8-2022

Cross-validation for autoregressive models.

Christina Han
University of Louisville

CROSS-VALIDATION FOR AUTOREGRESSIVE MODELS

By

Christina Han
B.A., Northland College, 2010
M.A., University of Louisville, 2020

A Dissertation
Submitted to the Faculty of the
College of Arts and Sciences of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in
Applied and Industrial Mathematics

Department of Mathematics
University of Louisville
Louisville, Kentucky

August 2022

CROSS-VALIDATION FOR AUTOREGRESSIVE MODELS

Submitted by

Christina Han

A Dissertation Approved on

Augus 3, 2022

by the Following Dissertation Committee:

Dr. Ryan Gill,
Dissertation Director

Dr. Cristina Tone

Dr. Dan Han

Dr. Karunarathna Kulasekera

DEDICATION

I am thankful to all of my friends and family who have believed in me even when I did not believe in myself. Their encouragement and support has made completing this paper possible.

ACKNOWLEDGEMENTS

I would like to thank the faculty and staff in the mathematics department; the foundational knowledge they gave me was invaluable. Thank you to my committee members for their time and guidance. Finally, I would like to thank to my advisor, Dr. Gill, for his tireless patience and help through this process.

ABSTRACT

CROSS-VALIDATION FOR AUTOREGRESSIVE MODELS

Christina Han

August 3, 2022

There are no set rules for choosing the lag order for autoregressive (AR) time series models. Currently, the most common methods employ AIC or BIC. However, AIC has been proven to be inconsistent and BIC is inefficient. Racine proposed an estimator based on Shao's work which he hypothesized would also be consistent, but left the proof as an open problem. We will show his claim does not follow immediately from Shao. However, Shao offered another consistent method for cross validation of linear models called APCV, and we will show that AR models satisfy Shao's conditions. Thus, APCV is a consistent method for choosing lag order. Simulations also show that APCV performs as well, and in some cases, performs better than AIC, AICc, and BIC.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Before getting into the formal math and definitions, we should keep in mind the big picture:

Our goal is to find a good way to determine how far into the past we need to look in order to predict the future.

In a modern world, forecasting is vital for everything from economic predictions to DDoS detection. An essential part of creating good models is using good model selection and evaluation techniques. Cross-validation (CV) is a standard procedure for model selection and evaluation with the benefit of making use of the entire data set. However, CV techniques are typically reserved for independent data where shuffling data will not impede the validity of the error measure. A natural extension of CV would be to apply it to time series data. However, time series observations are not independent, and keeping the temporal order of the data is important for model building and evaluation. Recent literature has applied modified cross-validation techniques to time series data [4] [5] [8] [11] [12] [21] [22] [28].

In particular, Racine uses a technique called hv-block cross-validation, a modified version of Shao's BICV (cross-validation based on balanced incomplete block design) and Burman et al.'s h-block cross-validation, in which the data is kept in temporal order and some data is removed around the test set in order to create independence between the training and test data [22]. Racine leaves proving that hv-block cross-validation is consistent as an open problem, but he conjectures that it should follow immediately from Shao's work on cross-validation for linear model selection. We will show that this is not true, as not being able to shuffle the data means that we cannot achieve a balanced incomplete block design except in trivial cases. However, Shao offers an alternative to BICV called APCV which does not depend on the BIBD, and it is designed specifically for linear models. While his proofs and examples are for deterministic predictors, he asserts that his results hold almost surely for random variables. Thus, we will use Shao's work as a framework to show that APCV is a consistent estimator for order selection in autoregressive models.

Chapter 2 introduces basic concepts primarily for time series data, as well as a few general concepts necessary to understand folowing chapters. Section 2.1 covers foundational definitions and notation, and in 2.2 we discuss and define current methods for determining the lag order of AR(p) models. We define commonly used penalized methods like Akaike information criterion (AIC) and Bayesian information criterion (BIC), and an algorithm called false nearest neighbors (FNN) [17]. All methods have been used for order selection with time series data. We briefly discuss some cross-validation based methods, APCV and HVCV, but leave thorough discussion of these methods in the literature review.

Our literature review is contained in Chapter 3. We cover the most relevant results from Shao [25], Burman et al [8], Racine [22], Cerqueira et al. [12], and Zeng [28]. We use these papers to help guide our work and show that APCV is consistent for choosing lag order in autoregressive models.

In 3.1.1 we cover Shao's results for three different methods for model selection [25]. The first is balanced incomplete cross-validation (BICV) which is based on his version of balanced incomplete block design (BIBD). BIBD is the set of rules used for choosing training and validation sets used in BICV. Shao's second method is Monte-Carlo cross-validation (MCCV) which is less computationally expensive than BICV. In MCCV we use a randomly selected subset of the training and validation sets for model selection rather than all sets determined through BIBD. Shao's last method, analytic approximate cross-validation (APCV) is the least computationally expensive of the three methods, but tends to require a larger data set to be competitive with BICV and MCCV. APCV is also limited to linear models, and expansion to other models is left as an open problem.

Section 3.1.2 covers Burman, Nolan, and Chow's modified leave-one-out cross-validation method for dependent data called h-block cross-validation. This method keeps the data in temporal order and removes h data points from training around the test point. Leave-one-out cross-validation is inconsistent and since h-block cross-validation is LOO when $h = 0$ it follows that it is also inconsistent.

In 3.1.3 we cover Racine's remedy for the inconsistency of h-block cross-validation with his method hv-block cross-validation (HVCV). Like h-block cross-validation HVCV keeps the data in temporal order, and removes h observations around the test set, but the test set size is larger than one observation. This follows the intution from Shao, that the validation set must be large in order to get an accurate prediction error. However, we show that HVCV's consistency does not follow immediately from Shao and we provide a simple explanation as to why HVCV generally cannot satisfy Shao's BIBD.

Cerqueira et al.'s work is covered in 3.1.4. They survey various cross-validation and out-of-sample methods for model evaluation. They determine that cross-validation methods do not work as well as out-of-sample methods with real-world data. We do note that their work focuses on model evaluation, whereas our work is in line with Shao and Racine and we focus on order selection. The subtle difference between model evaluation and order selection being that in order selection the goal is to find the correct size of the model, and in model evaluation we have presumably already determined the size of the model and are now gauging how the model performs on unseen data.

Finally, in 3.1.5 Zeng provides more examples and insight as to why HVCV does not follow from Shao, and provides a comprehensive list of papers that inherit the mistake from Racine. Zeng also provides a python package with different cross-validation and out-of-sample methods for temporal data.

Our main result can be found in Chapter 4. Using Shao's work as the framework, we show that APCV is a consistent method for choosing the lag order of an autoregressive model. We explain all necessary assumptions and notation. For Shao's results to hold for AR models we need to show that Shao's conditions are satisfied; so we adapt his conditions to the time series setting and prove them as lemmas. After proving the three lemmas hold, Shao's results follow for autoregressive models.

Simulations and examples with real world data in R can be found in Chapter 5. In 5.1 we provide a comparison of AIC, AICc, BIC, HVCV, and APCV with simulated data. We use the `arima.sim` function to create four different simulated data sets with 1000 observations for AR(2), AR(3), AR(4), and AR(5) models. Since we have the ground truth, we can easily evaluate each method's output. For the real-world example in 5.2, we follow Racine's example [22] and use G7 exchange rates data sets for the following six countries: Canada (CAD), Germany (DEM), France (FRF), Great Britain (GBP), Italy (ITL), and Japan (JPY) taken from [13]. We compare the selected lag order for the following methods: AIC, AICc, BIC, HVCV, and APCV. Then to evaluate each method we check the results against the PACF plots since, unlike the simulated data, we do not know the ground truth. In both the simulated and real-world examples we will see that APCV performs as well as other methods, and frequently outperforms other methods for order selection in autoregressive models.

The Appendix is reserved for two smaller results and detailed proofs for four of Shao's results. These proofs provide information and context that is not readily available, but is too dense to fit well elsewhere. We also use the appendix to write out all of the details of Shao's proofs where it is relevant to our work.

CHAPTER 2
DEFINITIONS

2.1 Definitions

2.1.1 Linear regression models

A commonly used, and relatively versatile, model is the linear model of the form

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e$$

where y is a response variable, $\boldsymbol{x}$ is a p-dimensional input vector, $\boldsymbol{\beta}$ is a p-dimensional vector of real valued parameters, and e is a random error with mean 0 and variance σ^2.

Typically, to estimate $\boldsymbol{\beta}$ we minimize the sum of squared error (SSE), which gives the distance of the data points to the regression line.

2.1.2 Consistency

For linear models, we want to estimate a parameter $\beta \in \mathbb{R}$. A sequence of estimators β_n, $n = 1, 2, ...$ of a parameter β is said to be consistent if it converges in probability to the true value of β. That is, for all $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\beta_n - \beta| \geq \varepsilon) = 0$$

or equivalently

$$\lim_{n\to\infty} P(|\beta_n - \beta| < \varepsilon) = 1.$$

2.1.3 O, O_p, and o_p

Suppose a_n and b_n are real-valued sequences. Then $a_n = O(b_n)$ if there exists a $C \in [0, \infty)$ such that $|a_n| \leq C|b_n|$ for all $n \in \mathbb{N}$.

Let X_n be a sequence of random vectors, then we say X_n is bounded in probability or tight, written $X_n = O_p(a_n)$, if $\forall \varepsilon > 0$ there exist constants c_ε and n_ε such that $P(|X_n| \leq c_\varepsilon a_n) \geq 1 - \varepsilon$ for all $n \geq n_\varepsilon$. We say $X_n = o_p(a_n)$ if for any $\varepsilon > 0$ there exists n_ε such that $P(|X_n| \leq \varepsilon a_n) \geq 1 - \varepsilon$ for all $n \geq n_\varepsilon$.

For fixed integers k and ℓ, a natural extension of this definition for real-valued sequences of $k \times \ell$ matrices $\boldsymbol{A}_n$ and real-valued sequence b_n is that $\boldsymbol{A}_n = O(b_n)$ if and only if there exists a $C \in [0, \infty)$ such that $|[\boldsymbol{A}_n]_{i,j}| \leq C|b_n|$ for all $n \in \mathbb{N}$ where $[\boldsymbol{A}_n]_{i,j}$ is the element in the ith row and jth column of $\boldsymbol{A}_n$.

2.1.4 Balanced incomplete block design (BIBD)

We will use Shao's simplified version of BIBD which only has two conditions. For BIBD to be satisfied, let $\mathscr{B}$ be a collection of b subsets of $\{1, ..., n\}$ that have size n_v. $\mathscr{B}$ is selected according to two conditions:

(a) for every i, $1 \leq i \leq n$ appears in the same number of subsets in $\mathscr{B}$

(b) for every pair (i, j), $1 \leq i < j \leq n$ appears in the same number of subsets in $\mathscr{B}$

We will give an example of how the BIBD may be employed. Let $n = 7$ so the observations in the data set are indexed $\{1, 2, 3, 4, 5, 6, 7\}$. We would like to create a set of 7 subsets with $n_v = 3$ that satisfies the conditions (a) and (b). Such a set $\mathscr{B}$ would be

$$\{\{1, 2, 4\}, \{2, 3, 5\}, \{3, 4, 6\}, \{4, 5, 7\}, \{5, 6, 1\}, \{6, 7, 2\}, \{7, 1, 3\}\}.$$

Upon investigation we can see that each number appears 3 times and appears with distinct numbers only once. Thus, (a) and (b) are satisfied and $\mathscr{B}$ is a BIBD.

2.1.5 k-fold cross-validation

Cross-validation is a procedure used for model selection and parameter tuning which makes full use of the data by not reserving a portion of the data only for testing. The most commonly used cross validation technique is k-fold cross validation, where k is the number of times we will "fold" or split the data set. Meaning, if we have a data set of size n and we are applying k-fold cross validation, we would randomly split the data set into k validation sets with n/k observations in each set (give or take a sample if n is not divisible by k). We run the algorithm k times, training the data on the $n - n/k$ samples not in the validation set, then test the generated model with the reserved samples. In this way, we make full use of the given data since we would have the option to not reserve a portion of the data only for testing. Using the entire data set for model selection is advantageous for small data sets in particular.

For consistency, we will use Shao's notation and vocabulary when discussing cross-validation. Let n be the number of observations in a data set, n_v be the size of the validation set, and $n_c = n - n_v$ be the training set size.

For a simple illustration, consider 5-fold cross-validation where $n = 100$ and $n_v = 20$, so $n_c = 80$. We would randomly select 20 observations five times without replacement to create the validation sets. Then we reserve the first validation set, run the algorithm with the remaining 80 data points, and test our model with the validation set. We repeat this procedure 5 times, once for each validation set. Figure 2.1 illustrates the procedure, with each blue block being the 20 samples in the validation set for that trial.

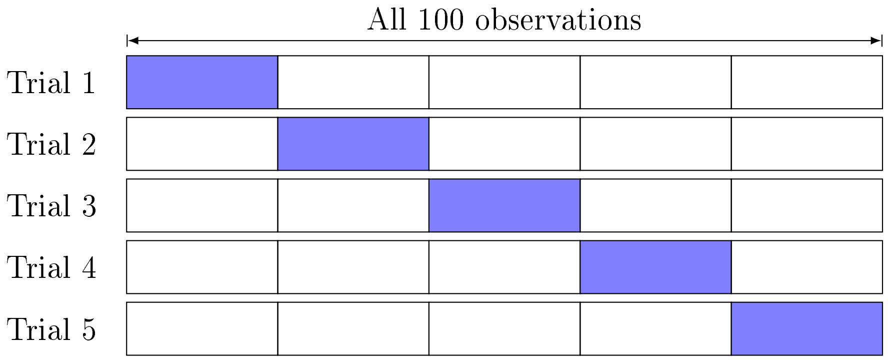

Figure 2.1: Illustration of 5-fold cross validation

2.1.6 Leave n_v-out cross-validation

Shao [25] presents a version of cross-validation that he calls leave "n_v-out cross-validation". It differs from $k-$fold cross-validation in a few ways: first, instead of choosing k for the number of folds, we choose n_v - the size of the training set. Then based on the BIBD conditions from section 2.3, we split the data into b blocks. This leads to the largest difference between $k-$fold and leave n_v-out cross-validation; in k-fold, the validation sets are disjoint, but in leave n_v-out, there are overlapping samples except in the case when $n_v = 1$ which is equivalent to n-fold cross-validation (LOO).

Only certain combinations of b, n and n_v lead to a BIBD. We can use the following to find if we can satisfy the BIBD conditions with fixed b, n and n_v. Let k_i be the number of times the ith observation appears in $\mathscr{B}$ and $k_{i,j}$ be the number of times the pair (i, j), $i \neq j$ appear together in $\mathscr{B}$. We claim that

$$k_i = \frac{n_v b}{n} \tag{2.1}$$

and

$$k_{i,j} = \frac{k^*(n_v - 1)}{(n-1)} = \frac{n_v b(n_v - 1)}{n(n-1)} \tag{2.2}$$

where k^* is the number of times an observation appears in $\mathscr{B}$. To see (2.1), note that the total number of positions overall is $\sum_{i=1}^{n} k_i = n_v b$, and for (a) to be satisfied, $k_1 = k_2 = \cdots = k_n$, so (2.1) follows.

Based on condition (b), we know that the ith observation appears in $k^*(n_v - 1)$ pairs since it appears in k^* subsets and can be paired with the $n_v - 1$ other observations in each subset. Then $\sum_{i=1}^{n} k^*(n_v - 1)$ and by (b) it follows that $k_{i,j} =$

$k_{i,1} = \cdots = k_{i,i-1} = k_{i,i+1} = \cdots = k_{i,n}$ and $k_{i,i} = 0$. Therefore (2.2) holds for $i \neq j$.

In many cases the only option is to construct $\mathscr{B}$ such that $|b| = \binom{n}{n_v}$, which could be exceptionally computationally expensive. To reduce the expense, Shao presents a solution he calls Monte Carlo cross-validation (MCCV(n_v)) which randomly selects a subset $\mathscr{R} \subset \mathscr{B}$ to use for the trials. Note that this is still different from $k-$fold cross-validation since the test sets are still not guaranteed to be disjoint.

2.1.7 Time series data

A *time series* $Y = \{y_1, y_2, ..., y_n\}$ is a set of observations y_t recorded at time t. Formally, a time series is a realization of a stochastic process, where a *stochastic process* is a family of random variables $\{y_t, t \in T\}$ defined on a probability space $(\Omega, \mathscr{F}, P)$. They can be used for regression (forecasting) or classification problems. Time series can be univariate or multivariate and discrete or continuous. We are focused on discrete univariate time series for regression problems.

A famous discrete time series data set is the "Lynx" data set that recorded the annual Canadian Lynx trappings from 1821 to 1934, shown in Figure 2.2 [10]. We can see that the number of trapped lynx is dependent upon time.

If $\{y_t, t \in T\}$ is such that $var(y_t) < \infty$ for all $t \in T$ then the *autocovariance function* for $r, s \in T$ is given by

$$\gamma_y(r, s) = Cov(y_r, y_s) = E[(y_r - E(y_r))(y_s - E(y_s))].$$

A time series (with index set $\mathbb{Z}$) is considered *stationary* if for all $r, s, t \in \mathbb{Z}$

1. $E|y_t|^2 < \infty$

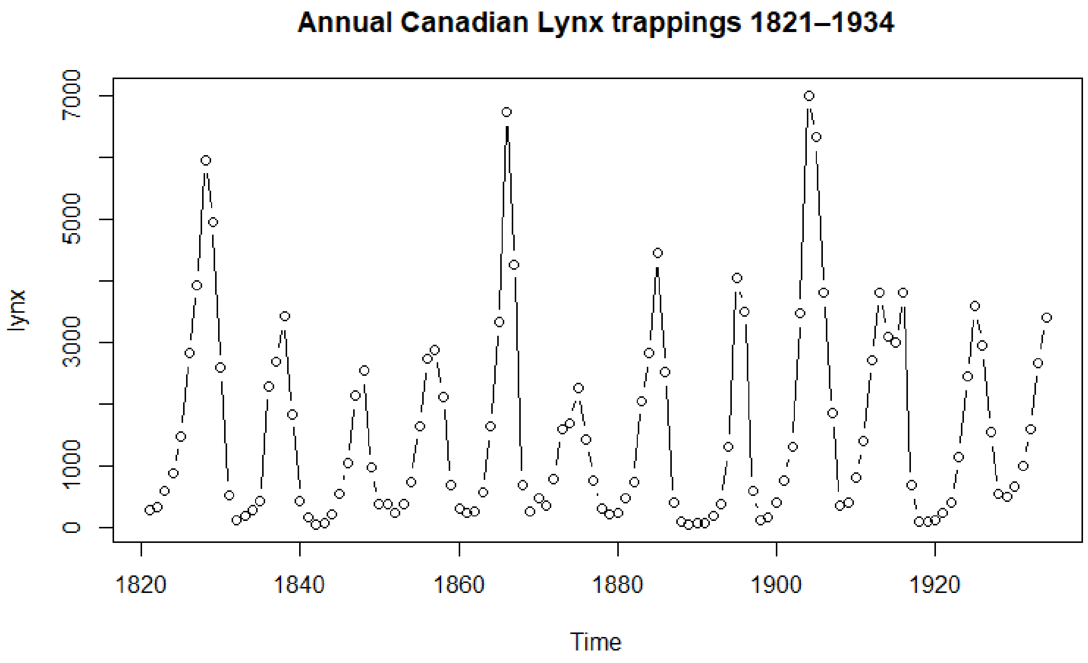

Figure 2.2: Canadian Lynx Plot

2. $E(y_t) = m$

3. $\gamma_y(r,s) = \gamma_y(r+t, s+t)$

If $\{y_t, t \in \mathbb{Z}\}$ is stationary, then we can write the autocovariance function as

$$\gamma_y(h) = Cov(y_{t+h}, y_t)\,\forall t, h \in \mathbb{Z}.$$

Then $\gamma_y(\cdot)$ is the autocovariance function of $\{y_t\}$, and $\gamma_X(h)$ is the value at lag h, where lag is the time difference.

We can visualize the difference between a non-stationary and stationary time series in Figure 2.3.

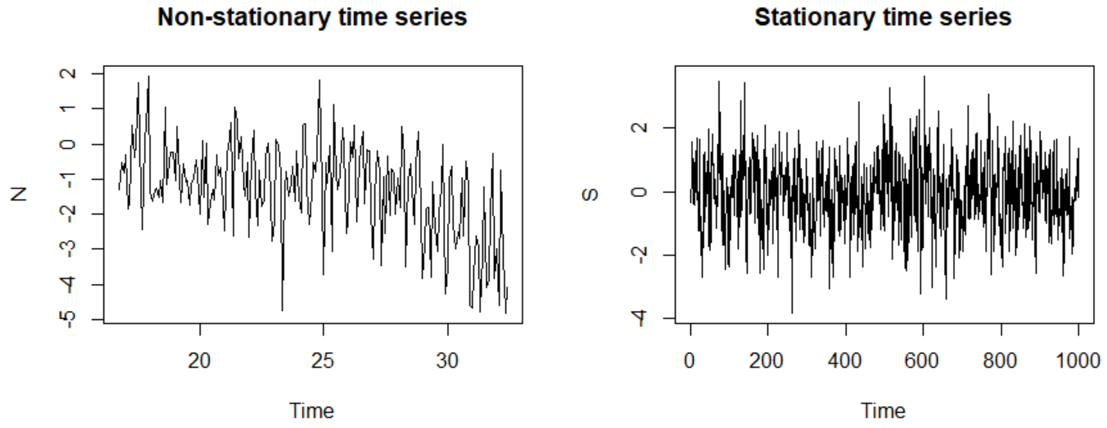

Figure 2.3: Non-stationary time series vs a stationary time series

An ARMA(p, q) process is causal if there exists a sequence of constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$, and

$$y_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}$$

where e_{t-j} is white noise with mean 0 and variance σ^2. Equivalently, we can define a causal process in terms of the characteristic equation,

$$\psi(z) = z^p - \psi_1 z^{p-1} - \ldots - \psi_p.$$

Note that in [6], the AR model is written as

$$\phi(B)y_t = e_t,$$

where B is the backshift operator such that $B(y_t) = y_{t-1}$ and $\phi(z) = 1 - \beta_1 z - \ldots - \beta_p z^p$. Then the condition that all roots of ψ lie inside the unit circle is equivalent to the condition that all roots of ϕ lie outside of the unit circle since

$$\phi\left(\frac{1}{e}\right) = \frac{e^p - \beta_1 e^{p-1} - \ldots - \beta_p}{e^p} = \frac{\psi(e)}{e^p}.$$

So

$$\psi(e) \neq 0 \text{ when } |e| \geq 1 \text{ if and only if } \phi(e) \neq 0 \text{ when } |e| \leq 1.$$

Causality is a statement about the relationship between the processes $\{y_t\}$ and $\{e_t\}$. We can also define stationarity as when all of the roots of the characteristic equation do not lie on the unit circle. Thus, causality implies stationarity.

The *autocorrelation* function of $\{y_t\}$ at lag h is defined as

$$\rho_y(h) = \gamma_y(h)/\gamma_y(0) = Corr(y_{t+h}, y_t).$$

As a naive method of determining the lag order for modeling time series we can look at the autocorrelation plot, an example is shown in figure 2.4.

