ESSAYS ON LEAST SQUARES MODEL AVERAGING

by

TIAN XIE

A thesis submitted to the
Department of Economics
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
July, 2013

Copyright © Tian Xie, 2013
Abstract

This dissertation adds to the literature on least squares model averaging by studying and extending current least squares model averaging techniques. The first chapter reviews existing literature and discusses the contributions of this dissertation.

The second chapter proposes a new estimator for least squares model averaging. A model average estimator is a weighted average of common estimates obtained from a set of models. I propose computing weights by minimizing a model average prediction criterion (MAPC). I prove that the MAPC estimator is asymptotically optimal in the sense of achieving the lowest possible mean squared error. For statistical inference, I derive asymptotic tests on the average coefficients for the “core” regressors. These regressors are of primary interest to researchers and are included in every approximation model.

In Chapter Three, two empirical applications for the MAPC method are conducted. I revisit the economic growth models in Barro (1991) in the first application. My results provide significant evidence to support Barro’s (1991) findings. In the second application, I revisit the work by Durlauf, Kourtellos and Tan (2008) (hereafter DKT). Many of my results are consistent with DKT’s findings and some of my results provide an alternative explanation to those outlined by DKT.

In the fourth chapter, I propose using the model averaging method to construct
optimal instruments for IV estimation when there are many potential instrument sets. The empirical weights are computed by minimizing the model averaging IV (MAIV) criterion through convex optimization. I propose a new loss function to evaluate the performance of the estimator. I prove that the instrument set obtained by the MAIV estimator is asymptotically optimal in the sense of achieving the lowest possible value of the loss function.

The fifth chapter develops a new forecast combination method based on MAPC. The empirical weights are obtained through a convex optimization of MAPC. I prove that with stationary observations, the MAPC estimator is asymptotically optimal for forecast combination in that it achieves the lowest possible one-step-ahead second-order mean squared forecast error (MSFE). I also show that MAPC is asymptotically equivalent to the in-sample mean squared error (MSE) and MSFE.
To my beautiful wife: I could not have done without you. Thank you for your priceless love and encouragement. Human language is not sufficient to describe my appreciation of you.
Acknowledgments

I am extremely grateful and truly indebted to James G. MacKinnon and Morten Ørregaard Nielsen for their valuable supervision, encouragement, and patience.

I am especially thankful to Frank Milne for his generous support and guidance.

I am also very grateful to Bruce E. Hansen, Jeffrey S. Racine, Joris Pinkse, Donald W. K. Andrews, Joon Y. Park, and Daniel Henderson for their comments and suggestions.

Helpful comments were received from participants at the Canadian Economics Association Conference held in Calgary (2012), the Canadian Econometrics Study Group Annual Meeting in Kingston (2012), and the American Economic Association Annual Conference in San Diego (2013).
## Contents

Abstract

Dedication

Acknowledgments

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation and Literature Review

1.2 Organization of Thesis

Chapter 2: Least Squares Model Averaging by Prediction Criterion

2.1 Introduction

2.2 Model Averaging Prediction Criterion

2.3 Asymptotic Properties

2.3.1 Asymptotic Optimality

2.3.2 Estimating the Variance-Covariance Matrix for $\hat{\beta}(\hat{w})$
Chapter 6: Conclusion and Future Work 85

Bibliography 87

Appendix A: 95

Appendix B: 109

Appendix C: 121
List of Tables

3.1 A Summary of Barro’s (1991) Findings ........................................... 35
3.2 Regressions for Per Capita Growth ............................................. 37
3.3 The Results of the MAPC Estimation ............................................. 39
3.4 Top Eight Models in the MAPC Estimation ................................. 40
3.5 Relative Out-of-sample Predictive Efficiency ............................... 41
3.6 MAPC Estimates with Asymptotic and Bootstrap $P$ Values .......... 44
# List of Figures

2.1 Simulation Results for \( n = 25 \) ............................................. 27
2.2 Estimated Risk without Normalization ................................. 28
2.3 Simulation Results for Various Sample Sizes ....................... 30
2.4 Rejection Frequencies for Three Tests ............................... 32

4.1 KO Estimators with Different Restrictions on \( w \) ................. 61
4.2 Risk Against \( R^2 \) for \( n = 25 \) ............................................. 62
4.3 Risk Against \( R^2 \) for Various Sample Sizes ...................... 63
4.4 Results of DN and KO for Different \( \lambda \) ......................... 65
4.5 Risk Against \( R^2 \) for \( n = 25 \) (Multiple Endogenous Variables) .... 66
4.6 Risk Against \( R^2 \) for Various Sample Sizes (Multiple Endogenous Variables) ........................................... 67

5.1 Simulation Results for Forecasting \( y_{T+1} \) with \( T = 25 \) ............ 81
5.2 Simulation Results for Forecasting \( y_{T+2} \) and \( y_{T+3} \) .................. 82
5.3 Simulation Results for Various Sample Sizes ......................... 83
Chapter 1

Introduction

1.1 Motivation and Literature Review

Model selection is a procedure through which the best model is selected from a set of approximation models. This procedure generally involves calculating a criterion function for all of the approximation models and ranking the criterion functions accordingly. One of the most widely used criterion functions is the Akaike information criterion (AIC) proposed by Akaike (1973). A popular alternative to AIC is the Bayesian information criterion (BIC) developed by Schwarz (1978). BIC is constructed similarly to AIC but with a stronger penalty for complexity. There are other methods based on various criteria. Examples of these methods include the Mallows Criterion (Mallows’ $C_p$) by Mallows (1973), the prediction criterion by Amemiya (1980), and the focused information criterion (FIC) by Claeskens and Hjort (2003). There is also a considerable amount of literature that concentrates on general-to-specific (GETS) modeling. The foundations of GETS modeling have been developed over the last several decades; see Hendry (1976, 1980, 1983), Gilbert (1986), Pagan (1987), Hoover and Perez (1999) and Hendry and Krolzig (1999). Campos, Ericsson, and Hendry
1.1. MOTIVATION AND LITERATURE REVIEW

(2005) provide an overview and a selected bibliography of GETS modeling.

Model averaging is an alternative to model selection. Instead of selecting a single “winning” model, model averaging calculates the weighted average of a set of approximation models. Barnard (1963) first mentioned the concept of “model combination” in a paper studying airline passenger data. Buckland, Burnham and Augustin (1997) suggested using exponential functions of AIC estimates as model weights and proposed the model average AIC (MA-AIC). Leamer (1978) proposed the basic paradigm for Bayesian model averaging (BMA). There is an increasing focus on BMA in current literature (Draper 1995, Raftery, Madigan and Hoeting 1997, Kass and Raftery 1995, etc.); for a literature review on this topic, see Hoeting, Madigan, Raftery and Volinsky (1999).

However, applying BMA can sometimes be difficult due to the “prior set-up” required: before using the BMA approach, researchers need to assign prior probability distributions to the parameters of each model and prior probabilities to each model. The correct and efficient assignment of these prior values can be controversial in the field of economics, although some recent applied works may provide guidance: see Sala-i-Martin, Doppelhofer and Miller (2004); Ley and Steel (2009); Liu and Maheu (2009); and Wright (2009).

Least squares model averaging is an alternative to BMA. Hansen (2007) proposed the Mallows model average (MMA) method based on the original Mallows criterion. An implementation of MMA in forecast combination was made in Hansen (2008). Hansen (2009) extended MMA to regressions with a possible structural break. Another extension was made to autoregression with a near unit root in Hansen (2010). Most of these works are based on homoskedastic error terms. Hansen and Racine
(2012) proposed a jackknife model averaging (JMA) technique for heteroskedastic regression. One limitation of MMA in Hansen (2007) is that the approximation models must be strictly nested in a way that depends on the ordering of regressors. In response to Hansen (2007), Wan, Zhang and Zou (2010) proved that the optimality of MMA holds for non-nested models.

One issue regarding the cost of computation may arise when we encounter a large number of approximation models. For least squares model averaging, to my knowledge, no significant progress has been made to overcome this issue. However, for Bayesian model averaging, this issue can be solved by, for example, using Markov chain Monte Carlo model composition (MC$^3$). The MC$^3$ approach uses a Markov chain Monte Carlo method to directly approximate the posterior distribution of the quantity of interest (such as an effect size, a future observable, or the utility of a course of action) based on given data. See Hoeting et al. (1999) for a detailed explanation on how to construct such a Markov chain. The stochastic process generated by MC$^3$ explores regions of the model space with high posterior model probabilities. Usually, we need to pre-specify the number of iterations in the MC$^3$ approach. The MC$^3$ approach greatly improves estimation efficiency. See Raftery, Madigan, and Hoeting (1997) for an application of MC$^3$ that averages across models with many predictors.


A natural extension of in-sample least squares model averaging is out-of-sample forecasting. Using model averaging to forecast out of sample is also known as forecast combination, which has a rich heritage in both statistics and econometrics. Bates
and Granger (1969) first introduced the idea of forecast combination, which was later extended by Granger and Ramanathan (1984). By compromising across a group of approximation models, forecast combination improves forecast accuracy. The performance of a forecast combination method crucially depends on the weights assigned to the different approximation models. Many existing methods have been proven to be versatile and successful in finding the appropriate weights. Hansen (2008) extended MMA to a forecast combination method, in which the forecast weights are computed by MMA. However, one limitation was inherited from Hansen’s (2007) work: the approximating models were required to be nested in a way that depended on the ordering of regressors.

The MMA method is designed to work with OLS models. In the presence of endogenous variables, the OLS estimator is no longer appropriate and we need to turn to other estimators, for example, the IV estimator. One important and frequently encountered problem in IV estimation is ensuring that the proper instruments are chosen. Donald and Newey (2001) proposed a selection criterion to select instruments in a way that balances higher order bias and efficiency. Sueishi (2012) proposed another IV selection criterion. Unlike Donald and Newey (2001), Sueishi’s (2012) method selects independent variables and instruments simultaneously. Kuersteiner and Okui (2010) extended the selection criterion of Donald and Newey (2001) to a model averaging criterion that constructs optimal instruments for IV estimation. Similar to Hansen’s (2007) treatment of regressors, Kuersteiner and Okui (2010) required the instrument sets to be strictly nested.
1.2 Organization of Thesis

In Chapter 2, I propose a model average estimator with empirical weights computed through numerical minimization of a model average prediction criterion (MAPC). Chapter 3 studies two empirical applications for the MAPC method. In Chapter 4, a new model averaging criterion is proposed for constructing optimal instruments for IV estimation. I extend MAPC to a forecast combination method in Chapter 5. Chapter 6 concludes this thesis. All proofs are relegated to Appendices A, B, and C.
Chapter 2

Least Squares Model Averaging by Prediction Criterion

2.1 Introduction

In this chapter, I propose a model average estimator with empirical weights computed through the numerical minimization of a model average prediction criterion (MAPC). This criterion can be seen as a model averaging version of the original prediction criterion proposed by Amemiya (1980). I prove that the MAPC estimator is asymptotically optimal in the sense of achieving the lowest possible mean squared error, which applies both to nested and non-nested approximation models. I divide the regressors into two groups: the “core” regressors, which are of primary interest to us and are included in every approximation model; and the “potential” regressors, which are of marginal interest to us and are included in some but not all approximation models. For statistical inference, I derive asymptotic tests for single hypotheses and joint hypotheses on the core average coefficients. To improve finite sample performance, I also consider bootstrap tests. In simulation experiments, the MAPC estimator is
shown to have significant efficiency gains over existing model selection and model averaging methods. I also show that the bootstrap tests have more reasonable rejection frequencies than the asymptotic tests in small samples.

This chapter continues with an introduction of the framework of MAPC in Section 2.2. Section 2.3 proves the asymptotic optimality of the MAPC estimator and derives asymptotic tests and bootstrap tests for single hypotheses and joint hypotheses on the average coefficients. Section 2.4 presents simulation experiments. Section 2.5 concludes this chapter.

### 2.2 Model Averaging Prediction Criterion

Let \((y_i, x_i) : i = 1, ..., n\) be a random sample, where \(y_i\) and \(x_i = [x_{i1}, x_{i2}, ...]\) are real-valued. Assume the data generating process is

\[
y_i = \mu_i + u_i, \tag{2.1}
\]

where \(\mu_i = \sum_{j=1}^{\infty} \beta_j x_{ij}\), \(\mathbb{E}(u_i|x_i) = 0\) and \(\mathbb{E}(u_i^2|x_i) = \sigma^2\).

Consider a sequence of linear approximation models \(m = 1, 2, ..., M\). The concept of “approximation model” can be vague. In this thesis, an approximation model \(m\) uses \(k^{(m)}\) regressors belonging to \(x_i\) such that

\[
y_i = \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)} + u_i^{(m)} \quad \text{for} \ i = 1, ..., n, \tag{2.2}
\]

where \(\beta_j^{(m)}\) is a coefficient in model \(m\) and \(x_{ij}^{(m)}\) is a regressor in model \(m\). Other forms of approximation models are beyond the scope of this thesis. The approximation error
for model $m$ is defined as

$$b_i^{(m)} \equiv \mu_i - \sum_{j=1}^{k^{(m)}} \beta_{ij}^{(m)} x_{ij} \quad \text{for } i = 1, \ldots, n.$$  \hspace{1cm} (2.3)$$

Therefore, as long as the approximation model is finite, it always contains a non-zero approximation error.

