n}\boldsymbol{\varepsilon}^T\mathbb{X}\boldsymbol{\beta}-\tfrac{2}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha(\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})$$

$$=\tfrac{2}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n-\mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}-\tfrac{2}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}.$$

Analogous as in Lemma 6, it could be shown that the third sum in (3.40) can be expressed as

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2=\Delta_{\alpha,n}+\tfrac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}.$$

Overall, one would obtain that

$$\xi_{\alpha,n}=\tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}+\Delta_{\alpha,n}+\tfrac{2}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n-\mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}-\tfrac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}. \tag{3.41}$$

By using Lemma 8, it holds that

$$\tfrac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}\xrightarrow[n\to\infty]{P}0\quad\text{and}\quad\tfrac{1}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n-\mathbb{P}_\alpha)\mathbb{X}\boldsymbol{\beta}\xrightarrow[n\to\infty]{P}0.$$

Therefore, one could simplify the equation (3.41) to

$$\xi_{\alpha,n}=\tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}+\Delta_{\alpha,n}+o_P(1). \tag{3.42}$$

Moreover, with the help of the law of large numbers, equation (3.42) can be rewritten as follows

$$\xi_{\alpha,n}=\sigma^2+\Delta_{\alpha,n}+o_P(1). \tag{3.43}$$

Bound the second term $\zeta_{\alpha,n}$ as follows

$$0\le\zeta_{\alpha,n}\le\frac{1}{n}\sum_{i=1}^{n}\Big[2w_{i,\alpha}^n+O(w_{i,\alpha}^n)\Big](Y_i-\boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2\le O\left(\max_{1\le i\le n}w_{i,\alpha}^n\right)\xi_{\alpha,n}$$

$$\overset{(3.35)}{\le}o(1)\xi_{\alpha,n}\overset{(3.43)}{=}o(1)\Big(\sigma^2+\Delta_{\alpha,n}+o_P(1)\Big)$$

$$\overset{(3.34)}{=}o(1)\Big(\sigma^2+O(1)+o_P(1)\Big)=o_P(1).$$

In other words it is proven that $\zeta_{\alpha,n}=o_P(1)$. Combining it with (3.39) and (3.43) one would get that

$$\widehat{\Gamma}_{\alpha,1}^{\mathrm{CV}}=\Gamma_{\alpha,n}+o_P(1).$$

□

Theorem 9 states that $\widehat{\Gamma}_{\alpha,1}^{\mathrm{CV}}$ is consistent for $\Gamma_{\alpha,n}$. In general, consistency does not guarantee that CV(1) works, i.e., it does not ensure that the condition (3.26) is satisfied. It will be discussed later that CV(1) actually does not work. However in the next theorem it is shown that CV(1) deals correctly with not favoring models from Category I.

Theorem 10. *Assume that conditions (3.13), (3.29) and (3.35) are satisfied. Then the following holds*

$$\lim_{n\to\infty} \mathrm{P}\left[M_{\widehat{\alpha}(1)} \text{ is in Category I}\right] = 0. \tag{3.44}$$

Proof. Let $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$ and the following conditions are satisfied

- $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$,
- $\forall \alpha \in \mathcal{A}_1 : \quad M_\alpha$ is a model in Category I,
- $\forall \gamma \in \mathcal{A}_2 : \quad M_\gamma$ is a model in Category II.

By fixing any $\eta \in \mathcal{A}_2$, one obtains the following

$$\begin{aligned}\left\{M_{\widehat{\alpha}(1)} \text{ is in Category I}\right\} &= \left\{\exists \alpha \in \mathcal{A}_1 \ \forall \gamma \in \mathcal{A} : \ \widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} \le \widehat{\Gamma}^{\mathrm{CV}}_{\gamma,1}\right\} \\ &\subseteq \left\{\exists \alpha \in \mathcal{A}_1 \ : \widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} \le \widehat{\Gamma}^{\mathrm{CV}}_{\eta,1}\right\} = \bigcup_{\alpha\in\mathcal{A}_1} \left\{\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} \le \widehat{\Gamma}^{\mathrm{CV}}_{\eta,1}\right\}. \end{aligned} \tag{3.45}$$

Therefore, it holds that

$$\begin{aligned}\mathrm{P}\left[M_{\widehat{\alpha}(1)} \text{ is in Category I}\right] &\overset{(3.45)}{\le} \sum_{\alpha\in\mathcal{A}_1} \mathrm{P}\left[\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} < \widehat{\Gamma}^{\mathrm{CV}}_{\eta,1}\right] \\ &\overset{\text{Thm. } 9}{=} \sum_{\alpha\in\mathcal{A}_1} \mathrm{P}\Big[\Gamma_{\alpha,n} < \Gamma_{\eta,n} + o_P(1)\Big] \overset{\text{Lem. } 6}{=} \sum_{\alpha\in\mathcal{A}_1} \mathrm{P}\Big[\tfrac{1}{n} d_\alpha \sigma^2 + \Delta_{\alpha,n} < \tfrac{1}{n} d_\eta \sigma^2 + o_P(1)\Big] \\ &= \sum_{\alpha\in\mathcal{A}_1} \mathrm{P}\Big[\Delta_{\alpha,n} < o_P(1)\Big]. \end{aligned}$$

Since $\liminf_{n\to\infty} \Delta_{\alpha,n} > 0$ for all $\alpha \in \mathcal{A}_1$ from the assumption (3.13), and $|\mathcal{A}_1| < 2^p$, one obtains the statement of the theorem. □

Asymptotic incorrectness of leave-one-out cross-validation

Recall from Lemma 6 that if M_α is in Category II, then

$$\Gamma_{\alpha,n} = \sigma^2 + \tfrac{\sigma^2}{n} d_\alpha \underset{n\to\infty}{\longrightarrow} \sigma^2. \tag{3.46}$$

In other words, as $n \to \infty$, it will be harder to distinguish models from Category II based on their prediction errors $\Gamma_{\alpha,n}$. To demonstrate that CV(1) is asymptotically incorrect, a more detailed asymptotic representation of $\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1}$ will be required for models in Category II. The following theorem provides such an asymptotic representation under stronger assumptions on moments of random errors.

Theorem 11. *Let the random errors $\varepsilon_1, \ldots, \varepsilon_n$, from the regression model (3.1), have finite fourth moments. Assume, as in Theorem 9, that conditions (3.29) and (3.35) are satisfied. If M_α is a model in Category* II*, then the following holds*

$$\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} = \tfrac{1}{n} \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} + \tfrac{2}{n} d_\alpha \sigma^2 - \tfrac{1}{n} \boldsymbol{\varepsilon}^T \mathbb{P}_\alpha \boldsymbol{\varepsilon} + o_P\left(\tfrac{1}{n}\right). \tag{3.47}$$

Proof. Because of M_α is a model in Category II, $\mathbb{X}\boldsymbol{\beta} = \mathbb{X}_\alpha \boldsymbol{\beta}_\alpha$ and hence

$$\Delta_{\alpha,n} = 0 \ \text{ and } \ \mathbb{P}_\alpha \mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta}.$$

Therefore, from (3.41) one gets

$$\xi_{\alpha,n} = \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} - \tfrac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}. \tag{3.48}$$

Now, it needs to be shown that the following asymptotic representation holds

$$\begin{aligned}\zeta_{\alpha,n} &= \underbrace{\frac{2}{n}\sum_{i=1}^n w_{i,\alpha}^n (Y_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2}_{A_n^{(\alpha)}=} + \underbrace{\frac{1}{n}\sum_{i=1}^n O\left((w_{i,\alpha}^n)^2\right)(Y_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2}_{B_n^{(\alpha)}=} \\ &= \tfrac{2}{n}d_\alpha\sigma^2 + o_P\left(\tfrac{1}{n}\right). \end{aligned} \tag{3.49}$$

The first sum $A_n^{(\alpha)}$ can be expressed as follows

$$\begin{aligned} A_n^{(\alpha)} &= \frac{2}{n}\sum_{i=1}^n w_{i,\alpha}^n (Y_i - \boldsymbol{x}_{i,\alpha}^T\boldsymbol{\beta}_\alpha + \boldsymbol{x}_{i,\alpha}^T\boldsymbol{\beta}_\alpha - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2 \\ &= \underbrace{\frac{2}{n}\sum_{i=1}^n w_{i,\alpha}^n\varepsilon_i^2}_{A_{n1}^{(\alpha)}=} + \underbrace{\frac{4}{n}\sum_{i=1}^n w_{i,\alpha}^n\varepsilon_i(\boldsymbol{x}_{i,\alpha}^T\boldsymbol{\beta}_\alpha - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)}_{A_{n2}^{(\alpha)}=} \\ &\quad + \underbrace{\frac{2}{n}\sum_{i=1}^n w_{i,\alpha}^n(\boldsymbol{x}_{i,\alpha}^T\boldsymbol{\beta}_\alpha - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha)^2}_{A_{n3}^{(\alpha)}=}. \end{aligned}$$

Gradually, the following will be shown

$$A_{n1}^{(\alpha)} = \tfrac{2}{n}d_\alpha\sigma^2 + o_P\left(\tfrac{1}{n}\right), \quad A_{n2}^{(\alpha)} = o_P\left(\tfrac{1}{n}\right), \quad A_{n3}^{(\alpha)} = o_P\left(\tfrac{1}{n}\right).$$

Showing that $A_{n1}^{(\alpha)} = \frac{2}{n}d_\alpha\sigma^2 + o_P\left(\frac{1}{n}\right)$.

It is sufficient to show that $\sum_{i=1}^n w_{i,\alpha}^n\varepsilon_i^2 = d_\alpha\sigma^2 + o_P(1)$. Define the following random variable $Z_n = \sum_{i=1}^n w_{i,\alpha}^n\varepsilon_i^2$, then

$$\begin{aligned} \mathbb{E}Z_n &= \sum_{i=1}^n w_{i,\alpha}^n\sigma^2 = \sigma^2\mathrm{tr}(\mathbb{P}_\alpha) = \sigma^2\mathrm{rank}(\mathbb{P}_\alpha) = \sigma^2 d_\alpha, \\ \mathrm{var}(Z_n) &= \sum_{i=1}^n (w_{i,\alpha}^n)^2\mathrm{var}\left(\varepsilon_i^2\right) = \mathrm{var}\left(\varepsilon_1^2\right)\cdot\sum_{i=1}^n (w_{i,\alpha}^n)^2 \\ &\le \underbrace{\mathrm{var}\left(\varepsilon_1^2\right)}_{<\infty}\cdot\underbrace{\left(\max_{1\le i\le n} w_{i,\alpha}^n\right)}_{\underset{n\to\infty}{\longrightarrow} 0 \text{ by (3.35)}}\cdot\underbrace{\sum_{i=1}^n w_{i,\alpha}^n}_{=d_\alpha} \underset{n\to\infty}{\longrightarrow} 0. \end{aligned}$$

Thus, by using Lemma A.2, it holds that $\sum_{i=1}^n w_{i,\alpha}^n\varepsilon_i^2 = d_\alpha\sigma^2 + o_P(1)$.

Dealing with $A_{n2}^{(\alpha)}$.

Define a matrix $\mathbb{D}_\alpha$ and a random vector $\boldsymbol{\varepsilon}_w$, as follows

$$\mathbb{D}_\alpha = \text{diag}(w_{1,\alpha}^n, \ldots, w_{n,\alpha}^n),$$

$$\boldsymbol{\varepsilon}_w = (w_{1,\alpha}^n \varepsilon_1, \ldots, w_{n,\alpha}^n \varepsilon_n)^T = \mathbb{D}_\alpha \boldsymbol{\varepsilon}.$$

Thus, one can express $A_{n2}^{(\alpha)}$ in the following way

$$A_{n2}^{(\alpha)} = \frac{4}{n} \sum_{i=1}^{n} w_{i,\alpha}^n \varepsilon_i (\boldsymbol{x}_{i,\alpha}^T \boldsymbol{\beta}_\alpha - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_\alpha) = \tfrac{4}{n} \boldsymbol{\varepsilon}_w^T (\mathbb{X}_\alpha \boldsymbol{\beta}_\alpha - \mathbb{X}_\alpha \widehat{\boldsymbol{\beta}}_\alpha)$$

$$= \tfrac{4}{n} \boldsymbol{\varepsilon}_w^T (\mathbb{X}_\alpha \boldsymbol{\beta}_\alpha - \mathbb{P}_\alpha \mathbb{X}_\alpha \boldsymbol{\beta}_\alpha - \mathbb{P}_\alpha \boldsymbol{\varepsilon}) = -\tfrac{4}{n} \boldsymbol{\varepsilon}^T \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon}.$$

Therefore one can see that $A_{n2}^{(\alpha)} = o_P(\frac{1}{n}) \iff \boldsymbol{\varepsilon}^T \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon} = o_P(1)$. By using Lemma A.1 together with $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}$, it holds that

$$\mathbb{E}\Big[\boldsymbol{\varepsilon}^T \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon}\Big] = \text{tr}(\sigma^2 \mathbb{D}_\alpha \mathbb{P}_\alpha) = \sigma^2 \sum_{i=1}^{n} (w_{i,\alpha}^n)^2 \leq \Big(\max_{1 \leq i \leq n} w_{i,\alpha}^n\Big) d_\alpha \sigma^2 \overset{(3.35)}{\underset{n\to\infty}{\longrightarrow}} 0. \quad (3.50)$$

Since $\mathbb{P}_\alpha$ is a positive-semidefinite matrix and $\mathbb{D}_\alpha$ consists of the diagonal elements of $\mathbb{P}_\alpha$, then it holds that $\mathbb{E}\Big[\boldsymbol{\varepsilon}^T \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon}\Big] \geq 0$, which together with (3.50) implies that $\boldsymbol{\varepsilon}^T \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon} = o_P(1)$.

Dealing with $A_{n3}^{(\alpha)}$.