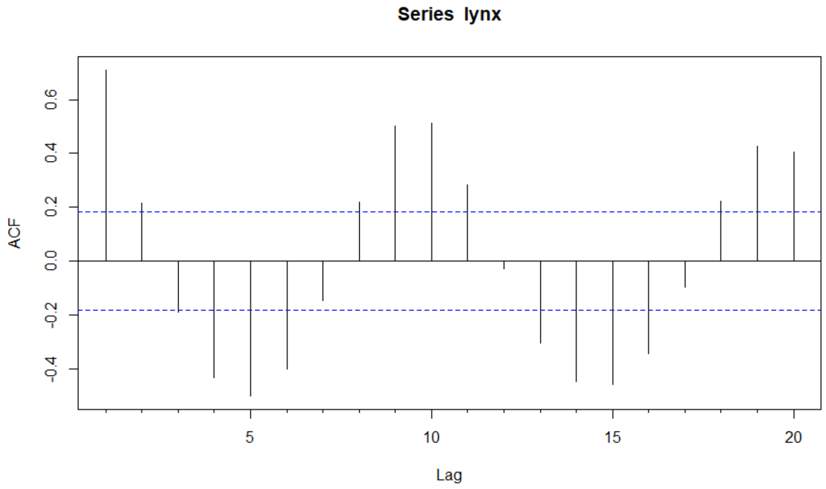

Figure 2.4: Autocorrelation plot of the lynx data set

The *partial autocorrelation* function of $\{y_t\}$ at lag 1, written $\alpha(1)$ is

$$\alpha(1) = Corr(y_{t+1}, y_t),$$

and at lag h, written $\alpha(h)$ is

$$\alpha(h) = Corr(y_{t+h} - P_{y_{t+k}}, y_t - P_{y_t}),$$

where $P_{y_{t+k}}$ is the linear combination of $\{y_{t+k}, y_{t+k-2}, ..., y_{t+1}\}$ that minimizes the mean square error, $E[y_{t+k} - P_{y_{t+k}}]^2$. Like with the autocorrelation plot, we may also use the partial autocorrelation plot to determine lag order. An example is given in figure 2.5.

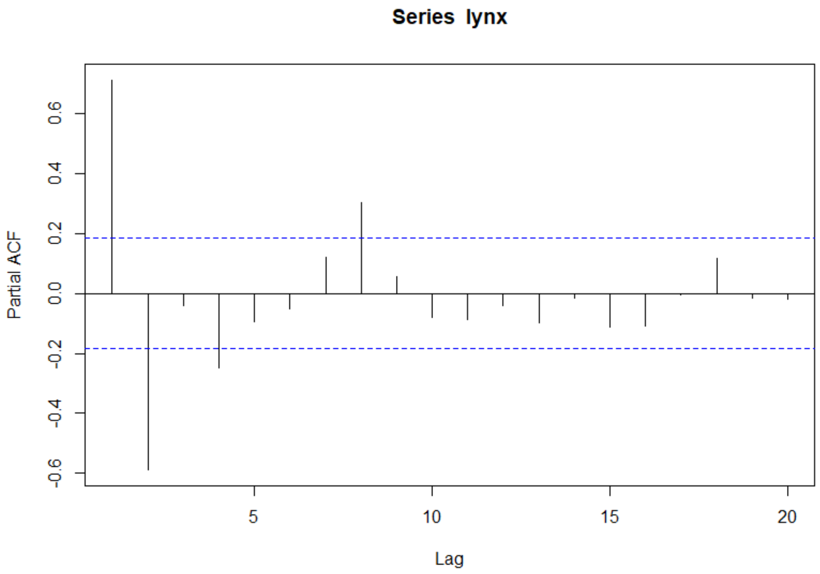

Figure 2.5: Partial autocorrelation plot of the lynx data set

2.1.8 Time series models

A stochastic process $\{y_t, t \in \mathbb{Z}\}$ is an autoregressive (AR(p)) process if $\{y_t\}$ is stationary and if $\forall t$

$$y_t - \beta_1 y_{t-1} - \cdots - \beta_p y_{t-p} = e_t$$

where e_t is white noise with mean 0 and variance σ^2.

A moving average (MA(q)) model can be written

$$y_t = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}.$$

In both MA(q) and AR(p) models, we call the size of the model the *lag order*, so if $p = 2$ then the AR(2) model is an autoregressive model of order 2:

$$y_t - \beta_1 y_{t-1} - \beta_2 y_{t-2} = e_t.$$

Determining the lag order for common time series models like moving average (MA), autoregressive (AR), autoregressive moving average (ARMA), and autoregressive moving average with differencing (ARIMA) is generally estimated with penalized methods like the Akaike information criterion (AIC), corrected Akaike information criterion (AICc) which are both defined in Section 2.2.1, and Bayesian information criterion (BIC) defined in Section 2.2.2.

Some modified versions of cross-validation have been proposed for time series data, but none have been proven to be consistent. Time series data is distinct from most data we consider since observations are not independent, and therefore the independence condition necessary for cross validation results to hold are violated. However, AR models are a linear model which have convenient properties which can be used to justify the use of cross-validation for determining lag order, which we address in 4.1. Currently, there is no definitive way to find the best p. The most common methods for order selection are discussed in the following section 2.2.

2.2 Determinimg lag order for time series models

2.2.1 Akaike information criterion (AIC)

Most commonly, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used to determine the size of the lag. Shibata [26] proposed using AIC for lag order selection in 1976. He assumes that $\{y_n\}$ is a Gaussian time series with mean zero such that

$$y_m = \beta_1 y_{m-1} + \beta_2 y_{m-2} + \cdots + \beta_k y_{m-p} + e_m$$

where $\beta_i \in \mathbb{R}$ such that

$$|y_p| = \left| \sum_{i=1}^{p} \beta_i y^{p-1} \right| < 1$$

and $\{e_n\}$ is a sequence of $N(0, \sigma^2)$ iid random variables. Let $y_1, ..., y_n$ be a set of n observations, p the order of the model, and K initial conditions. Then the MLE estimates $\hat{\beta}_i(p)$ of β_i, $i = 1, ..., p$ are defined by

$$\begin{bmatrix} \hat{R}(1,1) & ... & \hat{R}(1,p) \\ \vdots & & \vdots \\ \hat{R}(p,1) & ... & \hat{R}(p,p) \end{bmatrix} \begin{bmatrix} \hat{a}_1(p) \\ \vdots \\ \hat{a}_p(p) \end{bmatrix} = \begin{bmatrix} \hat{R}(0,1) \\ \vdots \\ \hat{R}(0,p) \end{bmatrix}$$

where

$$\hat{R}(i,j) = \frac{1}{n} \sum_{m=K+1}^{n} y_{m-i} y_{m-j}.$$

The MSE

$$\hat{\sigma}_e^2(p) = \frac{1}{n} \sum_{m=K+1}^{n} [y_m - \hat{\beta}_1(p) y_{m-1} - \cdots - \hat{\beta}_p(p) y_{m-p}]^2$$

is also the approximate MLE of σ_e^2. Then $\hat{\beta}_i(p) = 0$ for all $p < i \leq K$ so $\hat{\beta}'(p) = [\hat{\beta}_1(p), ..., \hat{\beta}_p(p), 0, ..., 0]$ is a K-dimensional vector and $p = 0, 1, ..., K$.

Let AIC be defined as a function of m where $m = 0, 1, ..., K$

$$AIC(m) = n \ln(\hat{\sigma}_e^2(m)) + 2m$$

so the selected lag order $\hat{p}$ is

$$\hat{p} = \arg\min_{m} AIC(m).$$

It has been shown that AIC is asymptotically inconsistent [20]. Meaning, AIC will sometimes pick the incorrect model, even with large datasets.

The corrected Akaike information criterion (AICc) [9] has an additional correction term which penalizes larger models. So the AICc is defined as

$$AICc(m) = n \ln(\hat{\sigma}_e^2(m)) + 2m(m+1)/(n-m-1)$$

so the selected lag is

$$\hat{p} = \arg\min_{m} AICc(m).$$

2.2.2 Bayesian information criterion (BIC)

The Bayesian information criterion is calculated similarly to the AIC, but it has a higher penalty on larger models. Then BIC is defined as a function of m where

$m = 0, 1, ..., K$

$$BIC(m) = (n-K)\ln(\hat{\sigma}_e^2(m)) + \ln(n-K)m$$

so the selected lag order $\hat{p}$ is

$$\hat{p} = \arg\min_m BIC(m).$$

Hannan (1980) [15] showed that BIC is consistent.

2.2.3 False nearest neighbors (FNN)

Kennel, Brown, Abarbanel [17] proposed an algorithm called false nearest neighbors to determine lag order. The false nearest neighbors algorithm examines the behavior of data points which are neighbors as the embedding dimension increases. If the embedding dimension is too low, many of the data points will be false neighbors, but as the embedding dimension increases, the nearest neighbors are real. Essentially, the algorithm considers the behavior of nearest neighbors as a function of dimension.

2.2.4 Analytic approximate cross-validation (APCV)

Shao [25] proposes an error measure for linear models (recall that the AR model is a linear model) called analytic approximate cross-validation (APCV) which is given by

$$\hat{\Gamma}_{\alpha,n}^{APCV} = \frac{1}{(n-K)}\|\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha\|^2 + \frac{(n-K)+n_c}{n_c((n-K)-1)}\sum_i w_{i\alpha}(y_i - \boldsymbol{x}'_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2.$$

It can be shown that this method is consistent for fixed and consistent almost surely for random y. However, in simulations, for small data sets APCV tends to select large models. This will be further discussed in section 3.1.1.

2.2.5 *hv*-block cross-validation (HVCV)

There have been two variations on cross-validation proposed for time series data which preserve the temporal order of the data and remove dependent data. The first was proposed by Burman, Nolan and Chow in "A cross-validatory method for dependent data" called h-block cross-validation, which is a variant of LOO CV. Later, Racine proposed hv-block cross-validation, which is h-block cross-validation with test sets that are larger than one observation. It is given by

$$HVCV = \frac{1}{(n-2v)n_v} \sum_{i=v}^{n-v} \|\boldsymbol{y}_{(i:v)} - \boldsymbol{X}_{(i:v)} \hat{\boldsymbol{\beta}}_{(-i:h,v)}\|^2$$

Where $n_v = 2v + 1$. These techniques will be discussed in sections 3.1.2 and 3.1.3, respectively.

2.2.6 AR model evaluation methods

There is some overlap in techniques for evaluating time series models as well as determining the lag order for a time series model. Testing a time series model is not as straightforward as it is for linear models with iid data. In this case, we have to decide if preserving the temporal order of the data is important and if we want to remove dependent data between training and testing sets.

Commonly, for time series data, we use out of sample methods (OOS), where the priority is preserving the temporal order of the data. OOS methods also never use future data to predict the past. The simplest OOS method is holdout, where we simply reserve the end of the data for testing and use the rest for training. Variants on this include iterating the holdout process and removing dependent data. Unlike cross-validation techniques, OOS methods typically do not make full use of the data. A more in-depth discussion of these methods will be in section 3.1.4.

CHAPTER 3
LITERATURE REVIEW

3.1 Literature Review

3.1.1 Model selection by cross-validation

Shao presents three different error measures and proves that under certain conditions all three are consistent and will choose the best linear model given enough data. The three methods are the balanced incomplete cross-validation (BICV), Monte-Carlo cross-validation (MCCV), and the analytic approximate cross-validation (APCV). The conditions that must be satisfied are

1. $\liminf_{n\to\infty} \Delta_{\alpha,n} > 0$ for $\mathscr{M}_\alpha$ in Category I
2. $\boldsymbol{X}'\boldsymbol{X} = O(n)$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = O\left(\frac{1}{n}\right)$
3. $\lim_{n\to\infty} \max_{i\leq n} w_{i\alpha} = 0 \, \forall \alpha \in \mathscr{A}$
4. $\lim_{n\to\infty} \max_{s\in\mathscr{B}} \left\| \frac{1}{n_v} \sum_{i\in s} \mathbf{x}_i\mathbf{x}_i' - \frac{1}{n_c} \sum_{i\in s^c} \mathbf{x}_i\mathbf{x}_i' \right\| = 0$ (not applicable for APCV)
5. $\frac{n_v}{n} \to 1$ and $n_c = n - n_v \to \infty$

where $\mathscr{A}$ is all nonempty subsets of $\{1, 2, ..., p\}$, $\mathscr{B}$ is the collection of sets described in 2.1.4, $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)'$ is a $n \times d_\alpha$ design matrix, $\mathscr{M}_\alpha$ is a model of size d_α where $\boldsymbol{X}_\alpha$ is a submatrix of $\boldsymbol{X}$ using only columns indexed by α, $w_{i\alpha}$ is the ith diagonal element of the projection matrix $\mathbf{P}_\alpha = \boldsymbol{X}_\alpha(\boldsymbol{X}_\alpha'\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}_\alpha'$, n is the size of the full data set, n_v is the size of the validation set, n_c is the size of the training set, and $\Delta_{\alpha,n} = \frac{1}{n}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta}$. Shao splits possible models into two categories: category I models which are too small and missing a non-zero component of β, and category II models which may be too large, but the best model is a category II model.

The first condition guarantees that the method will not choose a model which is too small. The second and third conditions set bounds on how fast the elements in the projection matrix can grow and that the elements of the diagonal of the projection matrix behave similarly. The fourth condition is specific to the BIBD, and it can be dropped when considering consistency of APCV. Condition 5 sets the size of the validation set to be large. Shao also only considers non-random $\mathbf{x}$ but states, without proof, that in the case of random $\mathbf{x}$, the results still hold almost surely.

The BICV method, which we discussed in section 2.1.6, selects a model by minimizing

$$\hat{\Gamma}_{\alpha,n}^{BICV} = \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \|y_s - \hat{y}_{\alpha,s^c}\|^2.$$

If it is too computationally expensive to use the BICV method, MCCV randomly selects a subset $\mathscr{R} \subset \mathscr{B}$ and selects a model by minimizing

$$\hat{\Gamma}_{\alpha,n}^{MCCV} = \frac{1}{n_v b} \sum_{s \in \mathscr{R}} \|y_s - \hat{y}_{\alpha,s^c}\|^2$$

where $\hat{y}_{\alpha,s^c} = \boldsymbol{X}_{\alpha,s^c}\hat{\boldsymbol{\beta}}_{\alpha,s^c}$ is the prediction of $\boldsymbol{y}_{s^c}$. Both methods work well when the data are independent and generalize to non linear models. The third method, APCV, selects a model by minimizing

$$\hat{\Gamma}_{\alpha,n}^{APCV} = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2 + \frac{n+n_c}{n_c(n-1)}\sum_i w_{i\alpha}(y_i - \boldsymbol{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)^2$$

and depends on the "special nature of linear models" [25]. In Shao's simulations APCV does not perform as well as MCCV and BICV, indicating that it requires more data to perform well. On small data sets APCV performs similarly to LOO.

3.1.2 *h*-block cross-validation

In "A Cross-Validatory Method for Dependent Data," Burman, Nolan, and Chow propose a modification for leave-one-out cross validation for dependent data called h-block cross validation [8]. For a data set with n samples, there would be n test points, but unlike regular leave-one-out cross validation there are two blocks of size h removed around the validation point for each trial so the training set size is $n_c = n - 2h - 1$. By removing the $2h$ data points, the test point is essentially independent from the training data. Figure 3.1 illustrates the procedure for iteration i. The $h-$block CV function is

$$HCV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i(-i:h)})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}_{(-i:h)})^2$$

where $\hat{\boldsymbol{\beta}}_{(-i:h)} = (\boldsymbol{X}_{(-i:h)}'\boldsymbol{X}_{(-i:h)})^{-1}\boldsymbol{X}_{(-i:h)}'\boldsymbol{y}_{(-i:h)}$ such that $\boldsymbol{X}_{(-i:h)}$ is the design matrix with the ith observation removed and h observations removed around the ith observation; $\boldsymbol{y}_{(-i:h)}$ is $\boldsymbol{y}$ with the ith observation removed and h observations removed around the ith observation; and $\hat{y}_{i(-i:h)}$ is the prediction for y_i.

The authors mention that it is ideal to have $\frac{h}{n} \to 0$. However, this is only reasonable for large n. So in the case of small sample sizes they set h as a fixed ratio $\frac{h}{n} \in \left(0, \frac{1}{2}\right)$. Then to correct for the under use of the sample, the authors propose a correction term and note that without it, the performance is equivalent to regular leave-one-out cross validation. The paper focuses on finding the optimal h and correction factor. They also note that conditions for asymptotic optimality is unsolved.

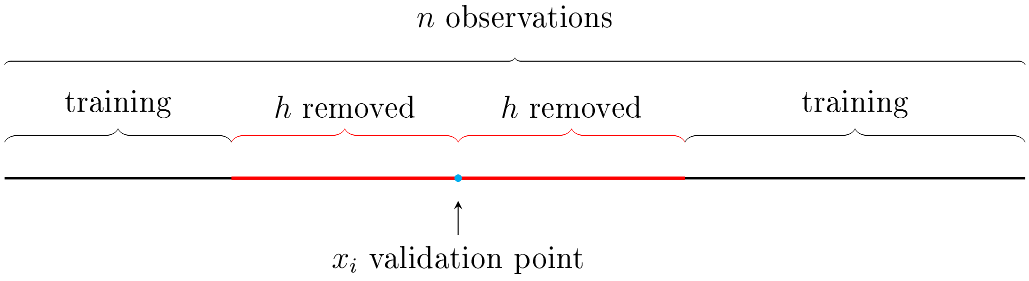

Figure 3.1: Illustration of h-block cross validation

3.1.3 hv-block cross-validation

Racine [22] proposes hv-block cross validation, a modification of the h-block cross validation proposed in Burman's paper [22]. Recall that Burman et.al. require a correction term since $\frac{h}{n}$ is not negligible, but Racine is only concerned with the case where $\frac{h}{n} \to 0$, so he ignores the correction term. Using Racine's notation, let

$\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X})$ be the matrix of n observations on the response and p predictors

$$\boldsymbol{Z} = \begin{bmatrix} y_1 & \mathbf{x_1'} \\ y_1 & \mathbf{x_2'} \\ \vdots & \vdots \\ y_n & \mathbf{x_n'} \end{bmatrix}$$

so we denote the removed data as $\boldsymbol{Z}_{(-i:h)} = (\boldsymbol{y}_{(-i:h)}, \boldsymbol{X}_{(-i:h)})$ and the remaining test set as $\boldsymbol{Z}_{(i:h)} = (\boldsymbol{y}_{(i:h)}, \boldsymbol{X}_{(i:h)})$. Then the $hv-$block CV function is defined as

$$HVCV = \frac{1}{(n-2v)n_v} \sum_{i=v}^{n-v} \|\boldsymbol{y}_{(i:v)} - \boldsymbol{X}_{(i:v)}\hat{\boldsymbol{\beta}}_{(-i:h,v)}\|^2$$

Note that HCV is HVCV where $n_v = 1$, and thus LOO and h-block CV are special cases of hv-block CV. Racine states that $HV-$Block CV satisfies the BIBD, and a proof similar to Shao's should follow proving that HVCV is consistent, but leaving the proof as an open question. Ignoring removing dependent data, achieving a BIBD and keeping the order of time series data is generally impossible except in trivial cases (block size 1, n, or $n-1$). Condition (ii) fails - every i, j pair appear in the same number of blocks - since data points appear more often with close neighbors than further data points. For example, let our data set have 6 observations, $\{1, 2, 3, 4, 5, 6\}$ and we want our block size to be 2. Then, the only acceptable sets are

$$\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}, \{5, 6\}, \{6, 1\}$$

but 1 never appears with 4. Similarly, if our block size is 3,

$$\{1, 2, 3\}, \{2, 3, 4\}, \{3, 4, 5\}, \{4, 5, 6\}, \{5, 6, 1\}, \{6, 1, 2\}$$

but 1 never appears with 4. We have a similar problem for block size 4.

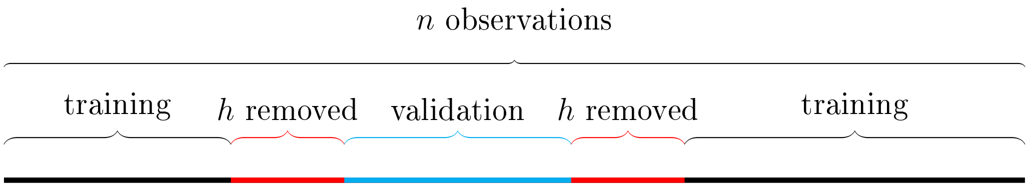

Figure 3.2: Illustration of hv-block cross validation

3.1.4 Evaluating time series forecasting models

Cerqueira et al. [12] build on the Bergmeir & Benítez paper [5] claiming the synthetic data results do not reflect real world data, and they do not test enough out of sample (OOS) methods of validation. The authors repeat the experiment with real-world data and more OOS methods. The 11 performance estimation methods are split into four cross-validation methods and seven OOS methods:

- k-fold cross-validation - described in 2.1.5
- blocked k-fold cross-validation - h-block cross validation described in 3.1.2
- modified k-fold cross-validation - k-fold cross-validation with dependent samples removed
- hv-blocked cross-validation - described in 3.1.3
- holdout - reserve the end of the sample for testing, and use the rest for training
- repeated holdout - iterative holdout where a sample is selected from the set which marks the end of the training data, then data past that point is for testing; then repeat.

- prequential blocks - split data into n blocks, use first block for training, second for testing, then use the first block and second block for training and third for testing, ..., use first through $n-1$th block for training and nth block for testing
- prequential sliding blocks - split data into n blocks, use first block for training, second for testing, then second block for training and third for testing, ..., $n-1$th block for training and nth block for testing
- prequential blocks with a gap - prequential blocks but leave out a block of data between training and testing
- prequential grow - use only one observation for testing in prequential blocks
- prequential sliding window - use only one observation for testing in prequential sliding blocks

They only consider an AR model using a rule based regression algorithm called Cubist, as well as LASSO and random forest. The authors use false nearest neighbors to estimate the length of the lag p. This is taken from Kennel and Brown [17]. The error estimators they use are predictive accuracy error (PAE), absolute predictive accuracy error (APAE). With L^m be the true error, and $\hat{g_i^m}$ be the estimated loss for model m so

$$PAE = \hat{g_i^m} - L^m$$

and

$$APAE = |\hat{g_i^m} - L^m|.$$

A decision tree is provided to show which validation technique gives better error estimation under given conditions. They come to the conclusion that for sythetic stationary data CV preforms competitively, but in the case of non-stationary real world data, validation techniques which preserve temporal order preform better with repeated holdout performing well across all data sets. Methods and data sets are here: `https://github.com/vcerqueira/performance_estimation` .

3.1.5 *hv*-block is not BIBD

As we noted in section 3.1.3, $hv-$block cross-validation does not satisfy the balanced incomplete block design (BIBD) defined in Shao's paper "Linear model selection by cross-validation." In "*hv*-Block Cross Validation is not a BIBD: a Note on the Paper by Jeff Racine (2000)," Zheng gives multiple points and reasons as to why *hv*-block is not BIBD, as well as compiles a list of 64 papers that reference Racine's paper without noting the mistake [28]. Therefore, theoretical proof of Racine's claim that *hv*-block CV is consistent is still an open question and does not follow from Shao's proofs.

Additionally, Zheng has created a Python package for time series cross validation. It has four variations on cross-validation for time series data: gap leave p out, gap k-fold, gap walk forward, and gap train test split. All four versions keep the data in temporal order, and the user sets the size of the data to be removed (gap) before and after the validation set. In gap leave p out, the validation sets are contiguous; this differs from gap k-fold since in gap k-fold the validation sets are disjoint. Gap walk forward is the same as a rolling window method, but it introduces a gap between the training and validation sets. Gap train test split is not actually a cross-validation method and simply splits the data into test, gap, and training sets.

CHAPTER 4
APCV FOR AR MODELS

4.1 APCV for AR time series models

We have already seen that an AR process cannot satisfy the BIBD condition necessary for Shao's proposed BICV and MCCV. However, APCV does not depend on the BIBD, only on least squares estimation for linear models. In the case of time series, our y_i are random and Shao assures that all of his statements hold if the conditions in 3.1.1 hold almost surely for random x_i. Thus, we need to show that the AR(p) process satisfies the conditions necessary for APCV to be a consistent estimator for the autoregressive time series model.