Hansen (2007) assumed that the regressors $x_i$ were an ordered set and an approximation model $m$ contained the first $k^{(m)}$ regressors from $x_i$. As a result, models with fewer regressors would always nest within larger models, which made the $k^{(m)}$ such that

$$0 \leq k^{(1)} < k^{(2)} < \cdots < k^{(m)} < \cdots < k^{(M)}.$$  

This nested model set-up was demonstrated to be unnecessary by Wan et al. (2010) and Hansen and Racine (2012). In this thesis, I place no restrictions on the orders of $x_i$ and the $k^{(m)}$. Approximation models in this thesis can be either nested or non-nested, which makes the methods in this thesis more widely applicable.

The DGP (2.1) and approximation model (2.2) can be represented in the following matrix forms:

$$y = \mu + u$$

and

$$y = X^{(m)} \beta^{(m)} + u^{(m)},$$

where $y$ is $n \times 1$, $\mu$ is $n \times 1$, $X^{(m)}$ is $n \times k^{(m)}$ with the $ij$th element being $x_{ij}^{(m)}$, $\beta^{(m)}$ is $k^{(m)} \times 1$ and $u^{(m)}$ is the error term for model $m$. For an approximation model $m$,
define the projection matrix

\[ P^{(m)} = X^{(m)} \left( X^{(m)^\top} X^{(m)} \right)^{-1} X^{(m)^\top}. \]  

(2.4)

Therefore, the least squares estimate of \( \mu \) from model \( m \) is \( \hat{\mu}^{(m)} = P^{(m)} y \).

Let \( w = [w^{(1)}, \ldots, w^{(M)}]^\top \) be a weight vector in the unit simplex in \( \mathbb{R}^M \),

\[ H_M \equiv \left\{ w \in [0,1]^M : \sum_{m=1}^{M} w^{(m)} = 1 \right\}, \]

which is a continuous set. Note that Hansen (2007) and Hansen and Racine (2012) assumed a discrete set \( H^*_M \) for \( w \), in which

\[ H^*_M(N) \equiv \left\{ w^{(m)} \in \left[0, \frac{1}{N}, \frac{2}{N}, \ldots, 1 \right] : \sum_{m=1}^{M} w^{(m)} = 1 \right\} \]

for some fixed integer \( N \). I will discuss this further in Section 2.3.

Define the model average estimator of \( \mu \) as

\[ \mu(w) \equiv \sum_{m=1}^{M} w^{(m)} \hat{\mu}^{(m)} = \sum_{m=1}^{M} w^{(m)} P^{(m)} y. \]  

(2.5)

Now define the weighted average projection matrix \( P(w) \) as

\[ P(w) \equiv \sum_{m=1}^{M} w^{(m)} P^{(m)}. \]

Accordingly, (2.5) can be simplified to

\[ \mu(w) = P(w)y. \]  

(2.6)
The effective number of parameters $k(w)$ in model averaging estimation is defined as

$$k(w) \equiv \sum_{m=1}^{M} w^{(m)}k^{(m)},$$  \hspace{1cm} (2.7)$$
which is a weighted sum of the $k^{(m)}$. Note that $k(w)$ is not necessarily an integer.

I propose the model average prediction criterion (MAPC):

$$\text{MAPC}_n(w) = (y - \mu(w))^\top (y - \mu(w)) \left( \frac{n + k(w)}{n - k(w)} \right),$$  \hspace{1cm} (2.8)$$
where $\mu(w)$ and $k(w)$ are defined in (2.6) and (2.7). MAPC can be understood as the model averaging version of the prediction criterion by Amemiya (1980). Like most model selection criteria and model averaging criteria, MAPC balances between the fit and the complexity of a model. MAPC can be used to calculate the empirical weight vector $\hat{w}$, in which

$$\hat{w} = \arg\min_{w \in H_M} \text{MAPC}_n(w).$$

The Mallows’ model average (MMA) criterion proposed by Hansen (2007) is

$$\text{MMA}_n(w) = (y - \mu(w))^\top (y - \mu(w)) + 2\sigma^2 k(w).$$  \hspace{1cm} (2.9)$$

The empirical weights $\hat{w}$ can be selected by minimizing (2.9) subject to $w \in H_M^*$. The MMA criterion is composed of an averaged sum of squared residuals and a penalty term for complexity. Note that the penalty term includes an unknown $\sigma^2$ that must be replaced by a sample estimate.

For convenience in calculations, we can rewrite both criteria. First, denote $\hat{u}^{(m)}$ as an $n \times 1$ estimated residual vector from model $m$. Let $\hat{U}$ be an $n \times M$ matrix
consisting of these residuals such that $\hat{U} \equiv [\hat{u}^{(1)}, \hat{u}^{(2)}, ..., \hat{u}^{(M)}]$. Define an $M \times 1$ vector $k$, which contains the number of parameters from each model, such that $k \equiv [k^{(1)}, k^{(2)}, ..., k^{(M)}]^\top$. Then, the MAPC in (2.8) can be written as

$$
\text{MAPC}_n(w) = w^\top \hat{U}^\top \hat{U} w \left( \frac{n + k^\top w}{n - k^\top w} \right).
$$

(2.10)

Likewise, the MMA criterion in (2.9) becomes

$$
\text{MMA}_n(w) = w^\top \hat{U}^\top \hat{U} w + 2\sigma^2 k^\top w.
$$

(2.11)

2.3 Asymptotic Properties

In this section, I first prove the asymptotic optimality of the MAPC estimator by showing that it achieves the lowest possible mean squared error as $n \to \infty$. Then, asymptotic tests are derived for single hypotheses and joint hypotheses on the core coefficients. I also recommend bootstrap tests for improved inference in finite samples. Proofs for this section are presented in Appendix A.

2.3.1 Asymptotic Optimality

We start by listing some properties of $P(w)$.

**Lemma 2.1.** Define $M(w) \equiv I - P(w)$. Then,

(i) $\text{Tr}(P(w)) = \sum_{m=1}^{M} w^{(m)} k^{(m)} = k(w)$,

(ii) $\lambda_{\text{max}}(P(w)) \leq 1$, where $\lambda_{\text{max}}(\cdot)$ denotes the largest eigenvalue of its argument,

(iii) $\|P(w^*) M(w) \mu\|^2 \leq \|M(w) \mu\|^2$ for any $w^*, w \in H_M$, 


(iv) $\text{Tr}[P(w)P(w^*)P(w)] \leq \text{Tr}[P(w)P(w)]$ for any $w^*, w \in H_M$.

The matrix $P(w)$ is not a traditional projection matrix. For example, it is not idempotent. Therefore, we cannot further simplify Lemma 2.1 (iv).

Define the average mean squared error as $L_n(w) \equiv (\mu(w) - \mu)^\top (\mu(w) - \mu)$ and the conditional average mean squared error as $R_n(w) \equiv \mathbb{E}(L_n(w)|X)$. The same definitions can be found in Li (1987), Hansen (2007) and Wan et al. (2010). We investigate the asymptotic properties of $R_n(w)$ in the next lemma.

**Lemma 2.2.**

(i) $R_n(w) \geq \|M(w)\mu\|^2$,

(ii) $R_n(w) \geq \sigma^2 \text{Tr}(P(w)P(w))$.

**Assumption 2.1.** For some fixed integer $1 \leq G < \infty$, we have $\mathbb{E}(|u_i|^{4G}|x_i) \leq \kappa < \infty$.

Assumption 2.1 is a bound condition on the conditional moments of the error term. It can be compared with the corresponding condition in Hansen (2007), in which

$$\mathbb{E}(|u_i|^{4(N+1)}|x_i) \leq \kappa < \infty. \quad (2.12)$$

Note that (2.12) depends on model weights $w$ through Hansen’s assumption that $w^{(m)}$ is restricted to the set $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$ for some integer $N$.

**Assumption 2.2.** As $n \to \infty$, $\xi_n^{-2G}M \sum_{m=1}^M (R_n(w_0^m))^G \to 0$, where $\xi_n = \inf_{w \in H_M} R_n(w)$ and $w_0^m$ is an $M \times 1$ vector of which the $m$th element is one and the others are zeros.

Assumption 2.2 is the convergence condition. It is the same as the convergence condition in Wan et al. (2010). A necessary condition for Assumption 2.2 to hold
is $\xi_n \to \infty$, which indicates that there is no finite approximating model for which the bias is zero. Moreover, I assume that as $n \to \infty$, $\xi_n^{2G}$ goes to infinity at a faster rate than $M \sum_{m=1}^{M} (R_n(w^0_m))^G$. This is a relatively stronger assumption than the corresponding condition required by Hansen (2007), which only needs $\xi_n \to \infty$ as $n \to \infty$. Note that Hansen’s (2007) assumptions are appropriate only for nested models and a discrete set for $w$; in contrast, our theorem is built on a more general set-up with non-nested models and continuous $H_M$. In practice, Assumption 2.2 can be easily sustained by removing poor models prior to estimation. See Wan et al. (2010) for two explicit examples under which Assumption 2.2 holds.

**Assumption 2.3.** As $n \to \infty$, $k^{(m)} \to \infty$ and $k^{(m)}/n \to 0$ for all $m$.

Assumption 2.3 states that as $n$ goes to infinity, $k^{(m)}$ goes to infinity at a slower rate for $m = 1, ..., M$. Similar assumptions can be found in other papers, such as Shibata (1981) and Hansen (2007). Based on Assumption 2.3, we have

**Lemma 2.3.** Let Assumption 2.3 hold. Then, $k(w)/n \to 0$ as $n \to \infty$.

With the above lemmas and assumptions, the following theorem demonstrates the asymptotic optimality of the MAPC estimator.

**Theorem 2.1.** Let Assumptions 2.1, 2.2, and 2.3 hold. Then, as $n \to \infty$

$$\frac{L_n(\hat{w})}{L_n(w_{opt})} \to 1,$$

where

$$w_{opt} = \arg\inf_{w \in H_M} L_n(w).$$
2.3. ASYMPTOTIC PROPERTIES

Theorem 2.1 states that by using the empirical weight vector $\hat{w}$, the mean squared error is asymptotically equivalent to the lowest possible mean squared error. This implies that the MAPC estimator is asymptotically optimal in the class of model average estimators (2.5) where the weight vector belongs to the set $H_M$.

Define the average estimate of $\sigma^2$ as $\hat{\sigma}^2(w)$:

$$\hat{\sigma}^2(w) \equiv \frac{(y - \hat{\mu}(w)) \top (y - \hat{\mu}(w))}{n - k(w)} = \frac{w \top \hat{U} \top \hat{U} w}{n - k \top w}.$$ 

By inserting $\hat{\sigma}^2(w)$ into (2.10) and rearranging the equation, we can rewrite MAPC as

$$\text{MAPC}_n(w) = w \top \hat{U} \top \hat{U} w + 2\hat{\sigma}^2(w)k \top w,$$

which is similar to MMA in (2.11) with $\sigma^2$ being replaced by an average estimate $\hat{\sigma}^2(w)$.

**Theorem 2.2.** Let Assumption 2.3 hold. Then $\hat{\sigma}^2(w) \xrightarrow{p} \sigma^2$ as $n \to \infty$.

Note that $\hat{\sigma}^2(w)$ is a consistent estimator of $\sigma^2$. Theorem 2.2 implies that $\hat{\sigma}^2(\hat{w}) \xrightarrow{p} \sigma^2$ as $n \to \infty$, where

$$\hat{\sigma}^2(\hat{w}) = \frac{\hat{w} \top \hat{U} \top \hat{U} \hat{w}}{n - k \top \hat{w}}.$$ 

(2.13)

The MMA criterion includes an unknown $\sigma^2$ that must be estimated. Hansen (2007, 2008) recommended using $\hat{\sigma}^2_L$ to replace the unknown $\sigma^2$, where

$$\hat{\sigma}^2_L = \frac{(y - \hat{\mu}^{(L)}) \top (y - \hat{\mu}^{(L)})}{n - k^{(L)}}$$

is the estimated $\sigma^2$ from a large approximation model $L$.\(^1\) As a result, the MMA

\(^1\)In fact, Hansen (2007, 2008) used the largest approximation model in his simulation experiments.
criterion in practice becomes

\[ \text{MMA}_n(w) = w^\top \hat{U}^\top \hat{U} w + 2 \hat{\sigma}_L^2 k^\top w. \]

The MMA estimator is a two-step estimator since a sample estimate of \( \sigma^2 \) must be provided prior to estimation. In contrast, the MAPC estimator is a continuous updating estimator that requires only one step of calculation. Estimating \( \hat{w} \) from the MMA criterion with constraints is a classic quadratic programming problem, while estimating \( \hat{w} \) by the MAPC estimator is a convex optimization problem.\(^2\)

### 2.3.2 Estimating the Variance-Covariance Matrix for \( \hat{\beta}(\hat{w}) \)

The average coefficient \( \hat{\beta}(\hat{w}) \) is not easy to compute. The number of regressors, \( k^{(m)} \), is usually not the same for different models; even if the number of regressors is the same for certain models, the regressors must be different from one model to another. Either scenario complicates the computation.