Similarly, one can express $A_{n3}^{(\alpha)}$ as

$$A_{n3}^{(\alpha)} = \frac{2}{n} \sum_{i=1}^{n} w_{i,\alpha}^n (\boldsymbol{x}_{i,\alpha}^T \boldsymbol{\beta}_\alpha - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_\alpha)^2 = \tfrac{2}{n} \Big[\mathbb{D}_\alpha \mathbb{X}_\alpha (\boldsymbol{\beta}_\alpha - \widehat{\boldsymbol{\beta}}_\alpha)\Big]^T \Big[\mathbb{X}_\alpha (\boldsymbol{\beta}_\alpha - \widehat{\boldsymbol{\beta}}_\alpha)\Big]$$

$$= \tfrac{2}{n} \boldsymbol{\varepsilon}^T \mathbb{P}_\alpha \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon}.$$

As previously, it will be sufficient to prove that $\boldsymbol{\varepsilon}^T \mathbb{P}_\alpha \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon} = o_P(1)$, which follows from

$$\mathbb{E}\Big[\boldsymbol{\varepsilon}^T \mathbb{P}_\alpha \mathbb{D}_\alpha \mathbb{P}_\alpha \boldsymbol{\varepsilon}\Big] \overset{\text{Lem. A.1}}{=} \sigma^2 \text{tr}(\mathbb{P}_\alpha \mathbb{D}_\alpha \mathbb{P}_\alpha) = \sigma^2 \text{tr}(\mathbb{D}_\alpha \mathbb{P}_\alpha^2) = \sigma^2 \text{tr}(\mathbb{D}_\alpha \mathbb{P}_\alpha) \overset{(3.50)}{\underset{n\to\infty}{\longrightarrow}} 0.$$

Overall, it is proven that

$$A_n^{(\alpha)} = \frac{2}{n} \sum_{i=1}^{n} w_{i,\alpha}^n (Y_i - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_\alpha)^2 = \tfrac{2}{n} d_\alpha \sigma^2 + o_P\left(\tfrac{1}{n}\right). \quad (3.51)$$

Showing that $B_n^{(\alpha)} = o_P\left(\frac{1}{n}\right)$.

To complete the proof of the entire theorem, one needs to show that

$$B_n^{(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} O\left((w_{i,\alpha}^n)^2\right) (Y_i - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_\alpha)^2 = o_P\left(\tfrac{1}{n}\right). \quad (3.52)$$

It holds that

$$B_n^{(\alpha)} = O(1)\ \frac{1}{n} \sum_{i=1}^{n} (w_{i,\alpha}^n)^2 (Y_i - \boldsymbol{x}_{i,\alpha}^T \widehat{\boldsymbol{\beta}}_\alpha)^2 \leq O(1) \max_{1 \leq i \leq n} w_{i,\alpha}^n\ A_n^{(\alpha)}, \quad (3.53)$$

which together with the assumption (3.35) and expression (3.51) implies that expression (3.52) is true. □

It is important to note that the residual term $o_P(\frac{1}{n})$ from Theorem 11 does indeed depend on α. However, since there are only finitely many such α, the dependence is not explicitly specified. Now, all necessary ingredients are prepared to demonstrate the asymptotic incorrectness of CV(1).

Theorem 12. *Suppose that the true regression parameter $\boldsymbol{\beta}$, as given by* (3.1)*, consists of at least one zero component, and $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} \mathsf{N}(0, \sigma^2)$. Also, assume that conditions (3.29) and (3.35) are satisfied, and* $\dim(M_*) \neq p$*. Then the following holds*

$$\liminf_{n\to\infty} \mathrm{P}\left[M_{\widehat{\alpha}(1)} \neq M_*\right] > 0. \tag{3.54}$$

Proof. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined similarly as in Theorem 10. Fix $\alpha \in \mathcal{A}_2$ such that $\alpha_* \subset \alpha$, $d_{\alpha_*} < d_\alpha$ and $M_* \neq M_\alpha$, where α_* is a subset of $\{1, \ldots, p\}$ corresponding to M_*. Note that, from the assumptions of the theorem, model M_α exists. Hence

$$\begin{aligned}
&\{M_\alpha \text{ is preferable to } M_* \text{ by the CV(1)}\} = \left\{\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} < \widehat{\Gamma}^{\mathrm{CV}}_{\alpha_*,1}\right\} \\
&\subseteq \left\{\exists \gamma \in \mathcal{A}\backslash\{\alpha_*\} :\ \widehat{\Gamma}^{\mathrm{CV}}_{\gamma,1} < \widehat{\Gamma}^{\mathrm{CV}}_{\alpha_*,1}\right\} = \left\{M_{\widehat{\alpha}(1)} \neq M_*\right\}.
\end{aligned}$$

Therefore

$$\mathrm{P}\left[M_{\widehat{\alpha}(1)} \neq M_*\right] \geq \mathrm{P}\Big[M_\alpha \text{ is preferable to } M_* \text{ by the CV(1)}\Big]. \tag{3.55}$$

The right hand side of the inequality (3.55) is equal to $\mathrm{P}\left[\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} < \widehat{\Gamma}^{\mathrm{CV}}_{\alpha_*,1}\right]$, where $\alpha_* \subset \{1, \ldots, p\}$ corresponding to M_*. Because M_α and M_* are models from Category II, one can use Theorem 11 to calculate estimators of prediction errors $\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1}$ and $\widehat{\Gamma}^{\mathrm{CV}}_{\alpha_*,1}$. Note that $\mathbb{P}_\alpha - \mathbb{P}_{\alpha_*}$ is symmetric and from Lemma A.6 it holds that $\mathrm{rank}(\mathbb{P}_\alpha - \mathbb{P}_{\alpha_*}) = d_\alpha - d_{\alpha_*}$. Hence, one can use Theorem A.1 to obtain that $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^T(\mathbb{P}_\alpha - \mathbb{P}_{\alpha_*})\boldsymbol{\varepsilon}$ is a positive random variable, which follows a $\chi^2_{d_\alpha - d_{\alpha_*}}$ distribution. Therefore, it holds that

$$\begin{aligned}
\mathrm{P}\left[\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1} < \widehat{\Gamma}^{\mathrm{CV}}_{\alpha_*,1}\right] &\overset{\text{Thm. 11}}{=} \mathrm{P}\left[2(d_\alpha - d_{\alpha_*}) < \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^T(\mathbb{P}_\alpha - \mathbb{P}_{\alpha_*})\boldsymbol{\varepsilon} + o_P(1)\right] \\
&\overset{\text{Lem. A.7}}{\geq} \mathrm{P}\left[2(d_\alpha - d_{\alpha_*}) < \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^T(\mathbb{P}_\alpha - \mathbb{P}_{\alpha_*})\boldsymbol{\varepsilon}\right] + o(1) \\
&\overset{\text{Thm. A.1}}{=} \underbrace{\mathrm{P}\left[2(d_\alpha - d_{\alpha_*}) < \chi^2_{d_\alpha - d_{\alpha_*}}\right]}_{=c \text{ for some } c>0} + o(1).
\end{aligned} \tag{3.56}$$

Finally, by combining (3.55) and (3.56) and taking the limit inferior, the statement of the theorem is proven. □

The previous theorem assumes normality of errors. However, a similar result can be expected for errors with the second finite moment. Under additional conditions on the regressor matrix, the previous theorem could be extended using equation (A.9) from Lemma A.6 and the central limit theorem for independent, non-identically distributed random variables.

Theorem 12 states that CV(1) is asymptotically incorrect and too conservative, as it may favor a model of a larger dimension over the optimal model M_*.

Now, it will be intuitively explained why CV(1) is asymptotically incorrect. Recall from (3.46) that for models from Category II, an important component of the prediction error, which can distinguish different models, is $\frac{1}{n}d_\alpha\sigma^2$. At the same time, from (3.47), the component in $\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1}$ influenced by the difference between models is

$$\tfrac{1}{n}d_\alpha\sigma^2 + \delta_{\alpha,n},$$

where

$$\delta_{\alpha,n} = \tfrac{1}{n}d_\alpha\sigma^2 - \tfrac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}$$

is the error in assessing the differences of the models in Category II by using CV(1). Note that the error $\delta_{\alpha,n}$ has the same order of magnitude as $\frac{1}{n}d_\alpha\sigma^2$, hence leave-one-out cross-validation can not distinguish models in Category II.

3.3.2 Balanced incomplete cross-validation

In this section, the deficiency of CV(1) will be rectified by employing CV(n_v), with a large n_v, satisfying $n_v/n \longrightarrow 1$ as $n \to \infty$. This is a quite surprising discovery, since it is totally opposite to the popular leave-one-out recipe in cross-validation. The intuition behind choosing a relatively large n_v is that cross-validation involves two steps: (1) fitting a model using $n_c = n - n_v$ data, and (2) validating the fitted model using n_v data. Thus, achieving a highly accurate model fit in step (1) of the cross-validation is not necessarily crucial. Instead, accurate assessment of the prediction error in step (2) is essential, which underscores the importance of selecting n_v to be sufficiently large. Also, note that for estimating $\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,n_v}$ from (3.23), one must first define the collection $\mathcal{B}_n$ of subsets of the set $\{1,\ldots,n\}$ that have size n_v. Naturally, one may choose $\mathcal{B}_n$ as the system of all subsets of the set $\{1,\ldots,n\}$ of size n_v, but it turns out to be impractical and unnecessary. It will be shown that it suffices to consider only special collections of subsets.

Definition 3. *Let $\mathcal{B}_n$ be a collection of b_n subsets of $\{1,\ldots,n\}$ that have size n_v. Also, assume that $\mathcal{B}_n$ is selected according to the following conditions:*

- *every $i \in \{1,\ldots,n\}$ appears in the same number of subsets in $\mathcal{B}_n$,*
- *every pair $\{i,j\}$, where $i,j \in \{1,\ldots,n\}$ and $i \neq j$, appears in the same number of subsets in $\mathcal{B}_n$.*

Then the collection $\mathcal{B}_n$ is called a Balanced Incomplete Block Design (BIBD).

Remark. Examples of BIBD can be found in John [1971], Chapter 13. It is often assumed that the cardinality b_n of the collection $\mathcal{B}_n$ satisfies $b_n \geq n$ and $b_n = O(n)$.

Assume that for all $n \in \mathbb{N}$, $\mathcal{B}_n$ is selected according to Definition 3, then the cross-validation estimate of $\Gamma_{\alpha,n}$ will be referred to as the balanced incomplete CV(n_v), abbreviated as BICV(n_v), and is defined as

$$\widehat{\Gamma}^{\mathrm{BICV}}_{\alpha,n_v} = \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\left\|\boldsymbol{Y}_{s_n} - \widehat{\boldsymbol{Y}}_{\alpha,s_n^c}\right\|_2^2.$$

The model selected by balanced incomplete cross-validation is $M_{\widehat{\alpha}_B(n_v)}$, where

$$\widehat{\alpha}_B(n_v) = \arg\min_{\alpha \in \mathcal{A}} \widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v}. \tag{3.57}$$

The following two theorems about asymptotic representations of $\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v}$ for models in different categories are stated without proofs. The proofs are technical and will be presented in the next subsection. Here, the focus is on the main result, Theorem 15, which justifies the asymptotic correctness of balanced incomplete cross-validation.

Theorem 13. *Suppose that conditions (3.29) and (3.35) hold. Also, suppose that for all $n \in \mathbb{N}$, $\mathcal{B}_n$ is a Balanced Incomplete Block Design (BIBD). If M_α is a model in Category* I*, then there exists nonnegative random variables R_n such that*

$$\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v} = \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \Delta_{\alpha,n} + o_P(1) + R_n.$$

Using the previous theorem and the law of large numbers, one obtains that

$$\begin{aligned}\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v} &= \sigma^2 + \Delta_{\alpha,n} + o_P(1) + R_n \\ &\overset{\text{Lem. } 6}{=} \Gamma_{\alpha,n} - \underbrace{\tfrac{1}{n} d_\alpha \sigma^2}_{=o(1)} + o_P(1) + R_n = \Gamma_{\alpha,n} + o_P(1) + R_n.\end{aligned}$$

Now, it can be seen that unlike $\widehat{\Gamma}^{\text{CV}}_{\alpha,1}$, the balanced incomplete cross-validation estimate $\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v}$ is not consistent for $\Gamma_{\alpha,n}$, unless R_n converges to zero in probability as $n \to \infty$. However, this is not a significant drawback of the BICV(n_v) method, as one is primarily interested in another type of consistency.

Theorem 14. *Let conditions (3.29) and (3.35) hold, and for all $n \in \mathbb{N}$, the collection $\mathcal{B}_n$ be a Balanced Incomplete Block Design (BIBD). Also, assume that*

$$\lim_{n\to\infty} \max_{s_n \in \mathcal{B}_n} \left\| \frac{1}{n_v} \sum_{i \in s_n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \frac{1}{n_c} \sum_{i \in s_n^c} \boldsymbol{x}_i \boldsymbol{x}_i^T \right\| = 0. \tag{3.58}$$

Additionally, suppose that n_v is selected in the following way

$$\frac{n_v}{n} \underset{n\to\infty}{\longrightarrow} 1 \quad \textit{and} \quad n_c = n - n_v \underset{n\to\infty}{\longrightarrow} \infty. \tag{3.59}$$

Then, if M_α is in Category II*, it holds that*

$$\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v} = \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \tfrac{1}{n_c} d_\alpha \sigma^2 + o_P\left(\tfrac{1}{n_c}\right).$$

Recall from the previous section that CV(1) is an asymptotically incorrect method, meaning that it does not consistently prefer the optimal model M_* over other models from Category II (see Theorem 12). The reason is that $\widehat{\Gamma}^{\text{CV}}_{\alpha,1}$ is an estimate for $\Gamma_{\alpha,n}$, which for the models in Category II can be expressed as follows

$$\Gamma_{\alpha,n} = \sigma^2 + \tfrac{\sigma^2}{n} d_\alpha.$$

The problem here is that for large n, $\Gamma_{\alpha,n}$ is almost flat as a function of α, which means that it is harder to find the minimum of $\Gamma_{\alpha,n}$. At the same time, Theorem 14 states that $\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v}$ estimates Γ_{α,n_c} rather than $\Gamma_{\alpha,n}$. From the assumptions of Theorem 14, one knows that n_c is relatively small compared to n. Therefore, it gives the intuition that BICV(n_v) can better recognize the minimum point of Γ_{α,n_c} than CV(1). The following theorem provides the main result of this section, stating that BICV(n_v) is an asymptotically correct method.