First we will define our notation and outline assumptions. We consider n observations, $\{y_1, y_2, ..., y_n\}$, from an AR(p) process with iid errors $e_t \sim WN(0, \sigma^2)$, $E(e_t^4) < \infty$ and $E(y_t) = 0 \, \forall t \in \mathbb{Z}$, which make up the design matrix

$$\boldsymbol{X} = \begin{bmatrix} y_K & \cdots & y_{K-p+1} \\ y_{K+1} & \cdots & y_{K-p+2} \\ \vdots & \vdots & \vdots \\ y_{n-1} & \cdots & y_{n-p} \end{bmatrix}$$

and projection matrix

$$\mathbf{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

The AR(p) model is given by

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + e_t.$$

Using notation from [2], the above can be written in matrix form as

$$\widetilde{\boldsymbol{y}}_t = \widetilde{\mathbf{e}}_t - \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{y}}_{t-1}$$

where

$$\widetilde{\boldsymbol{y}}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix}, \widetilde{\mathbf{e}}_t = \begin{bmatrix} e_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \widetilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{p-1} & \beta_p \\ -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix}$$

$E(\mathbf{e}_t\mathbf{e}_t') = \mathbf{\Sigma} = \sigma^2\boldsymbol{I}$, and that $E(\boldsymbol{y}_t\boldsymbol{y}_t') = \mathbf{\Gamma}$ where $\mathbf{\Gamma}$ is a positive definite matrix such that it is the solution to the equation

$$\mathbf{\Gamma} - \boldsymbol{\beta}'\mathbf{\Gamma}\boldsymbol{\beta} = \mathbf{\Sigma}. \tag{4.1}$$

Finally, we also assume that the process is causal, defined as in 2.1.7.

From the assumption that $E(y_t) = 0\,\forall t$, e_t are iid white noise with mean 0 and variance σ^2, $E(e_t^4) < \infty$, and $\boldsymbol{y}_t$ is causal then it can be shown that

$$E(y_t^4) < \infty. \tag{4.2}$$

Proof of (4.2) can be found in the appendix I.0.1. This is a necessary condition in following proofs.

We will prove three lemmas, and state two final conditions aligned with [25]:

1. **Lemma 1** $\liminf\limits_{n\to\infty} \Delta_{\alpha,n} > 0$ *for* $\mathscr{M}_\alpha$ *in Category I w.p.1*

2. **Lemma 2** $\boldsymbol{X}'\boldsymbol{X} = O(n)$ *and* $(\boldsymbol{X}'\boldsymbol{X})^{-1} = O\left(\frac{1}{n}\right)$ *w.p.1*

3. **Lemma 3** $\lim\limits_{n\to\infty} \max\limits_{i\leq n-K} w_{i\alpha} = 0\,\forall\alpha \in \mathscr{A}$ *w.p.1*

4. $\frac{n_v}{n} \to 1$ and $n_c = (n-K) - n_v \to \infty$

5. Let h be the amount of removed data on either side of the testing set, then $\frac{h}{n} \to 0$

Theorem 1 *Provided the above assumptions and conditions 4 and 5 are satisfied, the following statements from [25] hold for causal AR(p) processes.*

1. *If $\mathscr{M}_\alpha$ is in Category I, then there exists $R_n \geq 0$ such that*

$$\hat{\Gamma}^{APCV}_{\alpha,n} = \frac{1}{n-K}\boldsymbol{e}'\boldsymbol{e} + \Delta_{\alpha,n} + o_p(1) + R_n$$

2. *If $\mathscr{M}_\alpha$ is in Category II, then*

$$\hat{\Gamma}^{APCV}_{\alpha,n} = \frac{1}{n-K}\boldsymbol{e}'\boldsymbol{e} + \frac{1}{n_c}d_\alpha\sigma^2 + o_p\left(\frac{1}{n_c}\right)$$

3. *Consequently,*

$$\lim_{n\to\infty} P(\textit{the selected model is optimal}) = 1$$

In words, Theorem 1 guarantees that APCV is a consistent method for choosing lag order in causal AR models. We will show that Lemmas 1, 2 and 3 hold almost surely for AR models as defined above. Conditions 4 and 5 align with Shao's assumptions.

4.1.1 Proof of Lemma 1

First, we will show that Lemma 1 holds:

$$\liminf_{n\to\infty} \Delta_{\alpha,n} > 0 \text{ for } \mathscr{M}_\alpha \text{ in Category I} \tag{4.3}$$

where

$$\Delta_{\alpha,n} = \frac{1}{n-K}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}_{(n-K)} - \boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta} \text{ and } \boldsymbol{P}_\alpha = \boldsymbol{X}_\alpha(\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_\alpha. \tag{4.4}$$

Category I models are misspecified in that they are too small, i.e. $\boldsymbol{\beta}_\alpha$ is missing some non-zero components of the true $\boldsymbol{\beta}$. Shao gives the intuition as to why this is true by explaining that since $\boldsymbol{X}_\alpha$ is a submatrix of $\boldsymbol{X}$ we can see that (4.6) will be true. However, we will show this rigorously.

Proof:

Given (4.4),

$$\begin{aligned}\Delta_{\alpha,n} &= \frac{1}{n-K}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}_{n-K} - \boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta} \\ &= \frac{1}{n-K}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}_{n-K} - \boldsymbol{X}_\alpha(\boldsymbol{X}_\alpha'\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}_\alpha')\boldsymbol{X}\boldsymbol{\beta} \\ &= \frac{1}{n-K}(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}_\alpha(\boldsymbol{X}_\alpha'\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}_\alpha'\boldsymbol{X}\boldsymbol{\beta}). \qquad (4.5)\end{aligned}$$

So we need to show that

$$\frac{1}{n-K}(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}_\alpha(\boldsymbol{X}_\alpha'\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}_\alpha'\boldsymbol{X}\boldsymbol{\beta}) > 0. \qquad (4.6)$$

Under the assumption that $\mathscr{M}_\alpha$ is in Category I, we may define $\boldsymbol{X}$ a $(n-K)\times p$ matrix; $\boldsymbol{X}_{\tilde{\alpha}}$ a $(n-K)\times(p-m)$ matrix; and $\boldsymbol{X}_\alpha$ a $(n-K)\times m$ where $0 < m < p$ such that

$$\boldsymbol{X} = \begin{bmatrix} y_K & \cdots & y_{K-p+1} \\ y_{K+1} & \cdots & y_{K-p+2} \\ \vdots & \vdots & \vdots \\ y_{n-1} & \cdots & y_{n-p} \end{bmatrix}, \boldsymbol{X}_\alpha = \begin{bmatrix} y_K & \cdots & y_{K-m+1} \\ y_{K+1} & \cdots & y_{K-m+2} \\ \vdots & \vdots & \vdots \\ y_{n-1} & \cdots & y_{n-p+m} \end{bmatrix}, \boldsymbol{X}_{\tilde{\alpha}} = \begin{bmatrix} y_{K-m} & \cdots & y_{K-p+1} \\ y_{K-m+1} & \cdots & y_{K-p+2} \\ \vdots & \vdots & \vdots \\ y_{n-p+m-1} & \cdots & y_{n-p} \end{bmatrix}$$

so that we may write

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_\alpha & \boldsymbol{X}_{\tilde{\alpha}} \end{bmatrix}$$

Then

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{X} &= \begin{bmatrix}\boldsymbol{X}'_{\alpha}\\ \boldsymbol{X}'_{\tilde{\alpha}}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_{\alpha} & \boldsymbol{X}_{\tilde{\alpha}}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} & \boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\tilde{\alpha}}\\ \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\alpha} & \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}}\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{X}_{\alpha}(\boldsymbol{X}'\boldsymbol{X}_{\alpha}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\boldsymbol{X})\boldsymbol{X}'_{\alpha}\boldsymbol{X} &= \begin{bmatrix}\boldsymbol{X}'_{\alpha}\\ \boldsymbol{X}'_{\tilde{\alpha}}\end{bmatrix}\boldsymbol{X}_{\alpha}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\begin{bmatrix}\boldsymbol{X}_{\alpha} & \boldsymbol{X}_{\tilde{\alpha}}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} & \boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\tilde{\alpha}}\\ \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\alpha} & \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\alpha}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\tilde{\alpha}}\end{bmatrix}\end{aligned}$$

Since $\boldsymbol{\Gamma} = \begin{bmatrix}\boldsymbol{\Gamma}_{\alpha\alpha} & \boldsymbol{\Gamma}_{\alpha\tilde{\alpha}}\\ \boldsymbol{\Gamma}_{\tilde{\alpha}\alpha} & \boldsymbol{\Gamma}_{\tilde{\alpha}\tilde{\alpha}}\end{bmatrix}$ is positive definite, where $\boldsymbol{\Gamma}$ is defined in (4.1), $\boldsymbol{\beta}'\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\beta} > 0$ for any nonzero $\boldsymbol{\beta}$ [1]. Letting $\boldsymbol{\beta} = \begin{bmatrix}-\boldsymbol{\Gamma}^{-1}_{\alpha\alpha}\boldsymbol{\Gamma}_{\alpha\tilde{\alpha}}\boldsymbol{\beta}_{\tilde{\alpha}}\\ \boldsymbol{\beta}_{\tilde{\alpha}}\end{bmatrix}$ yields

$$\boldsymbol{\beta}'_{\tilde{\alpha}}\boldsymbol{\Gamma}_{\tilde{\alpha}\tilde{\alpha}}\boldsymbol{\beta}_{\tilde{\alpha}} - \boldsymbol{\beta}'_{\tilde{\alpha}}\boldsymbol{\Gamma}_{\tilde{\alpha}\alpha}\boldsymbol{\Gamma}^{-1}_{\alpha\alpha}\boldsymbol{\Gamma}_{\alpha\tilde{\alpha}}\boldsymbol{\beta}_{1} > 0. \tag{4.7}$$

With the above decomposition, we need to show that

$$\frac{1}{n}\boldsymbol{\beta}'_{\tilde{\alpha}}(\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}} - \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\alpha}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\tilde{\alpha}})\boldsymbol{\beta}_{\tilde{\alpha}} > 0 \tag{4.8}$$

for any nonzero $\boldsymbol{\beta}_{\tilde{\alpha}}$, which implies (4.3). Since

$$\begin{aligned}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{X}_{\alpha}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\boldsymbol{X})\boldsymbol{\beta} &= \begin{bmatrix}\boldsymbol{\beta}'_{\alpha} & \boldsymbol{\beta}'_{\tilde{\alpha}}\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}} - \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\alpha}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\tilde{\alpha}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_{\alpha}\\ \boldsymbol{\beta}_{\tilde{\alpha}}\end{bmatrix}\\ &= \boldsymbol{\beta}'_{\tilde{\alpha}}(\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}} - \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\alpha}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\tilde{\alpha}})\boldsymbol{\beta}_{\tilde{\alpha}}\end{aligned}$$

for $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_\alpha \\ \boldsymbol{\beta}_1 \end{bmatrix}$, it follows that

$$\Delta_{\alpha,n} = \frac{1}{n-K}\boldsymbol{\beta}'_{\tilde{\alpha}}(\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}} - \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_\alpha(\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_\alpha\boldsymbol{X}_{\tilde{\alpha}})\boldsymbol{\beta}_{\tilde{\alpha}}. \tag{4.9}$$

Using Lemma 2 from [1],

$$\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{\Gamma}\, a.s.$$

so

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} &= \lim_{n\to\infty}\begin{bmatrix} \frac{1}{n}\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha & \frac{1}{n}\boldsymbol{X}'_\alpha\boldsymbol{X}_{\tilde{\alpha}} \\ \frac{1}{n}\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_\alpha & \frac{1}{n}\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Gamma}_{\alpha\alpha} & \boldsymbol{\Gamma}_{\alpha\tilde{\alpha}} \\ \boldsymbol{\Gamma}_{\tilde{\alpha}\alpha} & \boldsymbol{\Gamma}_{\tilde{\alpha}\tilde{\alpha}} \end{bmatrix} a.s.\end{aligned}$$

Taking the limit of (4.9) results in

$$\lim_{n\to\infty}\frac{1}{n-K}\boldsymbol{\beta}'_{\tilde{\alpha}}(\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}} - \boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_\alpha(\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_\alpha\boldsymbol{X}_{\tilde{\alpha}})\boldsymbol{\beta}_{\tilde{\alpha}} \tag{4.10}$$

$$= \lim_{n\to\infty}\frac{1}{n-K}\boldsymbol{\beta}'_{\tilde{\alpha}}\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_{\tilde{\alpha}}\boldsymbol{\beta}_{\tilde{\alpha}} - \lim_{n\to\infty}\frac{1}{n-K}\boldsymbol{\beta}'_{\tilde{\alpha}}\boldsymbol{X}'_{\tilde{\alpha}}\boldsymbol{X}_\alpha(\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_\alpha\boldsymbol{X}_{\tilde{\alpha}}\boldsymbol{\beta}_{\tilde{\alpha}}$$

$$= \boldsymbol{\beta}'_{\tilde{\alpha}}\boldsymbol{\Gamma}_{\tilde{\alpha}\tilde{\alpha}}\boldsymbol{\beta}_{\tilde{\alpha}} - \boldsymbol{\beta}'_{\tilde{\alpha}}\boldsymbol{\Gamma}_{\tilde{\alpha}\alpha}\boldsymbol{\Gamma}^{-1}_{\alpha\alpha}\boldsymbol{\Gamma}_{\alpha\tilde{\alpha}}\boldsymbol{\beta}_{\tilde{\alpha}} > 0 \tag{4.11}$$

which must be positive when $\boldsymbol{\beta}_{\tilde{\alpha}}$ is nonzero by (4.7). Therefore, (4.8) holds.

□

4.1.2 Proof of Lemma 2

Next, we will show that Lemma 2 is true for AR models,

$$\boldsymbol{X}'_K\boldsymbol{X}_K = O(n) \text{ almost surely} \tag{4.12}$$

holds where

$$\boldsymbol{X}_K = \begin{bmatrix} y_K & \cdots & y_1 \\ y_{K+1} & \cdots & y_2 \\ \vdots & \vdots & \vdots \\ y_{n-1} & \cdots & y_{n-K} \end{bmatrix}$$

for the AR(K) model where $K \geq p$ for when p is the correct model size, such that

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-K} + e_t.$$

Proof:

When all of the roots of $\varphi(e)$ lie inside the unit circle (causal), it is shown in Theorem 3 of [18] that

$$\lambda_{\max}(\boldsymbol{X}_K'\boldsymbol{X}_K) = O(n) \text{ almost surely} \tag{4.13}$$

and

$$\liminf_{n\to\infty} \frac{1}{n}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K) > 0 \text{ almost surely} \tag{4.14}$$

where $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ are the largest and smallest eigenvalues of a matrix $\mathbf{A}$, respectively.

Recall "big O" notation from 2.1.3, defined as follows. Suppose a_n and b_n are real-valued sequences, and there exists a $C \in [0, \infty)$ such that $|a_n| \leq C|b_n|$ for all $n \in \mathbb{N}$ then $a_n = O(b_n)$.

For fixed integers k and ℓ, a natural extension of this definition for real-valued sequences of $k \times \ell$ matrices $\boldsymbol{A}_n$ and real-valued sequence b_n is that $\boldsymbol{A}_n = O(b_n)$ if and only if there exists a $C \in [0, \infty)$ such that $|[\boldsymbol{A}_n]_{i,j}| \leq C|b_n|$ for all $n \in \mathbb{N}$ where $[\boldsymbol{A}_n]_{i,j}$ is the element in the ith row and jth column of $\boldsymbol{A}_n$.

Now, we show that the first part of (4.12) holds: $\boldsymbol{X}_K'\boldsymbol{X}_K = O(n)$. By the Courant-Fischer Theorem, the smallest eigenvalue of a matrix $\boldsymbol{A}$ is the minimum value of $\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}$ for all $\boldsymbol{u}$ such that $\|\boldsymbol{u}\| = 1$ and the largest eigenvalue is the maximum value of $\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}$. If $\mathbf{1}_j$ is the unit vector such that its jth component is 1, then $\mathbf{1}_j'\boldsymbol{X}_K'\boldsymbol{X}_K\mathbf{1}_j$ is the jth diagonal element of $\boldsymbol{X}_K'\boldsymbol{X}_K$, say $[\boldsymbol{X}_K'\boldsymbol{X}_K]_{j,j}$. So, (4.13) implies that there exists a constant $C_j \in [0, \infty)$ such that

$$|[\boldsymbol{X}_K'\boldsymbol{X}_K]_{j,j}| = |\mathbf{1}_j'\boldsymbol{X}_K'\boldsymbol{X}_K\mathbf{1}_j| \leq \lambda_{\max}(\boldsymbol{X}_K'\boldsymbol{X}_K) \leq C_j n \text{ almost surely}$$

for all $n \in \mathbb{N}$. Furthermore, by the Cauchy-Schwartz inequality, we have

$$\begin{aligned} |[\boldsymbol{X}_K'\boldsymbol{X}_K]_{i,j}| &= |\mathbf{1}_i'\boldsymbol{X}_K'\boldsymbol{X}_K\mathbf{1}_j| \\ &\leq \sqrt{\mathbf{1}_i'\boldsymbol{X}_K'\boldsymbol{X}_K\mathbf{1}_i}\sqrt{\mathbf{1}_j'\boldsymbol{X}_K'\boldsymbol{X}_K\mathbf{1}_j} \\ &\leq \sqrt{C_i n}\sqrt{C_j n} = C_{i,j} n \text{ almost surely} \qquad (4.15) \end{aligned}$$

where $C_{i,j} = \sqrt{C_i C_j}$. Since (4.15) holds for all i and j (including the case when they are equal), this shows that $\boldsymbol{X}_K'\boldsymbol{X}_K = O(n)$ almost surely where the C in the definition is $C = \max_j C_j$.

Now, we show that the second part of (4.12) holds: $(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1} = O(n^{-1})$. First, note that $\boldsymbol{X}_K'\boldsymbol{X}_K$ is invertible since by (4.14) and by the properties of the limit inferior of a sequence (s_n) that $\forall \varepsilon > 0 \, \exists N \in \mathbb{N}$ such that

$$\liminf_{n\to\infty} s_n - \varepsilon < s_n, \, \forall n \geq N$$

So choose $\varepsilon < \liminf_{n\to\infty} \frac{1}{n}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K)$, then

$$\liminf_{n\to\infty} \frac{1}{n}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K) - \varepsilon < \frac{1}{n}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K) \text{ a.s. } \forall n \geq N$$

which implies that for all $n \geq N$

$$n\left(\liminf_{n\to\infty} \frac{1}{n}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K) - \varepsilon\right) < \lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K) \text{ a.s.}$$

Thus, the smallest eigenvalue of $\boldsymbol{X}_K'\boldsymbol{X}_K$ is positive, so all of the eigenvalues of $\boldsymbol{X}_K'\boldsymbol{X}_K$ is positive which means that $\boldsymbol{X}_K'\boldsymbol{X}_K$ is full rank. With the fact that $rank(\boldsymbol{X}_K) = rank(\boldsymbol{X}_K'\boldsymbol{X}_K)$ it follows that $\boldsymbol{X}_K'\boldsymbol{X}_K$ is full rank and therefore invertible.

Now, we can use the singular value decomposition $\boldsymbol{X}_K = \boldsymbol{U}_K\boldsymbol{D}_K\boldsymbol{V}_K'$ as described in [16]. Let $n_\star = n - K$ so that $\boldsymbol{X}_K$ is an $n_\star \times K$ matrix. In the "thin" version of the SVD shown in Figure 4.1 from [16], $\boldsymbol{U}_K$ is an $n_\star \times r$ matrix with orthogonal columns, $\boldsymbol{V}_K$ is a $K \times r$ matrix with orthogonal columns, and $\boldsymbol{D}_K$ is a square matrix of order r where r is the rank of $\boldsymbol{X}_K$.

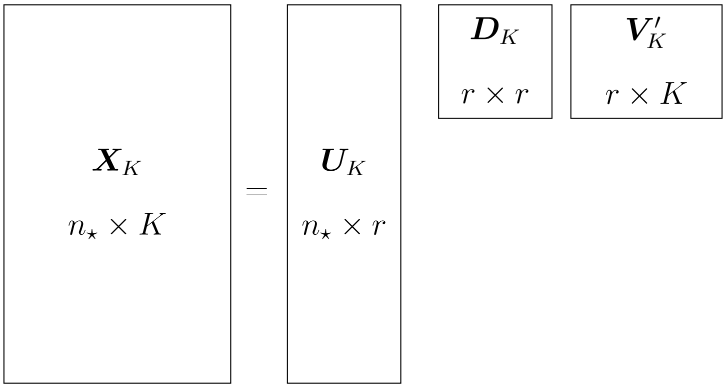

Figure 4.1: Illustration of the thin SVD described in [16]

We have $\boldsymbol{X}_K'\boldsymbol{X}_K = \boldsymbol{V}_K\boldsymbol{D}_K^2\boldsymbol{V}_K'$ so the diagonal elements of $\boldsymbol{D}_K$ are the square roots of the eigenvalues of $\boldsymbol{X}_K'\boldsymbol{X}_K$. If $\boldsymbol{X}_K$ is full rank when n is large, then $r = K$

so that $\boldsymbol{V}_K$ is a $K \times K$ orthogonal matrix and $(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1} = (\boldsymbol{V}_K')^{-1}(\boldsymbol{D}_K^2)^{-1}\boldsymbol{V}_K' = \boldsymbol{V}_K(\boldsymbol{D}_K^2)^{-1}\boldsymbol{V}_K'$. In this case, this shows that the eigenvalues of $(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1}$ are the reciprocals of the eigenvalues of $\boldsymbol{X}_K'\boldsymbol{X}_K$.

The SVD of $\boldsymbol{X}_K$ can also be written as

$$\boldsymbol{X}_K = \sum_{k=1}^{r} d_{n,k}\boldsymbol{u}_{n,k}\boldsymbol{v}_{n,k}'$$

where $\boldsymbol{u}_{n,k}$ is the kth column of $\boldsymbol{U}_K$, $\boldsymbol{v}_{n,k}$ is the kth column of $\boldsymbol{V}_K$, and $d_{n,k}$ is the kth diagonal element of $\boldsymbol{D}_K$. Then, when $\boldsymbol{X}_K'\boldsymbol{X}_K$ is full rank, we can write

$$\begin{aligned}\boldsymbol{X}_K'\boldsymbol{X}_K = \boldsymbol{V}_K(\boldsymbol{D}_K^2)\boldsymbol{V}_K' &= \sum_{k=1}^{K} d_{n,k}^2\boldsymbol{v}_{n,k}\boldsymbol{v}_{n,k}' \\ \text{and } (\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1} = \boldsymbol{V}_K(\boldsymbol{D}_K^2)^{-1}\boldsymbol{V}_K' &= \sum_{k=1}^{K} \frac{1}{d_{n,k}^2}\boldsymbol{v}_{n,k}\boldsymbol{v}_{n,k}'.\end{aligned}$$

Note that $d_{n,k}^2$ is the kth eigenvalue of $\boldsymbol{X}_K'\boldsymbol{X}_K$ and $\frac{1}{d_{n,k}^2}$ is the kth eigenvalue of $(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1}$. Then we see that

$$\begin{aligned}|[(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1}]_{i,j}| &= |\boldsymbol{1}_i'(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1}\boldsymbol{1}_j| \\ &= \left|\sum_{k=1}^{K}\frac{1}{d_{n,k}^2}\boldsymbol{1}_i'\boldsymbol{v}_{n,k}\boldsymbol{v}_{n,k}'\boldsymbol{1}_j\right| \\ &\leq \frac{1}{\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K)}\sum_{k=1}^{K}\boldsymbol{1}_i'\boldsymbol{v}_{n,k}\boldsymbol{v}_{n,k}'\boldsymbol{1}_j \\ &\leq \frac{1}{\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K)}\sum_{k=1}^{K} 1 \cdot 1 \\ &\leq \frac{K}{\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K)}. \end{aligned} \tag{4.16}$$

By (4.14), the smallest subsequential limit of $n^{-1}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K)$ is positive.