Assume that there exists a model \( L \) within which all of the approximation models are nested.\(^3\) The average coefficient \( \hat{\beta}(\hat{w}) \) can be computed by

\[ \hat{\beta}(\hat{w}) = \frac{1}{M} \sum_{m=1}^{M} \hat{w}^{(m)} \left( \Gamma^{(m)} \hat{\beta}^{(m)} \right), \]

where \( \Gamma^{(m)} \) is \( k^{(L)} \times k^{(m)} \) and plays the role of mapping the \( k^{(m)} \times 1 \) vector \( \hat{\beta}^{(m)} \) to a \( k^{(L)} \times 1 \) vector by padding \( \hat{\beta}^{(m)} \) with 0s. A convenient way to construct \( \Gamma^{(m)} \) is to

---

\(^2\)Note that convex optimization usually requires slightly more computation time than quadratic programming.

\(^3\)If no such model exists, one can be easily created by including all regressors within it.
use the following equation:

$$
\Gamma^{(m)} = \left( X^{(L)\top} X^{(L)} \right)^{-1} X^{(L)\top} X^{(m)}.
$$

(2.14)

In this case, the rank of $\Gamma^{(m)}$ is $k^{(m)}$ and each element in $\Gamma^{(m)}$ is either 1 or 0.

Although not shown for the sake of brevity, the variances and covariances in the rest of this subsection are in fact conditional on $X^{(L)}$. Therefore, $X^{(m)}$ can be assumed to be exogenous for all $m$. By straightforward algebra, the variance-covariance matrix of $\hat{\beta}(\hat{w})$ is

$$
\text{Var} \left( \hat{\beta}(\hat{w}) \right) = \text{Var} \left( \sum_{m=1}^{M} \hat{w}^{(m)} \left( \Gamma^{(m)} \hat{\beta}^{(m)} \right) \right)
= \sum_{m=1}^{M} \sum_{s=1}^{M} \hat{w}^{(m)} \hat{w}^{(s)} \text{Cov} \left( \Gamma^{(m)} \hat{\beta}^{(m)}, \Gamma^{(s)} \hat{\beta}^{(s)} \right).
$$

(2.15)

The right-hand-side of equation (2.15) is a linear combination of covariances of $\Gamma^{(m)} \hat{\beta}^{(m)}$ and $\Gamma^{(s)} \hat{\beta}^{(s)}$ for any $m, s$. When $s$ equals $m$, the covariance matrix of $\Gamma^{(m)} \hat{\beta}^{(m)}$ and $\Gamma^{(s)} \hat{\beta}^{(s)}$ becomes the variance-covariance matrix of $\Gamma^{(m)} \hat{\beta}^{(m)}$. Each $\Gamma^{(m)} \hat{\beta}^{(m)}$ includes possible model misspecification bias. I define the following $k^{(L)} \times 1$ vector

$$
\mathbf{d}^{(m)} = \mathbb{E} \left( \Gamma^{(m)} \hat{\beta}^{(m)} \right) - \beta(w)
$$

(2.16)

as the misspecification bias vector. I propose an estimator for $\text{Cov}(\Gamma^{(m)} \hat{\beta}^{(m)}, \Gamma^{(s)} \hat{\beta}^{(s)})$ in the following lemma.

**Lemma 2.4.** For any approximation models $m$ and $s$, in which $m$ and $s$ can represent
the same model, we have
\[
\widehat{\text{Cov}} \left( \Gamma^{(m)} \hat{\beta}^{(m)}, \Gamma^{(s)} \hat{\beta}^{(s)} \right) \\
= \sigma^2(\hat{w}) \Gamma^{(m)} \left( X^{(m)\top} X^{(m)} \right)^{-1} X^{(m)\top} X^{(s)} X^{(s)\top} X^{(s)} \left( X^{(s)} X^{(s)\top} \right)^{-1} \Gamma^{(s)} + \hat{d}^{(m)} \left( \hat{d}^{(s)} \right)^\top,
\]

where
\[
\hat{d}^{(m)} = \Gamma^{(m)} \hat{\beta}^{(m)} - \hat{\beta}(\hat{w}). \tag{2.17}
\]

For a particular average coefficient, for example \( \hat{\beta}_j(\hat{w}) \), the variance of \( \hat{\beta}_j(\hat{w}) \) is
\[
\text{Var}(\hat{\beta}_j(\hat{w})) = \left[ \text{Var}(\hat{\beta}(\hat{w})) \right]_{jj},
\]
which is the \( j^{th} \) element on the diagonal of the variance-covariance matrix \( \text{Var}(\hat{\beta}(\hat{w})) \).

Buckland et al. (1997) proposed other estimators for \( \text{Var}(\hat{\beta}_j(\hat{w})) \). Some of the estimators are based on a restrictive assumption that there is perfect correlation between each \( \beta^{(m)} \). They also proposed computing \( \text{Var}(\hat{\beta}_j(\hat{w})) \) via a pairs bootstrap.

In practice, we want to include certain regressors in every approximation model because these regressors are of primary interest to us. Models without these regressors provide no useful information and therefore are not of interest to us. Let us refer to these regressors as the core regressors. Let the \( n \times k_c \) matrix \( X_c \) represent the core regressors and let \( \hat{\beta}_c(\hat{w}) \) be the corresponding averaged core coefficients. Define \( \Gamma_c^{(m)} \) as a \( k^{(m)} \times k_c \) matrix that plays the role of extracting the \( n \times k_c \) matrix \( X_c \) out of any \( X^{(m)} \) such that \( X^{(m)} \Gamma_c^{(m)} = X_c \). Similar to \( \Gamma^{(m)} \) defined in (2.14), construct \( \Gamma_c^{(m)} \) as \( (X^{(m)\top} X^{(m)})^{-1} X^{(m)\top} X_c \). The variance-covariance matrix for \( \hat{\beta}_c(\hat{w}) \) is then
\[
\text{Var} \left( \hat{\beta}_c(\hat{w}) \right) = \Gamma_c^{(m)\top} \text{Var} \left( \hat{\beta}(\hat{w}) \right) \Gamma_c^{(m)}.
\]
2.3. ASYMPTOTIC PROPERTIES

We will refer to the regressors that are of marginal interest to us as the potential regressors. These potential regressors are not included in every approximation model. Let $\beta_p$ be the corresponding coefficient vector. The coefficient $\beta_p$ is a $k_p \times 1$ vector with $k_p \leq k^{(L)}$. Define the regressors that are not included in $X^{(m)}$ as $X^{(-m)}$ and the associated coefficient vector as $\beta_{-m}$. By definition, the set of $\beta_{-m}$ belongs to the set of $\beta_p$ for all $m$. Similar to the definition of $\Gamma^{(m)}$ in (2.14), let $\Gamma^{(-m)} = (X^{(L)}\top X^{(L)})^{-1}X^{(L)}\top X^{(-m)}$ play the role of mapping the $(k^{(L)} - k^{(m)}) \times 1$ vector $\beta_{-m}$ to a $k^{(L)} \times 1$ vector. How to distinguish potential regressors from core regressors is not the primary concern of this thesis. I leave that for future research.

2.3.3 Asymptotic Inference and Bootstrap Based Inference

We start by listing more assumptions.

Assumption 2.4.

(i) $(x_i, u_i)$ is an iid sequence,

(ii) $E(x_i u_i) = 0$, $E|x_i u_i|^2 < \infty$, and $\text{Var}(n^{-1/2}X\top u)$ is positive definite,

(iii) As $n \to \infty$, $n^{-1}X^{(m)}\top X^{(m)} \xrightarrow{p} S^{(m)}$ and $n^{-1}X^{(m)}\top X^{(s)} \xrightarrow{p} S^{(m,s)} \forall m, s$, where both $S^{(m)}$ and $S^{(m,s)}$ are finite, deterministic matrices and $S^{(m)}$ is also positive definite.

Assumption 2.5. $\beta_p = h_p/\sqrt{n}$, where $h_p$ is a fixed vector.

Assumption 2.6. The average coefficient $\beta(w)$ is a function of $\beta^{(m)}$ and $\beta_{-m}$ that can be written as $\beta(w) = f(\beta^{(m)}, \beta_{-m})$, where $\beta(w) = f(\beta^{(m)}, \beta_{-m})$ is twice differentiable in a neighborhood of $\beta_{-m}$ and $f(\beta^{(m)}, 0) = \Gamma^{(m)}\beta^{(m)}$. 
Assumption 2.4 is a standard assumption about regressors and error terms. Assumption 2.5 follows the logic that the potential regressors are only of marginal interest to us as they are weakly correlated with $y$. This idea of weak variables is similar to ideas put forth in the weak instruments literature. For example, Assumption $L_{II}$ in Staiger and Stock (1997) is almost identical to our Assumption 2.5. Assumption 2.6 is a standard assumption that allows us to use a Taylor expansion on $\beta(w)$.

Let $\hat{\beta}_c^{(m)} = \Gamma_c^{(m)} \hat{\beta}^{(m)}$ be the core part of $\hat{\beta}^{(m)}$. The asymptotic distribution of $\hat{\beta}_c^{(m)}$ is investigated in the following lemma:

**Lemma 2.5.** Let Assumptions 2.4, 2.5, and 2.6 hold. Then, as $n \to \infty$,

(i) $\beta_m = h_m / \sqrt{n}$ for all $m$, where $h_m$ is a fixed vector.

(ii) $\sqrt{n} \left( \hat{\beta}_c^{(m)} - \beta_c \right) \xrightarrow{d} N \left( \delta_1^{(m)}, \sigma^2 \Gamma_c^{(m)} (S^{(m)})^{-1} \Gamma_c^{(m)} \right)$, where

$$\delta_1^{(m)} \equiv \Gamma_c^{(m)} (S^{(m)})^{-1} S^{(m,-m)} h_m.$$

(iii) Define $F_{\beta_m} \equiv \partial f(\beta^{(m)}, \beta_m) / \partial \beta_m$. Then

$$\sqrt{n} \left( \beta_c - \beta_c(w) \right) \xrightarrow{p} \delta_2^{(m)},$$

where $\delta_2^{(m)} \equiv -\Gamma_c^{(L)} \left( F_{\beta_m} \bigg|_{\beta_m=0} \right) h_m$.

(iv) $\sqrt{n} d^{(m)} \xrightarrow{p} \delta^{(m)}$, where $\delta^{(m)} \equiv \Gamma^{(m)} (S^{(m)})^{-1} S^{(m,-m)} h_m - \left( F_{\beta_m} \bigg|_{\beta_m=0} \right) h_m$.

(v) The asymptotic distribution of $\hat{\beta}_c^{(m)}$ is

$$\sqrt{n} \left( \hat{\beta}_c^{(m)} - \beta_c(w) \right) \xrightarrow{d} \Lambda_c^{(m)} \sim N \left( \delta_c^{(m)}, V_c^{(m)} \right), \quad (2.18)$$
where \( \delta_c^{(m)} = \delta_1^{(m)} + \delta_2^{(m)} \) and \( V_c^{(m)} \equiv \sigma^2 \Gamma_c^{(m)} \top (S^{(m)})^{-1} \Gamma_c^{(m)} \).

Based on Lemma 2.5, the asymptotic distribution of the core average coefficient \( \hat{\beta}_c(\hat{w}) \) conditional on \( \hat{w} \) is derived in the following theorem:

**Theorem 2.3.** Let Assumptions 2.1–2.6 hold. Then, as \( n \to \infty \),

\[
\sqrt{n} \left( \hat{\beta}_c(\hat{w}) - \beta_c(w) \right) \mid \hat{w} \to_d \Lambda_c = \sum_{m=1}^{M} \hat{w}^{(m)} \Lambda_c^{(m)} \sim N \left( 0, \Gamma_c^{(m)} \top V(\hat{w}) \Gamma_c^{(m)} \right),
\]

where \( \Lambda_c^{(m)} \) is defined in (2.18) and

\[
V(\hat{w}) = \sum_{m=1}^{M} \sum_{s=1}^{M} \hat{w}^{(m)} \hat{w}^{(s)} \left( \sigma^2 \Gamma^{(m)} (S^{(m)})^{-1} S^{(m,s)} (S^{(s)})^{-1} \Gamma^{(s)} \top + \delta^{(m)} \delta^{(s)} \top \right).
\]

The terms \( S^{(m)}, S^{(m,s)}, \) and \( \delta^{(m)} \) are defined in Assumption 2.4 and Lemma 2.5.

There is joint convergence of \( \sqrt{n}(\hat{\beta}_c^{(m)} - \beta_c(w)) \) and the stochastic weights \( \hat{w} \). The conditional asymptotic distribution \( \Lambda_c \) is normal, which implies that the unconditional asymptotic distribution of \( \sqrt{n}(\hat{\beta}_c(\hat{w}) - \beta_c(w)) \) is a mixed normal distribution. Typically the mixed normal distribution has fatter tails than the normal distribution. Making inferences based on mixed normal distributions is difficult because the variance of the normal distribution in Theorem 2.3 depends on a random vector \( \hat{w} \). We explain how to make inferences about the averaged coefficients in the rest of this section.