Theorem 15. *Suppose that condition (3.13) and all assumptions of Theorem 14 are satisfied. Then the following holds*

$$\lim_{n\to\infty} \mathsf{P}\left[M_{\widehat{\alpha}_B(n_v)} \neq M_*\right] = 0,$$

which means that BICV(n_v) *is asymptotically correct.*

Proof. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined similarly as in Theorem 10 and α_* is a subset of $\{1,\dots,p\}$ corresponding to M_*.

$$\begin{aligned}
&\left\{M_{\widehat{\alpha}_B(n_v)} \neq M_*\right\} = \left\{\exists \eta \in \mathcal{A}\backslash\{\alpha_*\} :\ \widehat{\Gamma}^{\text{BICV}}_{\eta,n_v} < \widehat{\Gamma}^{\text{BICV}}_{\alpha_*,n_v}\right\} \\
&= \left\{\exists \gamma \in \mathcal{A}_1 :\ \widehat{\Gamma}^{\text{BICV}}_{\gamma,n_v} < \widehat{\Gamma}^{\text{BICV}}_{\alpha_*,n_v}\right\} \bigcup \left\{\exists \alpha \in \mathcal{A}_2\backslash\{\alpha_*\} :\ \widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v} < \widehat{\Gamma}^{\text{BICV}}_{\alpha_*,n_v}\right\}. \qquad (3.60)
\end{aligned}$$

By employing Theorem 13 and Theorem 14, one gets that for all $\gamma \in \mathcal{A}_1$ the following holds

$$\begin{aligned}
\mathsf{P}\left[\widehat{\Gamma}^{\text{BICV}}_{\gamma,n_v} < \widehat{\Gamma}^{\text{BICV}}_{\alpha_*,n_v}\right] &= \mathsf{P}\left[\tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \Delta_{\gamma,n} + o_P(1) + R_n < \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \tfrac{1}{n_c} d_{\alpha_*}\sigma^2 + o_P\left(\tfrac{1}{n_c}\right)\right] \\
&= \mathsf{P}\left[\Delta_{\gamma,n} + o_P(1) + R_n < \tfrac{1}{n_c} d_{\alpha_*}\sigma^2 + o_P\left(\tfrac{1}{n_c}\right)\right] \\
&\overset{R_n \text{ is nonneg.}}{\leq} \mathsf{P}\Big[\Delta_{\gamma,n} < \underbrace{\tfrac{1}{n_c} d_{\alpha_*}\sigma^2}_{=o(1)} + o_P(1)\Big]. \qquad (3.61)
\end{aligned}$$

Since $\liminf_{n\to\infty} \Delta_{\gamma,n} > 0$ for all $\gamma \in \mathcal{A}_1$ from the assumption (3.13), which together with (3.61) implies that for all $\gamma \in \mathcal{A}_1$ the following holds

$$\mathsf{P}\left[\widehat{\Gamma}^{\text{BICV}}_{\gamma,n_v} < \widehat{\Gamma}^{\text{BICV}}_{\alpha_*,n_v}\right] \underset{n\to\infty}{\longrightarrow} 0. \qquad (3.62)$$

Also, note that for all $\alpha \in \mathcal{A}_2\backslash\{\alpha_*\}$, it holds that $d_\alpha > d_{\alpha_*}$. Therefore

$$\begin{aligned}
&\mathsf{P}\left[\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v} < \widehat{\Gamma}^{\text{BICV}}_{\alpha_*,n_v}\right] \\
&\overset{\text{Thm. } 14}{=} \mathsf{P}\left[\tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \tfrac{1}{n_c} d_\alpha\sigma^2 + o_P\left(\tfrac{1}{n_c}\right) < \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \tfrac{1}{n_c} d_{\alpha_*}\sigma^2 + o_P\left(\tfrac{1}{n_c}\right)\right] \\
&= \mathsf{P}\left[(d_\alpha - d_{\alpha_*})\sigma^2 < o_P(1)\right] \underset{n\to\infty}{\longrightarrow} 0. \qquad (3.63)
\end{aligned}$$

Since the collection $\mathcal{A}$ contains only a finite number of sets, combining (3.60), (3.62) and (3.63) yields the statement of the theorem. □

3.3.3 Proofs of the Theorems from Subsection 3.3.2

Proof of Theorem 13

First, recall that the matrix $\mathbb{Q}_{\alpha,s_n}$ is defined as follows

$$\mathbb{Q}_{\alpha,s_n} = \mathbb{X}_{\alpha,s_n}(\mathbb{X}_{\alpha}^T\mathbb{X}_{\alpha})^{-1}\mathbb{X}_{\alpha,s_n}^T, \text{ where } s_n \in \mathcal{B}_n.$$

Note that $\forall n \in \mathbb{N}\ \forall s_n \in \mathcal{B}_n$, both $\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n}$ and $\mathbb{Q}_{\alpha,s_n}$ are symmetric and positive semidefinite matrices because they are $n_v \times n_v$ diagonal blocks of the symmetric and positive semidefinite matrices $\mathbb{I}_n - \mathbb{P}_\alpha$ and $\mathbb{P}_\alpha$, respectively. Therefore, for all $\boldsymbol{u} \in \mathbb{R}^{n_v}$ the following holds

$$0 \leq \boldsymbol{u}^T(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})\boldsymbol{u} \leq \boldsymbol{u}^T\boldsymbol{u}.$$

Analogous to inequalities (A.4) from Lemma A.3, one would get that

$$\begin{aligned} ||\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n}||_2^2 &= \sup_{||\boldsymbol{u}||_2=1} ||(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})\boldsymbol{u}||_2^2 = \sup_{||\boldsymbol{u}||_2=1} \boldsymbol{u}^T(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^2\boldsymbol{u} \\ &\leq \sup_{||\boldsymbol{u}||_2=1} \boldsymbol{u}^T(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})\boldsymbol{u} \leq \sup_{||\boldsymbol{u}||_2=1} \boldsymbol{u}^T\boldsymbol{u} = 1. \end{aligned} \tag{3.64}$$

Thus

$$\begin{aligned} &\left|\left|\boldsymbol{Y}_{s_n} - \mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2 \\ &= \left|\left|(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1}(\boldsymbol{Y}_{s_n} - \mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha)\right|\right|_2^2 \\ &\overset{\text{Lem. A.4}}{\leq} ||\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n}||_2^2 \ \cdot\ \left|\left|(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1}(\boldsymbol{Y}_{s_n} - \mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha)\right|\right|_2^2 \\ &\overset{(3.64)}{\leq} \left|\left|(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1}(\boldsymbol{Y}_{s_n} - \mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha)\right|\right|_2^2 \overset{\text{Lem. 7}}{=} \left|\left|\boldsymbol{Y}_{s_n} - \widehat{\boldsymbol{Y}}_{\alpha,s_n^c}\right|\right|_2^2. \end{aligned} \tag{3.65}$$

In other words, the following inequality is obtained

$$\widehat{\Gamma}_{\alpha,n_v}^{\text{BICV}} = \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\left|\left|\boldsymbol{Y}_{s_n} - \widehat{\boldsymbol{Y}}_{\alpha,s_n^c}\right|\right|_2^2 \overset{(3.65)}{\geq} \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}}\left|\left|\boldsymbol{Y}_{s_n} - \mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2. \tag{3.66}$$

Let m_n^i denote the number of occurrences of element $i \in \{1,\ldots,n\}$ in subsets $s_n \in \mathcal{B}_n$. Since $\mathcal{B}_n$ is a Balanced Incomplete Block Design (BIBD), it should hold that $m_n^1 = \ldots = m_n^n = m_n$. Also, because $|\mathcal{B}_n| = b_n$ and each subset $s_n \in \mathcal{B}_n$ satisfies $|s_n| = n_v$, hence $nm_n = n_v b_n$. Therefore, one obtains the following

$$\begin{aligned} &\frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\left|\left|\boldsymbol{Y}_{s_n} - \mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2 = \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\sum_{i_s\in s_n}\left(Y_{i_s} - \boldsymbol{x}_{i_s,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2 \\ &= \frac{1}{n_v b_n}\sum_{i=1}^n m_n^i\left(Y_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2 = \frac{m_n}{n_v b_n}\sum_{i=1}^n\left(Y_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n\left(Y_i - \boldsymbol{x}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_\alpha\right)^2 = \frac{1}{n}\left|\left|\boldsymbol{Y} - \mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2. \end{aligned} \tag{3.67}$$

By combining (3.66) and (3.67), one can define the random variable R_n as

$$R_n = \widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v} - \frac{1}{n}\left|\left|\boldsymbol{Y} - \mathbb{X}_\alpha \widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2 \geq 0. \tag{3.68}$$

Also, note that

$$\frac{1}{n}\left|\left|\boldsymbol{Y} - \mathbb{X}_\alpha \widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2 \overset{(3.39)}{=} \xi_{\alpha,n} \overset{(3.42)}{=} \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \Delta_{\alpha,n} + o_P(1). \tag{3.69}$$

Finally, from (3.68) and (3.69) one has

$$\widehat{\Gamma}^{\text{BICV}}_{\alpha,n_v} = \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \Delta_{\alpha,n} + o_P(1) + R_n.$$

□

To prove Theorem 14, one will need the following two lemmas.

Lemma 16. *Let $\mathcal{B}_n$ be a Balanced Incomplete Block Design (BIBD) and $\alpha \in \mathcal{A}$ such that M_α is a model in Category* II. *Define the following random vectors $\boldsymbol{r}_{\alpha,s_n} = \boldsymbol{Y}_{s_n} - \mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha$ for every $s_n \in \mathcal{B}_n$. Then the following equation holds*

$$\frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n} \boldsymbol{r}^T_{\alpha,s_n}\mathbb{Q}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n} = \left[\frac{1}{n} - \frac{n_v - 1}{n(n-1)}\right]\sum_{i=1}^n w^n_{i,\alpha} r^2_{i,\alpha},$$

where the matrix $\mathbb{Q}_{\alpha,s_n}$ is from Lemma 7, $w^n_{i,\alpha}$ is the i-th diagonal element of the projection matrix $\mathbb{P}_\alpha$, and $r_{i,\alpha} = Y_i - \boldsymbol{x}^T_{i,\alpha}\widehat{\boldsymbol{\beta}}_\alpha$.

Proof. Obviously, $(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1}$ is symmetric and positive definite matrix, hence there exists symmetric and positive definite matrix $(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1/2}$ such that

$$(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1} = (\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1/2}(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1/2}.$$

Define the following matrix $\widetilde{\mathbb{X}}_{\alpha,s_n} = \mathbb{X}_{\alpha,s_n}(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1/2}$. Then one can rewrite $\mathbb{Q}_{\alpha,s_n}$ as

$$\begin{aligned}\mathbb{Q}_{\alpha,s_n} &= \mathbb{X}_{\alpha,s_n}(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1}\mathbb{X}^T_{\alpha,s_n}\\ &= \mathbb{X}_{\alpha,s_n}(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1/2}(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1/2}\mathbb{X}^T_{\alpha,s_n} = \widetilde{\mathbb{X}}_{\alpha,s_n}\widetilde{\mathbb{X}}^T_{\alpha,s_n}.\end{aligned}$$

Further, it will be useful to express $\widetilde{\mathbb{X}}_{\alpha,s_n}$ as follows

$$\widetilde{\mathbb{X}}_{\alpha,s_n} = (\widetilde{\boldsymbol{x}}_{i,\alpha} : i \in s_n)^T. \tag{3.70}$$

Therefore, one obtains the following

$$\begin{aligned}&\frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n} \boldsymbol{r}^T_{\alpha,s_n}\mathbb{Q}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}\\ &= \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n} \boldsymbol{r}^T_{\alpha,s_n}\,\widetilde{\mathbb{X}}_{\alpha,s_n}\widetilde{\mathbb{X}}^T_{\alpha,s_n}\,\boldsymbol{r}_{\alpha,s_n} = \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\left(\sum_{i\in s_n} r_{i,\alpha}\widetilde{\boldsymbol{x}}^T_{i,\alpha}\right)\left(\sum_{j\in s_n} r_{j,\alpha}\widetilde{\boldsymbol{x}}_{j,\alpha}\right)\end{aligned}$$

$$= \frac{1}{n_v b_n} \sum_{s_n \in \mathcal{B}_n} \left(\sum_{i \in s_n} r_{i,\alpha}^2 \widetilde{\boldsymbol{x}}_{i,\alpha}^T \widetilde{\boldsymbol{x}}_{i,\alpha} \right) + \frac{1}{n_v b_n} \sum_{s_n \in \mathcal{B}_n} \left(\sum_{\substack{i \neq j \\ i,j \in s_n}} \sum r_{i,\alpha} r_{j,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha}^T \widetilde{\boldsymbol{x}}_{j,\alpha} \right). \tag{3.71}$$

Since $\mathcal{B}_n$ is a Balanced Incomplete Block Design (BIBD), there exist m_n and h_n satisfying the following conditions:

- m_n represents the number of occurrences of each element $i \in \{1, \ldots, n\}$ in subsets $s_n \in \mathcal{B}_n$,
- h_n represents the number of occurrences of each pair $\{i, j\}$, such that $i \neq j$ and $i, j \in \{1, \ldots, n\}$, in subsets $s_n \in \mathcal{B}_n$.