A number L is a subsequential limit of a sequence $\{\ell_n\}$ if, for every $\varepsilon > 0$, there exists some N and subsequence $\{n_i\}$ of integers greater than N such that $L-\varepsilon < \ell_{n_i} < L+\varepsilon$ for all n_i. Fix $\varepsilon > 0$ and let $L_{\min}$ denote the smallest subsequential limit. Then we can show that there is some N such that $\ell_n > L_{\min} - \varepsilon$ for all $n \geq N$. (Let $\{m_i\}$ be the set of all indices such that $\ell_{m_i} \leq L_{\min} - \varepsilon$. The set $\{m_i\}$ is finite; if it were not, then $L_{\min}$ is not the smallest subsequential limit. So, then take $N = \max m_i$.)

So, there is some N such that $n^{-1}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K) \geq C$ a.s. for $n \geq N$ where $C = \liminf_{n\to\infty} \frac{1}{n}\lambda_{\min}(\boldsymbol{X}_K'\boldsymbol{X}_K)$ is positive. From (4.16), it then follows that

$$|[(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1}]_{i,j}| \leq \frac{K}{Cn} \text{ almost surely}$$

when $n \geq N$. Hence, $(\boldsymbol{X}_K'\boldsymbol{X}_K)^{-1} = O(n^{-1})$ almost surely.

□

4.1.3 Proof of Lemma 3

Lemma 3 follows closely from Theorem 4 in [18]. In our case, we prove for the largest $\mathscr{M}_\alpha$ in Category II, which then implies the statement is true for any $\alpha \in \mathscr{A}$. We will provide a sketch of the proof. Lai & Wei's theorem in [18] is as follows:

Theorem 2 *Suppose that in the AR(p) model*

$$y_n = \beta_1 y_{n-1} + \cdots + \beta_p y_{n-p} + e_n,$$

$\{e_n\}$ is a martingale difference sequence with respect to an increasing sequence of $\sigma-$fields $\{\mathscr{F}_n\}$ such that $\liminf_{n\to\infty} E(e_n^2|\mathscr{F}_{n-1}) > 0\, a.s.$ holds. Assume that the roots e_j

of the characteristic polynomial $\varphi(e) = e^p - \beta_1 e^{p-1} - \cdots - \beta_p$ *lie on or inside the unit circle, i.e.* $|e_j| \leq 1$ *for* $j = 1, ..., p$. *Then*

$$\lim_{n\to\infty} \max_{p \leq j \leq n} \widetilde{\boldsymbol{y}}_j{}' \left(\sum_{i=p}^{n} \widetilde{\boldsymbol{y}}_i \widetilde{\boldsymbol{y}}_i{}' \right)^{-1} \widetilde{\boldsymbol{y}}_j = 0 \, a.s. \tag{4.17}$$

Before the proof of the theorem, Lai and Wei [18] outline five lemmas with proofs. The lemmas are as follows.

Lemma 4 *Let* $\{a_n\}$ *be a sequence of nonnegative numbers such that*

$$\sum_{i=1}^{n} a_i = o(n^{\delta}) \, \forall \, \delta > 0 \tag{4.18}$$

and there exist $C > 0$ *and* $\gamma > 0$ *such that*

$$a_{n+1} \leq a_n + Cn^{-\gamma} \text{ for all large } N \tag{4.19}$$

Then $\lim\limits_{n\to\infty} a_n = 0$.

Lemma 5 *Let* $g_1, ..., g_r$, $h_1, ..., h_s$ *be real numbers and let* $p = r + s$. *Define the* $s \times p$,

$r \times p$, and $p \times p$ matricies $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}$ by

$$\boldsymbol{M}_1 = \begin{pmatrix} 1 & g_1 & \cdots & g_r & 0 & \cdots & 0 \\ 0 & 1 & g_1 & \cdots & g_r & 0 & \vdots \\ \vdots & 0 & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 1 & g_1 & \cdots & \cdots & g_r \end{pmatrix},$$

$$\boldsymbol{M}_2 = \begin{pmatrix} 1 & h_1 & \cdots & h_s & 0 & \cdots & 0 \\ 0 & 1 & h_1 & \cdots & h_s & 0 & \vdots \\ \vdots & 0 & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 1 & h_1 & \cdots & \cdots & h_s \end{pmatrix},$$

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_1 \\ \boldsymbol{M}_2 \end{pmatrix} \tag{4.20}$$

Define the polynomials

$$P_1(z) = z^r + g_1 z^{r-1} + \cdots g_r,\ P_2(z) = z^s + h_1 z^{s-1} + \cdots h_s \tag{4.21}$$

(i) If P_1, P_2 are relatively prime (over the real field), then $\boldsymbol{M}$ is non-singluar.

(ii) Let $\varphi = P_1(z)P_2(z) = z^p - \beta_1 z^{p-1} - \cdots - \beta_p$. For a given sequence of real numbers $\{m\}$ and initial values $y_0, ..., y_{-p}$, define $y_n = \beta_1 y_{n-1} + \cdots \beta_p y_{n-p} + \varepsilon_n$, $n \geq 1$. Moreover, define

$$u_n = y_n + g_1 y_{n-1} + \cdots g_r y_{n-r},\ v_n = y_n + h_1 y_{n-1} + \cdots h_s y_{n-s}. \tag{4.22}$$

Then for $n \geq 1$,

$$u_n + h_1 u_{n-1} + \cdots h_s u_{n-s} = m_n = v_n + g_1 v_{n-1} + \cdots g_r v_{n-r}. \tag{4.23}$$

Lemma 6 *Let $\boldsymbol{A}$ be a $p \times p$ symmetric positive definite matrix.*

(i) If $\boldsymbol{A}^{-1} = \boldsymbol{I}_p + \boldsymbol{V} + \boldsymbol{W}$ where $\boldsymbol{V}$, $\boldsymbol{W}$ are symmetric $p \times p$ matrices such that $\boldsymbol{V}$ is nonnegative definite and $\|\boldsymbol{W}\| < 1$, then

$$\|\boldsymbol{A}\| \leq 1/(1 - \|\boldsymbol{W}\|). \tag{4.24}$$

(ii) If $\boldsymbol{A}$ is partitioned as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{P} & \boldsymbol{H} \\ \boldsymbol{H}' & \boldsymbol{Q} \end{pmatrix},$$

where $\boldsymbol{P}$, $\boldsymbol{Q}$ are respectively $r \times r$ and $s \times s$ matrices such that $p = r + s$, then for $u \in \mathbb{R}^r$,

$$\begin{pmatrix} \boldsymbol{u} \\ 0 \end{pmatrix}' \boldsymbol{A}^{-1} \begin{pmatrix} \boldsymbol{u} \\ 0 \end{pmatrix} \leq \boldsymbol{u}'\boldsymbol{P}^{-1}\boldsymbol{u}(1 + \|\boldsymbol{A}^{-1}\| tr(\boldsymbol{Q})). \tag{4.25}$$

Lemma 7 *Let $\boldsymbol{C}_n = \sum_{i=p}^{n} \widetilde{\boldsymbol{y}}_i \widetilde{\boldsymbol{y}}_i{}' = \boldsymbol{X}_{n+1}' \boldsymbol{X}_{n+1}$. Let $N = \inf\{n \geq p : \boldsymbol{C}_n$ in nonsingular$\}$. Then*

(i) $N < \infty$ a.s. and $\|\boldsymbol{C}_n^{-1/2}\| = O(n^{-1/2})$ a.s.,

(ii) $\widetilde{\boldsymbol{y}_n}{}' \boldsymbol{C}_n^{-1} \widetilde{\boldsymbol{y}_n} \leq 1$ for $n \geq N$ and

$$\sum_{i=N}^{n} \widetilde{\boldsymbol{y}}_i{}' \boldsymbol{C}_i^{-1} \widetilde{\boldsymbol{y}}_i = O(\log n) \, a.s., \tag{4.26}$$

(iii) $\|\boldsymbol{C}_n^{-1/2} \sum_{i=p}^{n} \widetilde{\boldsymbol{y}}_i \boldsymbol{e}_{i+1}\| = O((\log n)^{1/2})$ a.s.

Lemma 8 *Assume that B is nonsingular, where*

$$\boldsymbol{B} = \begin{pmatrix} \beta_1 & \cdots & \beta_{p-1} & \beta_p \\ \boldsymbol{I}_{p-1} & & & 0 \end{pmatrix}.$$

Define $\boldsymbol{C}_n$ and N as in Lemma 7. Then

(i) $\|\boldsymbol{C}_n^{1/2}\boldsymbol{B}'\boldsymbol{C}_{n+1}^{-1}\boldsymbol{B}\boldsymbol{C}_n^{1/2}\| \leq 1 + O(n^{-1/2}(\log n)^{1/2})a.s.$

(ii) Let $\rho > 1/\alpha$, where $\alpha > 2$. Then

$$\limsup_{n\to\infty} n^{1/2-\rho}(\widetilde{\boldsymbol{y}_{n+1}}'\boldsymbol{C}_{n+1}^{-1}\widetilde{\boldsymbol{y}_{n+1}} - \widetilde{\boldsymbol{y}_n}'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}_n}) \leq 0\, a.s. \tag{4.27}$$

Finally, we will include our own lemma which helps clarify the first part of the proof for Theorem 4.

Lemma 9 *Suppose the following conditions hold*

(i) $a_{nj} \geq 0$

(ii) $\lim\limits_{n\to\infty} a_{nj} = 0$ *for all fixed j*

(iii) $a_{jj} \geq a_{nj}$ *for all fixed j*

(iv) $\lim\limits_{n\to\infty} a_{nn} = 0$

Then $\lim\limits_{n\to\infty} \max\limits_{1\leq j\leq n} a_{nj} = 0.$

Proof:

By (ii), we have

$$\forall\varepsilon > 0\, \exists N_{\varepsilon,j} \text{ such that } a_{nj} < \varepsilon\, \forall n \geq N_{\varepsilon,j},\, j = 1, ..., n$$

and by (iii), we have

$$\forall \varepsilon > 0 \, \exists M_\varepsilon \text{ such that } a_{nn} < \varepsilon \, \forall n \geq M_\varepsilon.$$

We want to show that

$$\forall \varepsilon > 0 \, \exists R_\varepsilon \text{ such that } \max_{1 \leq j \leq n} a_{nj} < \varepsilon \, \forall n \geq R_\varepsilon \tag{4.28}$$

i.e. $a_{nj} < \varepsilon \, \forall n \geq R_\varepsilon$ and for all j. Fix $\varepsilon > 0$, and let $n \geq M_\varepsilon$, then by (iii) and (iv) we have

$$a_{nj} \leq a_{jj} < \varepsilon \, \forall j \geq M_\varepsilon. \tag{4.29}$$

So let $R_\varepsilon = \max\{N_{\varepsilon,1}, ..., N_{\varepsilon,M_\varepsilon - 1}, M_\varepsilon\}$, then

$$a_{nj} < \varepsilon \, \forall n > R_\varepsilon \text{ for } j = 1, ..., M_\varepsilon - 1. \tag{4.30}$$

Combining (4.29) and (4.30) we have (4.28).

□

Proof sketch of Theorem 2 (Lemma 3)

The proof of Theorem 2 can be broken down into four parts. The first part reduces 4.17 to

$$\lim_{j \to \infty} \widetilde{\boldsymbol{y}}_j{}' \boldsymbol{C}_n^{-1} \widetilde{\boldsymbol{y}}_j = 0 \, a.s. \tag{4.31}$$

where $\boldsymbol{C}_n = \sum_{i=p}^{n} \widetilde{\boldsymbol{y}}_i \widetilde{\boldsymbol{y}}_i{}'$ as in Lemma 7. Letting $a_{nj} = \widetilde{\boldsymbol{y}}_j{}' \boldsymbol{C}_n^{-1} \widetilde{\boldsymbol{y}}_j$ in Lemma 9, we can see that we need to check the four conditions. Condition (i) is satisfied since $\boldsymbol{C}_j$ is nonnegative definite, so $\widetilde{\boldsymbol{y}}_j{}' \boldsymbol{C}_n^{-1} \widetilde{\boldsymbol{y}}_j$ is nonnegative definite. Then (iii) is satisfied since $\boldsymbol{C}_j^{-1} - \boldsymbol{C}_n^{-1}$ is non-negative definite by Corollary 1 in the appendix it follows that

$$\widetilde{\boldsymbol{y}}_j{}' \boldsymbol{C}_j^{-1} \widetilde{\boldsymbol{y}}_j \geq \widetilde{\boldsymbol{y}}_j{}' \boldsymbol{C}_n^{-1} \widetilde{\boldsymbol{y}}_j \text{ for } n \geq j \geq N.$$

Next, Lemma 9 (ii) requires a little work to see. By Lemma 7 (i),

$$\|\boldsymbol{C}_n^{-1/2}\| = O(n^{-1/2})\, a.s.$$

where $\|\boldsymbol{C}_n^{-1/2}\| = \sup\limits_{\boldsymbol{x}\neq 0}\sqrt{\frac{\boldsymbol{x}'\boldsymbol{C}_n\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}}$. This implies that $\exists M > 0$ such that

$$\begin{aligned} P\left(\sqrt{\frac{n\boldsymbol{x}'\boldsymbol{C}_n\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}} \leq M\right) &= 1 \,\forall\, \boldsymbol{x} \neq 0 \\ \implies P\left(\sqrt{n\boldsymbol{x}'\boldsymbol{C}_n\boldsymbol{x}} \leq M\sqrt{\boldsymbol{x}'\boldsymbol{x}}\right) &= 1 \,\forall\, \boldsymbol{x} \\ \implies P\left(n\boldsymbol{x}'\boldsymbol{C}_n\boldsymbol{x} \leq M^2\boldsymbol{x}'\boldsymbol{x}\right) &= 1. \end{aligned}$$

We want to show that $P\left(\lim\limits_{n\to\infty} \widetilde{\boldsymbol{y}}_j{}'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}}_j = 0\right) = 1$ for all fixed j, equivalently, $\forall\varepsilon > 0$,

$$P\left(\widetilde{\boldsymbol{y}}_j{}'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}}_j > \varepsilon \text{ i.o.}\right) = 0. \tag{4.32}$$

Fix $\varepsilon > 0$. Then, by Markov's Inequality, for any random vector $\boldsymbol{x}$

$$P(\boldsymbol{x}'\boldsymbol{x} > c) = P((\boldsymbol{x}'\boldsymbol{x})^2 > c^2) \leq \frac{E[(\boldsymbol{x}'\boldsymbol{x})^2]}{c^2}.$$

With (4.2), $E[|y_{j,i}|^4] < \infty$, where $y_{j,i}$ is the ith component of $\widetilde{\boldsymbol{y}}_j$ so

$$E[(\tilde{\boldsymbol{y}}_i'\tilde{\boldsymbol{y}}_i)^2] = E[(y_{j,1}^2 + \cdots + y_{j,p}^2)^2] \leq \frac{\sum_{i=1}^p E[|y_{j,i}|^4]}{c^2}$$

Now,

$$\begin{aligned}
\sum_{i=1}^{\infty} P(\widetilde{\boldsymbol{y}}_j{}'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}}_j > \varepsilon) &\le \sum_{i=1}^{\infty}\left(\frac{M^2}{n}\widetilde{\boldsymbol{y}}_j{}'\widetilde{\boldsymbol{y}}_j > \varepsilon\right)\\
&= \sum_{i=1}^{\infty}\left(\widetilde{\boldsymbol{y}}_j{}'\widetilde{\boldsymbol{y}}_j > \frac{\varepsilon n}{M^2}\right)\\
&\le \sum_{i=1}^{\infty}\frac{\sum_{i=1}^{p} E[|y_{j,i}|^4]}{(\frac{\varepsilon n}{M^2})^2}\\
&= \frac{M^4\sum_{i=1}^{p} E[|y_{j,i}|^4]}{\varepsilon^2}\sum_{i=1}^{\infty}\frac{1}{n^2}\\
&< \infty.
\end{aligned}$$

So by the First Borel-Cantelli Lemma we have (4.32).

Now all that remains to be shown is condition (iv), which is precisely (4.31). The authors break down the proof of (4.31) into two cases based on whether or not $\boldsymbol{B}$ is invertible. We will note that the only time B is singular is when $\beta_p = 0$.

The first case examines if $\boldsymbol{B}$ is invertible and (4.31) follows from Lemmas 4, 7, and 8. From Lemma 8(ii) we have $\widetilde{\boldsymbol{y}}_{n+1}'\boldsymbol{C}_{n+1}^{-1}\widetilde{\boldsymbol{y}}_{n+1} \le \widetilde{\boldsymbol{y}}_n'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}}_n + o(n^{-1/2+\rho})\, a.s.$ for $\rho < 1/2$. By Lemma 7, we also have (4.26). Letting $a_i = \widetilde{\boldsymbol{y}}_i'\boldsymbol{C}_i^{-1}\widetilde{\boldsymbol{y}}_i$, then by Lemma 4, we have $\lim_{n\to\infty}\widetilde{\boldsymbol{y}}_n'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}}_n = 0\, a.s.$ which is (4.31).

The second case, where $\boldsymbol{B}$ is not invertible, i.e. 0 is a root of the characteristic polynomial $\varphi(e)$, so

$$\varphi(e) = e^r(e^s - \beta_1 e^{s-1} - \cdots - \beta_s),\ \beta_s \neq 0,\ \beta_{s+1} = \cdots = \beta_p = 0$$

where r is the multiplicity of the root 0. This case is then broken down into two sub cases: when $r < p$ and when $r = p$.

In the case where $r < p$, we let $g_1 = \cdots = g_r = 0$ and $h_1 = -\beta_1, ..., h_s = -\beta_s$ as in Lemma 5, so we have

$$\boldsymbol{M}_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \vdots \\ \vdots & 0 & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & \cdots & 0 \end{pmatrix},$$

$$\boldsymbol{M}_2 = \begin{pmatrix} 1 & -\beta_1 & \cdots & -\beta_s & 0 & \cdots & 0 \\ 0 & 1 & -\beta_1 & \cdots & -\beta_s & 0 & \vdots \\ \vdots & 0 & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 1 & -\beta_1 & \cdots & \cdots & -\beta_s \end{pmatrix},$$

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_1 \\ \boldsymbol{M}_2 \end{pmatrix}.$$

So

$$
\begin{aligned}
\boldsymbol{MY}_n &= \begin{pmatrix}
1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 & \vdots \\
\vdots & 0 & \cdots & \cdots & \cdots & \cdots & \vdots \\
0 & \cdots & 1 & 0 & \cdots & \cdots & 0 \\
1 & -\beta_1 & \cdots & -\beta_s & 0 & \cdots & 0 \\
0 & 1 & -\beta_1 & \cdots & -\beta_s & 0 & \vdots \\
\vdots & 0 & \cdots & \cdots & \cdots & \cdots & \vdots \\
0 & \cdots & 1 & -\beta_1 & \cdots & \cdots & -\beta_s
\end{pmatrix}
\begin{pmatrix}
y_n \\ y_{n-1} \\ \vdots \\ y_{n-p+1}
\end{pmatrix} \\
&= \begin{pmatrix}
y_n \\ y_{n-1} \\ \vdots \\ y_{n-s+1} \\ y_n - \beta_1 y_{n-1} - \cdots - \beta_s y_{n-s+1} = \varepsilon_n \\ \varepsilon_{n-1} \\ \vdots \\ \varepsilon_{n-r+1}
\end{pmatrix} \\
&= \begin{pmatrix}
\boldsymbol{U}_n \\ \boldsymbol{V}_n
\end{pmatrix}
\end{aligned}
$$

where $\boldsymbol{U}_n = (y_n \cdots y_{n-s+1})'$ and $\boldsymbol{V}_n = (\varepsilon_n \cdots y_{n-r+1})'$. It follows that

$$\boldsymbol{U}_n = \begin{pmatrix} y_n \\ y_{n-1} \\ \vdots \\ y_{n-s+1} \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_s \\ 1 & 0 & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Note that $\boldsymbol{B}_1 = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_s \\ 1 & 0 & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}$ is nonsingular, so

$$\boldsymbol{U}_n' \left(\sum_{i=p}^{n} \boldsymbol{U}_i \boldsymbol{U}_i' \right) \boldsymbol{U}_n \to 0 \, a.s. \tag{4.33}$$

Now define $\boldsymbol{A}_n = \boldsymbol{M}\boldsymbol{C}_n\boldsymbol{M'}$ so we have

$$\begin{aligned} \boldsymbol{A}_n &= \boldsymbol{M}\boldsymbol{C}_n\boldsymbol{M}' \\ &= \boldsymbol{M} \left(\sum_{i=p}^{n} \widetilde{\boldsymbol{y}}_i \widetilde{\boldsymbol{y}}_i{}' \right) \boldsymbol{M}' \\ &= \sum_{i=p}^{n} \boldsymbol{M} \widetilde{\boldsymbol{y}}_i \widetilde{\boldsymbol{y}}_i{}' \boldsymbol{M}' \\ &= \sum_{i=p}^{n} \boldsymbol{M} \widetilde{\boldsymbol{y}}_i (\boldsymbol{M} \widetilde{\boldsymbol{y}}_i)' \\ &= \sum_{i=p}^{n} \begin{pmatrix} \boldsymbol{U}_i \\ \boldsymbol{V}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_i & \boldsymbol{V}_i \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=p}^{n} \boldsymbol{U}_i \boldsymbol{U}_i' & \sum_{i=p}^{n} \boldsymbol{U}_i \boldsymbol{V}_i' \\ \sum_{i=p}^{n} \boldsymbol{V}_i \boldsymbol{U}_i' & \sum_{i=p}^{n} \boldsymbol{V}_i \boldsymbol{V}_i' \end{pmatrix}. \end{aligned} \tag{4.34}$$

Then

$$\begin{aligned}
\widetilde{\boldsymbol{y}}_n'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}}_n &= \widetilde{\boldsymbol{y}}_n'\boldsymbol{M}'\boldsymbol{M}^{-1}\boldsymbol{C}_n^{-1}\boldsymbol{M}'^{-1}\boldsymbol{M}\widetilde{\boldsymbol{y}}_n \\
&= (\boldsymbol{M}\widetilde{\boldsymbol{y}}_n)'\boldsymbol{A}_n^{-1}(\widetilde{\boldsymbol{M}\boldsymbol{y}_n}) \\
&= \begin{pmatrix} \boldsymbol{U}_n' & \boldsymbol{V}_n' \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{A}}_{n,1,1} & \tilde{\boldsymbol{A}}_{n,1,2} \\ \tilde{\boldsymbol{A}}_{n,2,1} & \tilde{\boldsymbol{A}}_{n,2,2} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_n \\ \boldsymbol{V}_n \end{pmatrix} \\
&= \boldsymbol{U}_n'\tilde{\boldsymbol{A}}_{n,1,1}\boldsymbol{U}_n + \boldsymbol{V}_n'\tilde{\boldsymbol{A}}_{n,1,2}\boldsymbol{U}_n + \boldsymbol{U}_n'\tilde{\boldsymbol{A}}_{n,2,1}\boldsymbol{V}_n + \boldsymbol{V}_n'\tilde{\boldsymbol{A}}_{n,2,2}\boldsymbol{V}_n \\
&= \begin{pmatrix} \boldsymbol{U}_n' & 0 \end{pmatrix} \boldsymbol{A}^{-1} \begin{pmatrix} \boldsymbol{U}_n \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & \boldsymbol{V}_n' \end{pmatrix} \boldsymbol{A}^{-1} \begin{pmatrix} 0 \\ \boldsymbol{V}_n \end{pmatrix} + 2\begin{pmatrix} \boldsymbol{U}_n' & 0 \end{pmatrix} \boldsymbol{A}^{-1} \begin{pmatrix} 0 \\ \boldsymbol{V}_n \end{pmatrix} \\
&= a_{uu} + a_{vv} + 2a_{uv}. \qquad (4.35)
\end{aligned}$$

By Lemma 7(i), $\|\boldsymbol{C}_n^{-1/2}\| = O(n^{-1/2})\,a.s.$ so $\|\boldsymbol{C}_n^{-1}\| = O(n^{-1})\,a.s.$, and by Lemma 5, we have that $\boldsymbol{M}$ is nonsingular so

$$\|\boldsymbol{A}_n^{-1}\| = \|(\boldsymbol{M}\boldsymbol{C}_n\boldsymbol{M}')^{-1}\| \leq \|\boldsymbol{M}^{-1}\|\|\boldsymbol{C}_n^{-1}\|\|(\boldsymbol{M}')^{-1}\| = O(n^{-1})\,a.s. \qquad (4.36)$$

where if $\boldsymbol{x}$ is a p-dimensional vector and $\boldsymbol{A}$ is a $p \times p$ matrix $\|\boldsymbol{x}\| = \boldsymbol{x}'\boldsymbol{x}$ and $\|\boldsymbol{A}\| = \sup\limits_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|$, i.e. the spectral norm.