We construct the \( t \)-statistic and the Wald statistic in the usual way. To test a single restriction, for example \( \beta_j(w) = \beta_{j0} \), the \( t \)-statistic for the average core coefficient \( \hat{\beta}_j(\hat{w}) \) is

\[
t_{\hat{\beta}_j(\hat{w})} \equiv \frac{\hat{\beta}_j(\hat{w}) - \beta_{j0}}{\text{Var}(\hat{\beta}_j(\hat{w}))}^{1/2},
\]
where the estimated variance $\hat{\text{Var}}(\hat{\beta}_j(\hat{w}))$ is the $j^{th}$ element on the diagonal of $\hat{\text{Var}}(\hat{\beta}(\hat{w}))$:

$$\hat{\text{Var}}(\hat{\beta}(\hat{w})) = \sum_{m=1}^{M} \sum_{s=1}^{M} \hat{\omega}^{(m)}(\hat{\omega})^{(s)} \left( \hat{d}^{(m)} \left( \hat{d}^{(s)} \right) \right)^\top$$

$$+ \hat{\sigma}^2(\hat{w}) \Gamma^{(m)} \left( X^{(m)} \right)^\top \left( X^{(m)} \right)^{-1} X^{(s)} \left( X^{(s)} \right)^\top \left( X^{(s)} \right)^{-1} \Gamma^{(s)} \left( \right)^\top,$$

where $\hat{d}^{(m)}$ and $\hat{d}^{(s)}$ are defined in (2.17) and $\hat{\sigma}^2(\hat{w})$ is defined in (2.13). To test the joint null hypothesis that $R\beta_c(\hat{w}) = r$, where $r$ is $k_r \times 1$, the Wald statistic is

$$W_r \equiv \left( R\hat{\beta}_c(\hat{w}) - r \right)^\top \left( R\hat{\text{Var}}(\hat{\beta}_c(\hat{w})) R^\top \right)^{-1} \left( R\hat{\beta}_c(\hat{w}) - r \right).$$

The asymptotic distributions of the $t$-statistic and of the Wald statistic are given in the following theorem.

**Theorem 2.4.** Let Assumptions 2.1–2.6 hold. Then

$$t_{\hat{\beta}_j(\hat{w})} \rightarrow d N(0, 1)$$

and

$$W_r \rightarrow d \chi^2(k_r).$$

The above theorem follows the findings in Johansen (1995, pp177–178). The intuition is that if the conditional distribution does not depend on $\hat{w}$, it also holds marginally. Note that we can use this theorem to test hypotheses by computing $P$ values. But we cannot construct confidence intervals in the conventional way by inverting the test statistics. This is because the upper and lower limits will include the random vector $\hat{w}$ through the standard error.

Since both test statistics are asymptotically pivotal, a semiparametric bootstrap test can be used to provide improved statistical inference in finite samples. Denote $\hat{t}$ as an estimated test statistic ($t$ or Wald). We first compute the estimated residuals $\hat{u}$ by
plugging in the estimated average coefficients $\tilde{\beta}(\tilde{w})$ under the null hypothesis. Then, the estimated residuals $\tilde{u}$ are resampled $B$ times. We use the resampled residuals to obtain $B$ bootstrap samples $y^*_l$ for $l = 1, \ldots, B$. The value of $B$ should satisfy the condition that $\alpha(1 + B)$ is an integer, where $\alpha$ is the desired level of significance. For each bootstrap sample $y^*_l$, we compute a simulated test statistic $\hat{\tau}^*_l$ in exactly the same way that $\hat{\tau}$ was computed from the original data.

Following MacKinnon (2009), we can use the following equation to compute the bootstrap $P$ value for a one-tail test that rejects in the upper tail, as is the case in the Wald test:

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B} \sum_{l=1}^{B} I(\hat{\tau}^*_l > \hat{\tau}),$$

where $I(\cdot)$ is the indicator function, which takes the value 1 when its argument is true and takes the value 0 otherwise. If we assume that $\tau$ is symmetrically distributed around zero, as is the case for some tests, we can use the symmetric bootstrap:

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B} \sum_{l=1}^{B} I(|\hat{\tau}^*_l| > |\hat{\tau}|).$$

If we are not willing to make the assumption that $\tau$ is symmetrically distributed around zero, we can instead use the equal-tail bootstrap:

$$\hat{p}^*(\hat{\tau}) = 2 \min \left( \frac{1}{B} \sum_{l=1}^{B} I(\hat{\tau}^*_l \leq \hat{\tau}), \frac{1}{B} \sum_{l=1}^{B} I(\hat{\tau}^*_l > \hat{\tau}) \right).$$

Bootstrap tests generally perform better than asymptotic tests, especially when working with a small sample size (see Section 2.4.2 as an example).

Computational cost can be a concern for model averaging estimation when the
2.4. FINITE SAMPLE PERFORMANCE

The total number of approximation models is large. In this situation, we do not want to assign a huge number for $B$. In fact, it is often possible to obtain reliable results without using a large value of $B$ by using the iterative procedure proposed in Davidson and MacKinnon (2000).

2.4 Finite Sample Performance

This section contains two parts. In the first part, I investigate the finite sample performance of the MAPC estimator in a simulation experiment. In the second part, I compare the finite sample performance of different tests under the same simulation design used in the first part. Contrary to the design in Hansen (2007), I propose a simulation experiment with non-nested models.

2.4.1 The MAPC Estimator Versus Other Estimators

The general unrestricted model is the regression model

$$ y = X\beta + u, \quad (2.19) $$

where $X$ is $n \times k$, $\beta$ is $k \times 1$ and $u$ is $n \times 1$. The number of regressors $k$ increases as $n$ increases but at a slower rate, in which $k = \text{round}(3 + n^{1/5})$. The first column of $X$, $x_1$, is the intercept; the remaining $x_i$ are assumed to be correlated with each other and are generated by $N(0, \Sigma)$, where $\Sigma$ is a $(k - 1) \times (k - 1)$ symmetric matrix with all diagonal terms equal to 1 and all off-diagonal terms equal to $\rho$. The coefficients of $\beta$ are determined by $\beta = [1/5, 1/5, 5/\sqrt{n}, 4/\sqrt{n}, ..., (8 - k)/\sqrt{n}]^\top$, where the first two regressors are core regressors and the remaining regressors are potential regressors. The error term $u$ is independent of the regressors $X$ and is distributed as $N(0, \sigma_u^2 I)$. 
The parameter $\sigma_u$ controls the population $R^2 = \beta_2^\top \Sigma \beta_2 / (\beta_2^\top \Sigma \beta_2 + \sigma_u^2)$ so as to vary on a grid between 0.01 and 0.99, where $\beta_2 = [\beta_2, ..., \beta_k]^\top$. I consider six different sample sizes $n = 25, 50, 100, 200, 400$ and 800. Other simulation results, which are not reported here, demonstrate that the findings are not sensitive to alternative distributions.

The total number of approximation models, $M$, is equal to the total number of combinations made by all the potential regressors. Therefore, the approximation models are clearly non-nested in our experiment.

I study seven methods: (1) the general-to-specific approach (GETS); (2) the Akaike information criterion (AIC); (3) model averaging by AIC (MA-AIC); (4) model averaging by Bayesian information criterion (MA-BIC); (5) Mallows model averaging (MMA); (6) jackknife model averaging (JMA); and (7) model averaging by prediction criterion (MAPC).

The GETS approach aims to modify the general unrestricted model by removing irrelevant variables according to pre-determined criteria. In this experiment, I adopt a simple GETS approach from Hendry and Nielsen (2007). We first estimate model (2.19) and conduct the $t$-test on each regressor that the corresponding coefficient equals to zero. Then, the regressor with the absolute value of the $t$-statistic smaller than 2 is eliminated. If multiple $t$-statistics are smaller than 2, only the smallest one is eliminated. The remaining regressors are retained and form a new model for the next-round test until no regressors can be eliminated.

The Akaike information criterion (AIC) for a model $m$ is defined as

$$AIC^{(m)} = n \log(\hat{\sigma}_m^2) + 2k^{(m)}.$$
The model that achieves the lowest value among all of the estimated AIC\(^{(m)}\) is selected. Following Buckland et al. (2007), the model average AIC technique (MA-AIC) makes use of the estimated AIC\(^{(m)}\) to compute the empirical weights, where

\[
\hat{w}_{AIC}^{(m)} = \exp\left(-\frac{1}{2}AIC^{(m)}\right) / \sum_{m=1}^{M} \exp\left(-\frac{1}{2}AIC^{(m)}\right).
\]

Buckland et al. (2007) also proposed the model averaging BIC (MA-BIC) technique which computes the empirical weights for its associated average estimator according to

\[
\hat{w}_{BIC}^{(m)} = \exp\left(-\frac{1}{2}BIC^{(m)}\right) / \sum_{m=1}^{M} \exp\left(-\frac{1}{2}BIC^{(m)}\right),
\]

where

\[
BIC^{(m)} = n \log(\hat{\sigma}_{m}^{2}) + \log(n)k^{(m)}.
\]

Jackknife model averaging (JMA) (Hansen and Racine 2012) is also known as leave-one-out cross-validation model averaging. As its name indicates, JMA requires the use of the jackknife residuals for the average estimator. The jackknife residual vector for model \(m\) can be conveniently written as \(\hat{u}_{J}^{(m)} = D^{(m)}\hat{u}^{(m)}\), where \(\hat{u}^{(m)}\) is the least squares residual vector and \(D^{(m)}\) is the \(n \times n\) diagonal matrix with the \(i^{th}\) diagonal element equal to \((1 - h_{i}^{(m)})^{-1}\). The term \(h_{i}^{(m)}\) is the \(i^{th}\) diagonal element of \(P^{(m)}\) defined in (2.4). Define an \(n \times M\) matrix that collects all the jackknife residuals, in which \(\hat{U}_{J} = [\hat{u}_{J}^{(1)}, ..., \hat{u}_{J}^{(M)}]\). The least squares cross-validation criterion for JMA is simply

\[
CV_n(\hat{w}) = \frac{1}{n} \hat{w}^{\top} \hat{U}_{J}^{\top} \hat{U}_{J} \hat{w} \quad \text{with} \quad \hat{w} = \arg\min_{\hat{w} \in \mathcal{H}_{M}} CV_n(\hat{w}).
\]

The MAPC estimator and the MMA estimator are presented in previous sections.
The unknown $\sigma^2$ in (2.11) is replaced by a sample estimate, $\hat{\sigma}_L^2$, from the largest approximation model (2.19).

Define the risk of an estimator as the mean squared error such that

$$\text{Risk} \equiv \frac{1}{n} (\mu(\hat{w}) - \mu_0)^\top (\mu(\hat{w}) - \mu_0),$$

where $\mu(\hat{w})$ is the average $\mu$ for the estimated $\hat{w}$ and $\mu_0$ is the true value of $\mu$ (feasible in simulation). We compute the risk for all seven estimators and average across 100,000 simulation draws. Risk for each estimator is normalized by dividing by the risk of the infeasible optimal least squares estimator (the OLS estimator of model (2.19)). Risk calculations for $n = 25$ with two different values of $\rho$ (0.1 and 0.5) are presented in Figures 2.1(a) and 2.1(b). The $R^2$ is presented on the horizontal axis and the risk is displayed on the vertical axis. The dash-dotted line, dashed line, solid line, star, cross, circle, and x-mark correspond to MMA, JMA, MAPC, GETS, AIC, MA-AIC, and MA-BIC, respectively.

Figures 2.1(a) and 2.1(b) show similar results. As $R^2 \to 1$, the calculated risk of each method tends to converge to 1, which suggests that all estimators are converging to the infeasible optimal least squares estimator. In many cases, the GETS method shows a much higher risk relative to other methods. When $R^2$ is close to 1, AIC generates the lowest risk. This means that the true model (2.19) is frequently selected by the AIC method when there is sufficient information. However, AIC has a relatively poor performance for other values of $R^2$. The MA-AIC estimator achieves lower risk than the MA-BIC estimator in most cases, except when $R^2$ is close to 0. In extreme cases, when $R^2$ is close to 0 or 1, MA-AIC and MA-BIC generate lower risk than MMA, JMA, and MAPC. However, they are outperformed by MMA, JMA, and
MAPC for values of $R^2$ that are more reasonable in practice. It is instructive to compare the performance of MMA, JMA, and MAPC. In most cases, these three methods have better performance when compared to other methods. I provide Figures 2.1(c) and 2.1(d) to show only these three methods. For all values of $R^2$, the MAPC estimator achieves lower risk than the MMA estimator and the

Figure 2.1: Simulation Results for $n = 25$
JMA estimator. This suggests that, in finite samples, MAPC is more efficient than MMA and JMA.

Notice that all the risk curves are hump-shaped. As previously mentioned, all the risk curves are normalized by dividing by the risk of the infeasible optimal OLS estimator. The hump shape is caused by this normalization. To demonstrate the reason for this, I use the MAPC estimator as an example and present its estimated risk without normalization (denoted by solid lines) in Figure 2.2, along with the estimated risk of the optimal OLS estimator (denoted by dashed lines).