Also, because $|\mathcal{B}_n| = b_n$ and each subset $s_n \in \mathcal{B}_n$ satisfies $|s_n| = n_v$, then the following holds

$$\begin{aligned} & m_n n = b_n n_v \ \text{ and } \ h_n \binom{n}{2} = b_n \binom{n_v}{2} \\ \iff & m_n = \frac{b_n n_v}{n} \ \text{ and } \ h_n = \frac{b_n n_v (n_v - 1)}{n(n-1)}. \end{aligned} \tag{3.72}$$

Thus, it holds that

$$\frac{1}{n_v b_n} \sum_{s_n \in \mathcal{B}_n} \left(\sum_{i \in s_n} r_{i,\alpha}^2 \widetilde{\boldsymbol{x}}_{i,\alpha}^T \widetilde{\boldsymbol{x}}_{i,\alpha} \right) = \frac{m_n}{n_v b_n} \sum_{i=1}^n r_{i,\alpha}^2 \widetilde{\boldsymbol{x}}_{i,\alpha}^T \widetilde{\boldsymbol{x}}_{i,\alpha} \overset{(3.72)}{=} \frac{1}{n} \sum_{i=1}^n r_{i,\alpha}^2 w_{i,\alpha}^n. \tag{3.73}$$

Now, it will be proven that $\sum_{i=1}^n r_{i,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha} = 0$. Denote $\boldsymbol{r}_\alpha = (r_{1,\alpha}, \ldots, r_{n,\alpha})^T$. Since M_α is a model in Category II, one has

$$\boldsymbol{r}_\alpha = \boldsymbol{Y} - \mathbb{X}_\alpha \widehat{\boldsymbol{\beta}}_\alpha \overset{(3.2)}{=} \mathbb{X}_\alpha \boldsymbol{\beta}_\alpha + \boldsymbol{\varepsilon} - \mathbb{X}_\alpha \widehat{\boldsymbol{\beta}}_\alpha \overset{(3.3)}{=} (\mathbb{I}_n - \mathbb{P}_\alpha) \boldsymbol{\varepsilon}. \tag{3.74}$$

Similar to (3.70), one can define $\widetilde{\mathbb{X}}_\alpha = (\widetilde{\boldsymbol{x}}_{i,\alpha} : i = 1, \ldots, n)^T$. Therefore

$$\sum_{i=1}^n r_{i,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha} = \widetilde{\mathbb{X}}_\alpha^T \boldsymbol{r}_\alpha \overset{(3.74)}{=} (\mathbb{X}_\alpha^T \mathbb{X}_\alpha)^{-1/2} \mathbb{X}_\alpha^T (\mathbb{I}_n - \mathbb{P}_\alpha) \boldsymbol{\varepsilon} = \boldsymbol{0}. \tag{3.75}$$

Finally, the following holds

$$\begin{aligned} & \frac{1}{n_v b_n} \sum_{s_n \in \mathcal{B}_n} \left(\sum_{\substack{i \neq j \\ i,j \in s_n}} \sum r_{i,\alpha} r_{j,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha}^T \widetilde{\boldsymbol{x}}_{j,\alpha} \right) = \frac{h_n}{n_v b_n} \sum_{i \neq j}^n \sum^n r_{i,\alpha} r_{j,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha}^T \widetilde{\boldsymbol{x}}_{j,\alpha} \\ & \overset{(3.72)}{=} \frac{n_v - 1}{n(n-1)} \sum_{i \neq j}^n \sum^n r_{i,\alpha} r_{j,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha}^T \widetilde{\boldsymbol{x}}_{j,\alpha} \\ & = \frac{n_v - 1}{n(n-1)} \sum_{i=1}^n r_{i,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha}^T \Bigg[\underbrace{\sum_{j=1}^n r_{j,\alpha} \widetilde{\boldsymbol{x}}_{j,\alpha}}_{\overset{(3.75)}{=} 0} - r_{i,\alpha} \widetilde{\boldsymbol{x}}_{i,\alpha} \Bigg] = -\frac{n_v - 1}{n(n-1)} \sum_{i=1}^n r_{i,\alpha}^2 w_{i,\alpha}. \end{aligned} \tag{3.76}$$

By combining (3.71), (3.73) and (3.76), it holds that

$$\frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}^T_{\alpha,s_n}\mathbb{Q}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}=\left[\frac{1}{n}-\frac{n_v-1}{n(n-1)}\right]\sum_{i=1}^{n}r^2_{i,\alpha}w^n_{i,\alpha}.$$

□

Lemma 17. *Suppose that conditions (3.29) and (3.59) hold. Let $\forall n\in\mathbb{N}$, the collection $\mathcal{B}_n$ be a Balanced Incomplete Block Design (BIBD), and let condition (3.58) also be satisfied. Then, for all $s_n\in\mathcal{B}_n$, the following holds*

$$\mathbb{Q}_{\alpha,s_n}=\left[\tfrac{n_v}{n}+o\left(\tfrac{n_c}{n}\right)\right]\mathbb{P}_{\alpha,s_n},$$

where

$$\mathbb{Q}_{\alpha,s_n}=\mathbb{X}_{\alpha,s_n}(\mathbb{X}^T_\alpha\mathbb{X}_\alpha)^{-1}\mathbb{X}^T_{\alpha,s_n}\quad\textit{and}\quad\mathbb{P}_{\alpha,s_n}=\mathbb{X}_{\alpha,s_n}\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}\mathbb{X}^T_{\alpha,s_n}.$$

Proof. For all $n\in\mathbb{N}$ and $s_n\in\mathcal{B}_n$ the following holds

$$\begin{aligned}
&\tfrac{1}{n}\mathbb{X}^T_\alpha\mathbb{X}_\alpha-\tfrac{1}{n_v}\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}=\tfrac{1}{n}\left(\mathbb{X}^T_{\alpha,s_n}\ \mathbb{X}^T_{\alpha,s^c_n}\right)\begin{pmatrix}\mathbb{X}_{\alpha,s_n}\\ \mathbb{X}_{\alpha,s^c_n}\end{pmatrix}-\tfrac{1}{n_v}\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\\
&=\tfrac{1}{n}\mathbb{X}^T_{\alpha,s^c_n}\mathbb{X}_{\alpha,s^c_n}+\left(\tfrac{1}{n}-\tfrac{1}{n_v}\right)\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}=\tfrac{1}{n}\mathbb{X}^T_{\alpha,s^c_n}\mathbb{X}_{\alpha,s^c_n}-\tfrac{n_c}{nn_v}\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\\
&=\tfrac{n_c}{n}\underbrace{\left[\tfrac{1}{n_c}\mathbb{X}^T_{\alpha,s^c_n}\mathbb{X}_{\alpha,s^c_n}-\tfrac{1}{n_v}\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right]}_{\overset{(3.58)}{=}o(1)}=o\left(\tfrac{n_c}{n}\right).
\end{aligned}\tag{3.77}$$

It holds that

$$\begin{aligned}
&\tfrac{1}{n}\mathbb{X}^T_\alpha\mathbb{X}_\alpha\left[\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}-\tfrac{n}{n_v}\left(\mathbb{X}^T_\alpha\mathbb{X}_\alpha\right)^{-1}\right]=\tfrac{1}{n}\mathbb{X}^T_\alpha\mathbb{X}_\alpha\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}-\tfrac{1}{n_v}\mathbb{I}_{d_\alpha}\\
&=\left[\tfrac{1}{n}\mathbb{X}^T_\alpha\mathbb{X}_\alpha-\tfrac{1}{n_v}\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right]\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}\overset{(3.77)}{=}o\left(\tfrac{n_c}{n}\right)\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}.
\end{aligned}$$

Therefore one obtains

$$\begin{aligned}
&\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}-\tfrac{n}{n_v}\left(\mathbb{X}^T_\alpha\mathbb{X}_\alpha\right)^{-1}=o\left(\tfrac{n_c}{n}\right)\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{i,\alpha}\boldsymbol{x}^T_{i,\alpha}\right]^{-1}\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}\\
&\overset{(3.29)}{=}o\left(\tfrac{n_c}{n}\right)O(1)\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}=o\left(\tfrac{n_c}{n}\right)\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}.
\end{aligned}$$

Finally, one gets the following equation

$$\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}-\tfrac{n}{n_v}\left(\mathbb{X}^T_\alpha\mathbb{X}_\alpha\right)^{-1}=o\left(\tfrac{n_c}{n}\right)\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}.\tag{3.78}$$

Denote $\mathbb{P}_{\alpha,s_n}=\mathbb{X}_{\alpha,s_n}\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}\mathbb{X}^T_{\alpha,s_n}$ as the projection matrix onto the column space of $\mathbb{X}_{\alpha,s_n}$, where $s_n\in\mathcal{B}_n$. Then, by using (3.78), one can express $\mathbb{P}_{\alpha,s_n}$ as follows

$$\begin{aligned}
\mathbb{P}_{\alpha,s_n}&\overset{(3.78)}{=}\mathbb{X}_{\alpha,s_n}\left[o\left(\tfrac{n_c}{n}\right)\left(\mathbb{X}^T_{\alpha,s_n}\mathbb{X}_{\alpha,s_n}\right)^{-1}+\tfrac{n}{n_v}\left(\mathbb{X}^T_\alpha\mathbb{X}_\alpha\right)^{-1}\right]\mathbb{X}^T_{\alpha,s_n}\\
&=o\left(\tfrac{n_c}{n}\right)\mathbb{P}_{\alpha,s_n}+\tfrac{n}{n_v}\mathbb{Q}_{\alpha,s_n}.
\end{aligned}\tag{3.79}$$

From (3.79) and condition (3.59),

$$\mathbb{Q}_{\alpha,s_n}=\left[\tfrac{n_v}{n}+o\left(\tfrac{n_c}{n}\right)\right]\mathbb{P}_{\alpha,s_n}.\tag{3.80}$$

□

Note that the preceding lemma states that the matrix $\mathbb{Q}_{\alpha,s_n}$ is actually a rescaled projection matrix $\mathbb{P}_{\alpha,s_n}$ in the asymptotic sense. Now, everything is ready to prove Theorem 14.

Proof of Theorem 14

Define the sequence of constants $c_n = \frac{n_v(n+n_c)}{n_c^2}$ and suppose that the random vectors $\boldsymbol{r}_{\alpha,s_n}$ are as defined in Lemma 16. Thus, one gets the following

$$\begin{aligned}
\frac{c_n}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}||\mathbb{P}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}||_2^2 &= \frac{c_n}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}_{\alpha,s_n}^T\mathbb{P}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}\\
&\overset{\text{Lem. 17}}{=} \frac{c_n}{n_v b_n}\left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]^{-1}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}_{\alpha,s_n}^T\mathbb{Q}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}\\
&= \frac{c_n n}{n_v}\left[1+o\left(\frac{n_c}{n}\right)\right]^{-1}\frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}_{\alpha,s_n}^T\mathbb{Q}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}\\
&\overset{\text{Lem. 16}}{=} \frac{c_n n}{n_v}\left[1+o\left(\frac{n_c}{n}\right)\right]^{-1}\left[\frac{1}{n}-\frac{n_v-1}{n(n-1)}\right]\sum_{i=1}^n w_{i,\alpha}^n r_{i,\alpha}^2\\
&= \frac{c_n n}{n_v}\left[1+o\left(\frac{n_c}{n}\right)\right]^{-1}\frac{n_c}{n(n-1)}\sum_{i=1}^n w_{i,\alpha}^n r_{i,\alpha}^2\\
&= \frac{n_v(n+n_c)}{n_c^2}\frac{n}{n_v}\left[1+o\left(\frac{n_c}{n}\right)\right]^{-1}\frac{n_c}{n(n-1)}\sum_{i=1}^n w_{i,\alpha}^n r_{i,\alpha}^2\\
&\overset{\text{Lem. A.8}}{=} \left[1+o\left(\frac{n_c}{n}\right)\right]\frac{n+n_c}{n_c(n-1)}\sum_{i=1}^n w_{i,\alpha}^n r_{i,\alpha}^2. \qquad (3.81)
\end{aligned}$$

Define

$$\mathbb{U}_{\alpha,s_n} = (\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})(\mathbb{I}_{n_v}+c_n\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n}),$$

$$\begin{aligned}
A_n^{(\alpha)} &= \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}_{\alpha,s_n}^T(\mathbb{I}_{n_v}+c_n\mathbb{P}_{\alpha,s_n})\boldsymbol{r}_{\alpha,s_n}\\
&= \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}_{\alpha,s_n}^T(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})^{-1}\mathbb{U}_{\alpha,s_n}(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})^{-1}\boldsymbol{r}_{\alpha,s_n},
\end{aligned}$$

and

$$B_n^{(\alpha)} = \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}_{\alpha,s_n}^T(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})^{-1}(\mathbb{I}_{n_v}-\mathbb{U}_{\alpha,s_n})(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})^{-1}\boldsymbol{r}_{\alpha,s_n}.$$

Therefore

$$\begin{aligned}
\widehat{\Gamma}_{\alpha,n_v}^{\text{BICV}} &= \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\left|\left|\boldsymbol{Y}_{s_n}-\widehat{\boldsymbol{Y}}_{\alpha,s_n^c}\right|\right|_2^2\\
&\overset{\text{Lem. 7}}{=} \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\left|\left|(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})^{-1}\left(\boldsymbol{Y}_{s_n}-\mathbb{X}_{\alpha,s_n}\widehat{\boldsymbol{\beta}}_\alpha\right)\right|\right|_2^2\\
&= \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n}\boldsymbol{r}_{\alpha,s_n}^T(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})^{-1}(\mathbb{I}_{n_v}-\mathbb{Q}_{\alpha,s_n})^{-1}\boldsymbol{r}_{\alpha,s_n} = A_n^{(\alpha)}+B_n^{(\alpha)}.
\end{aligned}$$

To complete the proof, it is sufficient to show that

$$A_n^{(\alpha)} = \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \tfrac{1}{n_c}d_\alpha\sigma^2 + o_P\left(\tfrac{1}{n_c}\right) \quad \text{and} \quad B_n^{(\alpha)} = o_P\left(\tfrac{1}{n_c}\right).$$

Showing that $A_n^{(\alpha)} = \frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \frac{1}{n_c}d_\alpha\sigma^2 + o_P\left(\frac{1}{n_c}\right)$.