By our assumptions, we have e_i are iid random variables with mean 0 and variance σ^2, a stronger condition than originally used in [18]. So $\boldsymbol{V}_n = (e_n, ..., e_{n-r+1})'$ and then

$$\frac{e_n^2 + \cdots + e_{n-r+1}^2 + e_{n-r}^2 + \cdots + e_1^2}{n} \to \sigma^2 a.s.$$

Then rewriting the above we can see that

$$\frac{e_n^2 + \cdots + e_{n-r+1}^2 + e_{n-r}^2 + \cdots + e_1^2}{n} = \frac{e_n^2 + \cdots + e_{n-r+1}^2}{n} + \frac{e_{n-r}^2 + \cdots + \varepsilon_1^2}{n-r}\left(\frac{n-r}{n}\right)$$

and

$$\frac{e_{n-r}^2 + \cdots + e_1^2}{n-r}\left(\frac{n-r}{n}\right) \to \sigma^2 \, a.s.$$

then deduce that

$$\frac{e_n^2 + \cdots + e_{n-r+1}^2}{n} \to 0$$

as $n \to \infty$. Therefore,

$$\sqrt{e_n^2 + \cdots + e_{n-r+1}^2} = \|\boldsymbol{V}_n\| = o(n^{1/2}). \tag{4.37}$$

Note that $o(a_n)o(b_n) = o(a_n b_n)$ and $o(a_n)O(b_n) = o(a_n b_n)$ [24] so (4.36) and (4.37) result in

$$\begin{aligned} 0 \le a_{vv} &= \begin{pmatrix} 0 & \boldsymbol{V}_n' \end{pmatrix} \boldsymbol{A}^{-1} \begin{pmatrix} 0 \\ \boldsymbol{V}_n \end{pmatrix} \\ &\le \left\| \begin{pmatrix} 0 & \boldsymbol{V}_n' \end{pmatrix} \right\| \|\boldsymbol{A}^{-1}\| \left\| \begin{pmatrix} 0 \\ \boldsymbol{V}_n \end{pmatrix} \right\| \\ &= \|\boldsymbol{A}_n^{-1}\| \|\boldsymbol{V}_n\|^2 \\ &= O(n^{-1}) o(n^{1/2})^2 \, a.s. \\ &= o(1) \, a.s.. \end{aligned} \tag{4.38}$$

Now, with $\boldsymbol{A}_n$ defined as in (4.34), and $\boldsymbol{U}_n$ a $1 \times s$ dimensional vector, Lemma 6(ii) gives us

$$a_{uu} = \begin{pmatrix} \boldsymbol{U}_n' & 0 \end{pmatrix} \boldsymbol{A}^{-1} \begin{pmatrix} \boldsymbol{U}_n \\ 0 \end{pmatrix} \le \boldsymbol{U}_n' \left(\sum_{i=p}^{n} \boldsymbol{U}_i \boldsymbol{U}_i' \right) \boldsymbol{U}_n \left(1 + \|\boldsymbol{A}_n^{-1}\| tr \left(\sum_{i=p}^{n} \boldsymbol{V}_i \boldsymbol{V}_i' \right) \right).$$

Then with (4.33), (4.36), (4.37)

$$a_{uu} \to 0\, a.s. \tag{4.39}$$

Finally, by the Cauchy-Schwartz inequality, we have

$$a_{uv} \leq \sqrt{a_{uu}a_{vv}} \to 0\, a.s.. \tag{4.40}$$

Therefore, putting (4.35), (4.38), (4.39), and (4.40) together we see that

$$\widetilde{\boldsymbol{y}_n}'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}_n} = a_{uu} + a_{vv} + 2a_{u_v} \to 0\, a.s.$$

so (4.31) follows.

Now, consider the case where $\boldsymbol{B}$ is singular and $r = p$. Then we have p zero roots of $\varphi(e)$, which means that $\widetilde{\boldsymbol{y}_n} = \boldsymbol{V}_n = (e_n, ..., e_{n-p+1})$, so

$$\widetilde{\boldsymbol{y}_n}'\boldsymbol{C}_n^{-1}\widetilde{\boldsymbol{y}_n} = \|\boldsymbol{C}_n^{-1}\|\|\boldsymbol{V}_n\|^2 = o(1)$$

follows from (4.38) and (4.31) holds for the largest $\mathscr{M}_\alpha$.

Suppose $\mathscr{M}_\alpha$ is in Category I. If $\alpha \subset \tilde{\alpha}$, then $w_{j\alpha} \leq w_{j\tilde{\alpha}}$ for all j. Choosing $\tilde{\alpha}$ such that $\mathscr{M}_{\tilde{\alpha}}$ is in Category II, we have

$$\lim_{n\to\infty} \max_{j\leq n_\star} w_{j\alpha} \leq \lim_{n\to\infty} \max_{j\leq n_\star} w_{j\tilde{\alpha}} = 0\ w.p.1$$

by Corollary 2 in the Appendix. Therefore (4.31) holds for any $\alpha \in \mathscr{A}$.

CHAPTER 5
SIMULATIONS

We will present some simple experiments, the first with simulated data, and the second with a real-world dataset.

5.1 Simulated data

We used the `arima.sim` function in `R` to create four AR models with lags 2, 3, 4, and 5 with $n = 100$ and 1000. We used the `auto.arima` function to determine lag order which uses AIC, AICc, and BIC. To limit the `auto.arima` function to AR models, we set the integrated and moving average parameters to 0 and the maximum autoregressive parameter to be 7. Then we compared it with our APCV and HVCV functions with options $p = 1, 2, ..., 7$.

Shao does not give any set rules for choosing the testing and training set size, so we use δ as an extra parameter such that

$$n_c = \lfloor n^{\delta} \rfloor$$

and let $\delta = 0.75$ as a default, which tends to do well in general.

The first table shows the results for true models size 2, 3, 4, and 5 with 100 observations, and we saw that all methods did not do well with a small sample size.

In following simulations, we set $n = 1000$ and repeated the experiment 1000 times. Tables entries represent the proportion of times each method chooses $\hat{p} = 1, 2, 3, 4, 5$. We saw that APCV was competitive with AIC, AICc, and BIC, whereas HVCV tended to pick the largest models.

Selected Model Size ($n = 100$)

True Model	AIC	AICc	BIC	APCV	HVCV
$\beta = (0.4, -0.3)'$	2	2	2	7*	7
$\beta = (0.4, -0.3, 0.3)'$	2	2	0**	3	6
$\beta = (0.4, -0.3, 0.3, 0.3)'$	4	4	4	4	7
$\beta = (0.4, -0.3, 0.3, 0.3, 0.2)'$	5***	3***	3	5	7

Table 5.1: Simulation: $n = 100$

* If $K = 6$ then $\hat{p} = 2$

** Indicates white noise

*** Non-zero mean specified

Selected Model Size ($n = 1000$, repeated 1000 times)

True Model	$\hat{p}$	AIC	AICc	BIC	APCV	HVCV
$\beta = \begin{pmatrix} 0.4 \\ -0.3 \end{pmatrix}$	1	0.000	0.000	0.000	0.000	0.000
	2	0.738	0.740	0.975	0.948	0.026
	3	0.189	0.190	0.025	0.045	0.037
	4	0.051	0.049	0.000	0.006	0.051
	5	0.015	0.015	0.000	0.001	0.084
	6	0.006	0.005	0.000	0.000	0.164
	7	0.001	0.001	0.000	0.000	0.638

Table 5.2: Simulation: true model size = 2, $n = 1000$

Selected Model Size ($n = 1000$, repeated 1000 times)

True Model	$\hat{p}$	AIC	AICc	BIC	APCV	HVCV
$\beta = \begin{pmatrix} 0.4 \\ -0.3 \\ 0.3 \end{pmatrix}$	1	0.000	0.000	0.000	0.000	0.000
	2	0.000	0.000	0.000	0.000	0.000
	3	0.773	0.777	0.971	0.938	0.037
	4	0.175	0.172	0.029	0.051	0.053
	5	0.042	0.041	0.000	0.008	0.071
	6	0.007	0.007	0.000	0.003	0.174
	7	0.003	0.003	0.000	0.000	0.665

Table 5.3: Simulation: true model size = 3, $n = 1000$

Selected Model Size ($n = 1000$, repeated 1000 times)

True Model	$\hat{p}$	AIC	AICc	BIC	APCV	HVCV
$\beta = \begin{pmatrix} 0.4 \\ -0.3 \\ 0.3 \\ 0.3 \end{pmatrix}$	1	0.000	0.000	0.001	0.000	0.000
	2	0.000	0.000	0.000	0.000	0.000
	3	0.000	0.000	0.000	0.000	0.000
	4	0.779	0.780	0.976	0.949	0.075
	5	0.166	0.167	0.022	0.038	0.088
	6	0.043	0.042	0.001	0.012	0.176
	7	0.012	0.011	0.000	0.001	0.661

Table 5.4: Simulation: true model size = 4, $n = 1000$

Selected Model Size ($n = 1000$, repeated 1000 times)

True Model	$\hat{p}$	AIC	AICc	BIC	APCV	HVCV
$\beta = \begin{pmatrix} 0.4 \\ -0.3 \\ 0.3 \\ 0.3 \\ 0.2 \end{pmatrix}$	1	0.186	0.186	0.424	0.000	0.000
	2	0.000	0.000	0.000	0.000	0.000
	3	0.000	0.000	0.000	0.000	0.000
	4	0.000	0.000	0.000	0.000	0.000
	5	0.614	0.616	0.570	0.958	0.136
	6	0.148	0.147	0.006	0.033	0.180
	7	0.052	0.051	0.000	0.009	0.684

Table 5.5: Simulation: true model size = 5, $n = 1000$

5.2 Real-world data

Racine used a G7 exchange rates dataset in [22] for Canada (CAD), Germany (DEM), France (FRF), Great Britain (GBP), Italy (ITL), and Japan (JPY). For this experiment we also used a G7 exchange rates dataset for the same countries taken from [13], for the years 1995, 1996, 1997, and 1998. For each country we had approximately 250 daily observations for a total of approximately 1000 observations for each country. Plots of the data are shown in Figure 5.1. We also visualized the PACF plots shown in Figure 5.2, we will use these plots as a naive reference for determining lag order. We can see that using this method, we would hope that the order selection method should choose $p = 1$. In [22], the HVCV method aligned with the PACF plots and selected a AR(1) for each country.

We used a augmented Dickey–Fuller test (ADF) test in R to check for stationarity, which is relatively common for determining whether or not a time series is stationary. The ADF test uses a p-value in order to determine stationarity with the null hypothesis that the data is not stationary. Thus, we want a p-value less than 0.01. We saw that each dataset was not stationary, and used differencing once on each dataset in order to achieve stationarity. In Table 5.6, we have the results of the ADF test before and after differencing once.

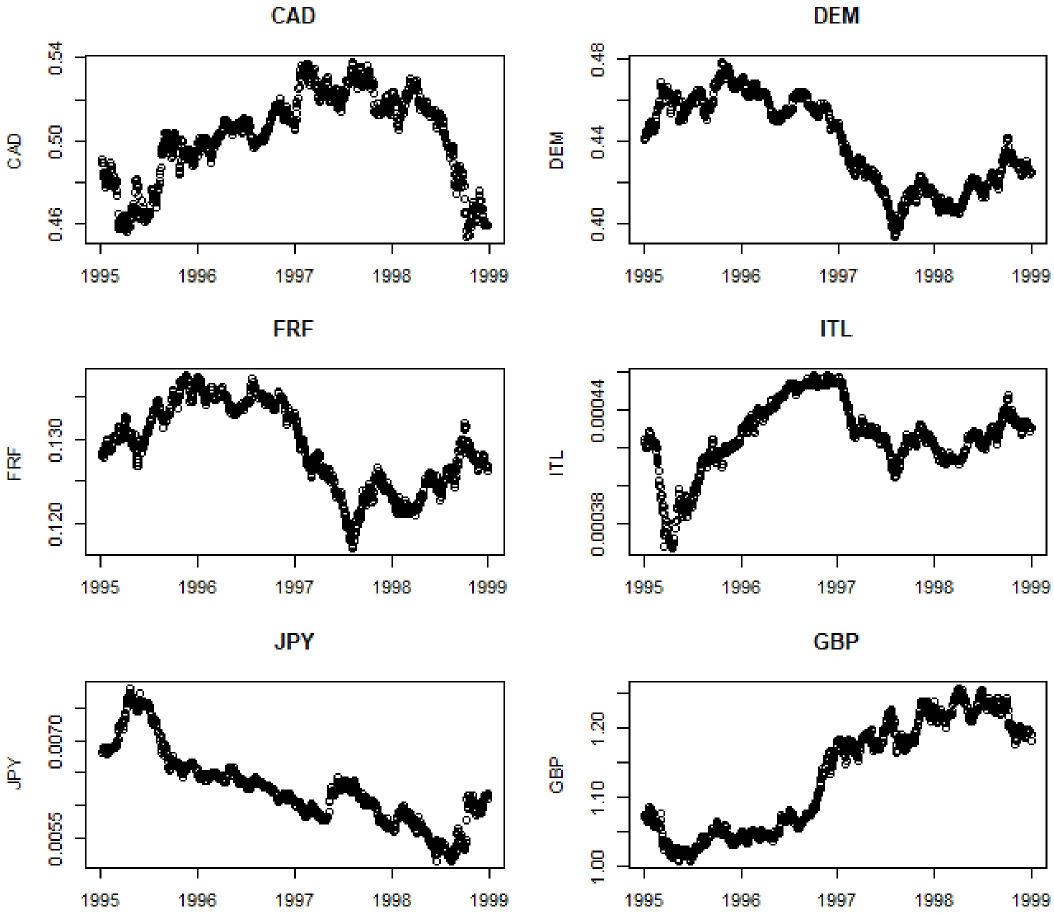

Figure 5.1: Plots of the G7 exchange rates data sets

Country	p-value before	p-value after
CAD	0.99	<0.01
DEM	0.52	<0.01
FRF	0.54	<0.01
GBP	0.66	<0.01
ITL	0.70	<0.01
JPY	0.68	<0.01

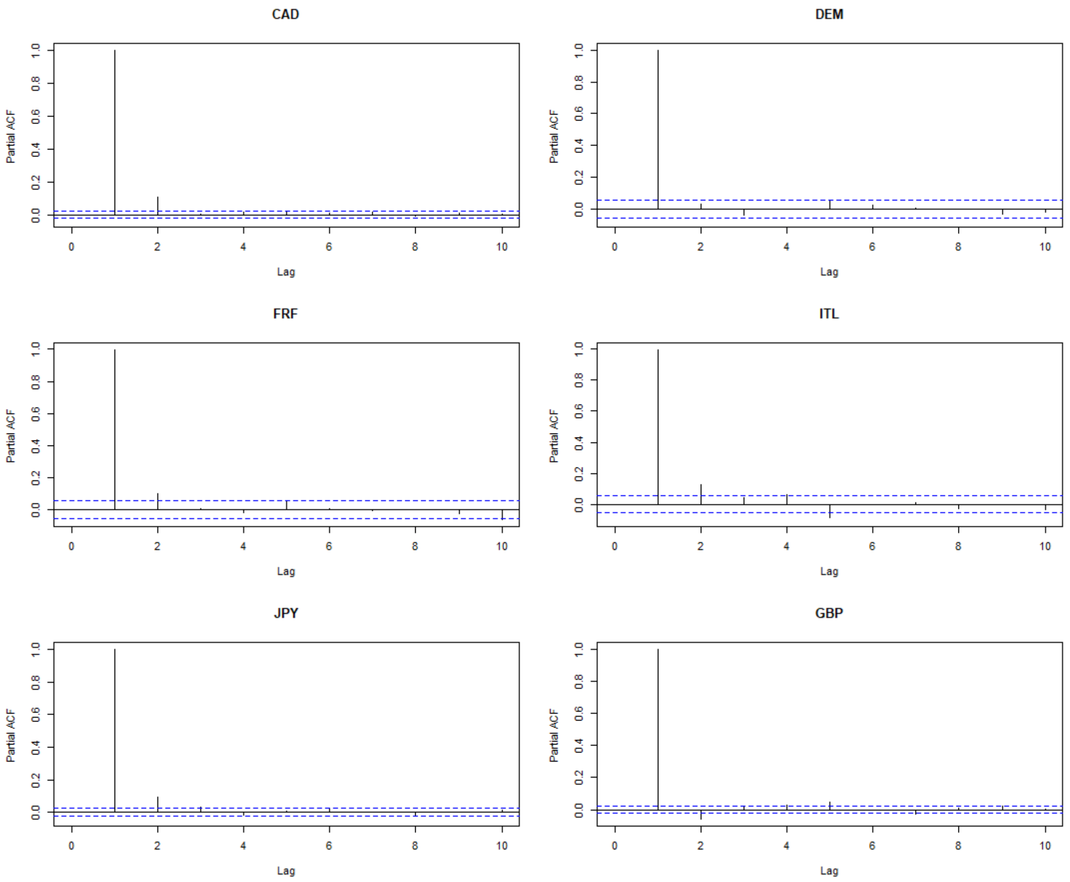

Figure 5.2: Partial autocorrelation plots of the G7 exchange rates data sets

Table 5.6: ADF test p-values before and after differencing

Similar to 5.1, we used the `auto.arima` function in R, with a maximum $p = 7$, $d = 0$, and $q = 0$, to determine the AR lag order selected by AIC, AICc, and BIC. We will compare these results to HVCV and APCV function for time series. We used the same parameters for APCV given in 5.1. We repeated the experiment 100 times, and each model selected the same lag order in every repetition. The selected lag order by each method are shown in Table 5.7. In most cases, AIC, AICc, and

BIC select a model of white noise. We can see that APCV is competitive with AIC, AICc, BIC, and HVCV, and reflects the inference we made from the PACF plots and selects an AR(1) model for each country. We will also note that the results for HVCV in [22] do not match our version of HVCV, and Racine's version selected an AR(1) model for each country.

Selected Model Size ($n \sim 1000$)

Country	AIC	AICc	BIC	APCV	HVCV
CAD	0	0	0	1	7
DEM	0	0	0	1	3
FRF	1	1	0	1	4
GBP	0	0	0	1	7
ITL	1	1	1	1	7
JPY	0	0	0	1	7

Table 5.7: AR lag order for G7 data for selected countries

CHAPTER 6
CONCLUSIONS

It is natural to use APCV for linear models, so the extension to AR(p) models is logical. Using Shao's work as a framework, we showed that causal AR(p) models satisfy the conditions almost surely necessary for APCV to be a consistent estimator for order selection. Thus, we offer an alternative to arbitrary penalized methods such as AIC and BIC.

In simulations and with real-world data, we have also shown that APCV is a competitive estimator with standard methods. For larger models, we saw that APCV outperformed AIC, AICc, BIC, and HVCV. However, we saw that APCV requires a large dataset to work well with simulated or real-world data.

We limited this research to autoregressive models, so future work should expand to non-linear time series models and non-causal AR models that doe not satisfy Shao's conditions that guarantee consistency. Additionally, we did not prove that HVCV is not consistent, simply that such a proof does not follow from Shao, so proof that HVCV is or is not consistient remains an open problem.

REFERENCES

[1] T. W. Anderson and John B. Taylor, *Strong consistency of least squares estimates in dynamic models*, The Annals of Statistics **7** (1979), no. 3, 484–489.

[2] T.W. Anderson, *The statistical analysis of time series*, Wiley Classics Library, Wiley, 1994.

[3] Sylvain Arlot and Alain Celisse, *A survey of cross-validation procedures for model selection*, Statistics Surveys **4** (2010), no. none.

[4] Christoph Bergmeir and José M. Benítez, *On the use of cross-validation for time series predictor evaluation*, Information Sciences **191** (2012), 192–213 (English).

[5] Christoph Bergmeir, Rob Hyndman, and Bonsoo Koo, *A note on the validity of cross-validation for evaluating autoregressive time series prediction*, Computational Statistics & Data Analysis **120** (2018), no. C, 70–83.

[6] P.J. Brockwell and R.A. Davis, *Time series: Theory and methods*, Springer, 1991.

[7] ———, *Introduction to time series and forecasting*, Springer, 1996.

[8] P. Burman, E. Chow, and D. Nolan, *A cross-validatory method for dependent data*, Biometrika **81** (1994), no. 2, 351–358.

[9] Kenneth P. Burnham and David R. Anderson, *Model selection and multimodel inference: A practical information-theoretic approach*, Springer, 2002.

[10] M. J. Campbell and A. M. Walker, *A survey of statistical work on the mackenzie river series of annual canadian lynx trappings for the years 1821-1934 and a new analysis*, Journal of the Royal Statistical Society. Series A (General) **140** (1977), no. 4, 411–431.

[11] Patrick S. Carmack, William R. Schucany, Jeffrey S. Spence, Richard F. Gunst, Qihua Lin, and Robert W. Haley, *Far casting cross-validation*, Journal of Computational and Graphical Statistics **18** (2009), no. 4, 879–893.

[12] Vitor Cerqueira, Luis Torgo, and Igor Mozetič, *Evaluating time series forecasting models: an empirical study on performance estimation methods*, Machine Learning **109** (2020), no. 11, 1997–2028.

[13] International Monetary Fund, *Imf exchange rates.*

[14] G. H. Golub and C. F. Van Loan, *Matrix computations, 3rd ed.*, The Johns Hopkins University Press, 1996.

[15] E. J. Hannan, *The Estimation of the Order of an ARMA Process*, The Annals of Statistics **8** (1980), no. 5, 1071 – 1081.

[16] J. Hopcroft and R. Kannan, *Foundations of data science*, Cambridge University Press, 2018.

[17] Matthew B. Kennel, Reggie Brown, and Henry D. I. Abarbanel, *Determining embedding dimension for phase-space reconstruction using a geometrical construction*, Phys. Rev. A **45** (1992), 3403–3411.

[18] T. L. Lai and C. Z. Wei, *Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters*, Journal of Multivariate Analysis **13** (1983), no. 1, 1–23.

[19] Zhe Liu and Xiangfeng Yang, *Cross validation for uncertain autoregressive model*, Communications in Statistics - Simulation and Computation **0** (2020), no. 0, 1–12.

[20] Ryuei Nishii, *Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression*, The Annals of Statistics **12** (1984), no. 2, 758 – 765.

[21] JEFF RACINE, *Feasible cross-validatory model selection for general stationary processes*, Journal of Applied Econometrics **12** (1997), no. 2, 169–179.