In Figure 2.2, part (a) represents the estimated risk for the MAPC estimator and the OLS estimator. When $R^2$ is low, both estimators generate high risk. A low $R^2$ means that the model does not fit the data very well. As a consequence, the estimated risk for the OLS estimator is high. The MAPC estimator, however,

Figure 2.2: Estimated Risk without Normalization
2.4. FINITE SAMPLE PERFORMANCE

considers other approximation models as opposed to limiting itself to the unrestricted model (2.19). Therefore, the MAPC estimator generates smaller risk when $R^2$ is low. The corresponding normalized risk for MAPC is smaller than 1. As $R^2$ increases, the risk decreases for both estimators and the gap between the two curves shrinks. This implies that when $R^2$ is low, the normalized risk for the MAPC estimator increases as $R^2$ increases from 0.

Part (b) of Figure 2.2 is the truncated version of part (a). The vertical axis has the range of 0 to 2. The horizontal axis has the range of 0.2 to 1. As $R^2$ increases, the estimated risk for the OLS estimator decreases and eventually becomes smaller than the MAPC risk. Therefore, the corresponding normalized risk for MAPC is greater than 1. The gap between the two curves first expands and then shrinks. The two methods yield almost identical risk when $R^2 = 0.99$. This implies that the normalized risk curve is hump-shaped and that the normalized risk for the MAPC estimator is very close to 1 when $R^2 = 0.99$.

Risk calculations for various sample sizes ($n = 25, 50, 100, 200, 400,$ and 800) are displayed in Figure 2.3. In this case, I set $\rho = 0.1$. To keep the figure uncluttered, only MMA, JMA, and MAPC are displayed and are represented by the dash-dotted line, dashed lines, and solid line, respectively. In each panel, the MAPC estimator has a better performance than the MMA and JMA estimator. As $n$ increases, the index on the vertical axis shrinks and MMA and JMA merge with MAPC. This implies that as $n \to \infty$ these three estimators converge to the infeasible optimal least squares estimator. Notice that as $n$ increases, the hump shifts from the right to the left. Again, this is caused by the normalization. As $n$ becomes larger, the performance of the OLS estimator also gets better and generates lower risk. Therefore, the normalized
2.4. FINITE SAMPLE PERFORMANCE

Figure 2.3: Simulation Results for Various Sample Sizes

![Simulation Results for Various Sample Sizes](image)
risk increases with $n$ when $R^2$ is low.

### 2.4.2 Rejection Frequencies of Tests

Buckland *et al.* (1997) proposed using a pairs bootstrap method to estimate the variance of an average coefficient. To implement this method, we first resample the pairs of regressand and regressors $(y_i, x_i)$ and obtain $B$ bootstrap resamples. Then, we apply model averaging to each resample and obtain $B$ bootstrap estimates. Finally, we estimate the sample variance of the $B$ bootstrap estimates as the estimated variance of the average coefficient. We can conduct an asymptotic $t$-test using the pairs bootstrap sample variance. In this subsection, I refer to this test as “pairs-test”.

In Section 2.3.3, I derived the asymptotic $t$ test and the bootstrap test for the average core coefficient. In this section, I compare the finite sample performance of these tests with that of the pairs-test via the rejection frequency. I use the same simulation design in Section 2.4.1. Tests are conducted on the second coefficient, $\beta_2$, and the null hypothesis is that $\beta_2 = 1/5$, which is the true value of $\beta_2$. I generate 10,000 simulation draws for 15 sample sizes: $n = 25, 30, 35, ..., 95$. For each simulation draw, the $P$ values are computed from the asymptotic $t$ test, the semiparametric bootstrap test and the pairs-test for each sample size. In order to reduce the costs of doing the Monte Carlo experiments, we set $B = 199$ for each bootstrap test and pairs-test. I record the frequencies with which tests based on the three $P$ values reject at the 0.05 level. Four different $R^2$ are considered: 0.25, 0.5, 0.75, and 0.99. The results of this simulation are presented in Figure 2.4.

In Figure 2.4, the asymptotic $t$ test, the bootstrap test, and the pairs-test are represented by the dashed line, solid line, and dash-dotted line, respectively. The
results suggest that both the asymptotic test and the pairs-test tend to overreject severely when $n$ is small, although their performance improves quickly as $n$ increases. In contrast, the bootstrap test overrejects only very slightly on average for all values of $R^2$. The rejection frequencies are always very close to 0.05. Overall, the bootstrap test outperforms the asymptotic test and the pairs-test, especially when the sample size is small. One drawback of the bootstrap test is that it requires more computation.
time than the asymptotic test.

Simulation results also suggest that the asymptotic test and the pairs-test have similar performances. The performance of both tests are sensitive to different values of $R^2$. When $R^2 = 0.25$, rejection frequencies generated by the pairs-test are closer to 0.05 compared with the asymptotic test. As $R^2$ increases, the performance of both tests improves. When $R^2 = 0.75$, both tests generate similar rejection frequencies for most values of $n$. In the extreme case of $R^2 = 0.99$, the asymptotic test has a better performance than the pairs test by generating rejection frequencies closer to 0.05 for most values of $n$.

2.5 Conclusion

The MAPC estimator computes the weighted average across all approximation models. In doing so, it reduces the chance of a poor model being selected. I prove that the MAPC estimator is asymptotically optimal in the sense of achieving the lowest possible mean squared error. For statistical inference, I derived asymptotic tests for single hypotheses and joint hypotheses on the average coefficients for the core regressors. The simulations suggested that the MAPC estimator is more efficient than many existing estimators. The improvement is very significant when sample sizes are small. The MAPC, MMA, and JMA estimators tend to converge to the same limit asymptotically. The asymptotic tests work well in large sample sizes and bootstrap tests are highly recommended for small sample sizes.
Chapter 3

Empirical Applications

3.1 Introduction

In this chapter, I use the MAPC estimator to examine model uncertainty in cross-country economic growth regressions. Model uncertainty has been recognized as a problem in cross-country growth analysis for the past few decades. Levine and Renelt (1992) used model selection methods to analyze the robustness of the key economic, political, and institutional variables in long-run economic growth analyses.

Recent literature applies model averaging methods, most frequently Bayesian model averaging, in growth regressions; see, for example, Brock and Durlauf (2001) and Ley and Steel (2009). One important aspect regarding BMA’s sensitivity is the choice of priors. Recent work by Pesaran et al. (2009) and Eicher et al. (2011) has shown that BMA growth applications can be highly sensitive to the choice of priors.

Two empirical applications are considered in this chapter. In Section 3.2, I revisit the economic growth models in Barro (1991). In Section 3.3, I apply the MAPC estimator to the models in Durlauf, Kourtellos, and Tan (2008). Section 3.4 concludes this chapter.
3.2 Cross-Country Economic Growth Models

In this section, I apply the MAPC estimator to the economic growth models in Barro (1991), which proposed a group of non-nested approximation models to analyze the relation between economic growth rate and a number of variables in a cross section of countries. The models considered by Barro (1991) are a small subset of the models I consider in this subsection. The main findings in Barro (1991) are summarized in the following table.

Table 3.1: A Summary of Barro’s (1991) Findings

- The growth rate of real per capita GDP (GR6085) is positively related to the initial human capital$^a$
- GR6085 is negatively related to the initial (1960) level of real per capita GDP (GDP60)
- GR6085 is negatively related to the share of government consumption in GDP ($g^c/y$)
- GR6085 is positively related to the share of total investment on GDP ($i/y$)
- GR6085 is negatively related to the measures of political instability$^b$
- GR6085 is negatively related to the proxy for market distortions$^c$

$^a$ Proxied by 1960 primary (PRIM60) and secondary (SEC60) school enrollment rates.
$^b$ Proxied by the number of revolutions and coups per year (REV) and the number per million population of political assassinations per year (AS).
$^c$ Proxied by the magnitude of PPPI60 deviation (PPI60DEV), where PPPI60 is the 1960 PPP ratio based on the investment deflator.

I create my data set based on the Barro and Lee (1994) updated growth data for 98 countries. The dependent variable is the growth rate of real per capita GDP from 1960 to 1985 (GR6085). There are 15 independent variables:

1. the constant term;
3.2. CROSS-COUNTRY ECONOMIC GROWTH MODELS

2. GDP60: 1960 value of real per capita GDP;

3. SEC60: 1960 secondary-school enrollment rate;

4. PRIM60: 1960 primary-school enrollment rate;

5. $g^c/y$: average from 1970 to 1985 of the ratio of real government consumption (exclusive of defense and education) to real GDP;

6. REV: number of revolutions and coups per year (1960-1985);

7. AS: number of assassinations per million population per year (1960-1985);

8. PPI60DEV: magnitude of the deviation of PPI60 from the sample mean, where PPI60 is the 1960 PPP value for the investment deflator (U.S. = 1.0);

9. GDP60SQ: square of GDP60;

10. RPRI: student-teacher ratio in primary schools in 1960;

11. RSEC: student-teacher ratio in secondary schools in 1960;

12. AFRICA: dummy variable for sub-Saharan Africa;

13. LAT.AMER: dummy variable for Latin America;

14. $i/y$: average from 1960-1985 of the ratio of real domestic investment to real GDP;

15. FERTNET: FERT $\times (1-MORT01)$, where FERT is the total fertility rate, average of 1965 and 1985 and MORT01 is the mortality rate for age 0 through 1, average of 1965 and 1985.
There are eight variables that are of primary interest (core variables, including a constant term) and the other seven variables are only of marginal interest (potential variables).

I first revisit Barro’s (1991) approach using my updated data set. There are

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Const.</td>
<td>0.0287</td>
<td>0.0304</td>
<td>0.0368</td>
<td>0.0355</td>
<td>0.0332</td>
<td>0.0218</td>
<td>0.0449</td>
<td>0.0438</td>
</tr>
<tr>
<td>GDP60</td>
<td>–0.0473</td>
<td>–0.0826</td>
<td>–0.0495</td>
<td>–0.0490</td>
<td>–0.0435</td>
<td>–0.0489</td>
<td>–0.0554</td>
<td>–0.0501</td>
</tr>
<tr>
<td>SEC60</td>
<td>0.0331</td>
<td>0.0352</td>
<td>0.0321</td>
<td>0.0306</td>
<td>0.0107</td>
<td>0.0213</td>
<td>0.0127</td>
<td>0.0013</td>
</tr>
<tr>
<td>PRIM60</td>
<td>0.0202</td>
<td>0.0237</td>
<td>0.0196</td>
<td>0.0187</td>
<td>0.0249</td>
<td>0.0071</td>
<td>0.0037</td>
<td>0.0110</td>
</tr>
<tr>
<td>$g_c / y$</td>
<td>–0.1071</td>
<td>–0.1102</td>
<td>–0.1060</td>
<td>–0.1047</td>
<td>–0.0800</td>
<td>–0.1104</td>
<td>–0.1071</td>
<td>–0.0853</td>
</tr>
<tr>
<td>REV</td>
<td>–0.0185</td>
<td>–0.0183</td>
<td>–0.0197</td>
<td>–0.0202</td>
<td>–0.0144</td>
<td>–0.0129</td>
<td>–0.0123</td>
<td>–0.0111</td>
</tr>
<tr>
<td>AS</td>
<td>–0.0441</td>
<td>–0.0430</td>
<td>–0.0440</td>
<td>–0.0428</td>
<td>–0.0242</td>
<td>–0.0401</td>
<td>–0.0375</td>
<td>–0.0254</td>
</tr>
<tr>
<td>PPI60DEV</td>
<td>–0.0074</td>
<td>–0.0080</td>
<td>–0.0071</td>
<td>–0.0073</td>
<td>–0.0082</td>
<td>–0.0045</td>
<td>–0.0051</td>
<td>–0.0061</td>
</tr>
<tr>
<td>GDP60SQ</td>
<td>–0.0377</td>
<td>0.1929</td>
<td>0.2267</td>
<td>0.02452</td>
<td>–0.0141</td>
<td>–0.0114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPRI</td>
<td>–0.1929</td>
<td>–0.2637</td>
<td>0.1770</td>
<td>–0.2452</td>
<td>–0.0114</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSEC</td>
<td>–2.267</td>
<td>0.036</td>
<td>0.0033</td>
<td>0.0033</td>
<td>0.0033</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFRICA</td>
<td>–0.0164</td>
<td>–0.0143</td>
<td>–0.0171</td>
<td>–0.0126</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAT.AMER</td>
<td>–0.0135</td>
<td>–0.0164</td>
<td>–0.0126</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i / y$</td>
<td>0.0986</td>
<td>0.0846</td>
<td>0.0640</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FERTNET</td>
<td>–0.0035</td>
<td>–0.0025</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$R^2$          | 0.4679  | 0.4780  | 0.4760  | 0.4829  | 0.6015  | 0.5624  | 0.5888  | 0.6578  |

* Values in parentheses are standard errors.
eight models that are summarized from Barro (1991). Table 3.2 shows the regression results for these models, which are labeled from [1] to [8]. The smallest model is model [1], which only contains the eight core variables, while the largest is model [8], which contains twelve variables (eight core variables and four potential variables). The table also reports the standard error associated with each estimated coefficient and the centered $R^2$ for each model.