Since, M_α is a model from Category II, equation (3.51) can be rewritten as follows

$$\sum_{i=1}^{n} w_{i,\alpha} r_{i,\alpha}^2 = d_\alpha\sigma^2 + o_P(1). \tag{3.82}$$

From the balance property of $\mathcal{B}_n$, one gets the following

$$\begin{aligned}
A_n^{(\alpha)} &= \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n} \boldsymbol{r}_{\alpha,s_n}^T(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n})\boldsymbol{r}_{\alpha,s_n} \\
&= \frac{1}{n_v b_n}\sum_{s_n\in\mathcal{B}_n} ||\boldsymbol{r}_{\alpha,s_n}||_2^2 + \frac{c_n}{n_v b_n}\sum_{s_n\in\mathcal{B}_n} ||\mathbb{P}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}||_2^2 \\
&\overset{\text{balance}}{=} \frac{1}{n}\left|\left|\boldsymbol{Y} - \mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2 + \frac{c_n}{n_v b_n}\sum_{s_n\in\mathcal{B}_n} ||\mathbb{P}_{\alpha,s_n}\boldsymbol{r}_{\alpha,s_n}||_2^2 \\
&\overset{(3.81)}{=} \frac{1}{n}\left|\left|\boldsymbol{Y} - \mathbb{X}_\alpha\widehat{\boldsymbol{\beta}}_\alpha\right|\right|_2^2 + \left[1 + o\left(\frac{n_c}{n}\right)\right]\frac{n+n_c}{n_c(n-1)}\sum_{i=1}^{n} w_{i,\alpha}^n r_{i,\alpha}^2 \\
&\overset{(3.74)}{=} \tfrac{1}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\boldsymbol{\varepsilon} + \left[1 + o\left(\frac{n_c}{n}\right)\right]\frac{n+n_c}{n_c(n-1)}\sum_{i=1}^{n} w_{i,\alpha}^n r_{i,\alpha}^2 \\
&\overset{(3.82)}{=} \tfrac{1}{n}\boldsymbol{\varepsilon}^T(\mathbb{I}_n - \mathbb{P}_\alpha)\boldsymbol{\varepsilon} + \left[1 + o\left(\frac{n_c}{n}\right)\right]\frac{n+n_c}{n_c(n-1)}\left[d_\alpha\sigma^2 + o_P(1)\right].
\end{aligned} \tag{3.83}$$

From Lemma A.1, it holds that $\mathbb{E}\left[\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon}\right] = \sigma^2 d_\alpha$. Moreover, as $\frac{n_c}{n} \underset{n\to\infty}{\longrightarrow} 0$, it follows that $\frac{n_c}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon} \xrightarrow[n\to\infty]{L_1} 0$. Therefore, one obtains

$$\tfrac{1}{n}\boldsymbol{\varepsilon}^T\mathbb{P}_\alpha\boldsymbol{\varepsilon} = o_P\left(\tfrac{1}{n_c}\right). \tag{3.84}$$

Also, note the following

$$\frac{n+n_c}{n_c(n-1)} = \frac{1+\frac{n_c}{n}}{n_c\left(1-\frac{1}{n}\right)} = \tfrac{1}{n_c}(1+o(1))^2 = \tfrac{1}{n_c} + o\left(\tfrac{1}{n_c}\right). \tag{3.85}$$

By combining (3.83), (3.84) and (3.85), it holds that

$$\begin{aligned}
A_n^{(\alpha)} &= \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + o_P\left(\tfrac{1}{n_c}\right) + \left[1 + o\left(\tfrac{n_c}{n}\right)\right]\cdot\left[\tfrac{1}{n_c} + o\left(\tfrac{1}{n_c}\right)\right]\cdot\left[d_\alpha\sigma^2 + o_P(1)\right] \\
&= \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + o_P\left(\tfrac{1}{n_c}\right) + \left[\tfrac{1}{n_c} + o\left(\tfrac{1}{n_c}\right) + o\left(\tfrac{1}{n}\right)\right]\cdot\left[d_\alpha\sigma^2 + o_P(1)\right] \\
&= \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + o_P\left(\tfrac{1}{n_c}\right) + \left[\tfrac{1}{n_c} + o\left(\tfrac{1}{n_c}\right)\right]\cdot\left[d_\alpha\sigma^2 + o_P(1)\right] \\
&= \tfrac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \tfrac{1}{n_c}d_\alpha\sigma^2 + o_P\left(\tfrac{1}{n_c}\right).
\end{aligned} \tag{3.86}$$

Showing that $B_n^{(\alpha)} = o_P\left(\frac{1}{n_c}\right)$.

Firstly, note from Lemma 17

$$\begin{aligned}
&(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})\mathbb{P}_{\alpha,s_n}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n}) \\
&= (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})\mathbb{P}^2_{\alpha,s_n}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n}) = (\mathbb{P}_{\alpha,s_n} - \mathbb{Q}_{\alpha,s_n}\mathbb{P}_{\alpha,s_n})(\mathbb{P}_{\alpha,s_n} - \mathbb{P}_{\alpha,s_n}\mathbb{Q}_{\alpha,s_n}) \\
&\overset{\text{Lem. } 17}{=} \left(\mathbb{P}_{\alpha,s_n} - \left[\tfrac{n_v}{n} + o\left(\tfrac{n_c}{n}\right)\right]\mathbb{P}_{\alpha,s_n}\right)\left(\mathbb{P}_{\alpha,s_n} - \left[\tfrac{n_v}{n} + o\left(\tfrac{n_c}{n}\right)\right]\mathbb{P}_{\alpha,s_n}\right) \\
&= \left[1 - \tfrac{n_v}{n} + o\left(\tfrac{n_c}{n}\right)\right]^2\mathbb{P}_{\alpha,s_n} = \left[\tfrac{n_c}{n} + o\left(\tfrac{n_c}{n}\right)\right]^2\mathbb{P}_{\alpha,s_n}.
\end{aligned}$$

Hence

$$\begin{aligned}
\left(\tfrac{n}{n_c}\right)^2(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})\mathbb{P}_{\alpha,s_n}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n}) &= \left(\tfrac{n}{n_c}\right)^2\left[\tfrac{n_c}{n} + o\left(\tfrac{n_c}{n}\right)\right]^2\mathbb{P}_{\alpha,s_n} \\
&= [1 + o(1)]^2\mathbb{P}_{\alpha,s_n} = \mathbb{P}_{\alpha,s_n} + o(1)\mathbb{P}_{\alpha,s_n}.
\end{aligned} \tag{3.87}$$

Obviously, for sufficiently large n, the matrix $\frac{1}{2}\mathbb{P}_{\alpha,s_n} + o(1)\mathbb{P}_{\alpha,s_n}$ is positive semidefinite, i.e., $\forall \widetilde{\boldsymbol{u}} \in \mathbb{R}^{n_v}$

$$\begin{aligned}
&\widetilde{\boldsymbol{u}}^T\left(\tfrac{1}{2}\mathbb{P}_{\alpha,s_n} + o(1)\mathbb{P}_{\alpha,s_n}\right)\widetilde{\boldsymbol{u}} \\
&\overset{(3.87)}{=} \widetilde{\boldsymbol{u}}^T\left(\left(\tfrac{n}{n_c}\right)^2(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})\mathbb{P}_{\alpha,s_n}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n}) - \tfrac{1}{2}\mathbb{P}_{\alpha,s_n}\right)\widetilde{\boldsymbol{u}} \geq 0.
\end{aligned}$$

Consider $\widetilde{\boldsymbol{u}} = (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1}\boldsymbol{u}$ for $\boldsymbol{u} \in \mathbb{R}^{n_v}$ then

$$\boldsymbol{u}^T\left(\left(\tfrac{n}{n_c}\right)^2\mathbb{P}_{\alpha,s_n} - \tfrac{1}{2}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1}\mathbb{P}_{\alpha,s_n}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1}\right)\boldsymbol{u} \geq 0.$$

The last inequality implies that the matrix

$$2\left(\tfrac{n}{n_c}\right)^2\mathbb{P}_{\alpha,s_n} - (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1}\mathbb{P}_{\alpha,s_n}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s_n})^{-1} \geq \mathbf{0} \tag{3.88}$$

is positive semidefinite (where the symbol $\geq$ in the matrix context indicates positive semidefiniteness). Also, by Lemma 17,

$$\begin{aligned}
\mathbb{U}_{\alpha,s_n} &\overset{\text{Lem. } 17}{=} \left(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n} + o\left(\tfrac{n_c}{n}\right)\mathbb{P}_{\alpha,s_n}\right)(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n}) \\
&\qquad \times \left(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n} + o\left(\tfrac{n_c}{n}\right)\mathbb{P}_{\alpha,s_n}\right) \\
&= (\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n}) \\
&\quad + o\left(\tfrac{n_c}{n}\right)(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n})\mathbb{P}_{\alpha,s_n} \\
&\quad + o\left(\tfrac{n_c}{n}\right)\mathbb{P}_{\alpha,s_n}(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n}) \\
&\quad + \left[o\left(\tfrac{n_c}{n}\right)\right]^2\mathbb{P}_{\alpha,s_n}(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n})\mathbb{P}_{\alpha,s_n} \\
&= (\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n}) \\
&\quad + o\left(\tfrac{n_c}{n}\right)(1 - \tfrac{n_v}{n})(1 + c_n)\mathbb{P}_{\alpha,s_n} + \left[o\left(\tfrac{n_c}{n}\right)\right]^2(1 + c_n)\mathbb{P}_{\alpha,s_n}.
\end{aligned} \tag{3.89}$$

Since $\frac{n_c}{n} \underset{n\to\infty}{\longrightarrow} 0$ and $c_n \left(\frac{n_c}{n}\right)^2 = \frac{n_v}{n}\left(2 - \frac{n_v}{n}\right) \underset{n\to\infty}{\longrightarrow} 1$, it follows that

$$\begin{aligned} & o\left(\tfrac{n_c}{n}\right)\left(1 - \tfrac{n_v}{n}\right)(1+c_n)\mathbb{P}_{\alpha,s_n} = o\left(\tfrac{n_c}{n}\right)\tfrac{n_c}{n}(1+c_n)\mathbb{P}_{\alpha,s_n} \\ & = \left[o(1)\left(\tfrac{n_c}{n}\right)^2 + o(1)\left(\tfrac{n_c}{n}\right)^2 c_n\right]\mathbb{P}_{\alpha,s_n} = \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)\mathbb{P}_{\alpha,s_n}. \end{aligned} \tag{3.90}$$

By combining (3.89) and (3.90), one obtains

$$\begin{aligned} \mathbb{U}_{\alpha,s_n} &= (\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} + c_n\mathbb{P}_{\alpha,s_n})(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n}) + \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)\mathbb{P}_{\alpha,s_n} \\ &= \left(\mathbb{I}_{n_v} - \tfrac{n_v}{n}\mathbb{P}_{\alpha,s_n}\right)^2 + c_n\left(1 - \tfrac{n_v}{n}\right)^2\mathbb{P}_{\alpha,s_n} + \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)\mathbb{P}_{\alpha,s_n} \\ &= \mathbb{I}_{n_v} - \tfrac{2n_v}{n}\mathbb{P}_{\alpha,s_n} + \left(\tfrac{n_v}{n}\right)^2\mathbb{P}_{\alpha,s_n} + c_n\left(\tfrac{n_c}{n}\right)^2\mathbb{P}_{\alpha,s_n} + \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)\mathbb{P}_{\alpha,s_n} \\ &= \mathbb{I}_{n_v} + \tfrac{1}{n^2}\underbrace{\left(n_v^2 - 2n_v n + c_n n_c^2\right)}_{=0}\mathbb{P}_{\alpha,s_n} + \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)\mathbb{P}_{\alpha,s_n} \\ &= \mathbb{I}_{n_v} + \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)\mathbb{P}_{\alpha,s_n}. \end{aligned} \tag{3.91}$$

Multiplying the matrix from (3.88) by the factor $\left[o\left(\frac{n_c}{n}\right)\right]^2 (1+c_n)$, one easily obtains that the matrix

$$o(1)(1+c_n)\mathbb{P}_{\alpha,s} - \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{P}_{\alpha,s}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1} \geq \mathbf{0} \tag{3.92}$$

is positive semidefinite. Also by (3.91), the following holds

$$\begin{aligned} & (\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}(\mathbb{I}_{n_v} - \mathbb{U}_{\alpha,s})(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1} \\ & \overset{(3.91)}{=} \left[o\left(\tfrac{n_c}{n}\right)\right]^2 (1+c_n)(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\mathbb{P}_{\alpha,s}(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}. \end{aligned} \tag{3.93}$$

By combining (3.92) and (3.93), one can infer that $\forall \boldsymbol{u} \in \mathbb{R}^{n_v}$

$$o(1)(1+c_n)\boldsymbol{u}^T\mathbb{P}_{\alpha,s}\boldsymbol{u} \geq \boldsymbol{u}^T(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}(\mathbb{I}_{n_v} - \mathbb{U}_{\alpha,s})(\mathbb{I}_{n_v} - \mathbb{Q}_{\alpha,s})^{-1}\boldsymbol{u}.$$

Therefore

$$B_n^{(\alpha)} \leq o(1)(1+c_n)\left(\frac{1}{n_v b}\sum_{s\in\mathcal{B}} ||\mathbb{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s}||_2^2\right) = o_P\left(\tfrac{1}{n_c}\right),$$

where the last equation follows from the fact that in (3.83) and (3.86) it was shown that

$$\frac{c_n}{n_v b}\sum_{s\in\mathcal{B}} ||\mathbb{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s}||_2^2 = \tfrac{1}{n_c}d_\alpha\sigma^2 + o_P\left(\tfrac{1}{n_c}\right).$$

□

3.3.4 Monte Carlo cross-validation

Using the BICV(n_v) requires having a balanced incomplete block design $\mathcal{B}_n$ for all $n \in \mathbb{N}$, which can be hard to obtain in real applications. Therefore, a Monte Carlo alternative of BICV(n_v) may be used. For all $n \in \mathbb{N}$ randomly draw without replacement a collection $\mathcal{T}_n$ of b_n subsets of $\{1, \ldots, n\}$ that have size n_v. Then the cross-validation estimate of $\Gamma_{\alpha,n}$ will be referred to as the Monte Carlo CV(n_v), abbreviated as MCCV(n_v), and is defined as

$$\widehat{\Gamma}^{\text{MCCV}}_{\alpha,n_v} = \frac{1}{n_v b_n} \sum_{s_n \in \mathcal{T}_n} \left|\left| \boldsymbol{Y}_{s_n} - \widehat{\boldsymbol{Y}}_{\alpha,s_n^c} \right|\right|_2^2 .$$

The model selected by Monte Carlo cross-validation is $M_{\widehat{\alpha}_M(n_v)}$, where

$$\widehat{\alpha}_M(n_v) = \arg\min_{\alpha \in \mathcal{A}} \widehat{\Gamma}^{\text{MCCV}}_{\alpha,n_v}. \tag{3.94}$$

The following result is similar to Theorems 13, 14 and 15. The probability statements in Theorem 18 are with respect to the joint probability corresponding to $\boldsymbol{Y}$ and the Monte Carlo selection of the subsets.