[22] Jeffrey Racine, *Consistent cross-validatory model-selection for dependent data: hv-block cross-validation*, Journal of Econometrics **99** (2000), no. 1, 39–61.

[23] G.A.F. Seber and A.J. Lee, *Linear regression analysis*, Wiley Series in Probability and Statistics, Wiley, 2012.

[24] Cosma Shalizi, *Big o and little o notation.*

[25] Jun Shao, *Linear model selection by cross-validation*, Journal of the American Statistical Association **88** (1993), no. 422, 486–494.

[26] R. Shibata, *Selection of the order of an autoregressive model by akaike's information criterion*, Biometrika **63** (1976), no. 1, 117–126.

[27] Cun-Hui Zhang, *Strong law of large numbers for sums of products*, The Annals of Probability **24** (1996), no. 3, 1589 – 1615.

[28] Wenjie Zheng, *hv-block cross validation is not a bibd: a note on the paper by jeff racine (2000)*, 2019.

APPENDIX I
Proofs

We present the following proofs:

- I.0.1 If $E(y_t) = 0 \, \forall t$, e_t are iid white noise with mean 0 and variance σ^2, $E(e_t^4) < \infty$, and $\boldsymbol{y}_t$ is causal then $E[y_t^4] < \infty$.
- I.0.2 If $\boldsymbol{A}$ is an $n \times K$ matrix with $n \geq K$ and $\boldsymbol{B}$ is an $m \times K$ matrix with $m \geq K$, then all eigenvalues of $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{A}'$ are in the interval $[0, 1]$
- I.0.3 $\sum_i w_{i\alpha} r_{i\alpha}^2 = d_\alpha \sigma^2 + o_p(1)$ from [25]
- I.0.4 $n_v^{-1} \|\mathbf{y}_s - \hat{\mathbf{y}}_{\alpha,s^c}\|^2 = n_v^{-1} \|(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}(\mathbf{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\|^2$ from [25]
- I.0.5 BICV is consistent from [25]
- I.0.6 APCV is consistent from [25]

I.0.1 Proof of (4.2)

Since $\{y_t\}$ is a causal process, we can write $y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ where $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Then

$$\begin{aligned} y_t^4 &= \left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \right)^4 \\ &= \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \sum_{j_3=0}^{\infty} \sum_{j_4=0}^{\infty} \psi_{j_1}\psi_{j_2}\psi_{j_3}\psi_{j_4}\varepsilon_{t-j_1}\varepsilon_{t-j_2}\varepsilon_{t-j_3}\varepsilon_{t-j_4} \\ &= \sum_{M_1=-\infty}^{t} \sum_{M_2=-\infty}^{t} \sum_{M_3=-\infty}^{t} \sum_{M_4=-\infty}^{t} \psi_{t-M_1}\psi_{t-M_2}\psi_{t-M_3}\psi_{t-M_4}\varepsilon_{M_1}\varepsilon_{M_2}\varepsilon_{M_3}\varepsilon_{M_4}. \end{aligned}$$

So, we obtain

$$\begin{aligned} E[y_t^4] &= \sum_{M_1=-\infty}^{t} \sum_{M_2=-\infty}^{t} \sum_{M_3=-\infty}^{t} \sum_{M_4=-\infty}^{t} \psi_{t-M_1}\psi_{t-M_2}\psi_{t-M_3}\psi_{t-M_4} E[\varepsilon_{M_1}\varepsilon_{M_2}\varepsilon_{M_3}\varepsilon_{M_4}] \\ &= \sum_{M=-\infty}^{t} \psi_{t-M}^4 \mu_4' + 3 \sum_{M_1=-\infty}^{t} \sum_{\substack{M_2=-\infty \\ M_2 \neq M_1}}^{t} \psi_{t-M_1}^2 \psi_{t-M_2}^2 \sigma^4 \\ &= \sum_{M=-\infty}^{t} \psi_{t-M}^4 (\mu_4' - 3\sigma^4) + 3 \left(\sum_{M=-\infty}^{t} \psi_{t-M}^2 \right)^2 \sigma^4 \end{aligned}$$

where $\mu_4' = E[\varepsilon_t^4]$ and $\sigma^2 = E[\varepsilon_t^2]$ for all t.

Since $\sum |\psi_j|$ converges, then by the limit comparison test $\sum \psi_j^2$ and $\sum \psi_j^4$ converges, and it follows that $E[y_t^4] < \infty$.

I.0.2 Singular value decomposition and generalized singular value decomposition

Theorem 3 *(Singular Value Decomposition) [14]: If $\boldsymbol{A}$ is an $n \times K$ matrix with $n \geq K$, then there exists orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ such that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{V}'$*

where $\boldsymbol{\Sigma}_A$ is a diagonal $n \times K$ matrix with diagonal elements $\sigma_1 \geq \cdots \geq \sigma_K \geq 0$. Moreover, the eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$ are $\sigma_1^2, \ldots, \sigma_K^2$.

Theorem 4 *(Generalized Singular Value Decomposition) [14]: If $\boldsymbol{A}$ is an $n \times K$ matrix with $n \geq K$ and $\boldsymbol{B}$ is an $m \times K$ matrix with $m \geq K$, then there exists an orthogonal $n \times n$ matrix $\boldsymbol{U}$, an orthogonal $m \times m$ matrix $\boldsymbol{V}$, and an invertible $K \times K$ matrix $\boldsymbol{Z}$ such that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{Z}^{-1}$ and $\boldsymbol{B} = \boldsymbol{\Sigma}_B\boldsymbol{Z}^{-1}$ where $\boldsymbol{\Sigma}_A$ is a diagonal $n \times K$ matrix with diagonal elements $\sigma_{A,1}, \ldots, \sigma_{A,K}$ and $\boldsymbol{\Sigma}_B$ is a diagonal $m \times K$ matrix with diagonal elements $\sigma_{B,1}, \ldots, \sigma_{B,K}$.*

Lemma 10 *If $\boldsymbol{A}$ is an $n \times K$ matrix with $n \geq K$ and $\boldsymbol{B}$ is an $m \times K$ matrix with $m \geq K$, then all eigenvalues of $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{A}'$ are in the interval* $[0, 1]$.

Proof:

By Theorem 2, there exists an orthogonal $n \times n$ matrix $\boldsymbol{U}$, an orthogonal $m \times m$ matrix $\boldsymbol{V}$, an invertible $K \times K$ matrix $\boldsymbol{Z}$, and diagonal $n \times K$ matrix $\boldsymbol{\Sigma}_A$ with diagonal elements $\sigma_{A,1}, \ldots, \sigma_{A,K}$, and a diagonal $m \times K$ matrix $\boldsymbol{\Sigma}_B$ with diagonal elements $\sigma_{B,1}, \ldots, \sigma_{B,K}$ such that

$$\begin{aligned}
\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{A}' &= \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{Z}^{-1}\left((\boldsymbol{Z}^{-1})'\boldsymbol{\Sigma}_A'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{Z}^{-1} + (\boldsymbol{Z}^{-1})'\boldsymbol{\Sigma}_B'\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Sigma}_B\boldsymbol{Z}^{-1}\right)^{-1} \\
&\quad (\boldsymbol{Z}^{-1})'\boldsymbol{\Sigma}_A'\boldsymbol{U}' \\
&= \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{Z}^{-1}\left((\boldsymbol{Z}^{-1})'\boldsymbol{\Sigma}_A'\boldsymbol{\Sigma}_A\boldsymbol{Z}^{-1} + (\boldsymbol{Z}^{-1})'\boldsymbol{\Sigma}_B'\boldsymbol{\Sigma}_B\boldsymbol{Z}^{-1}\right)^{-1}(\boldsymbol{Z}^{-1})'\boldsymbol{\Sigma}_A'\boldsymbol{U}' \\
&= \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{Z}^{-1}\left((\boldsymbol{Z}^{-1})'\left(\boldsymbol{\Sigma}_A'\boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_B'\boldsymbol{\Sigma}_B\right)\boldsymbol{Z}^{-1}\right)^{-1}(\boldsymbol{Z}^{-1})'\boldsymbol{\Sigma}_A'\boldsymbol{U}' \\
&= \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{Z}^{-1}\boldsymbol{Z}\left(\boldsymbol{\Sigma}_A'\boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_B'\boldsymbol{\Sigma}_B\right)^{-1}\boldsymbol{Z}'(\boldsymbol{Z}')^{-1}\boldsymbol{\Sigma}_A'\boldsymbol{U}' \qquad \text{(I.1)} \\
&= \boldsymbol{U}\left(\boldsymbol{\Sigma}_A\left(\boldsymbol{\Sigma}_A'\boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_B'\boldsymbol{\Sigma}_B\right)^{-1}\boldsymbol{\Sigma}_A'\right)\boldsymbol{U}'.
\end{aligned}$$

Note that $\boldsymbol{\Sigma}_A \left(\boldsymbol{\Sigma}_A'\boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_B'\boldsymbol{\Sigma}_B\right)^{-1} \boldsymbol{\Sigma}_A'$ is an $n \times n$ square diagonal matrix since it is a product of diagonal matrices, and its diagonal elements are

$$\frac{\sigma_{A,j}^2}{\sigma_{A,j}^2 + \sigma_{B,j}^2} \tag{I.2}$$

for $j = 1, \ldots, K$ and $n - K$ zeros. Since (I.1) is the singular value decomposition of $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{A}'$, Theorem 1 implies that (I.2) are its eigenvalues, and all of these eigenvalues are in the interval $[0, 1]$.

□

I.0.3 Proof that $\sum_i w_{i\alpha} r_{i\alpha}^2 = d_\alpha \sigma^2 + o_p(1)$

Here we show that $\sum_i w_{i\alpha} r_{i\alpha}^2 = d_\alpha \sigma^2 + o_p(1)$; that is, $\sum_i w_{i\alpha} r_{i\alpha}^2 \xrightarrow{p} d_\alpha \sigma^2$. Where $w_{i\alpha}$ is the ith diagonal element of the projection matrix $\boldsymbol{P}_\alpha$, so we may define $diag(\boldsymbol{P}_\alpha) = \boldsymbol{W}_\alpha$ such that $w_{i\alpha}$ is the ith diagonal element of $\boldsymbol{W}_\alpha$. Also let $\boldsymbol{r}_\alpha$ be the n-dimensional vector with ith component $r_i = y_i - \boldsymbol{x}_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha$, where $\boldsymbol{x}_{i\alpha}$ is the ith row of $\boldsymbol{X}_\alpha$. So $\boldsymbol{r}_\alpha = \boldsymbol{y} - \boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha = (\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{y}$. First, we will show that $E[\sum_i w_{i\alpha} r_{i\alpha}^2] \to d_\alpha \sigma^2$ as $n \to \infty$, then that $var[\sum_i w_{i\alpha} r_{i\alpha}^2] \to 0$ as $n \to \infty$, and the result will follow.

Now we compute

$$\begin{aligned} E\left[\sum_i w_{i\alpha} r_{i\alpha}^2\right] &= E[\boldsymbol{r}_\alpha'\boldsymbol{W}_\alpha\boldsymbol{r}_\alpha] = E[tr(\boldsymbol{r}_\alpha'\boldsymbol{W}_\alpha\boldsymbol{r}_\alpha)] \\ &= E[tr(\boldsymbol{W}_\alpha\boldsymbol{r}_\alpha\boldsymbol{r}_\alpha')] = tr(E[\boldsymbol{W}_\alpha\boldsymbol{r}_\alpha\boldsymbol{r}_\alpha']) = tr(\boldsymbol{W}_\alpha E[\boldsymbol{r}_\alpha\boldsymbol{r}_\alpha']). \end{aligned}$$

Since $E[\boldsymbol{r}_\alpha\boldsymbol{r}_\alpha'] = var[\boldsymbol{r}_\alpha] = var[(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{y}] = (\boldsymbol{I}_n - \boldsymbol{P}_\alpha)(\sigma^2\boldsymbol{I}_n)(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)' = \sigma^2(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)$, it follows that

$$E\left[\sum_i w_{i\alpha} r_{i\alpha}^2\right] = tr(\boldsymbol{W}_\alpha\sigma^2(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)) = \sigma^2 tr(\boldsymbol{W}_\alpha - \boldsymbol{W}_\alpha\boldsymbol{P}_\alpha) = \sigma^2(tr(\boldsymbol{W}_\alpha) - tr(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha)).$$

The ith diagonal element of $\boldsymbol{W}_\alpha \boldsymbol{P}_\alpha$ is $w_{i\alpha} w_{i\alpha} = w_{i\alpha}^2$ so $tr(\boldsymbol{W}_\alpha \boldsymbol{P}_\alpha) = \sum_i w_{i\alpha}^2$ and we have

$$E\left[\sum_i w_{i\alpha} r_{i\alpha}^2\right] = \sigma^2 \left(\sum_i w_{i\alpha} - \sum_i w_{i\alpha}^2\right).$$

Since $\sum_i w_{i\alpha} = tr(\boldsymbol{W}_\alpha) = tr(\boldsymbol{P}_\alpha) = tr(\boldsymbol{X}_\alpha(\boldsymbol{X}'_\alpha \boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_\alpha) = tr((\boldsymbol{X}'_\alpha \boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_\alpha \boldsymbol{X}_\alpha) = tr(\boldsymbol{I}_{d_\alpha}) = d_\alpha$, it follows that

$$E\left[\sum_i w_{i\alpha} r_{i\alpha}^2\right] = \sigma^2 (d_\alpha - \sum_i w_{i\alpha}^2).$$

Next we show that $\sum_{i=1}^n w_{i\alpha}^2 \to 0$. We have

$$\lim_{n\to\infty} \sum_{i=1}^n w_{i\alpha}^2 \le \lim_{n\to\infty} \max_i w_{i\alpha} \sum_{i=1}^n w_{i\alpha} = \lim_{n\to\infty} \max_i w_{i\alpha} d_\alpha = d_\alpha \lim_{n\to\infty} \max_i w_{i\alpha}$$

equals 0 by (3.4) in [25]. It follows that $E[\sum_i w_{i\alpha} r_{i\alpha}^2] \to \sigma^2 d_\alpha$.

Now we will show that $var[\sum_i w_{i\alpha} r_{i\alpha}^2] \to 0$. We write it as weighted sum of squared errors to use the fact that the errors are independent. Since $\boldsymbol{r}_\alpha = (\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{P}_\alpha)(\boldsymbol{X}_\alpha \boldsymbol{\beta} + \boldsymbol{e}_\alpha)$, it follows that

$$\begin{aligned}
\sum_i w_{i\alpha} r_{i\alpha}^2 &= \boldsymbol{r}'_\alpha \boldsymbol{W}_\alpha \boldsymbol{r}_\alpha = (\boldsymbol{X}_\alpha \boldsymbol{\beta} + \boldsymbol{e}_\alpha)'(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)(\boldsymbol{X}_\alpha \boldsymbol{\beta} + \boldsymbol{e}_\alpha) \\
&= \boldsymbol{\beta}'\boldsymbol{X}'_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{X}_\alpha \boldsymbol{\beta} + 2\boldsymbol{e}'_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{X}_\alpha \boldsymbol{\beta} + \\
&\quad \boldsymbol{e}'_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e}_\alpha \\
&= \boldsymbol{e}'_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e}_\alpha
\end{aligned}$$

since $(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{X}_\alpha = \boldsymbol{X}_\alpha - \boldsymbol{P}_\alpha \boldsymbol{X}_\alpha = \boldsymbol{O}$.

Let $\boldsymbol{A}_{n,\alpha} = (\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)$. Then

$$tr(\boldsymbol{A}_{n,\alpha}^2) = tr((\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha))$$

$$= tr((\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)) = tr(\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)^2)$$

$$= tr(\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)) = tr(\boldsymbol{W}_\alpha^2 - \boldsymbol{W}_\alpha^2\boldsymbol{P}_\alpha - \boldsymbol{W}_\alpha\boldsymbol{P}_\alpha\boldsymbol{W}_\alpha + \boldsymbol{W}_\alpha\boldsymbol{P}_\alpha\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha)$$

$$= tr(\boldsymbol{W}_\alpha^2) - tr(\boldsymbol{W}_\alpha^2\boldsymbol{P}_\alpha) - tr(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha\boldsymbol{W}_\alpha) + tr(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha)$$

$$= tr(\boldsymbol{W}_\alpha^2) - tr(\boldsymbol{W}_\alpha^2\boldsymbol{P}_\alpha) - tr(\boldsymbol{W}_\alpha^2\boldsymbol{P}_\alpha) + tr((\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha)^2).$$

Since $\boldsymbol{W}_\alpha$ is diagonal, $\boldsymbol{W}_\alpha^2$ is also diagonal with diagonal elements $w_{1\alpha}^2, \ldots, w_{n\alpha}^2$ and we already showed that $\sum_i w_{i\alpha}^2 \to 0$.

The ith diagonal element of $\boldsymbol{W}_\alpha^2\boldsymbol{P}_\alpha$ is $w_{i\alpha}^2 w_{i\alpha} = w_{i\alpha}^3$ so $tr(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha) = \sum_i w_{i\alpha}^3$. Then we have

$$\lim_{n\to\infty}\sum_{i=1}^n w_{i\alpha}^3 \leq \lim_{n\to\infty}\max_i w_{i\alpha}\sum_{i=1}^n w_{i\alpha}^2 = 0.$$

Finally, consider $(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha)^2$. Let $\lambda_{n,1}, \ldots, \lambda_{n,n}$ denote the eigenvalues of $\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha$. Then the eigenvalues of $(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha)^2$ are $\lambda_{n,1}^2, \ldots, \lambda_{n,n}^2$. Since $tr(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha) = \sum_{i=1}^n \lambda_{n,i} \to 0$ and $\lambda_{n,i} \geq 0$, we see that $\max_i \lambda_{n,i} \to 0$. Thus, for sufficiently large n, $\lambda_{n,i} \in [0, 1]$ for all i so that $\lambda_{n,i}^2 \leq \lambda_{n,i}$ for all i. Thus $tr((\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha)^2) = \sum_{i=1}^n \lambda_{n,i}^2 \leq \sum_{i=1}^n \lambda_{n,i} = tr(\boldsymbol{W}_\alpha\boldsymbol{P}_\alpha) \to 0$.

So, we have

$$tr(\boldsymbol{A}_{n,\alpha}^2) \to 0 - 0 - 0 + 0 = 0.$$

Letting a_{ij} denote the element in the ith row and jth column of $\boldsymbol{A}_{n,\alpha}$, we also have $tr(\boldsymbol{A}_{n,\alpha}^2) = \sum_i\sum_j a_{ij}^2 = \|\boldsymbol{A}_{n,\alpha}\|_F^2$ (where $\|\boldsymbol{A}_{n,\alpha}\|_F$ is the *Frobenius norm*). Clearly, $0 \leq \sum_i a_{ii}^2 \leq \|\boldsymbol{A}_{n,\alpha}\|_F^2$ so $\sum_i a_{ii}^2 \to 0$.

Using the formula for the variance of a quadratic form in Theorem 1.6 of [23] with $\theta = 0$, we have

$$\begin{aligned} var(\boldsymbol{e}'_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e}_\alpha) &= var(\boldsymbol{e}'_\alpha\boldsymbol{A}_{n,\alpha}\boldsymbol{e}_\alpha) \\ &= (\mu_4 - 3(\sigma^2)^2)\sum_i a_{ii}^2 + 2(\sigma^2)^2 tr(A_{n,\alpha}^2) \to 0. \end{aligned}$$

So then

$$\boldsymbol{e}'_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{W}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e}_\alpha = \sum_i w_{i\alpha} r_{i\alpha}^2 \xrightarrow{p} d_\alpha\sigma^2$$

since $E[\sum_i w_{i\alpha} r_{i\alpha}^2] \to d_\alpha\sigma^2$ and $var[\sum_i w_{i\alpha} r_{i\alpha}^2] \to 0$ as $n \to \infty$.

I.0.4 Proof of 3.1 from Shao

From Shao's paper, we have the average squared prediction error is

$$n_v^{-1}\|\mathbf{y}_s - \hat{\mathbf{y}}_{\alpha,s^c}\|^2 = n_v^{-1}\|(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}(\mathbf{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\|^2 \tag{I.3}$$

where the data set of size n is split into two parts, a subset s which is a validation set of size n_v and s^c which is a training set of size $n_c = n - n_v$ used to fit the model $\mathscr{M}_\alpha$. Furthermore, we define

$$\begin{aligned} \hat{\mathbf{y}}_{\alpha,s^c} &= \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_{\alpha,s^c} = \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}_{\alpha,s^c}\mathbf{y}_{s^c} \\ \boldsymbol{Q}_{\alpha,s} &= \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_{\alpha,s} \\ \hat{\boldsymbol{\beta}}_\alpha &= (\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha)^{-1}\boldsymbol{X}'_\alpha\mathbf{y} \end{aligned} \tag{I.4}$$

Using the Woodbury matrix identity,

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$$

if we set

$$\begin{aligned}
\boldsymbol{A} &= \mathbf{I}_{n_v} \\
\boldsymbol{U} &= -\boldsymbol{X}_{\alpha,s} \\
\boldsymbol{C} &= (\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1} \\
\boldsymbol{V} &= \boldsymbol{X}'_{\alpha,s}
\end{aligned}$$

then

$$(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1} = \mathbf{I}_{n_v} + \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} - \boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\boldsymbol{X}'_{\alpha,s} \tag{I.5}$$

With equations (I.4) and (I.5), this implies that the right hand side of (I.3) inside the norm is

$$(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}(\mathbf{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_{\alpha}) \tag{I.6}$$

$$\begin{aligned}
&= (\mathbf{I}_{n_v} + \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} - \boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\boldsymbol{X}'_{\alpha,s})(\mathbf{y}_s - \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}) \\
&= \mathbf{y}_s - \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y} + \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} - \boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\boldsymbol{X}'_{\alpha,s}\mathbf{y}_s \\
&- \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} - \boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y} && \text{(I.7)} \\
&= \mathbf{y}_s - \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y} + \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha,s}\mathbf{y}_s \\
&- \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y} + \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y} && \text{(I.8)} \\
&= \mathbf{y}_s + \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha,s}\mathbf{y}_s - \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y} \\
&= \mathbf{y}_s - \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}[\boldsymbol{X}'_{\alpha}\mathbf{y} - \boldsymbol{X}'_{\alpha,s}\mathbf{y}_s] \\
&= \mathbf{y}_s - \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha,s^c}\mathbf{y}_{s^c} \\
&= \mathbf{y}_s - \hat{\mathbf{y}}_{\alpha,s^c}.
\end{aligned}$$

Thus we have shown (I.3). To see how we get from (I.7) to (I.8) observe

$$\begin{aligned}
&-\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha}-\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}\\
&=-\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha}-\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}\\
&=[-\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha}+\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c}](\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}\\
&=-\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}+\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}\\
&=-\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}+\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}'_{\alpha}\mathbf{y}.
\end{aligned}$$