One problem with Barro’s (1991) approach is the huge fluctuation of estimates (coefficients and $P$ values) for the core variables across different models using different potential variables. Take SEC60 as an example. The estimated coefficient in regression [2] is 0.0352 with a standard error of 0.0109, which means that SEC60 is highly significant. However, the estimated coefficient of the same variable drops to 0.0013 in regression [8] with a standard error of 0.0102, which implies that SEC60 is highly insignificant. By using different models (potential variables), one obtains contradictory results.

To solve this model uncertainty problem, I conduct model average estimation using the MAPC estimator. Following Barro (1991), I include the constant term, GDP60, SEC60, PRIM60, $g^c/y$, REV, AS, and PPI60DEV in every approximation model as the core variables and consider every possible combination of the potential variables. Since there are seven such variables, the total number of combinations is $2^7 = 128$. Therefore, 128 approximation models are included in the estimation.

In Table 3.3, I report the estimated average coefficients for the core variables and their standard errors. I also estimate the $P$ value for each coefficient equal to zero using the $t$-test derived in Chapter 2. The bootstrap $P$ value is also calculated for each average coefficient. I set $B = 9,999$ for all variables. The top eight models with
The results of the MAPC estimation provide significant evidence to support Barro’s (1991) findings from the perspective of model average estimation. All estimates yield the same signs as Barro’s, which implies similar relations between the growth rate and the core variables. The estimated coefficient for GDP60 is $-0.0713$, which indicates a negative relationship between the growth rate and the initial per capita GDP. Note that the magnitude of this estimate is larger than those estimated by Barro’s (1991) models (except model [2]). Both SEC60 and PRIM60 are proxies for the initial human capital. Positive signs on these proxies indicate a positive relation between the growth rate and the initial human capital. The estimate on $g^c/y$ is negative, which indicates a negative relation between the growth rate and the share of government consumption in GDP. Political instability, which is proxied by REV and AS, has a negative effect on the growth rate.

I also compare the relative out-of-sample predictive efficiency of the MAPC estimator with other estimators. The original data sample $n = 98$ is split into a training set of $n_1$ and an evaluation set of size $n_2 = n - n_1$. The seven methods (GETS, AIC,
Table 3.4: Top Eight Models in the MAPC Estimation

<table>
<thead>
<tr>
<th>Assigned Weights</th>
<th>0.0814</th>
<th>0.0795</th>
<th>0.0792</th>
<th>0.0626</th>
<th>0.0547</th>
<th>0.0522</th>
<th>0.0518</th>
<th>0.0504</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const.</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GDP60</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SEC60</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>PRIM60</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>$g^c/y$</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>REV</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>AS</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>PPI60DEV</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GDP60SQ</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RPRI</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RSEC</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AFRICA</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LAT.AMER</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>$i/y$</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>FERTNET</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

The “+” indicates that the corresponding parameters are included in the model and the “–” indicates the opposite.

MA-AIC, MA-BIC, MMA, JMA, and MAPC) in Chapter 2 are applied to the training set. The results are evaluated using the mean squared prediction error (MSPE):

$$\text{MSPE} = \frac{1}{n_2} (y_2 - X_2 \hat{\beta}_1)\top (y_2 - X_2 \hat{\beta}_1),$$

where $(y_2, X_2)$ is the evaluation set, $n_2$ is the number of observations of the evaluation set, and $\hat{\beta}_1$ is the vector of the estimated coefficients by a particular method based on the training set. For each parameterization, the result is normalized by the MSPE by the MAPC estimator. I repeat the entire procedure 10,001 times and report the median. I vary $n_1$ and consider $n_1 = 20, 30, ..., 80$. Table 3.5 reports the relative out-of-sample predictive efficiency. Entries larger than one indicate inferior performance relative to the MAPC estimator.
Table 3.5 suggests that the MAPC estimator delivers models that have better out-of-sample predictive efficiency than those of the six existing methods. As $n_1$ increases, the relative performance for all six methods improves. The MA-AIC method works better than MMA. As mentioned in Hansen and Racine (2012), the MMA method is sensitive to the preliminary estimate of $\sigma^2$ needed for its computation and relying on a “large” approximating model may not be sufficient to deliver efficient results in finite samples. In contrast, the MAPC estimator does not require a preliminary estimate.

### 3.3 The Robustness of Economic Growth Theories

Durlauf, Kourtellos and Tan (2008), hereafter DKT, investigated the strength of empirical evidence for various growth theories when there is model uncertainty with respect to the correct growth model. Using Bayesian model averaging, DKT’s results showed little empirical evidence that fundamental growth theories play an important role in explaining aggregate growth. In contrast, DKT’s results showed strong empirical evidence for macroeconomic policy effects and regional heterogeneity. Therefore, DKT concluded that the ability of cross-country growth regressions to adjudicate the relative importance of alternative growth theories is limited.
I apply the MAPC estimator to the DKT data set. It is an unbalanced panel data set over three periods: 1960-74 (53 countries), 1975-84 (54 countries), and 1985-94 (57 countries). Time dummy variables for each of these periods are also included. The dependent variable is the average growth rate of real per worker GDP corresponding to the three periods. The independent variables are divided into eight categories, each of which represents one growth theory:

1. **Neoclassical**: initial income, population growth rate, proxy for schooling, and investment to GDP ratio;

2. **Demography**: proxy for mortality rate and fertility rate;

3. **Macro Policy**: proxy for openness, government consumption, and inflation;

4. **Regional Heterogeneity**: dummy variables for East Asia, Africa, and Latin America;

5. **Religion**: Eastern, Hindu, Jewish, Muslim, Orthodox, Protestant, and other;

6. **Geography**: a geographic accessibility/isolation (LCR100km) and a climate variable (KGATRSTR);

7. **Fractionalisation**: language and ethnicity;

8. **Institutions**: risk of expropriation of private investments, constraints on executive power institute, an index for the quality of governance in 1996 (KKZ96), and an index of legal formalism (CHEQUE).

To examine if one theory is particularly important, I include the theory and, consequently, its included parameters in all approximation models as the core regressors.
3.3. THE ROBUSTNESS OF ECONOMIC GROWTH THEORIES

All other variables are considered as potential regressors. I conduct model averaging estimation using the MAPC estimator. The asymptotic and bootstrap $P$ values are computed for each included parameter. The total number of bootstraps is $B = 999$. This procedure is conducted for each growth theory. The results are collected in Table 3.6.

Table 3.6 shows the results of a series of estimations (time dummies are unreported). Columns 1 to 5 correspond to variables, coefficients, standard errors, asymptotic $P$ values, and bootstrap $P$ values, respectively. Many of the results are consistent with DKT’s findings. The neoclassical and macro policy theories are very significant as most of their parameters have very small $P$ values for both asymptotic and bootstrap values. Although the schooling parameter from the neoclassical theory is highly insignificant, this is probably caused by the fact that schooling is a weak proxy for human capital. Note that the same result regarding schooling was also concluded from DKT’s estimation. The demography, religion, geography, and institution theories are insignificant, as most of their parameters have high $P$ values.

Some of the results provide an alternative explanation to that outlined by DKT in regards to the significance of particular growth theories. The regional heterogeneity theory is highly significant according to DKT. In fact, the asymptotic $P$ values in the MAPC estimation show the same result. However, the bootstrap $P$ values show that the parameters from regional heterogeneity are either insignificant or only relatively significant. As I show in Section 2.4, bootstrap $P$ values are more robust with small sample sizes. Therefore, I believe that the bootstrap $P$ values are more convincing given the sample size of the DKT data set is 164, which is not large. On the other hand, both the asymptotic and the bootstrap $P$ values for fractionalisation
Table 3.6: MAPC Estimates with Asymptotic and Bootstrap $P$ Values

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Asymptotic $P$</th>
<th>Bootstrap $P$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neoclassical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial In.</td>
<td>-0.0169</td>
<td>0.0027</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Pop.Rate</td>
<td>-0.0300</td>
<td>0.0122</td>
<td>0.0138</td>
<td>0.0601</td>
</tr>
<tr>
<td>Schooling</td>
<td>0.0022</td>
<td>0.0029</td>
<td>0.4506</td>
<td>0.5696</td>
</tr>
<tr>
<td>Investments</td>
<td>0.0123</td>
<td>0.0028</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td><strong>Demography</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/Life Expec.</td>
<td>-0.0066</td>
<td>0.0145</td>
<td>0.6508</td>
<td>0.7878</td>
</tr>
<tr>
<td>Fertility</td>
<td>-0.0079</td>
<td>0.0072</td>
<td>0.2722</td>
<td>0.6116</td>
</tr>
<tr>
<td><strong>Macro Policy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Openness</td>
<td>0.0136</td>
<td>0.0072</td>
<td>0.0571</td>
<td>0.0801</td>
</tr>
<tr>
<td>G.C.</td>
<td>-0.1224</td>
<td>0.0279</td>
<td>0.0000</td>
<td>0.0170</td>
</tr>
<tr>
<td>Inflation</td>
<td>-0.0122</td>
<td>0.0050</td>
<td>0.0145</td>
<td>0.0280</td>
</tr>
<tr>
<td><strong>Regional Hetero.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East Asia</td>
<td>0.0141</td>
<td>0.0050</td>
<td>0.0051</td>
<td>0.1031</td>
</tr>
<tr>
<td>Africa</td>
<td>-0.0133</td>
<td>0.0060</td>
<td>0.0253</td>
<td>0.1552</td>
</tr>
<tr>
<td>Latin</td>
<td>-0.0037</td>
<td>0.0042</td>
<td>0.3792</td>
<td>0.6216</td>
</tr>
<tr>
<td><strong>Religion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East</td>
<td>0.0030</td>
<td>0.0075</td>
<td>0.6927</td>
<td>0.8769</td>
</tr>
<tr>
<td>Hindu</td>
<td>0.0149</td>
<td>0.0126</td>
<td>0.2361</td>
<td>0.3784</td>
</tr>
<tr>
<td>Jewish</td>
<td>0.0098</td>
<td>0.0114</td>
<td>0.3890</td>
<td>0.4434</td>
</tr>
<tr>
<td>Muslim</td>
<td>-0.0060</td>
<td>0.0056</td>
<td>0.2819</td>
<td>0.5135</td>
</tr>
<tr>
<td>Ortho.</td>
<td>-0.0061</td>
<td>0.0094</td>
<td>0.5182</td>
<td>0.6106</td>
</tr>
<tr>
<td>Prote.</td>
<td>-0.0096</td>
<td>0.0053</td>
<td>0.0703</td>
<td>0.1161</td>
</tr>
<tr>
<td>Other</td>
<td>-0.0173</td>
<td>0.0085</td>
<td>0.0407</td>
<td>0.1301</td>
</tr>
<tr>
<td><strong>Geography</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCR100km</td>
<td>0.0018</td>
<td>0.0045</td>
<td>0.6900</td>
<td>0.7878</td>
</tr>
<tr>
<td>KGATRSTR</td>
<td>-0.0089</td>
<td>0.0049</td>
<td>0.0703</td>
<td>0.3333</td>
</tr>
<tr>
<td><strong>Fractionalisation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>-0.0212</td>
<td>0.0063</td>
<td>0.0008</td>
<td>0.0030</td>
</tr>
<tr>
<td>Ethnic</td>
<td>-0.0144</td>
<td>0.0060</td>
<td>0.0163</td>
<td>0.0531</td>
</tr>
<tr>
<td><strong>Institutions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>0.0041</td>
<td>0.0125</td>
<td>0.7427</td>
<td>0.7658</td>
</tr>
<tr>
<td>Constraints</td>
<td>-0.0097</td>
<td>0.0047</td>
<td>0.0402</td>
<td>0.0470</td>
</tr>
<tr>
<td>KKZ96</td>
<td>0.0046</td>
<td>0.0038</td>
<td>0.2212</td>
<td>0.2703</td>
</tr>
<tr>
<td>CHEQUE</td>
<td>-0.0132</td>
<td>0.0093</td>
<td>0.1568</td>
<td>0.1842</td>
</tr>
</tbody>
</table>

low, which implies that the fractionalisation theory is supported by strong empirical evidence when there is model uncertainty with respect to the correct growth model.
3.4 Conclusion

I considered two empirical applications in this chapter. I revisited the economic growth models in Barro (1991) in the first application and Durlauf, Kourtellos, and Tan (2008) in the second application. The eight models from Barro (1991) are a small subset of the 128 models in the MAPC estimation. The approach of considering every possible approximation model helps avoid the risk of omitting important variables or retaining irrelevant variables. As for the DKT case, I applied the MAPC estimator as an alternative method to the same data set. Many of my results are consistent with DKT's findings, and some of my results provide an alternative explanation to those outlined by DKT.
Chapter 4

Constructing Optimal Instruments for IV Estimation by Model Averaging

4.1 Introduction

This chapter considers the situation in which there are many potential instruments for IV estimation. I propose using model averaging to construct optimal instruments for IV estimation when there are many potential instrument sets. A classic model averaging estimator is a weighted average of common estimates obtained from approximating models. The performance of such an estimator necessarily depends on the choice of weights. I construct an averaged projection matrix, which is a weighted sum of the projection matrices by all potential instrument sets, in which the weights are computed by the model averaging IV (MAIV) criterion through convex optimization. Then, I compute the IV estimator based on the averaged projection matrix of the instruments. The instrument sets in this method can be nested or non-nested.