Theorem 18. *Suppose that conditions (3.13), (3.29), (3.35) and (3.59) are satisfied. Also, assume that*

$$\max_{s_n \in \mathcal{T}_n} \left|\left| \frac{1}{n_v} \sum_{i \in s_n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \frac{1}{n_c} \sum_{i \in s_n^c} \boldsymbol{x}_i \boldsymbol{x}_i^T \right|\right| = o_P(1), \tag{3.95}$$

where $\mathcal{T}_n$ contains b_n subsets selected randomly with b_n satisfying

$$\frac{n^2}{b_n n_c^2} \underset{n\to\infty}{\longrightarrow} 0. \tag{3.96}$$

This yields the following conclusions:

- *If M_α is a model in Category* I*, then there exists nonnegative random variables R_n such that*

$$\widehat{\Gamma}^{\text{MCCV}}_{\alpha,n_v} = \frac{1}{n_v b_n} \sum_{s_n \in \mathcal{T}_n} \boldsymbol{\varepsilon}_{s_n}^T \boldsymbol{\varepsilon}_{s_n} + \Delta_{\alpha,n} + o_P(1) + R_n,$$

  where $\boldsymbol{\varepsilon}_{s_n} = \boldsymbol{Y}_{s_n} - \mathbb{X}_{s_n}\boldsymbol{\beta}$ and $\mathbb{X}_{s_n}$ is the $n_v \times p$ matrix containing the rows of $\mathbb{X}$ indexed by $i \in s_n$.

- *If M_α is a model in Category* II*, it holds that*

$$\widehat{\Gamma}^{\text{MCCV}}_{\alpha,n_v} = \frac{1}{n_v b_n} \sum_{s_n \in \mathcal{T}_n} \boldsymbol{\varepsilon}_{s_n}^T \boldsymbol{\varepsilon}_{s_n} + \tfrac{1}{n_c} d_\alpha \sigma^2 + o_P\left(\tfrac{1}{n_c}\right).$$

- *Consequently,*

$$\lim_{n\to\infty} \mathsf{P}\left[M_{\widehat{\alpha}_M(n_v)} \neq M_* \right] = 0,$$

  which means that MCCV(n_v) *is asymptotically correct.*

The proof is given in the appendix of paper Shao [1993], p. 494.

Note that condition (3.95) is the probabilistic analog of assumption (3.58). Condition (3.96) imposes certain restrictions on b_n and n_c, indicating that when fewer data are used for model construction, more splits are needed. In addition to the Monte Carlo method CV(n_v), there is another alternative to BICV(n_v) called an analytic approximate cross-validation, details of which can be found in the paper Shao [1993], Chapter 3.3.2.

3.4 Real Data

In this chapter, the cross-validation techniques will be compared using real data from parliamentary and presidential elections in the Czech Republic held in the years 2021 and 2023. In 2021, elections were held for the Chamber of Deputies of the Czech Parliament, with the participation of 22 political parties. In 2023, presidential elections took place, with the participation of 9 candidates. The presidential election system is such that initially, the first round is held, where two candidates with the highest number of votes are selected. These candidates then advance to the second round, where the election is held again, and the president is chosen. On the other hand, the parliamentary election system is conducted in just one round. Voting in parliamentary and presidential elections took place in each district of every municipality in the country. The goal of this chapter will be to make predictions of the percentage of votes obtained by the two candidates in the first round (the candidates from the presidential elections who advanced to the second round) based on information from the parliamentary elections. Predictions will be made using leave-one-out cross-validation and Monte Carlo cross-validation, and the results will then be compared.

3.4.1 Data Structure

All data used in this chapter are available on the following website Czech Statistical Office.

The data from the parliamentary elections include the following information: $\mathrm{id_M}$ (identification number of the municipality), $\mathrm{id_D}$ (identification number of the district), P^1, ... , P^{22} (number of votes for individual political parties), IE (total number of issued envelopes), RV (total number of registered voters).

Similarly, the data from the first round of the presidential elections include the following information: $\mathrm{id_M}$ (identification number of the municipality), $\mathrm{id_D}$ (identification number of the district), V (total number of valid votes in the first round) and RV^1 (total number of registered voters in the first round). Additionally, it includes the number of votes received by the candidates who advanced to the second round: Y (Petr Pavel) and Z (Andrej Babiš).

Both datasets are arranged according to the processing time. Note that each district is characterized by a pair of numbers, $\mathrm{id_M}$ and $\mathrm{id_D}$, and will be treated as an individual observation.

3.4.2 Used Linear Models

It is important to note that since the data are arranged according to processing time, the data structure dictates that smaller districts will appear first in the

datasets, followed by larger ones. Therefore, one would expect that homoscedasticity assumption would not be met in a dataset of this nature. To approach the assumption of homoscedasticity, all variables were transformed using square root transformations. Recall that in a linear model with p predictors, the cross-validation criterion for model selection involves all $2^p - 1$ possible submodels being explored, and the one with the smallest estimate of $\Gamma_{\alpha,n}$ is selected. Therefore, to ensure computational manageability, only the seven largest political parties (in terms of the number of votes they received) were considered as predictors when constructing models to predict the number of votes for candidates Y and Z. These political parties are P^4 (Svoboda a přímá demokracie), P^5 (Česká strana sociálně demokratická), P^{12} (PŘÍSAHA Roberta Šlachty), P^{13} (SPOLU), P^{17} (PIRÁTI a STAROSTOVÉ), P^{18} (Komunistická strana Čech a Moravy), and P^{20} (ANO 2011). Overall, the following linear models were considered:

- M_1 is a linear model whose response is $\sqrt{\mathrm{Y}}$ and its predictors are $\sqrt{\mathrm{RV}^1}$, $\sqrt{\mathrm{RV}}$, $\sqrt{\mathrm{IE}}$ and $\sqrt{\mathrm{P}^k}$ for $k = 4, 5, 12, 13, 17, 18, 20$.
- M_2 is a linear model whose response is $\sqrt{\mathrm{Z}}$ and its predictors are $\sqrt{\mathrm{RV}^1}$, $\sqrt{\mathrm{RV}}$, $\sqrt{\mathrm{IE}}$ and $\sqrt{\mathrm{P}^k}$ for $k = 4, 5, 12, 13, 17, 18, 20$.

Since one wants to predict the percentage of votes obtained by the candidates Y and Z, it is necessary to have a model to predict the total number of valid votes in the first round V. Therefore, the following model is also considered:

- M_3 is a linear model whose response is $\sqrt{\mathrm{V}}$ and its predictors are $\sqrt{\mathrm{RV}^1}$, $\sqrt{\mathrm{RV}}$ and $\sqrt{\mathrm{IE}}$.

Using cross-validation techniques, the optimal models will be chosen from M_1 and M_2, and then predictions will be made using those optimal models. For predictions of the number of valid votes in the first round V, the entire model M_3 will be employed.

3.4.3 Notation

In total, there are $N = 14\,844$ observations available. To construct and validate the model, the initial $n = q \cdot N$ observations will be used, where $q \in [0, 1]$. Recall that observations are arranged according to the processing time, so it's crucial to use the earliest observations in the dataset for model construction and validation. The motivation behind this is the desire to make predictions about the overall impact of elections once, for example, the first 15% of results are known. The remaining $N - n$ data points will be reserved as test data. For model construction, $n_c = n^{\frac{3}{4}}$ data points will be employed, while $n_v = n - n_c$ data points will be used for model validation. As mentioned earlier, CV(1) and Monte Carlo CV(n_v) methods will be used to select the optimal submodels of M_1 and M_2. It's worth noting that the estimate $\widehat{\Gamma}^{\mathrm{CV}}_{\alpha,1}$ can be directly computed from the expression (3.28), while for computing $\widehat{\Gamma}^{\mathrm{MCCV}}_{\alpha,n_v}$, $b_n = 2n$ selections for Monte Carlo sampling were used. Therefore, it can be observed that with the chosen b_n and n_c, condition (3.96) is satisfied.

Summarize the notation that will be used here:

- $\{1, \ldots, N\}$ is the set of all available data.
- $\{1, \ldots, n\}$ is the set of data used for model construction and validation, where $n = q \cdot N$ and $q \in [0, 1]$. For model validation, a subset $s \subset \{1, \ldots, n\}$ is used such that $|s| = n_v$, while for model construction, the complement subset $s^c \subset \{1, \ldots, n\} \setminus s$ is used such that $|s^c| = n_c$.
- Y_i represents the number of votes received by Petr Pavel in the i-th district, while Z_i represents the number of votes received by Andrej Babiš in the i-th district, for $i = 1, \ldots, N$. At the same time, V_i represents the total number of valid votes in the first round in the i-th district.
- $\widehat{\mathrm{Y}}_{i,1}$ denotes the prediction of the number of votes received by Petr Pavel in the i-th district, for $i = n+1, \ldots, N$, where this prediction was made based on leave-one-out cross-validation applied to model M_1. Similarly, $\widehat{\mathrm{Y}}_{i,n_v}$ can be understood as the prediction based on Monte Carlo CV(n_v) applied to model M_1, while $\widehat{\mathrm{Y}}_i$ denotes the prediction based on the full model M_1.
- $\widehat{\mathrm{Z}}_{i,1}$ denotes the prediction of the number of votes received by Andrej Babiš in the i-th district, for $i = n+1, \ldots, N$, where this prediction was made based on leave-one-out cross-validation applied to model M_2. Similarly, $\widehat{\mathrm{Z}}_{i,n_v}$ can be understood as the prediction based on Monte Carlo CV(n_v) applied to model M_2, while $\widehat{\mathrm{Z}}_i$ denotes the prediction based on the full model M_2.
- $\widehat{\mathrm{V}}_i$ denotes the prediction of the total number of valid votes in the first round in the i-th district, for $i = n+1, \ldots, N$, which was made by model M_3.

3.4.4 Assessment of the Prediction

To predict the percentage of votes obtained by the two candidates in the first round, the following quantities will be needed:

- $R_{\mathrm{Y}}(n) = \sum_{i=1}^{n} \mathrm{Y}_i / \sum_{i=1}^{n} \mathrm{V}_i$ and $R_{\mathrm{Z}}(n) = \sum_{i=1}^{n} \mathrm{Z}_i / \sum_{i=1}^{n} \mathrm{V}_i$ show the percentage ratio of how many votes each of the candidates actually received based on the data $\{1, \ldots, n\}$.
- $\widehat{R}_{\mathrm{Y}} = \left(\sum_{i=1}^{n} \mathrm{Y}_i + \sum_{j=n+1}^{N} \widehat{\mathrm{Y}}_j\right) / \left(\sum_{i=1}^{n} \mathrm{V}_i + \sum_{j=n+1}^{N} \widehat{\mathrm{V}}_j\right)$ represents the estimation of $R_{\mathrm{Y}}(N)$ based on the full model M_1, and similarly, $\widehat{R}_{\mathrm{Z}}$ can be defined.
- $\widehat{R}_{\mathrm{Y}}(1) = \left(\sum_{i=1}^{n} \mathrm{Y}_i + \sum_{j=n+1}^{N} \widehat{\mathrm{Y}}_{j,1}\right) / \left(\sum_{i=1}^{n} \mathrm{V}_i + \sum_{j=n+1}^{N} \widehat{\mathrm{V}}_j\right)$ represents the CV(1) estimation of $R_{\mathrm{Y}}(N)$, and similarly, $\widehat{R}_{\mathrm{Z}}(1)$ can be defined.
- $\widehat{R}_{\mathrm{Y}}(n_v) = \left(\sum_{i=1}^{n} \mathrm{Y}_i + \sum_{j=n+1}^{N} \widehat{\mathrm{Y}}_{j,n_v}\right) / \left(\sum_{i=1}^{n} \mathrm{V}_i + \sum_{j=n+1}^{N} \widehat{\mathrm{V}}_j\right)$ represents the Monte Carlo CV(n_v) estimation of $R_{\mathrm{Y}}(N)$, and similarly, $\widehat{R}_{\mathrm{Z}}(n_v)$ can be defined.

Another measure that describes the quality of the prediction is the average squared prediction error:

- $\mathrm{PE_Y} = \frac{1}{N-n}\sum_{j=n+1}^{N}\left(\mathrm{Y}_j - \widehat{\mathrm{Y}}_j\right)^2$, and similarly, $\mathrm{PE_Z}$ can be defined.
- $\mathrm{PE_Y}(1) = \frac{1}{N-n}\sum_{j=n+1}^{N}\left(\mathrm{Y}_j - \widehat{\mathrm{Y}}_{j,1}\right)^2$, and similarly, $\mathrm{PE_Z}(1)$ can be defined.
- $\mathrm{PE_Y}(n_v) = \frac{1}{N-n}\sum_{j=n+1}^{N}\left(\mathrm{Y}_j - \widehat{\mathrm{Y}}_{j,n_v}\right)^2$, and similarly, $\mathrm{PE_Z}(n_v)$ can be defined.

Additionally, the normalized average squared prediction errors will be needed:

- $\mathrm{RE_Y}(1) = \mathrm{PE_Y}(1)/\min\left\{\mathrm{PE_Y}(1), \mathrm{PE_Y}(n_v), \mathrm{PE_Y}\right\}$, and similarly, $\mathrm{RE_Z}(1)$ can be defined.
- $\mathrm{RE_Y}(n_v) = \mathrm{PE_Y}(n_v)/\min\left\{\mathrm{PE_Y}(1), \mathrm{PE_Y}(n_v), \mathrm{PE_Y}\right\}$, and similarly, $\mathrm{RE_Z}(n_v)$ can be defined.
- $\mathrm{RE_Y} = \mathrm{PE_Y}/\min\left\{\mathrm{PE_Y}(1), \mathrm{PE_Y}(n_v), \mathrm{PE_Y}\right\}$, and similarly, $\mathrm{RE_Z}$ can be defined.

All the above-mentioned quantities will be used to assess the quality of prediction in the following subsection.

3.4.5 Discussion of results

From Table 3.1, it can be observed that from the beginning, the difference between the estimates $\widehat{R}_\mathrm{Y}$, $\widehat{R}_\mathrm{Y}(1)$, and $\widehat{R}_\mathrm{Y}(n_v)$ appears at the third decimal place. It is also evident that if more than 17% of the data were used for model construction and validation, differences between the estimates begin to emerge at the fourth decimal place, which is almost negligible. Overall, it can be seen from the results of Table 3.1 that it almost does not matter whether the overall model M_1 is used for prediction or some submodel selected via CV(1) or Monte Carlo CV(n_v) is used. A nearly analogous situation arises in Table 3.2, where unlike Table 3.1, differences at the third decimal place appear even for larger percentages of data used for model construction and validation.