Furthermore, we note that for $\mathbf{M}_s$ of size $n_v \times n$ obtained from $\boldsymbol{I}_n$ using the rows indexed by s and $\mathbf{M}_{s^c}$ of size $n_c \times n$ obtained from $\boldsymbol{I}_n$ using the rows indexed by s^c we have

$$\begin{aligned}
\boldsymbol{X}_{\alpha,s} &= \mathbf{M}_s\boldsymbol{X}_{\alpha}\\
\boldsymbol{X}_{\alpha,s^c} &= \mathbf{M}_{s^c}\boldsymbol{X}_{\alpha}\\
\mathbf{M}'_{s^c}\mathbf{M}_{s^c} &= \mathbf{I}_n-\mathbf{M}'_s\mathbf{M}_s\\
\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} &= \boldsymbol{X}'_{\alpha}\mathbf{I}_n\boldsymbol{X}_{\alpha}\\
\boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c} &= \boldsymbol{X}'_{\alpha}(\mathbf{M}'_{s^c}\mathbf{M}_{s^c})\boldsymbol{X}_{\alpha}\\
\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s} &= \boldsymbol{X}'_{\alpha}(\mathbf{M}'_s\mathbf{M}_s)\boldsymbol{X}_{\alpha}\\
\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha}-\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s} &= \boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha}-\boldsymbol{X}'_{\alpha}(\mathbf{M}'_s\mathbf{M}_s)\boldsymbol{X}_{\alpha}\\
&= \boldsymbol{X}'_{\alpha}(\mathbf{I}_n-\mathbf{M}'_s\mathbf{M}_s)\boldsymbol{X}_{\alpha}\\
&= \boldsymbol{X}'_{\alpha}(\mathbf{M}'_{s^c}\mathbf{M}_{s^c})\boldsymbol{X}_{\alpha}\\
&= \boldsymbol{X}'_{\alpha,s^c}\boldsymbol{X}_{\alpha,s^c}\\
\implies \boldsymbol{X}'_{\alpha}\mathbf{y}-\boldsymbol{X}'_{\alpha,s}\mathbf{y}_s &= \boldsymbol{X}'_{\alpha,s^c}\mathbf{y}_{s^c}
\end{aligned}$$

I.0.5 Proof of Theorem 1 from Shao

Suppose that the following conditions hold, where $\boldsymbol{X}$ is the largest design matrix,

1. $\liminf_{n\to\infty} \Delta_{\alpha,n} > 0$ for $\mathscr{M}_\alpha$ in Category I
2. $\boldsymbol{X}'\boldsymbol{X} = O(n)$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = O\left(\frac{1}{n}\right)$
3. $\lim_{n\to\infty} \max_{i\leq n} w_{i\alpha} = 0\,\forall\alpha \in \mathscr{A}$
4. $\lim_{n\to\infty} \max_{s\in\mathscr{B}} \left\| \frac{1}{n_v} \sum_{i\in s} \mathbf{x}_i\mathbf{x}_i' - \frac{1}{n_c} \sum_{i\in s^c} \mathbf{x}_i\mathbf{x}_i \right\| = 0$

Suppose also that n_v is selected so that

$$n_v/n \to 1 \text{ and } n_c = n - n_v \to \infty \tag{I.9}$$

Then we have the following conclusions:

(a) If $\mathscr{M}_\alpha$ is in Category I, then there exists $R_n \geq 0$ such that

$$\hat{\Gamma}_{\alpha,n}^{BICV} = n^{-1}\mathbf{e}'\mathbf{e} + \Delta_{\alpha,n} + o_p(1) + R_n. \tag{I.10}$$

(b) If $\mathscr{M}_\alpha$ is in Category II, then

$$\hat{\Gamma}_{\alpha,n}^{BICV} = n^{-1}\mathbf{e}'\mathbf{e} + n_c^{-1}d_\alpha\sigma^2 + o_p(n_c^{-1}). \tag{I.11}$$

(c) Consequently,

$$\lim_{n\to\infty} P(\text{the selected model is } \mathscr{M}_*) = 1. \tag{I.12}$$

First we show (I.10):

By definition,

$$\begin{aligned}
\hat{\Gamma}^{BICV}_{\alpha,n} &= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \|\boldsymbol{y}_s - \hat{\boldsymbol{y}}_{\alpha,s^c}\|^2 \\
&= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \|(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\|^2 \\
&\geq \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \|\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha\|^2. \qquad \text{(I.13)}
\end{aligned}$$

Since $\boldsymbol{Q}_{\alpha,s}$ has the form $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A}+\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{A}'$, all of its eigenvalues are in the interval $[0,1]$. If λ is an eigenvalue of $\boldsymbol{Q}_{\alpha,s}$, then $1-\lambda$ is an eigenvalue of $\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}$ so all of the eigenvalues of $\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}$, and consequently $(\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^2$ are in $[0,1]$. So

$$\begin{aligned}
\|\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}\| &= \sup_{\|\boldsymbol{x}\|\neq 0} \frac{\|(\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \\
&\leq 1
\end{aligned}$$

and thus, for all $\boldsymbol{x}$

$$\|(\boldsymbol{I} - \boldsymbol{Q}_{\alpha,s})\boldsymbol{x}\| \leq \|\boldsymbol{x}\|.$$

Letting $\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)$ we obtain (I.13).

For each $s \in \mathscr{B}$

$$\begin{aligned}
\|\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha\|^2 &= (\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)'(\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha) \\
&= \boldsymbol{y}_s'\boldsymbol{y}_s - \boldsymbol{y}_s'\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha - (\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)'\boldsymbol{y}_s + (\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)'(\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)
\end{aligned}$$

Looking at $\boldsymbol{y}_s'\boldsymbol{y}_s$.

$$\begin{aligned}\sum_{s\in\mathscr{B}}\boldsymbol{y}_s'\boldsymbol{y}_s &= \sum_{s\in\mathscr{B}}[y_{s_1}^2+y_{s_2}^2+\cdots+y_{s_{n_v}}^2]\\ &= \frac{n_v b}{n}[y_1^2+\cdots+y_n^2]\\ &= \frac{n_v b}{n}\boldsymbol{y}'\boldsymbol{y}\end{aligned}$$

where $\frac{n_v b}{n}$ is the number of times each observation appears in $\mathscr{B}$. Similarly,

$$\begin{aligned}\sum_{s\in\mathscr{B}}\boldsymbol{y}_s'\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha &= \frac{n_v b}{n}\boldsymbol{y}'\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\\ \sum_{s\in\mathscr{B}}(\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\boldsymbol{y}_s &= \frac{n_v b}{n}(\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha)\boldsymbol{y}\\ \sum_{s\in\mathscr{B}}\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)'(\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha) &= \frac{n_v b}{n}(\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha)'(\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha).\end{aligned}$$

The above implies that

$$\begin{aligned}\frac{1}{n_v b}\sum_{s\in\mathscr{B}}\|\boldsymbol{y}_s-\boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha\|^2 &= \left(\frac{1}{n_v b}\right)\left(\frac{n_v b}{n}\right)[\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{y}'\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha-(\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha)'\boldsymbol{y}+(\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha)'(\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha)]\\ &= \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2.\end{aligned}$$

So then

$$\begin{aligned}\hat{\Gamma}_{\alpha,n}^{BICV} &= \frac{1}{n_v b}\sum_{s\in\mathscr{B}}\|\boldsymbol{y}_s-\hat{\boldsymbol{y}}_{\alpha,s^c}\|^2\\ &= \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2+R_n\\ &= \frac{1}{n}\boldsymbol{e}'\boldsymbol{e}+\Delta_{\alpha,n}+\left[-\frac{1}{n}\boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e}+\frac{2}{n}\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta}\right]+R_n.\end{aligned}$$

Since $E[\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta}]=0$ and $var(\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta})=E[(\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta})^2]=\sigma^2\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}_n-\boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta}$ it follows that

$$\frac{2}{n}\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{P}_\alpha)\boldsymbol{X}\boldsymbol{\beta}=o_p(1).$$

Also, since $E[\boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e}] = d_\alpha\sigma^2$, then for every $\varepsilon > 0$, we have

$$P\left(\left|\frac{1}{n}\boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e} - 0\right| > \varepsilon\right) = P(\boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e} > n\varepsilon) \leq \frac{E[\boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e}]}{n\varepsilon} \to 0$$

as $n \to \infty$. So, $\frac{1}{n}\boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e} \to_p 0$ (that is, $\frac{1}{n}\boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e} = o_P(1)$).

So, it follows that

$$\begin{aligned}\hat{\Gamma}^{BICV}_{\alpha,n} &= \frac{1}{n}\boldsymbol{e}'\boldsymbol{e} + \Delta_{\alpha,n} + [o_P(1) + o_P(1)] + R_n \\ &= \frac{1}{n}\boldsymbol{e}'\boldsymbol{e} + \Delta_{\alpha,n} + R_n + o_P(1).\end{aligned}$$

Where we let

$$R_n = \hat{\Gamma}^{BICV}_{\alpha,n} - \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2.$$

Thus we have shown (I.10)

Now we will show (I.11) and (I.12). By condition (4), for $s \in \mathscr{B}$

$$\begin{aligned}\frac{1}{n}\boldsymbol{X}_\alpha'\boldsymbol{X}_\alpha - \frac{1}{n_v}\boldsymbol{X}_{\alpha,s}'\boldsymbol{X}_{\alpha,s} &= \frac{1}{n}(\boldsymbol{X}_{\alpha,s}'\boldsymbol{X}_{\alpha,s} + \boldsymbol{X}_{\alpha,s^c}'\boldsymbol{X}_{\alpha,s^c}) - \frac{1}{n_v}\boldsymbol{X}_{\alpha,s}'\boldsymbol{X}_{\alpha,s} \\ &= \frac{1}{n}\boldsymbol{X}_{\alpha,s^c}'\boldsymbol{X}_{\alpha,s^c} - \left(\frac{1}{n} - \frac{1}{n_v}\right)\boldsymbol{X}_{\alpha,s}'\boldsymbol{X}_{\alpha,s} \\ &= \frac{n_c}{n_c n}\boldsymbol{X}_{\alpha,s^c}'\boldsymbol{X}_{\alpha,s^c} + \left(\frac{n_c n_v - n_c n}{n_c n n_v}\right)\boldsymbol{X}_{\alpha,s}'\boldsymbol{X}_{\alpha,s} \\ &= \frac{n_c}{n_c n}\boldsymbol{X}_{\alpha,s^c}'\boldsymbol{X}_{\alpha,s^c} + \left(\frac{n_c(-n_c)}{n_c n n_v}\right)\boldsymbol{X}_{\alpha,s}'\boldsymbol{X}_{\alpha,s} \\ &= \frac{n_c}{n}\left[\frac{1}{n_c}\boldsymbol{X}_{\alpha,s^c}'\boldsymbol{X}_{\alpha,s^c} + \frac{1}{n_v}\boldsymbol{X}_{\alpha,s}'\boldsymbol{X}_{\alpha,s}\right] \\ &= o\left(\frac{n_c}{n}\right)\end{aligned}$$

With (2) we have that

$$
\begin{aligned}
(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1} - \frac{n}{n_v}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1} \qquad\qquad & \\
&= \left(\boldsymbol{I}_{d_\alpha} - \frac{n}{n_v}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})\right)(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1} \\
&= \left((\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha}) - \frac{n}{n_v}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})\right)(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1} \\
&= n(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\left(\frac{1}{n}\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha} - \frac{1}{n_v}(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})\right)(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1} \\
&= nO\left(\frac{1}{n}\right)o\left(\frac{n_c}{n}\right)(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1} \\
&= o\left(\frac{n_c}{n}\right)(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}
\end{aligned}
\tag{I.14}
$$

Therefore

$$
\begin{aligned}
\boldsymbol{P}_{\alpha,s} &= \boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\boldsymbol{X}'_{\alpha,s} \\
&= \boldsymbol{X}_{\alpha,s}\left(\frac{n}{n_v}(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1} + o\left(\frac{n_c}{n}\right)(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\right)\boldsymbol{X}'_{\alpha,s} \\
&= \frac{n}{n_v}\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha}\boldsymbol{X}_{\alpha})^{-1}\boldsymbol{X}_{\alpha,s} + o\left(\frac{n_c}{n}\right)\boldsymbol{X}_{\alpha,s}(\boldsymbol{X}'_{\alpha,s}\boldsymbol{X}_{\alpha,s})^{-1}\boldsymbol{X}'_{\alpha,s}
\end{aligned}
\tag{I.15}
$$

$$
= \frac{n}{n_v}\boldsymbol{Q}_{\alpha,s} + o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s} \tag{I.16}
$$

By condition (I.9) we have $\frac{n_v}{n} \to 1$ and $n_c = n - n_v \to \infty$ then

$$
\begin{aligned}
\boldsymbol{Q}_{\alpha,s} &= \frac{n_v}{n}\left(\boldsymbol{P}_{\alpha,s} + o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s}\right) \\
&= \frac{n_v}{n}\boldsymbol{P}_{\alpha,s} + o\left(\frac{n_c}{n}\right)\frac{n_v}{n}\boldsymbol{P}_{\alpha,s} \\
&= \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\frac{n_v}{n}\right]\boldsymbol{P}_{\alpha,s} \\
&= \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]\boldsymbol{P}_{\alpha,s}
\end{aligned}
\tag{I.17}
$$

Note that $\boldsymbol{P}_{\alpha,s}$ is symmetric and idempotent

$$\begin{aligned}
\frac{c_n}{n_v b}\sum_{s\in\mathscr{B}}\|\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s}\|^2 &= \frac{c_n}{n_v b}\sum_{s\in\mathscr{B}}(\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,})'(\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s}) \\
&= \frac{c_n}{n_v b}\sum_{s\in\mathscr{B}}\boldsymbol{r}'_{\alpha,s}\boldsymbol{P}'_{\alpha,s}\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s} \\
&= \frac{c_n}{n_v b}\sum_{s\in\mathscr{B}}\boldsymbol{r}'_{\alpha,s}\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s} \\
&= \frac{c_n}{n_v b}\sum_{s\in\mathscr{B}}\boldsymbol{r}'_{\alpha,s}\left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]^{-1}\boldsymbol{Q}_{\alpha,s}\boldsymbol{r}_{\alpha,s} \\
&= \left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]^{-1}\frac{c_n}{n_v b}\sum_{s\in\mathscr{B}}\boldsymbol{r}'_{\alpha,s}\boldsymbol{Q}_{\alpha,s}\boldsymbol{r}_{\alpha,s} \\
&= \left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]^{-1}\frac{c_n}{n_v b}(n_v b)\left(\frac{1}{n}-\frac{n_v-1}{n(n-1)}\right)\sum_i w_{i\alpha}r_{i\alpha}^2 \\
&= \left[1+o\left(\frac{n_c}{n}\right)\right]\frac{n+n_c}{n_c(n-1)}\sum_i w_{i\alpha}r_{i\alpha}^2 \qquad \text{(I.18)}
\end{aligned}$$

where

$$c_n = n_v(n+n_c)n_c^{-2} \tag{I.19}$$

To see how we arrived at the coefficients for (I.18)

$$\begin{aligned}
&\left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]^{-1}\frac{c_n}{n_v b}\frac{c_n}{n_v b}(n_v b)\left(\frac{1}{n}-\frac{n_v-1}{n(n-1)}\right) \\
&\qquad = \left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]\frac{n_v(n+n_c)}{n_c^2}\left(\frac{1}{n}-\frac{n_v-1}{n(n-1)}\right) \\
&\qquad = \left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]\frac{n_v(n+n_c)}{n_c^2}\left(\frac{n_c-1}{n(n-1)}\right) \\
&\qquad = \left[1+o\left(\frac{n_c}{n}\right)\right]\frac{n+n_c}{n_c(n-1)}.
\end{aligned}$$

Define

$$\begin{aligned}
\boldsymbol{U}_{\alpha,s} &= (\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})(\boldsymbol{I}_{n_v} - c_n\boldsymbol{P}_{\alpha,s})(\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})\\
A_\alpha &= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \boldsymbol{r}'_{\alpha,s}(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{U}_{\alpha,s}(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{r}_{\alpha,s}\\
B_\alpha &= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \boldsymbol{r}'_{\alpha,s}(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{I}_{n_v}-\boldsymbol{U}_{\alpha,s})(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{r}_{\alpha,s}
\end{aligned}$$

Then by (I.3),

$$\begin{aligned}
\hat{\Gamma}^{BICV}_{\alpha,n} &= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \|\boldsymbol{y}_s - \hat{\boldsymbol{y}}_{\alpha,s^c}\|^2\\
&= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \|(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\|^2\\
&= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} [(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)]'[(\mathbf{I}_{n_v}-\mathbf{Q}_{\alpha,s})^{-1}(\mathbf{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)]\\
&= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} (\boldsymbol{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)'(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}(\mathbf{y}_s - \boldsymbol{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\\
&= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \boldsymbol{r}'_{\alpha,s}(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{r}_{\alpha,s}\\
&= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \boldsymbol{r}'_{\alpha,s}(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{I}_{n_v}(\mathbf{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{r}_{\alpha,s}\\
&= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \boldsymbol{r}'_{\alpha,s}(\mathbf{I}_{n_v}-\mathbf{Q}_{\alpha,s})^{-1}(\mathbf{U}_{\alpha,s}+(\boldsymbol{I}_{n_v}-\mathbf{U}_{\alpha,s}))(\boldsymbol{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\mathbf{r}_{\alpha,s}\\
&= A_\alpha + B_\alpha
\end{aligned}\tag{I.20}$$

From the balance property of $\mathscr{B}$ and (I.18),

$$\begin{aligned}
A_\alpha &= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{U}_{\alpha,s} (\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} (\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})(\boldsymbol{I}_{n_v} - c_n \boldsymbol{P}_{\alpha,s})(\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})(\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\boldsymbol{I}_{n_v} + c_n \boldsymbol{P}_{\alpha,s}) \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} \boldsymbol{r}_{\alpha,s} + c_n \boldsymbol{r}'_{\alpha,s} \boldsymbol{P}_{\alpha,s} \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \|\boldsymbol{r}_{\alpha,s}\|^2 + \frac{c_n}{n_v b} \sum_{s \in \mathscr{B}} \|\boldsymbol{P}_{\alpha,s} \boldsymbol{r}_{\alpha,s}\|^2 \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \|\boldsymbol{y}_{\alpha,s} - \boldsymbol{X}_{\alpha,s} \hat{\boldsymbol{\beta}}_\alpha\|^2 + \frac{c_n}{n_v b} \sum_{s \in \mathscr{B}} \|\boldsymbol{P}_{\alpha,s} \boldsymbol{r}_{\alpha,s}\|^2 \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \|\boldsymbol{y}_{\alpha,s} - \boldsymbol{X}_{\alpha,s} \hat{\boldsymbol{\beta}}_\alpha\|^2 + \left[1 + o\left(\frac{n_c}{n}\right)\right] \frac{n + n_c}{n_c(n-1)} \sum_i w_{i\alpha} r_{i\alpha}^2 \qquad \text{(I.21)}
\end{aligned}$$

Assume that $\mathscr{M}_\alpha$ is in Category II. Then by (I.21) and the fact that $\sum_i w_{i\alpha} r_{i\alpha}^2 = d_\alpha \sigma^2 + o_p(1)$ we have

$$\begin{aligned}
A_\alpha &= \frac{1}{n} \boldsymbol{e}'(\boldsymbol{I} - \boldsymbol{P}_\alpha)\boldsymbol{e} + \left[1 = o\left(\frac{n_c}{n}\right)\right] \frac{n + n_c}{n_c(n-1)} [d_\alpha \sigma^2 + o_p(1)] \\
&= \frac{1}{n} \boldsymbol{e}'\boldsymbol{e} + \frac{d_\alpha \sigma^2}{n_c} + o_p\left(\frac{1}{n_c}\right)
\end{aligned}$$

Now we need to show that $B_\alpha = o_p(n_c^{-1})$. From (I.17)

$$\begin{aligned}
(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) \boldsymbol{P}_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) &= \left(\mathbf{I}_{n_v} - \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right] \boldsymbol{P}_{\alpha,s}\right) \boldsymbol{P}_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) \\
&= \left(\boldsymbol{P}_{\alpha,s} - \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right] \boldsymbol{P}_{\alpha,s}\right) (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) \\
&= \left(1 - \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]\right) (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) \\
&= \left[\frac{n_c}{n} + o\left(\frac{n_c}{n}\right)\right]^2 \boldsymbol{P}_{\alpha,s}
\end{aligned}$$

which implies that

$$\begin{aligned}\left(\frac{n}{n_c}\right)^2(\mathbf{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})\boldsymbol{P}_{\alpha,s}(\mathbf{I}_{n_v}-\boldsymbol{Q}_{\alpha,s}) &= \left(\frac{n}{n_c}\right)^2\left[\frac{n_c}{n}+o\left(\frac{n_c}{n}\right)\right]^2\boldsymbol{P}_{\alpha,s}\\ &= [1+o(1)]^2\boldsymbol{P}_{\alpha,s}\\ &\geq \frac{1}{2}\boldsymbol{P}_{\alpha,s}\end{aligned}$$

for $s \in \mathscr{B}$ when n is sufficiently large. Then

$$(\mathbf{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{P}_{\alpha,s}(\mathbf{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})^{-1} \leq 2\left(\frac{n}{n_c}\right)^2\boldsymbol{P}_{\alpha,s} \tag{I.22}$$

Also by (I.17),

$$\begin{aligned}\boldsymbol{U}_{\alpha,s} &= (\mathbf{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})(\mathbf{I}_{n_v}-\boldsymbol{Q}_{\alpha,s})\\ &= \left(\mathbf{I}_{n_v}-\left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]\boldsymbol{P}_{\alpha,s}\right)(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})\left(\mathbf{I}_{n_v}-\left[\frac{n_v}{n}+o\left(\frac{n_c}{n}\right)\right]\boldsymbol{P}_{\alpha,s}\right)\\ &= \left(\left(\mathbf{I}_{n_v}-\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)-o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s}\right)(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})\left(\left(\mathbf{I}_{n_v}-\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)-o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s}\right)\\ &= \left[\left(\mathbf{I}_{n_v}-\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})-o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s}(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})\right] \qquad (I.23)\\ &\left(\left(\mathbf{I}_{n_v}-\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)-o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s}\right)\\ &= \left(\mathbf{I}_{n_v}-\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})\left(\mathbf{I}_{n_v}-\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) \qquad (I.24)\\ &- o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s}(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})\left(\mathbf{I}_{n_v}+\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) \qquad (I.25)\\ &- \left(\mathbf{I}_{n_v}-\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s} \qquad (I.26)\\ &+ o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s}(\mathbf{I}_{n_v}+c_n\boldsymbol{P}_{\alpha,s})o\left(\frac{n_c}{n}\right)\boldsymbol{P}_{\alpha,s} \qquad (I.27)\end{aligned}$$

For (I.24),

$$
\begin{aligned}
(I.24) &= \left[\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) + \left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) c_n\boldsymbol{P}_{\alpha,s}\right]\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) \\
&= \left[\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) + \left(c_n\boldsymbol{P}_{\alpha,s} - c_n\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)\right]\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) \\
&= \left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)^2 + \left(c_n\boldsymbol{P}_{\alpha,s} - c_n\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) \qquad \text{(I.28)}
\end{aligned}
$$

Looking at the latter part of (I.28),

$$
\begin{aligned}
\left(c_n\boldsymbol{P}_{\alpha,s} - c_n\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) &= c_n\boldsymbol{P}_{\alpha,s} - c_n\frac{n_v}{n}\boldsymbol{P}_{\alpha,s} - c_n\frac{n_v}{n}\boldsymbol{P}_{\alpha,s} + c_n\left(\frac{n_v}{n}\right)^2\boldsymbol{P}_{\alpha,s} \\
&= c_n\left(1 - 2\frac{n_v}{n} + \left(\frac{n_v}{n}\right)^2\right)\boldsymbol{P}_{\alpha,s} \\
&= c_n\left(1 - \frac{n_v}{n}\right)^2\boldsymbol{P}_{\alpha,s}
\end{aligned}
$$

For (I.26),

$$
\begin{aligned}
(I.26) &= -o\left(\frac{n_c}{n}\right)\left[(\boldsymbol{P}_{\alpha,s} - c_n\boldsymbol{P}_{\alpha,s})\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right) + \left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)(\boldsymbol{P}_{\alpha,s} - c_n\boldsymbol{P}_{\alpha,s})\right] \\
&= o\left(\frac{n_c}{n}\right)\left[2\boldsymbol{P}_{\alpha,s} - 2\frac{n_v}{n}\boldsymbol{P}_{\alpha,s} - 2c_n\boldsymbol{P}_{\alpha,s} + 2c_n\frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right] \\
&= 2o\left(\frac{n_c}{n}\right)\left[1 - \frac{n_v}{n} - c_n + c_n\frac{n_v}{n}\right]\boldsymbol{P}_{\alpha,s} \\
&= 2o\left(\frac{n_c}{n}\right)\left(1 - \frac{n_v}{n}\right)(1 - c_n)\,\boldsymbol{P}_{\alpha,s}
\end{aligned}
$$

For (I.27),

$$
\begin{aligned}
(I.27) &= \left[o\left(\frac{n_c}{n}\right)\right]^2(\boldsymbol{P}_{\alpha,s} + c_n\boldsymbol{P}_{\alpha,s})\boldsymbol{P}_{\alpha,s} \\
&= \left[o\left(\frac{n_c}{n}\right)\right]^2(1 + c_n)\boldsymbol{P}_{\alpha,s}
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\boldsymbol{U}_{\alpha,s} &= (I.24) + (I.26) + (I.27) \\
&= \left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)^2 + c_n\left(1 - \frac{n_v}{n}\right)^2 \boldsymbol{P}_{\alpha,s} \\
&+ 2o\left(\frac{n_c}{n}\right)\left(1 - \frac{n_v}{n}\right)\left(1 - c_n\right)\boldsymbol{P}_{\alpha,s} \\
&+ \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)\boldsymbol{P}_{\alpha,s} \\
&= \left(\mathbf{I}_{n_v} - \frac{n_v}{n}\boldsymbol{P}_{\alpha,s}\right)^2 + c_n\left(1 - \frac{n_v}{n}\right)^2 \boldsymbol{P}_{\alpha,s} + \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)\boldsymbol{P}_{\alpha,s} \\
&= \mathbf{I}_{n_v} - 2\frac{n_v}{n}\boldsymbol{P}_{\alpha,s} + \left(\frac{n_v}{n}\right)^2 \boldsymbol{P}_{\alpha,s} + c_n\left(1 - \frac{n_v}{n}\right)^2 \boldsymbol{P}_{\alpha,s} + \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)\boldsymbol{P}_{\alpha,s} \\
&= \mathbf{I}_{n_v} + \left(-\frac{n_v}{n}\left(2 - \frac{n_v}{n}\right) + c_n\left(1 - \frac{n_v}{n}\right)^2\right)\boldsymbol{P}_{\alpha,s} + \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)\boldsymbol{P}_{\alpha,s} \\
&= \mathbf{I}_{n_v} + \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)\boldsymbol{P}_{\alpha,s}
\end{aligned}
$$

since $c_n\left(1 - \frac{n_v}{n}\right)^2 = \frac{n_v}{n}\left(2 - \frac{n_v}{n}\right)$.