The rest of the chapter is organized as follows. Section 4.2 sets up the basic framework for IV estimation and reviews existing methods. Section 4.3 proposes the MAIV criterion and introduces a computational algorithm for easy calculation. In
Section 4.4, I first introduce a new loss function for evaluating the performance of the IV estimator by different instruments. I then prove the asymptotic optimality of the MAIV estimator by demonstrating that it achieves the lowest possible value of the loss function asymptotically. Section 4.5 presents the results of the simulation experiments. Section 4.6 concludes the chapter.

4.2 Model and Estimation

Let \((y_i, x_i)\) be a random sample for \(i = 1, \ldots, n\). Assume the data generating process is

\[ y_i = \mu_i + u_i, \]

where \(\mu_i = \sum_{j=1}^{k} \beta_j x_{ij}\) and \(k\) is a finite number. Unlike those from OLS models, the error terms are not innovations. In this model, the error terms are correlated with some of the regressors for the same observation. Assume that an information set \(\Omega_i\) exists such that

\[ \mathbb{E}(u_i|\Omega_i) = 0 \quad \text{and} \quad \mathbb{E}(u_i^2|\Omega_i) = \sigma^2. \]

Any regressor of which the value in period \(i\) is correlated with \(u_i\) cannot belong to the information set \(\Omega_i\).

The structural equation can be written as

\[ y = Z\beta_1 + Y\beta_2 + u = X\beta + u = \mu + u, \quad (4.1) \]

where \(X \equiv [Z \ Y]\) is an \(n \times k\) matrix with \(Z\) being an \(n \times k_1\) matrix of observations on exogenous variables, \(Y\) being an \(n \times k_2\) matrix of observations on endogenous variables, and \(\beta \equiv [\beta_1^{\top} \ \beta_2^{\top}]^{\top}\). The rest of the system is treated as an unrestricted
reduced form. For an element in $Y$, for example $Y_{ij}$, we have

$$Y_{ij} = \sum_{s=1}^{g} W_{is} \Pi_{sj} + V_{ij},$$

where $g$ is a finite number. In matrix form, we have

$$Y = W\Pi + V,$$

(4.2)

where the instruments $W \equiv [Z \ W_2]$. The errors in (4.1) and (4.2) are assumed I.I.D., with expectation zero conditional on $W$, and nonzero correlation between the errors for the same observation.

For model (4.2), consider a group of $q = 1, 2, ..., Q$ approximation models, where the $q^{th}$ model uses instruments $W^{(q)}$. Let the rank of $W^{(q)}$ be $l^{(q)}$ for $q = 1, 2, ..., Q$. The IV estimator using instruments $W^{(q)}$ is

$$\hat{\beta}_{IV}^{(q)} = (X^\top P_{W^{(q)}} X)^{-1} X^\top P_{W^{(q)}} y,$$

(4.3)

where $P_{W^{(q)}} = W^{(q)} \left(W^{(q)^\top} W^{(q)}\right)^{-1} W^{(q)^\top}$. The IV estimator (4.3) is commonly known as the two-stage least squares (2SLS) estimator. As its name indicates, the 2SLS estimator can be calculated in two stages using OLS regressions. In the first stage, each column $x_i$ of $X$ is regressed on $W^{(q)}$ such that

$$x_i = W^{(q)} \pi_i^{(q)} + \text{error},$$

(4.4)

where $\pi_i^{(q)}$ is the $l^{(q)} \times 1$ coefficient vector corresponding to instruments $W^{(q)}$ and regressors $x_i$. The fitted values are collected from model (4.4) to form the matrix
Then the second-stage regression

\[ y = P_{W^{(q)}} X \beta^{(q)} + \text{error} \]

is used to obtain (4.3).

Donald and Newey (2001), henceforth DN, derived a criterion based on the first-stage regression model of 2SLS to select the optimal instruments. The selection is based on minimizing a loss function that takes the form of the approximate MSE of a linear combination \( \lambda^\top \hat{\beta}_{IV} \), where \( \lambda \) is a vector of predetermined linear combination parameters.\(^1\) Let \( \hat{\beta}_{IV} \) be a preliminary IV estimator using instruments \( \hat{W} \), \( \hat{u} = y - X \hat{\beta}_{IV} \) and \( \hat{H} = X^\top P_{\hat{W}} X / n \). Also, let \( \hat{\epsilon} = (I - P_{\hat{W}}) X \) denote a preliminary reduced form residual estimator. Then, let \( \hat{\epsilon}_\lambda = \hat{\epsilon} \hat{H}^{-1} \lambda \) and we have

\[ \hat{\sigma}_u^2 = \hat{u}^\top \hat{u} / n, \quad \hat{\sigma}_\lambda^2 = \hat{\epsilon}_\lambda^\top \hat{\epsilon}_\lambda / n, \quad \hat{\sigma}_{\lambda u}^2 = \hat{\epsilon}_\lambda^\top \hat{u} / n. \]

It is important that none of these preliminary estimators depend on \( q \). Recall that \( l^{(q)} \) denotes the rank of the instruments \( W^{(q)} \). The DN criterion can be written as

\[ \hat{S}_{\lambda}^{(q)} = \hat{\sigma}_{\lambda u}^2 \frac{(l^{(q)})^2}{n} + \hat{\sigma}_u^2 \left( \hat{H}_{\lambda}^{(q)} - \hat{\sigma}_\lambda^2 \frac{l^{(q)}}{n} \right), \quad (4.5) \]

\(^1\)If we set \( \lambda \) to be a special vector whereby the \( i^{th} \) element equals 1 and all other elements equal 0, the loss function is equivalent to the MSE of the \( i^{th} \) coefficient.

\(^2\)For example, \( \hat{\beta}_{IV} \) can be an IV estimator with only as many instruments as right-hand-side variables.
where $\hat{R}_\lambda^{(q)}$ is the Mallows criterion\footnote{In Donald and Newey (2001). The authors also proposed a cross-validation criterion for $\hat{R}_\lambda(q)$. The differences caused by the two criteria are negligible in our simulation. Therefore, in this chapter, only the Mallows criterion is presented in detail.}

$$
\hat{R}_\lambda^{(q)} = \frac{(\hat{\epsilon}_\lambda^{(q)})^\top (\hat{\epsilon}_\lambda^{(q)})}{n} + \hat{\sigma}_\lambda^2 \left( \frac{2l^{(q)}}{n} \right)
$$

with $\hat{\epsilon}_\lambda^{(q)}$ being the estimated $\epsilon_\lambda$ by the $q^{th}$ auxiliary instrument set. The $q$ that minimizes the criterion (4.5) corresponds to the selected optimal instrument sets.

Kuersteiner and Okui (2010), henceforth KO, extended the selection criterion (4.5) to a model averaging criterion. Let $\mathbf{w} = [w^{(1)}, ..., w^{(Q)}]^\top$ be a weight vector in the unit simplex in $\mathbb{R}^Q$. One should not confuse the weight vector $\mathbf{w}$ with the instruments $\mathbf{W}$, for the latter is denoted by capital $\mathbf{W}$. If the weights are restricted on the unit simplex,

$$
H_Q^w \equiv \left\{ \mathbf{w} \in [0, 1]^Q : \sum_{q=1}^Q w^{(q)} = 1 \right\},
$$

then the KO criterion function is

$$
\hat{S}_\lambda(\mathbf{w}) = \hat{\sigma}_u^2 \mathbf{w}^\top \mathbf{l}^2/n + \hat{\sigma}_u^2 \left( \frac{\mathbf{w}^\top \hat{U}_\lambda \mathbf{w} - \hat{\sigma}_\lambda^2 (Q - 2\mathbf{w}^\top \mathbf{l} + \mathbf{w}^\top \Gamma \mathbf{w})}{n} \right).
$$

The $\hat{\mathbf{w}}^{opt}$ that minimizes (4.7) with constraints (4.6) is the estimated optimal weight vector for auxiliary instrument sets. KO also considered negative weights on each auxiliary instrument set. Define $H_Q^u \equiv \left\{ \mathbf{w} \in [-\infty, \infty]^Q : \sum_{q=1}^Q w^{(q)} = 1 \right\}$. The set $H_Q^u$ is unconstrained and, therefore, not compact. Following KO, from a finite sample point of view, it may be useful to constrain the weights $\mathbf{w}$ to lie in the compact set:

$$
H_Q^c \equiv \left\{ \mathbf{w} \in [-1, 1]^Q : \sum_{q=1}^Q w^{(q)} = 1 \right\}. \quad \text{Let } \hat{\mathbf{u}}_\lambda^{(q)} = (P_{W^{(u)}} - P_{W^{(q)}})X \hat{H}_\lambda^{-1} \lambda
$$
\[ \hat{U}_\lambda = [\hat{u}^{(1)}_\lambda, ..., \hat{u}^{(Q)}_\lambda]^\top [\hat{u}^{(q)}_\lambda, ..., \hat{u}^{(Q)}_\lambda], \]

where \( P_{W^{(\ell)}} \) is the projection matrix with all available instruments. Let \( \Gamma \) denote the \( Q \times Q \) matrix of which the \( ij^{th} \) element is \( \min(i, j) \), and let \( l = [l^{(1)}, l^{(2)}, ..., l^{(Q)}]^\top \). The KO model averaging criterion is

\[
\hat{S}_\lambda(w) = \frac{\hat{\sigma}_{\lambda u}^2 (w^\top l)^2}{n} + \left( \hat{\sigma}_{\lambda u}^2 + \hat{\sigma}_{\lambda}^2 \right) \frac{w^\top \Gamma w}{n} - \frac{w^\top \Gamma l}{n} \lambda^\top \tilde{H}^{-1} \hat{B}_n \tilde{H}^{-1} \lambda \\
+ \frac{\hat{\sigma}_{\lambda}^2}{n} \left( w^\top \hat{U}_\lambda w - \hat{\sigma}_{\lambda}^2 (Q - 2w^\top l + w^\top \Gamma w) \right),
\]

(4.8)

where \( \hat{B}_n \) is a complicated estimator:

\[
\hat{B}_n = 2 \left( \hat{\sigma}_{\epsilon u}^2 \hat{\Sigma}_\epsilon + k \hat{\sigma}_{\epsilon u} \hat{\sigma}_{\epsilon u}^\top + \frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{\sigma}_{\epsilon u} \tilde{H}^{-1} \hat{\sigma}_{\epsilon u} \hat{g}_i^\top \\
+ \frac{1}{n} \sum_{i=1}^n \left( \hat{g}_i \hat{\sigma}_{\epsilon u} \tilde{H}^{-1} \hat{g}_i + \hat{\sigma}_{\epsilon u} \hat{g}_i \tilde{H}^{-1} \hat{\sigma}_{\epsilon u} \hat{g}_i^\top \right) \right).
\]

Here \( \hat{g}_i = (P_{\tilde{W}} X)_i \), \( \hat{\Sigma}_\epsilon \) is an estimator for \( \Sigma_\epsilon = E(\epsilon_i \epsilon_i^\top) \) and \( \hat{\sigma}_{\epsilon u} = \hat{\epsilon}^\top \hat{u} / n \).

Both DN and KO depend on many preliminary estimates. These estimates are determined before calculating the criteria, which may result in arbitrariness. However, both papers showed the consistency of their estimators. Therefore, the effect of the preliminary estimates on both criteria should fade away as \( n \to \infty \). Another limitation of KO is that the instruments \( W^{(q)} \) must be an ordered set. Instrument sets with fewer elements should always nest within larger instrument sets.

Another issue with both methods is how to assign the predetermined linear combination parameters \( \lambda \) when researchers are not interested in one particular coefficient (or combinations of some coefficients) but in the prediction of the whole model. The parameters \( \lambda \) arise because both methods are designed to minimize an approximation
4.3. MODEL AVERAGING CRITERION FOR IV ESTIMATION

In this section, I propose a new method for constructing optimal instruments for IV estimation. For the first-stage regression (4.4), it may be convenient to use the
4.3. MODEL AVERAGING CRITERION FOR IV ESTIMATION

Following model instead of the reduced form (4.2)

\[
\begin{align*}
X &= W^{(q)}\Pi^{(q)} + V_0, \\
[Z \ Y] &= [Z \ W_2^{(q)}] \Pi^{(q)} + [0 \ V], \\
&= F^{(q)} + V_0, \\
\end{align*}
\]

(4.11)

where \(\Pi^{(q)} = [\pi_1^{(q)}, \pi_2^{(q)}, ..., \pi_l^{(q)}]\) is a \(l^{(q)} \times k\) matrix and \(F^{(q)} \equiv W^{(q)}\Pi^{(q)}\). Since the exogenous part of \(X\) is included in \(W\), the corresponding error terms in \(V_0\) must be zero. Assume

\[
E(V_0V_0^\top) = \Sigma_0 = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma \end{bmatrix},
\]

where \(E(VV^\top) = \Sigma\) and \(\Sigma\) is a finite, deterministic, positive definite \(l_2 \times l_2\) matrix.