Note that in Table 3.1, $\widehat{R}_\mathrm{Y}(1)$ and $\widehat{R}_\mathrm{Y}(n_v)$ are almost the same as the estimate $\widehat{R}_\mathrm{Y}$, which indicates that both methods CV(1) and Monte Carlo CV(n_v) try to select models close to the full model M_1 (analogously in Table 3.2). This may indicate that the majority of the considered predictors in models M_1 and M_2 are important.

The overall percentage ratios of votes received by candidates Y and Z are $R_\mathrm{Y}(N) = 0.3538$ and $R_\mathrm{Z}(N) = 0.3502$ respectively. Note that these numbers slightly differ from those stated in Czech Statistical Office because in the presidential elections of 2023, there were 13 newly added districts that were not present in the parliamentary elections of 2021. Since the number of such districts is relatively small, such districts were excluded from the dataset. An interesting observation is that from Tables 3.1 and 3.2, it is already evident, using only 5% of the data for model construction and validation, that all three methods estimate the overall vote percentage ratios approximately as 0.35. Meanwhile, the current percentage ratios of votes $R_\mathrm{Y}(n) = 0.3038$ underestimate, and $R_\mathrm{Z}(n) = 0.3981$ overestimate the percentage ratios of votes each candidate will receive for $n = 0.05 \cdot N$.

On the other hand, one does not see in Tables 3.1 and 3.2 any clear advantage of the Monte Carlo CV(n_v) method over CV(1). One of the reasons could be that the theory mentioned above assesses prediction quality by considering the average squared prediction error, while here, the ratios R_{Y} and R_{Z} were considered. Those quantities were used here because it was hoped that minimizing the average squared prediction error would also reflect in the quality of estimating R_{Y} and R_{Z}.

However, since the differences in estimates of R_{Y} and R_{Z} are small, let's look at Tables 3.3 and 3.4, where average squared prediction errors were used to assess the quality of prediction. Fitting the models M_1 and M_2 on data $\{1, \ldots, 0.05 \cdot N\}$ and then conducting the formal statistical test confirmed heteroscedasticity (Koenker's studentized version of the Breusch-Pagan Test was used). Since the assumption of homoscedasticity is violated, one wouldn't expect the results to align with the theory. However, as shown, for example, in Table 3.4, in almost all cases, Monte Carlo CV(n_v) performs better than CV(1). The only exceptions are when 20%, 22%, and 40% of the total data were used for model construction and validation. Nonetheless, from this table, it's apparent that quite often, the normalized values of prediction error computed from CV(1) differ from the normalized values of the optimal model by the first or second decimal place. This could be interpreted as CV(1) differing from the optimal model within units or tens of percent. Whereas, in all cases, the normalized values of prediction error from Monte Carlo CV(n_v) differ from the normalized values of the optimal model by the third decimal place. This could be interpreted as Monte Carlo CV(n_v) differing from the optimal model by less than 1 percent. Therefore, despite the violation of homoscedasticity, Table 3.4 indicates that Monte Carlo CV(n_v) outperforms CV(1) in terms of prediction error. Additionally, it can be seen from this table that the full model M_1 performs optimally in terms of prediction error with larger percentages of used data for model construction and validation.

Nearly an analogous situation arises in Table 3.4. When less than 15% of all data is used for model construction and validation, CV(1) differs more from the optimal model than Monte Carlo CV(n_v). In other cases, CV(1) selects the optimal model, but Monte Carlo CV(n_v) differs from it at the third decimal place, which may be deemed negligible.

Another reason why CV(1) may not be easily surpassed by Monte Carlo CV(n_v) in practice could be that for large datasets, all reasonably selected predictors are statistically significant. This implies that both methods tend to select the full model.

Percentage of used data	$R_{\mathrm{Y}}(n)$	$\widehat{R}_{\mathrm{Y}}$	$\widehat{R}_{\mathrm{Y}}(1)$	$\widehat{R}_{\mathrm{Y}}(n_v)$
5 %	0.3038	0.3505	0.3510	0.3510
7 %	0.3087	0.3532	0.3540	0.3514
10 %	0.3117	0.3546	0.3548	0.3528
12 %	0.3118	0.3534	0.3534	0.3514
15 %	0.3124	0.3514	0.3514	0.3496
17 %	0.3124	0.3523	0.3523	0.3529
20 %	0.3135	0.3520	0.3520	0.3520
22 %	0.3144	0.3519	0.3519	0.3519
25 %	0.3146	0.3518	0.3519	0.3519
27 %	0.3152	0.3512	0.3513	0.3513
30 %	0.3156	0.3506	0.3506	0.3506
40 %	0.3189	0.3504	0.3505	0.3505

Table 3.1: The table illustrates the percentage ratio of the votes Petr Pavel actually received, along with various estimations of this percentage ratio obtained by the full model M_1, leave-one-out, and Monte Carlo cross-validations. The overall percentage ratio of votes received by Petr Pavel is $R_{\mathrm{Y}}(N) = 0.3538$.

Percentage of used data	$R_{\mathrm{Z}}(n)$	$\widehat{R}_{\mathrm{Z}}$	$\widehat{R}_{\mathrm{Z}}(1)$	$\widehat{R}_{\mathrm{Z}}(n_v)$
5 %	0.3981	0.3508	0.3511	0.3508
7 %	0.3937	0.3486	0.3487	0.3469
10 %	0.3913	0.3477	0.3465	0.3465
12 %	0.3909	0.3473	0.3472	0.3452
15 %	0.3915	0.3479	0.3479	0.3460
17 %	0.3932	0.3485	0.3484	0.3476
20 %	0.3913	0.3490	0.3491	0.3491
22 %	0.3907	0.3490	0.3491	0.3491
25 %	0.3914	0.3491	0.3492	0.3486
27 %	0.3908	0.3497	0.3498	0.3493
30 %	0.3906	0.3502	0.3503	0.3498
40 %	0.3870	0.3501	0.3501	0.3501

Table 3.2: The table illustrates the percentage ratio of the votes Andrej Babiš actually received, along with various estimations of this percentage ratio obtained by the full model M_2, leave-one-out, and Monte Carlo cross-validations. The overall percentage ratio of votes received by Andrej Babiš is $R_{\mathrm{Z}}(N) = 0.3502$.

Percentage of used data	RE_Y	$RE_Y(1)$	$RE_Y(n_v)$
5 %	1.0000	1.0000	1.0060
7 %	1.0000	1.0425	1.0161
10 %	1.0000	1.0417	1.0044
12 %	1.0000	1.0474	1.0001
15 %	1.0000	1.0571	1.0021
17 %	1.0117	1.0000	1.0113
20 %	1.0000	1.0000	1.0005
22 %	1.0000	1.0000	1.0004
25 %	1.0000	1.0000	1.0020
27 %	1.0000	1.0000	1.0019
30 %	1.0000	1.0000	1.0005
40 %	1.0000	1.0000	1.0014

Table 3.3: The table shows the normalized average squared prediction errors computed on the test data for Petr Pavel. These errors are calculated using three methods: employing the full model M_1 , using CV(1), and Monte Carlo CV(n_v). It's important to note that normalization is performed based on the smallest prediction error. Thus, a value of 1.0000 in the table indicates that the method has the smallest prediction error. The higher the value in the table, the worse the method is in terms of prediction error relative to the other methods.

Percentage of used data	RE_Z	$RE_Z(1)$	$RE_Z(n_v)$
5 %	1.0014	1.1275	1.0000
7 %	1.0000	1.1335	1.0004
10 %	1.1047	1.1047	1.0000
12 %	1.0010	1.1579	1.0000
15 %	1.0004	1.1485	1.0000
17 %	1.0004	1.0309	1.0000
20 %	1.0000	1.0000	1.0009
22 %	1.0000	1.0000	1.0014
25 %	1.0000	1.0291	1.0020
27 %	1.0000	1.0278	1.0016
30 %	1.0000	1.0262	1.0019
40 %	1.0000	1.0000	1.0013

Table 3.4: The table shows the normalized average squared prediction errors computed on the test data for Andrej Babiš. These errors are calculated using three methods: employing the full model M_2 , using CV(1), and Monte Carlo CV(n_v). As in the caption of Table 3.3, the higher the value in the table, the worse the method performs in terms of prediction error relative to the other methods.

Conclusion

In this work, we focused on the use of cross-validation methods in various contexts. The first two chapters showed the application of leave-one-out cross-validation for kernel density estimation and kernel regression. The structure of these chapters was based on lecture notes Nagy and Omelka [2024], with a greater emphasis on methods for bandwidth selection relying on cross-validation. In Chapter 1, theoretical results ensuring the optimality of cross-validation were introduced. For instance, one of the theoretical contributions is a rigorous proof of Lemma 3. In comparison to the orginal paper Scott and Terrell [1987], it seems that additional technical assumptions (1.11) and (1.12) are needed. A possible extension of the results from Chapter 1 would be to derive the asymptotical variances of bandwidths selected by cross-validation based on the paper Hall and Marron [1987]. Theoretical results of the first two chapters were illustrated using simulated data.

Chapter 3, based on the paper Shao [1993], can be considered the main part of the thesis. In this chapter, in comparison to the original paper, proofs of certain claims were elaborated in greater detail. Many claims used in the argumentation of the paper are not obvious, and some of them required nontrivial derivation. The methods introduced in this chapter were then applied to real data from parliamentary and presidential elections in the Czech Republic in 2021 and 2023. It became clear that homoscedasticity is a crucial assumption for highlighting the advantages of the CV(n_v) method over CV(1). However, it was demonstrated that even in our data, where this assumption is not met, if CV(1) outperforms CV(n_v), the difference in prediction errors is not large. Conversely, in cases where CV(n_v) outperforms CV(1), the difference in prediction errors is more pronounced.

Another crucial assumption throughout Chapter 3 was that the number of predictors remained constant as $n \to \infty$. It could potentially be studied how the aforementioned methods behave when the number of predictors also increases. It would be also interesting to study various extensions of the theory from Chapter 3 to generalized linear models or heteroscedastic linear regression.

A. Appendix

Theorem A.1. *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)^T$*, where* $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$*. Let* $Q = \sigma^{-2}\boldsymbol{X}^T\mathbb{A}\boldsymbol{X}$ *for a symmetric matrix* $\mathbb{A}$ *with rank* r*. Then* Q *has a* χ^2_r *distribution if and only if* $\mathbb{A}$ *is idempotent.*

The proof is given in Hogg et al. [2018], Theorem 9.8.4.

Lemma A.1. *Let* $\boldsymbol{Z}$ *be a random vector of length* n *with the expectation* $\boldsymbol{\mu}$ *and the finite covariance matrix* $\boldsymbol{\Sigma}$*. Let* $\mathbb{B}$ *be any matrix of size* $n \times n$*. Then*

$$\mathbb{E}\left[\boldsymbol{Z}^T\mathbb{B}\boldsymbol{Z}\right] = \boldsymbol{\mu}^T\mathbb{B}\boldsymbol{\mu} + \operatorname{tr}(\mathbb{B}\boldsymbol{\Sigma}).$$

The proof can be found in the course notes Omelka, Lemma 2.5.

Lemma A.2. *Assume that* $\{Z_n\}_{n=1}^{\infty}$ *is a sequence of random variables such that* $\operatorname{var}(Z_n) \underset{n\to\infty}{\longrightarrow} 0$*. Then*

$$Z_n - \mathbb{E}Z_n \underset{n\to\infty}{\xrightarrow{P}} 0.$$

Proof. The proof follows simply from Chebyshev's inequality as

$$\mathrm{P}\left[|Z_n - \mathbb{E}Z_n| > \varepsilon\right] \overset{\text{Cheb.}}{\leq} \frac{\operatorname{var}(Z_n)}{\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0,$$

for each $\varepsilon > 0$. □

Definition A.1. *The spectral radius* $\rho(\mathbb{A})$ *of a square matrix* $\mathbb{A}$ *is defined by*

$$\rho(\mathbb{A}) = \max|\lambda|, \quad \text{where } \lambda \text{ is an eigenvalue of } \mathbb{A}.$$

If $\lambda = \alpha + \beta i$ *is complex, then* $|\lambda| = (\alpha^2 + \beta^2)^{\frac{1}{2}}$*.*

Definition A.2. *A matrix norm on the set of all* $n \times m$ *matrices is a real-valued function,* $||\cdot||$*, defined on this set, satisfying for all* $n \times m$ *matrices* $\mathbb{A}$ *and* $\mathbb{B}$ *and all real numbers* α*:*

1. $||\mathbb{A}|| \geq 0$*;*
2. $||\mathbb{A}|| = 0$*, if and only if* $\mathbb{A}$ *is the matrix with all* 0 *entries;*
3. $||\alpha\mathbb{A}|| = |\alpha|\ ||\mathbb{A}||$*;*
4. $||\mathbb{A} + \mathbb{B}|| \leq ||\mathbb{A}|| + ||\mathbb{B}||$*;*
5. $||\mathbb{A}\mathbb{B}|| \leq ||\mathbb{A}||\ ||\mathbb{B}||$*.*

Definition A.3. *Let* $||\cdot||_v$ *is a vector norm. Then,*

$$||\mathbb{A}|| = \sup_{||\boldsymbol{x}||_v = 1} ||\mathbb{A}\boldsymbol{x}||_v$$

is called the natural matrix norm associated with the vector norm $||\cdot||_v$*.*

Remark. It can be easily shown that the natural matrix norm satisfies Definition A.2. The most commonly used natural matrix norms are

$$||\mathbb{A}||_2 = \sup_{||\boldsymbol{x}||_2=1} ||\mathbb{A}\boldsymbol{x}||_2, \qquad \text{the } l_2 \text{ norm}, \tag{A.1}$$

$$||\mathbb{A}||_\infty = \sup_{||\boldsymbol{x}||_\infty=1} ||\mathbb{A}\boldsymbol{x}||_\infty, \qquad \text{the } l_\infty \text{ norm}, \tag{A.2}$$

where

$$||\boldsymbol{x}||_2 = \left\{\sum_{i=1}^{m} x_i^2\right\}^{\frac{1}{2}} \quad \text{and} \quad ||\boldsymbol{x}||_\infty = \max_{1\le i\le m} |x_i|.$$