Then, by (I.22),

$$
\begin{aligned}
&(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1}(\mathbf{I}_{n_v} - \boldsymbol{U}_{\alpha,s})(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \\
&\qquad = (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1}\left(\mathbf{I}_{n_v} - \mathbf{I}_{n_v} - \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)\boldsymbol{P}_{\alpha,s}\right)(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \\
&\qquad = \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{P}_{\alpha,s}(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \\
&\qquad \leq \left[o\left(\frac{n_c}{n}\right)\right]^2 (1 + c_n)2\left(\frac{n^2}{n_c}\right)\boldsymbol{P}_{\alpha,s} \\
&\qquad = 2o(1)(1 + c_n)\boldsymbol{P}_{\alpha,s} \\
&\qquad \leq o(1)(1 + c_n)\boldsymbol{P}_{\alpha,s}
\end{aligned}
$$

so

$$\begin{aligned}
B_\alpha &= \frac{1}{n_v b}\sum_{s\in\mathscr{B}} \boldsymbol{r}'_{\alpha,s}(\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{I}_{n_v} - \boldsymbol{U}_{\alpha,s})(\boldsymbol{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1}\boldsymbol{r}_{\alpha,s} \\
&\le o(1)(1+c_n)\left(\frac{1}{n_v b}\sum_{s\in\mathscr{B}} \boldsymbol{r}'_{\alpha,s}\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s}\right) \\
&= o(1)(1+c_n)\left(\frac{1}{n_v b}\sum_{s\in\mathscr{B}} \|\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s}\|^2\right) \qquad \text{(I.29)}
\end{aligned}$$

because from the proof of (3.5) and (3.6) in [25] $\frac{c_n}{n_v b}\sum_{s\in\mathscr{B}} \|\boldsymbol{P}_{\alpha,s}\boldsymbol{r}_{\alpha,s}\|^2 = O_p\left(\frac{1}{n_c}\right)$. Thus we have shown (I.11) and (I.12) follows.

I.0.6 Proof that APCV is consistent from Shao

Here we show that APCV can be derived from BICV. For readability, define the following:

$$\begin{aligned}
\frac{1}{n_v}\|\boldsymbol{y}_s - \boldsymbol{y}_{\hat{\alpha,s^c}}\|^2 &= \frac{1}{n_v}\|(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1}(\boldsymbol{y}_s - \boldsymbol{x}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\|^2 \\
\boldsymbol{U}_{\alpha,s} &= (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})(\mathbf{I}_{n_v} + c_n\boldsymbol{P}_{\alpha,s})(\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) \\
\boldsymbol{Q}_{\alpha,s} &= \left[\frac{n_n}{n} + o\left(\frac{n_c}{n}\right)\right]\boldsymbol{P}_{\alpha,s} \\
c_n &= n_v(n+n_c)n_c^{-2}
\end{aligned}$$

Recall that

$$\hat{\Gamma}^{BICV}_{\alpha,n} = A_\alpha + B_\alpha$$

where

$$A_\alpha = \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{U}_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{r}_{\alpha,s}$$

$$B_\alpha = \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} (\mathbf{I}_{n_v} - \boldsymbol{U}_{\alpha,s}) (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{r}_{\alpha,s}$$

Through some calculations we can show that $A_\alpha \to \Gamma^{A\hat{P}CV}_{\alpha,n}$

$$\begin{aligned}
A_\alpha &= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{U}_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) (\mathbf{I}_{n_v} + c_n \boldsymbol{P}_{\alpha,s}) (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s}) (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\mathbf{I}_{n_v} + c_n \boldsymbol{P}_{\alpha,s}) \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} \boldsymbol{r}_{\alpha,s} + c_n \boldsymbol{r}'_{\alpha,s} \boldsymbol{P}_{\alpha,s} \boldsymbol{r}_{\alpha,s} \\
&= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \|\boldsymbol{y}_{\alpha,s} - \boldsymbol{X}_{\alpha,s} \hat{\boldsymbol{\beta}}_\alpha\|^2 + \frac{c_n}{n_v b} \sum_{s \in \mathscr{B}} \|\boldsymbol{P}_{\alpha,s} \boldsymbol{r}_{\alpha,s}\|^2 \\
&= \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha\|^2 + \left[1 + o\left(\frac{n_c}{n}\right)\right] \frac{n + n_c}{n_c(n-1)} \sum_i w_{i\alpha} (\boldsymbol{y}_i - \boldsymbol{x}'_{i\alpha} \hat{\boldsymbol{\beta}}_\alpha)^2 \\
&= \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha\|^2 + \left[1 + o\left(\frac{n_c}{n}\right)\right] \frac{n + n_c}{n_c(n-1)} \sum_i w_{i\alpha} (\boldsymbol{y}_i - \boldsymbol{x}'_{i\alpha} \hat{\boldsymbol{\beta}}_\alpha)^2 \\
&= \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha\|^2 + \frac{n + n_c}{n_c(n-1)} \sum_i w_{i\alpha} (\boldsymbol{y}_i - \boldsymbol{x}'_{i\alpha} \hat{\boldsymbol{\beta}}_\alpha)^2 \\
&= \hat{\Gamma}^{APCV}_{\alpha,n}
\end{aligned}$$

Similarly, we can show that $B_\alpha \to 0$

$$\begin{aligned}
B_\alpha &= \frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} (\mathbf{I}_{n_v} - \boldsymbol{U}_{\alpha,s}) (\mathbf{I}_{n_v} - \boldsymbol{Q}_{\alpha,s})^{-1} \boldsymbol{r}_{\alpha,s} \\
&\leq o(1)(1 + c_n) \left(\frac{1}{n_v b} \sum_{s \in \mathscr{B}} \boldsymbol{r}'_{\alpha,s} \boldsymbol{P}_{\alpha,s} \boldsymbol{r}_{\alpha,s} \right) \\
&= o(1)(1 + c_n) \left(\frac{1}{n_v b} \sum_{s \in \mathscr{B}} \| \boldsymbol{P}_{\alpha,s} \boldsymbol{r}_{\alpha,s} \|^2 \right) \\
&= o_p \left(\frac{1}{n_c} \right).
\end{aligned}$$

Since it can be shown that

$$\boldsymbol{U}_{\alpha,s} = \mathbf{I}_{n_v} \left[o \left(\frac{n_c}{n} \right) \right]^2 (1 + c_n) \boldsymbol{P}_{\alpha,s}.$$

We will show that if $\mathscr{M}$ is in Category II, and conditions 1, 2, 3, and I.9 are met then

$$\hat{\Gamma}^{BICV}_{\alpha,n} = \frac{1}{n} \boldsymbol{e}' (\boldsymbol{I}_n - \boldsymbol{P}_\alpha) \boldsymbol{e} + \frac{n + n_c}{n_c (n-1)} (d_\alpha \sigma^2 + o_p(1))$$

The APCV selects a model by minimizing

$$\hat{\Gamma}^{APCV}_{\alpha,n} = \frac{1}{n} \| \boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha \|^2 + \frac{n + n_c}{n_c (n-1)} \sum_i w_{i\alpha} (y_i - \mathbf{x}'_{i\alpha} \hat{\boldsymbol{\beta}}_\alpha)^2$$

Looking at the first term,

$$\begin{aligned}
\| \boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha \|^2 &= (\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha)' (\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha) \\
&= (\boldsymbol{y}' - \hat{\boldsymbol{\beta}}'_\alpha \boldsymbol{X}'_\alpha) (\boldsymbol{y} - \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha) \\
&= \boldsymbol{y}' \boldsymbol{y} - \hat{\boldsymbol{\beta}}'_\alpha \boldsymbol{X}'_\alpha \boldsymbol{y} - \boldsymbol{y}' \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha + \hat{\boldsymbol{\beta}}'_\alpha \boldsymbol{X}'_\alpha \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha.
\end{aligned}$$

Through straightforward computation, we find

$$
\begin{aligned}
\boldsymbol{y}'\boldsymbol{y} &= \boldsymbol{\beta}\boldsymbol{X}_\alpha\boldsymbol{X}'_\alpha\boldsymbol{\beta} + \boldsymbol{e}'\boldsymbol{X}'_\alpha\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_\alpha\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e} \\
\hat{\boldsymbol{\beta}}'_\alpha\boldsymbol{X}'_\alpha\boldsymbol{y} &= \boldsymbol{y}'\boldsymbol{P}_\alpha\boldsymbol{y} \\
\boldsymbol{y}'\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha &= \boldsymbol{y}'\boldsymbol{P}_\alpha\boldsymbol{y} \\
\hat{\boldsymbol{\beta}}'_\alpha\boldsymbol{X}'_\alpha\boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha &= \boldsymbol{y}'\boldsymbol{P}_\alpha\boldsymbol{y}
\end{aligned}
$$

then since $\mathscr{M}_\alpha$ is in Category II

$$
\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2 &= (\boldsymbol{\beta}_\alpha\boldsymbol{X}_\alpha\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{e}'\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{\beta}'_\alpha\boldsymbol{X}_\alpha\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e}) - \boldsymbol{y}'\boldsymbol{P}_\alpha\boldsymbol{y} \\
&= (\boldsymbol{\beta}_\alpha\boldsymbol{X}_\alpha\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{e}'\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{\beta}'_\alpha\boldsymbol{X}_\alpha\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e}) \\
&- (\hat{\boldsymbol{\beta}}_\alpha\boldsymbol{X}_\alpha\boldsymbol{P}_\alpha\boldsymbol{X}'_\alpha\hat{\beta}_\alpha + \boldsymbol{e}'\boldsymbol{P}'_\alpha\boldsymbol{X}'_\alpha\hat{\boldsymbol{\beta}}_\alpha + \hat{\boldsymbol{\beta}}'_\alpha\boldsymbol{X}_\alpha\boldsymbol{P}_\alpha\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e}) \\
&= (\boldsymbol{\beta}_\alpha\boldsymbol{X}_\alpha\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{e}'\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{\beta}'_\alpha\boldsymbol{X}_\alpha\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e}) \\
&- (\boldsymbol{\beta}_\alpha\boldsymbol{X}_\alpha\boldsymbol{P}_\alpha\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{e}'\boldsymbol{P}'_\alpha\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{\beta}'_\alpha\boldsymbol{X}_\alpha\boldsymbol{P}_\alpha\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{P}_\alpha\boldsymbol{e}) \\
&= \boldsymbol{\beta}_\alpha\boldsymbol{X}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha + \boldsymbol{e}'(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{X}'_\alpha\boldsymbol{\beta}_\alpha \\
&+ \boldsymbol{\beta}'_\alpha\boldsymbol{X}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e} + \boldsymbol{e}'(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e} \\
&= \boldsymbol{e}'(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e} \\
\implies \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2 &= \frac{1}{n}\boldsymbol{e}'(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e}
\end{aligned}
$$

since $\boldsymbol{X}_\alpha(\boldsymbol{I}_n - \boldsymbol{P}_\alpha) = (\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{X}'_\alpha = 0$. Now we will look at the second term in $\hat{\Gamma}^{APCV}_{\alpha,n}$,

$$\begin{aligned}\sum_i w_{i\alpha}(\boldsymbol{y}_i - \boldsymbol{x}'_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2 &= \sum_i w_{i\alpha} r^2_{i\alpha} \\ &= d_\alpha \sigma^2 + o_p(1) \\ \Longrightarrow \frac{n+n_c}{n_c(n-1)} \sum_i w_{i\alpha}(y_i - \boldsymbol{x}'_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2 &= \frac{n+n_c}{n_c(n-1)}(d_\alpha\sigma^2 + o_p(1)).\end{aligned}$$

Therefore

$$\hat{\Gamma}^{APCV}_{\alpha,n} = \frac{1}{n}\boldsymbol{e}'(\boldsymbol{I}_n - \boldsymbol{P}_\alpha)\boldsymbol{e} + \frac{n+n_c}{n_c(n-1)}(d_\alpha\sigma^2 + o_p(1)).$$

I.0.7 Properties of the hat matrix

Theorem 5 *If $\boldsymbol{A}_n$ is a full rank matrix and $\boldsymbol{A}_{n+1}$ is a matrix formed by appending an extra row $\boldsymbol{x}'$ to $\boldsymbol{A}_n$, then $(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1} - (\boldsymbol{A}'_{n+1}\boldsymbol{A}_{n+1})^{-1}$ is nonnegative definite.*

Proof:

Here $\boldsymbol{A}_{n+1} = \begin{bmatrix} \boldsymbol{A}_n \\ \boldsymbol{x}' \end{bmatrix}$. Then $\boldsymbol{A}'_{n+1}\boldsymbol{A}_{n+1} = \boldsymbol{A}'_n\boldsymbol{A}_n + \boldsymbol{x}\boldsymbol{x}'$ and, using the Woodbury matrix identity, it follows that

$$(\boldsymbol{A}'_{n+1}\boldsymbol{A}_{n+1})^{-1} = (\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1} - (\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}\boldsymbol{x}\boldsymbol{x}'(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}/(1+\boldsymbol{x}'(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}\boldsymbol{x}).$$

For any nonzero vector $\boldsymbol{u}$ with dimension equal to the order of $\boldsymbol{A}'_n\boldsymbol{A}_n$,

$$\begin{aligned}\boldsymbol{u}'((\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1} - (\boldsymbol{A}'_{n+1}\boldsymbol{A}_{n+1})^{-1})\boldsymbol{u} &= \boldsymbol{u}'(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}\boldsymbol{x}\boldsymbol{x}'(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}\boldsymbol{u}/(1+\boldsymbol{x}'(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}\boldsymbol{x}) \\ &= (\boldsymbol{u}'(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}\boldsymbol{x})^2/(1+\boldsymbol{x}'(\boldsymbol{A}'_n\boldsymbol{A}_n)^{-1}\boldsymbol{x})\end{aligned}$$

is nonnegative since $\boldsymbol{x}'(\boldsymbol{A}_n'\boldsymbol{A}_n)^{-1}\boldsymbol{x} > 0$ because $(\boldsymbol{A}_n'\boldsymbol{A}_n)^{-1}$ is full rank and thus positive definite.

□

Corollary 1 *If $\boldsymbol{A}_n$ is a full rank matrix and $\boldsymbol{A}_{n+r}$ is a matrix formed by appending r extra rows to $\boldsymbol{A}_n$, then $(\boldsymbol{A}_n'\boldsymbol{A}_n)^{-1} - (\boldsymbol{A}_{n+r}'\boldsymbol{A}_{n+r})^{-1}$ is nonnegative definite.*

Theorem 6 *Suppose that $\boldsymbol{A}_d$ is an $n \times d$ full rank matrix, $\boldsymbol{x}$ is an n-dimensional column vector which is linearly independent of the columns of $\boldsymbol{A}_d$, and $\boldsymbol{A}_{d+1}$ is a matrix formed by appending the column $\boldsymbol{x}$ to $\boldsymbol{A}_d$. Let $h_i^{(j)}$ be the ith diagonal element of $\boldsymbol{H}_j = \boldsymbol{A}_j(\boldsymbol{A}_j'\boldsymbol{A}_j)^{-1}\boldsymbol{A}_j'$. Then $h_i^{(d+1)} \geq h_i^{(d)}$ for $i = 1, \ldots, n$.*

Proof:

Here $\boldsymbol{A}_{d+1} = \begin{bmatrix} \boldsymbol{A}_d, & \boldsymbol{x} \end{bmatrix}$. Then

$$\boldsymbol{A}_{d+1}'\boldsymbol{A}_{d+1} = \begin{bmatrix} \boldsymbol{A}_d'\boldsymbol{A}_d & \boldsymbol{A}_d'\boldsymbol{x} \\ \boldsymbol{x}'\boldsymbol{A}_d & \boldsymbol{x}'\boldsymbol{x} \end{bmatrix}$$

and

$$(\boldsymbol{A}_{d+1}'\boldsymbol{A}_{d+1})^{-1} = \frac{1}{b}\begin{bmatrix} b(\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1} + (\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1}\boldsymbol{A}_d'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{A}_d(\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1} & -(\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1}\boldsymbol{A}_d'\boldsymbol{x} \\ -\boldsymbol{x}'\boldsymbol{A}_d(\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1} & 1 \end{bmatrix}$$

where $b = \boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{H}_d\boldsymbol{x}$. Then

$$(\boldsymbol{A}_{d+1}'\boldsymbol{A}_{d+1})^{-1}\boldsymbol{A}_{d+1}' = \frac{1}{b}\begin{bmatrix} b(\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1}\boldsymbol{A}_d' + (\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1}\boldsymbol{A}_d'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{H}_d - (\boldsymbol{A}_d'\boldsymbol{A}_d)^{-1}\boldsymbol{A}_d'\boldsymbol{x}\boldsymbol{x}' \\ -\boldsymbol{x}'\boldsymbol{H}_d + \boldsymbol{x}' \end{bmatrix}$$

and

$$\boldsymbol{H}_{d+1} = \boldsymbol{H}_d + \frac{1}{b}\left(\boldsymbol{H}_d\boldsymbol{x}\boldsymbol{x}'\boldsymbol{H}_d - \boldsymbol{H}_d\boldsymbol{x}\boldsymbol{x}' - \boldsymbol{x}\boldsymbol{x}'\boldsymbol{H}_d + \boldsymbol{x}\boldsymbol{x}'\right).$$

Then the ith diagonal element of $\boldsymbol{H}_{d+1}$ is

$$\begin{aligned}
\mathbf{1}_i'\boldsymbol{H}_{d+1}\mathbf{1}_i &= \mathbf{1}_i'\boldsymbol{H}_d\mathbf{1}_i + \frac{1}{b}\left(\mathbf{1}_i'\boldsymbol{H}_d\boldsymbol{x}\boldsymbol{x}'\boldsymbol{H}_d\mathbf{1}_i - \mathbf{1}_i'\boldsymbol{H}_d\boldsymbol{x}\boldsymbol{x}'\mathbf{1}_i - \mathbf{1}_i'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{H}_d\mathbf{1}_i + \mathbf{1}_i'\boldsymbol{x}\boldsymbol{x}'\mathbf{1}_i\right) \\
&= \mathbf{1}_i'\boldsymbol{H}_d\mathbf{1}_i + \frac{1}{b}\left(\mathbf{1}_i'\boldsymbol{H}_d\boldsymbol{x}\boldsymbol{x}'\boldsymbol{H}_d\mathbf{1}_i - \mathbf{1}_i'\boldsymbol{H}_d\boldsymbol{x}\boldsymbol{x}'\mathbf{1}_i - \mathbf{1}_i'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{H}_d\mathbf{1}_i + \mathbf{1}_i'\boldsymbol{x}\boldsymbol{x}'\mathbf{1}_i\right) \\
&= \mathbf{1}_i'\boldsymbol{H}_d\mathbf{1}e_i + (\mathbf{1}_i'\boldsymbol{x} - \mathbf{1}_i'\boldsymbol{H}_d\boldsymbol{x})^2/b \\
&= \mathbf{1}_i'\boldsymbol{H}_d\mathbf{1}_i + \frac{(\mathbf{1}_i'(\boldsymbol{I} - \boldsymbol{H}_d)\boldsymbol{x})^2}{\boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{H}_d)\boldsymbol{x}}.
\end{aligned}$$

Since $\boldsymbol{I}_n - \boldsymbol{H}_d$ is positive definite (it is an idempotent matrix with rank $n - d$), $\boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{H}_d)\boldsymbol{x} > 0$ and thus $h_i^{(d+1)} \geq h_i^{(d)}$.

□

Corollary 2 *Suppose that $\boldsymbol{A}_d$ is an $n \times d$ full rank matrix, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_c$ are n-dimensional column vectors such that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_c$ and the columns of $\boldsymbol{A}_d$ are linearly independent, and $\boldsymbol{A}_{d+c}$ is a matrix formed by appending the column vectors to $\boldsymbol{A}_d$. Let $h_i^{(j)}$ be the ith diagonal element of $\boldsymbol{H}_j = \boldsymbol{A}_j(\boldsymbol{A}_j'\boldsymbol{A}_j)^{-1}\boldsymbol{A}_j'$. Then $h_i^{(d+c)} \geq h_i^{(d)}$ for $i = 1, \ldots, n$.*

CURRICULUM VITAE

Christina Han

Education

University of Louisville, Louisville, KY
Ph.D. in Applied and Industrial Mathematics **Expected Summer 2022**
Graduate data mining certificate **May 2021**
M.A. Mathematics **May 2020**
Areas of Concentration: Statistics, time series, machine learning, data science.
Title of Dissertation: “Cross-validation for autoregressive models”
Advisor: Dr. Ryan Gill

Northland College, Ashland, WI
Bachelors of Arts **May 2010**
Major: Studio art

Experience

August 2021 - present, ConsumerAffairs - Data scientist
June 2021 - August 2021 - Cognitive Scale - Algorithmic science intern
January 2017 - June 2021 - University of Louisville - Graduate teaching assistant

Leadership and Volunteer Experience

2020-2021 American Mathematical Society UofL Chapter Vice president

Technical Skills and Other Abilities

Proficient in R, Python, LaTeX, SQL

Honors, Recognitions, and Awards

2020, University of Louisville, School of Graduate Studies - Outstanding graduate in mathematics