Lemma A.3. *Let $\mathbb{A}_1$ and $\mathbb{A}_2$ be square matrices of order n that are symmetric, positive definite and $\mathbb{A}_1 - \mathbb{A}_2$ is also positive definite, then $\rho(\mathbb{A}_1^{-1}\mathbb{A}_2) < 1$.*

Proof. First, note that because $\mathbb{A}_1$ is symmetric and positive definite, there exists a symmetric and positive definite matrix $\mathbb{A}_1^{1/2}$ such that $\mathbb{A}_1 = \mathbb{A}_1^{1/2}\mathbb{A}_1^{1/2}$. From the assumptions of the lemma, it holds that $\mathbb{A}_1 - \mathbb{A}_2$ is a positive definite matrix. Therefore, $\mathbb{A}_1^{-1/2}(\mathbb{A}_1 - \mathbb{A}_2)\mathbb{A}_1^{-1/2}$ is also positive definite. Denote $\mathbb{B} = \mathbb{A}_1^{-1/2}\mathbb{A}_2\mathbb{A}_1^{-1/2}$, which is also positive definite. Then, it holds that

$$\mathbb{A}_1^{-1/2}(\mathbb{A}_1 - \mathbb{A}_2)\mathbb{A}_1^{-1/2} = \mathbb{I}_n - \mathbb{A}_1^{-1/2}\mathbb{A}_2\mathbb{A}_1^{-1/2} = \mathbb{I}_n - \mathbb{B} > \mathbf{0},$$

where the symbol $>$, in the context of matrices, indicates positive definiteness. Therefore, for all nonzero vectors $\boldsymbol{x} \in \mathbb{R}^n$ one would get that

$$\boldsymbol{x}^T(\mathbb{I}_n - \mathbb{B})\boldsymbol{x} > 0.$$

Equivalently, for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$, it holds that

$$0 < \boldsymbol{x}^T\mathbb{B}\boldsymbol{x} < \boldsymbol{x}^T\boldsymbol{x}, \tag{A.3}$$

where the first inequality holds due to the positive definiteness of $\mathbb{B}$. Note that because $\mathbb{B}$ is positive definite and symmetric, there exists a positive definite and symmetric matrix $\mathbb{B}^{1/2}$ such that $\mathbb{B} = \mathbb{B}^{1/2}\mathbb{B}^{1/2}$. Therefore, it could also be shown that for all $\boldsymbol{x} \in \mathbb{R}^n$, the following holds

$$0 \le (\mathbb{B}\boldsymbol{x})^T(\mathbb{B}\boldsymbol{x}) \overset{\text{sym.}}{=} \boldsymbol{x}^T\mathbb{B}^2\boldsymbol{x} = \left(\mathbb{B}^{1/2}\boldsymbol{x}\right)^T \mathbb{B} \left(\mathbb{B}^{1/2}\boldsymbol{x}\right) \overset{\text{(A.3)}}{<} \boldsymbol{x}^T\mathbb{B}\boldsymbol{x} < \boldsymbol{x}^T\boldsymbol{x}. \tag{A.4}$$

Overall, one obtains that

$$\sup_{||\boldsymbol{x}||_2=1} \boldsymbol{x}^T\mathbb{B}^2\boldsymbol{x} < 1.$$

Finally, bound the l_2 matrix norm as follows

$$||\mathbb{B}||_2 = \sup_{||\boldsymbol{x}||_2=1} ||\mathbb{B}\boldsymbol{x}||_2 = \sup_{||\boldsymbol{x}||_2=1} \sqrt{\boldsymbol{x}^T\mathbb{B}^T\mathbb{B}\boldsymbol{x}} = \sup_{||\boldsymbol{x}||_2=1} \sqrt{\boldsymbol{x}^T\mathbb{B}^2\boldsymbol{x}} < 1.$$

After using Theorem 7.15 from Burden and Faires [2010], one can conclude that

$$\rho(\mathbb{A}_1^{-1/2}\mathbb{A}_2\mathbb{A}_1^{-1/2}) = \rho(\mathbb{B}) \le ||\mathbb{B}||_2 < 1.$$

The last thing to note is that $\mathbb{A}_1^{-1/2}\mathbb{A}_2\mathbb{A}_1^{-1/2}$ and $\mathbb{A}_1^{-1/2}\mathbb{A}_1^{-1/2}\mathbb{A}_2 = \mathbb{A}_1^{-1}\mathbb{A}_2$ have the same eigenvalues, which means that

$$\rho(\mathbb{A}_1^{-1}\mathbb{A}_2) = \rho(\mathbb{A}_1^{-1/2}\mathbb{A}_2\mathbb{A}_1^{-1/2}) < 1.$$

□

Lemma A.4. *For any vector $\boldsymbol{x} \neq 0$, matrix $\mathbb{A}$, and any natural norm $||\cdot||$, it holds that*

$$||\mathbb{A}\boldsymbol{x}|| \leq ||\mathbb{A}|| \cdot ||\boldsymbol{x}||.$$

The proof can be found in Burden and Faires [2010], Corollary 7.10.

Lemma A.5. *If $\mathbb{A} = (a_{i,j})$ is an $n \times m$ matrix, then*

$$||\mathbb{A}||_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{m} |a_{ij}|.$$

The proof is given in Burden and Faires [2010], Theorem 7.11.

Theorem A.2. *For any two matrix norms $||\cdot||_\gamma$ and $||\cdot||_\beta$, there exist positive constants r and s such that for all matrices $\mathbb{A}$ the following holds*

$$r||\mathbb{A}||_\gamma \leq ||\mathbb{A}||_\beta \leq s||\mathbb{A}||_\gamma.$$

The proof is provided in Horn and Johnson [2012], Corollary 5.4.5.

Lemma A.6. *Let $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ and $\mathbb{X}_* = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d_*})$ be $n \times d$ and $n \times d_*$ matrices, respectively, such that* $\text{rank}(\mathbb{X}) = d$, $\text{rank}(\mathbb{X}_*) = d_*$, *and $d_* < d$. Then* $\text{rank}(\mathbb{P} - \mathbb{P}_*) = d - d_*$, *where $\mathbb{P}$ and $\mathbb{P}_*$ are projection matrices onto the column spaces* $\text{Im}(\mathbb{X})$ *and* $\text{Im}(\mathbb{X}_*)$, *respectively.*

Proof. It is known from the $\mathbb{QR}$-factorization (see Horn and Johnson [2012], Theorem 2.1.14) that there exist an $n \times d$ matrix $\mathbb{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d)$ with orthonormal columns and an upper triangular $d \times d$ matrix $\mathbb{R}$ with positive elements on its diagonal, such that $\mathbb{X} = \mathbb{QR}$. Since $\mathbb{X}_*$ is a submatrix of $\mathbb{X}$ that contains the first d_* columns of $\mathbb{X}$, there should also exist an upper triangular $d_* \times d_*$ matrix $\mathbb{R}_*$ with positive elements on its diagonal, such that $\mathbb{X}_* = \mathbb{Q}_*\mathbb{R}_*$, where $\mathbb{Q}_* = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{d_*})$. Thus, one can express the projection matrices in the following way

$$\mathbb{P} = \mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T = \mathbb{Q}\left[\mathbb{R}\left(\mathbb{R}^T\mathbb{R}\right)^{-1}\mathbb{R}^T\right]\mathbb{Q}^T, \tag{A.5}$$

$$\mathbb{P}_* = \mathbb{X}_*\left(\mathbb{X}_*^T\mathbb{X}_*\right)^{-1}\mathbb{X}_*^T = \mathbb{Q}_*\left[\mathbb{R}_*\left(\mathbb{R}_*^T\mathbb{R}_*\right)^{-1}\mathbb{R}_*^T\right]\mathbb{Q}_*^T. \tag{A.6}$$

Since $\mathbb{R}$ is an upper triangular matrix with positive elements on its diagonal, it implies that both $\mathbb{R}$ and $\mathbb{R}^T$ are invertible (similarly for matrix $\mathbb{R}_*$). Therefore, the following holds

$$\mathbb{R}\left(\mathbb{R}^T\mathbb{R}\right)^{-1}\mathbb{R}^T = \mathbb{R}\mathbb{R}^{-1}\left(\mathbb{R}^T\right)^{-1}\mathbb{R}^T = \mathbb{I}_d. \tag{A.7}$$

$$\mathbb{R}_*\left(\mathbb{R}_*^T\mathbb{R}_*\right)^{-1}\mathbb{R}_*^T = \mathbb{R}_*\mathbb{R}_*^{-1}\left(\mathbb{R}_*^T\right)^{-1}\mathbb{R}_*^T = \mathbb{I}_{d_*}. \tag{A.8}$$

Finally, by combining (A.5), (A.6), (A.7) and (A.8), one obtains

$$\mathbb{P} - \mathbb{P}_* = \mathbb{Q}\mathbb{Q}^T - \mathbb{Q}_*\mathbb{Q}_*^T = \sum_{i=1}^{d} \boldsymbol{q}_i\boldsymbol{q}_i^T - \sum_{j=1}^{d_*} \boldsymbol{q}_j\boldsymbol{q}_j^T = \sum_{i=d_*+1}^{d} \boldsymbol{q}_i\boldsymbol{q}_i^T. \tag{A.9}$$

The statement of the lemma follows from equation (A.9) and the fact that the set of vectors $(q_{d_*+1}, \ldots, q_d)$ is also orthonormal. □

Lemma A.7. *Let $Y \geq 0$ be a nonnegative random variable with a continuous cumulative distribution function F_Y, and $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables such that $X_n \xrightarrow[n\to\infty]{P} 0$. Then for all $a > 0$, it holds that*

$$\mathrm{P}(X_n + Y > a) \geq \mathrm{P}(Y > a) + o(1).$$

Proof. Since $X_n \xrightarrow[n\to\infty]{P} 0$, there exists a sequence of positive real numbers $\{\varepsilon_n\}_{n=1}^{\infty}$ such that $\varepsilon_n \underset{n\to\infty}{\longrightarrow} 0$ and $\mathrm{P}(|X_n| < \varepsilon_n) \underset{n\to\infty}{\longrightarrow} 1$. Therefore

$$\begin{aligned}
&\mathrm{P}(X_n + Y > a) \\
&\geq \mathrm{P}(X_n + Y > a,\ |X_n| \leq \varepsilon_n) \geq \mathrm{P}(Y > a + \varepsilon_n,\ |X_n| \leq \varepsilon_n) \\
&\geq 1 - \mathrm{P}(Y \leq a + \varepsilon_n) - \mathrm{P}(|X_n| > \varepsilon_n) = \underbrace{\mathrm{P}(Y > a + \varepsilon_n)}_{=\mathrm{P}(Y>a)+o(1)} - \underbrace{\mathrm{P}(|X_n| > \varepsilon_n)}_{=o(1)} \\
&= \mathrm{P}(Y > a) + o(1).
\end{aligned}$$

□

Lemma A.8. *Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of positive real numbers such that $x_n \underset{n\to\infty}{\longrightarrow} 0$. Then it holds that*

$$\frac{1}{1 + o(x_n)} = 1 + o(x_n).$$

Proof. Since the function $\frac{1}{1+x}$ is defined on the open interval $\left(-\frac{1}{2}, 1\right)$. Therefore, by employing the Lagrange form of the remainder (see Bartle and Sherbert [2011], Theorem 6.4.1), then there exists $c \in \left(-\frac{1}{2}, 1\right)$ such that for any $x \in \left(-\frac{1}{2}, 1\right)$

$$\frac{1}{1+x} = 1 - x + \frac{1}{(1+c)^3} x^2.$$

Thus, for sufficiently large n

$$\begin{aligned}
\frac{1}{1 + o(x_n)} &= 1 - o(x_n) + \frac{1}{(1+c)^3} \left[o(x_n)\right]^2 \\
&= 1 - o(x_n) + \frac{1}{(1+c)^3} o(x_n) = 1 + o(x_n).
\end{aligned}$$

□

Bibliography

R. G. Bartle and D. R. Sherbert. *Introduction to Real Analysis*. Fourth Edition. John Wiley Sons, 2011. ISBN 9780471433316.

R. L. Burden and J. D. Faires. *Numerical Analysis*. Ninth Edition. Cengage Learning, 2010. ISBN 9780538733519.

Jorge Luis Ojeda Cabrera and Bastiaan Quast. *locpol: Kernel Local Polynomial Regression*, 2022. `https://cran.r-project.org/web/packages/locpol/` (Version 0.8.0).

Czech Statistical Office. VOLBY.CZ. `https://www.volby.cz/`.

Isha Dewan and B. L. S. Prakasa Rao. Remarks on the strong law of large numbers for a triangular array of associated random variables. *Metrika*, 45: 225–234, 1997.

J. Fan and I. Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20:2008–2036, 1992.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. First Edition. Chapman Hall, London, 1996. ISBN 0412983214.

P. Hall and J. S. Marron. Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields*, 74:567–581, 1987.

W. Hardle and J. S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, 13:1465–1481, 1985.

R. V. Hogg, J. W. McKean, and A. T. Craig. *Introduction to Mathematical Statistics*. Eighth Edition. Pearson, 2018. ISBN 9780134686998.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Second Edition. Cambridge University Press, 2012. ISBN 9780521548236.

P. W. M. John. *Statistical Design and Analysis of Experiments*. Macmillan Company, 1971. ISBN 9781124050287.

J. S. Marron and D. Nolan. Canonical kernels for density estimation. *Statistics Probability Letters*, 7:195–199, 1988.

S. Nagy and M. Omelka. Mathematical statistics 4. `http://www.karlin.mff.cuni.cz/~nagy/NMST545/NMST545.pdf`, 2024. Last updated: February 22, 2024.

M. Omelka. Mathematická statistika 1. `https://www.karlin.mff.cuni.cz/~komarek/vyuka/2022_23/nmsa331/ms1.pdf`. Last updated: August 13, 2022.

J. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 1962.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023a. `https://www.R-project.org/` (Version 4.3.2).

R Core Team. *stats: The R Stats Package*, 2023b. `https://www.R-project.org/` (Version 4.3.2).

D. W. Scott and G. R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82:1131–1146, 1987.

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. First Edition. Chapman Hall, London, 1986. ISBN 0412246201.

C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12:1285–1297, 1984.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. First Edition. Chapman Hall, London, 1995. ISBN 0412552701.

D. Wied and R. Weißbach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 51:1–21, 2012.

Y. Xia and W. K. Li. Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of Multivariate Analysis*, 83:265–287, 2002.