The OLS estimate of \(\Pi^{(q)}\) is \(\hat{\Pi}^{(q)} = (W^{(q)}W^{(q)}\top)^{-1}W^{(q)}\top X\). Therefore,

\[
F^{(q)} = W^{(q)}\Pi^{(q)} = W^{(q)}\left(W^{(q)}\top W^{(q)}\right)^{-1}W^{(q)}\top X = P_{W^{(q)}X}.
\]

Model (4.11) is an approximation of the true model (4.2). Let \(w = [w^{(1)}, ..., w^{(Q)}]^\top\) be a weight vector in the unit simplex in \(\mathbb{R}^Q\). Since the weights are put on approximation models, the weight vector \(w\) cannot be negative in this method.

Define the model average estimator of \(F\) as

\[
F(w) = \sum_{q=1}^Q w^{(q)}W^{(q)}\hat{\Pi}^{(q)} = \sum_{q=1}^Q w^{(q)}P_{W^{(q)}X}.
\]
4.3. MODEL AVERAGING CRITERION FOR IV ESTIMATION

Following the definition of the weighted averaged projection matrix

\[ P(w) \equiv \sum_{q=1}^{Q} w^{(q)} P_{W^{(q)}}, \]

we have \( F(w) = P(w) X \). The averaged fitted value from model (4.11) is \( P(w) X \).

Define the averaged second-stage regression as

\[ P(w)y = P(w)X\beta + \text{error}. \]

By simple OLS, we obtain the averaged IV coefficient \( \hat{\beta}(w) \) as

\[ \hat{\beta}(w) = \left( X^\top P(w)P(w)X \right)^{-1} X^\top P(w)P(w)y. \] (4.12)

Note that the averaged projection matrix is not idempotent, which means that equation (4.12) cannot be further simplified.

To estimate the weight vector \( w \), I propose the model averaging IV criterion (MAIV):

\[ \text{MAIV}_n(w) \equiv \tilde{\beta}^\top \left( X - F(w) \right)^\top \left( X - F(w) \right) \tilde{\beta} \left( \frac{n + l(w)}{n - l(w)} \right), \] (4.13)

where \( \tilde{\beta} \equiv \left( X^\top P_{W^{(L)}} X \right)^{-1} X^\top P_{W^{(L)}} y \) is a preliminary estimator of \( \beta \) that uses all available instruments. The term \( l(w) \) is the effective number of parameters and is defined as \( l(w) \equiv \sum_{m=1}^{M} w^{(q)}l^{(q)} \). The empirical weight vector \( \hat{w} \) can be estimated
through minimizing MAIV such that

$$\hat{w} = \arg\min_{w \in H_Q} MAIV_n(w),$$  \hspace{1cm} (4.14)

where $H_Q$ is defined in (4.6). Since there is no closed-form solution to (4.14), we need to compute $\hat{w}$ numerically. I present the following algorithm for doing so:

1. Estimate $\tilde{\beta}$ using all available instruments.
2. Compute the $n \times 1$ vector $\hat{e}^{(q)} \equiv (X - \hat{F}^{(q)})\tilde{\beta}$ for each $q = 1, \ldots, Q$.
3. Put all $\hat{e}^{(q)}$ for $q = 1, \ldots, Q$ in the $n \times Q$ matrix $\hat{E} = \left[\hat{e}^{(1)}, \hat{e}^{(2)}, \ldots, \hat{e}^{(Q)}\right]$.
4. Construct a $Q \times 1$ vector $l = \left[l^{(1)}, l^{(2)}, \ldots, l^{(Q)}\right]^\top$, where $l^{(q)}$ is the rank of the $q^{th}$ instrument set.
5. Minimize the following function for $\hat{w}$ subject to the restriction (4.6):

$$MAIV_n(w) \equiv w^\top \hat{E}^\top \hat{E} w \left(\frac{n + l^\top w}{n - l^\top w}\right).$$  \hspace{1cm} (4.15)

This is a convex optimization problem.

### 4.4 Asymptotic Optimality

Before I prove the optimality of the MAIV estimator, a loss function is needed to evaluate its performance. A commonly used loss function for the OLS estimator, as in Li (1987) and Hansen (2007), is the mean squared error of $\hat{\mu} = X\hat{\beta}$:

$$L_n(\hat{\beta}) = \|\mu - \hat{\mu}\|^2 = \|y - \hat{\mu}\|^2 + \|u\|^2 - 2u^\top(y - \hat{\mu}).$$  \hspace{1cm} (4.16)
Since $\mathbb{E} \|u\|^2 = n\sigma^2$ and $\mathbb{E} (u^\top (y - \hat{\mu})) = 0$ for the OLS estimator, minimizing (4.16) is asymptotically equivalent to minimizing the sum of squared residuals. However, this is not true for the IV estimator, as it does not minimize the sum of squared residuals and the condition $\mathbb{E} (u^\top (y - \hat{\mu})) = 0$ does not hold. Instead, the IV estimator minimizes the sum of squared residuals of the second-stage regression:

$$y = P_w X \beta + \text{error} = \mu_2 + \text{error},$$

where $\mu_2 \equiv P_w X \beta$. Following (4.16), I propose the following loss function for the IV estimator:

$$L_n(w) = \|\mu_2 - \hat{\mu_2}\|^2 = \left\|P_w X \beta - P(w) X \hat{\beta}(w)\right\|^2. \quad (4.17)$$

The loss function (4.17) follows the pattern of (4.16), which is the squared norm of the difference between the true value $P_w X \beta$ and its estimate. Let $w_q^0$ be a special weight vector such that its $q^{th}$ element equals one and the other elements equal zero. Then, the loss function (4.10) from Sueishi (2012) can be written as

$$L_n(w_q^0) = \frac{1}{n} \left\|P(w_q^0) X \beta - P(w_q^0) X \hat{\beta}(w_q^0)\right\|^2,$$

where the multiplicative term $1/n$ does not matter. This loss function can be problematic, since the true value $P(w_q^0) X \beta$ depends on $w_q^0$. Therefore, the NS estimator may not be efficient under the loss function (4.17).

Before I show the optimality of the MAIV estimator, I first duplicate the results
of Lemma 2.1 from Chapter 2 about the averaged projection matrix $P(w)$. Proofs in this section are presented in Appendix B.

**Lemma 4.1.** Define $M(w) \equiv I - P(w)$. Then,

(i) $\text{Tr}(P(w)) = \sum_{q=1}^{Q} w^{(q)} k^{(q)} = k(w),$

(ii) $\lambda_{\text{max}}(P(w)) \leq 1$, where $\lambda_{\text{max}}$ is the largest eigenvalue of $P(w),$

(iii) $\|P(w^*) M(w) \mu\|^2 \leq \|M(w) X \beta\|^2$ for any $w^*, w \in H^c_Q$.

(iv) $\text{Tr}[P(w) P(w^*) P(w)] \leq \text{Tr}[P(w) P(w)]$ for any $w^*, w \in H^c_Q$.

The averaged projection matrix $P(w)$ is not a traditional projection matrix. It is symmetric but not idempotent. In contrast, even though the matrix $P_{P(w)X}$ includes $P(w)$, it is a traditional projection matrix.

**Assumption 4.1.** As $n \to \infty$, $l^{(q)} \to \infty$ and $l^{(q)}/n \to 0$ for all $q$.

Assumption 4.1 indicates that $l^{(q)}$ goes to infinity at a slower rate than $n$. Similar assumptions can be found in other papers, such as Hansen (2007) and Wan et al. (2010). Based on Assumption 4.1,

**Lemma 4.2.** Let Assumption 4.1 hold. Then, as $n \to \infty$,

$$\frac{n + l(w)}{n - l(w)} = 1 + o(1).$$

Define the expected loss function as $R_n(w) = \mathbb{E}(L_n(w)|W)$. Then,

**Lemma 4.3.**

(i) $R_n(w) \geq \|P - P_{P(w)X} X \beta\|^2,$
4.4. ASYMPTOTIC OPTIMALITY

(ii) \( R_n(w) \geq \sigma^2 \text{Tr}(P(w)P(w)) \).

Assumption 4.2. For some fixed integer \( 1 \leq G < \infty \), \( \mathbb{E}(|u_i|^4G|w_i) \leq \kappa < \infty \), where \( \kappa \) are the bounding constraints on the conditional moments of the error term.

Assumption 4.3. As \( n \to \infty \), \( \xi_n^{-2G}Q \sum_{q=1}^{Q} (R_n(w_0))^G \to 0 \), where \( \xi_n = \inf_{w \in H_Q} R_n(w) \) and \( w_0 \) is a \( Q \times 1 \) vector of which the \( q^{th} \) element is one and the other elements are zeros.

Assumption 4.2 is a bounding constraint on the conditional moments of the error term. Assumption 4.3 is the convergence condition. Similar assumptions can be found in Wan et al. (2010).

The optimality of the MAIV estimator is presented in the following theorem.

Theorem 4.1. Let Assumptions 4.1, 4.2, and 4.3 hold. Then, as \( n \to \infty \)

\[
\frac{L_n(\hat{w})}{L_n(w_{opt})} \to_p 1,
\]

where

\[
\hat{w} = \arg \min_{w \in H_Q^{p}} \text{MAIV}_n(w), \\
w_{opt} = \arg \inf_{w \in H_Q^{p}} L_n(w).
\]

Theorem 4.1 shows that \( L_n(\hat{w}) \) is asymptotically equivalent to the optimal \( L_n(w_{opt}) \). This implies that the MAIV estimator is asymptotically optimal in the class of model average estimators where the weight vector belongs to the set \( H_Q^{p} \).
4.5 Monte Carlo Simulation

In this section, I compare the finite sample performance of the MAIV estimator with other estimators using Monte Carlo experiments. Two cases are considered. The first case assumes only one endogenous variable on the right-hand-side (RHS), and the second case considers multiple endogenous variables on the RHS.

4.5.1 The Case of One Endogenous Variable

The design is similar to the one in Davidson and MacKinnon (2006). The dependent variable $y$ is generated by the equation

$$y = \beta_1 \iota + \beta_2 y_1 + u. \tag{4.18}$$

Here $\iota$ is a vector of ones and $y_1$ is the endogenous variable, which is generated as a function of the instruments in the $n \times 5$ matrix $W$ by

$$y_1 = W \pi + v.$$ 

The first column of $W$ is $\iota$. All of the columns $W$, except the first column, are I.I.D. standard normal random variables. The elements of $u$ and $v$ have variances $\sigma_u^2$ and $\sigma_v^2$, respectively, and correlation $\rho$. I set $\beta = [1 1/2]^\top$, $\rho = 0.5$ and $\sigma_u^2 = \sigma_v^2 = 1$ in all experiments.

The ratio of $||\pi||^2$ to $\sigma_v^2$ can be interpreted as the signal-to-noise ratio. Since $n^{-1}W^T W$ tends to an identity matrix as $n \to \infty$, we see that

$$R_{\infty}^2 = 1/(1 + \sigma_u^2/||\pi||^2) = ||\pi||^2/(||\pi||^2 + \sigma_v^2).$$
In addition, the values of the \( \pi_j \) for \( j \geq 2 \) are constrained to be equal, with \( \pi_1 = 0 \). I generate \( \pi \) to control \( R^2 \) so as to vary on a grid between 0.01 and 0.80.

In total, there are 15 sets of auxiliary instruments. I compare four methods: DN, KO, NS, and MAIV. The estimators for these methods were introduced in Sections 4.2 and 4.3. The arbitrary parameter \( \lambda \) in DN and KO does not matter in this experiment since I only have one RHS endogenous variable. I consider five sample sizes: \( n = 25, 50, 100, 200, \) and 400.

I evaluate the performance of the above estimators using the following risk function:

\[
\text{Risk} \equiv \frac{1}{n} \left\| P_W X \beta - P(w) X \hat{\beta}(w) \right\|^2,
\]

which is \( 1/n \) times the loss function in (4.17). For model selection methods like DN and NS, \( \hat{w} \) is a \( Q \times 1 \) vector with one element equal to 1 and others equal to zero.

Kuerstainer and Okui (2010) proposed three estimators based on different restrictions on \( w \): unbounded set \( H^u_Q \), compact set \( H^c_Q \), and positive set \( H^p_Q \). The KO criterion functions are introduced in (4.8) for unbounded or compact sets and in (4.7) for positive sets. I plot the risk of the three KO estimators against \( R^2 \) for \( n = 25 \) in Figure 4.1. All experiments in this section are based on the average of 10,000 simulation draws. The dashed line, dots, and dash-dotted line correspond to the unbounded set, compact set, and positive set, respectively. As we see in Figure 4.1(a), the unbounded set has the worst performance and is unstable. The compact set and the positive set are indistinguishable. To compare the compact set with the positive set, I normalize the estimators with respect to the compact set in Figure 4.1(b). Note that the normalized unbounded estimator is out of range. As we can see, when \( R^2 \) is low, the positive set generates lower risk and when \( R^2 \) is high, the compact set has a
better performance. In the rest of this section, I use the estimator with $w \in H_Q^c$ to represent the KO method.

I plot the risk of the four estimators against $R^2$ for $n = 25$ in Figure 4.2. The $R^2$ is presented on the horizontal axis and the risk is displayed on the vertical axis. The dashed line, dots, solid line, and dot-solid line correspond to DN, KO, MAIV, and NS, respectively. As $R^2$ increases, NS has the worst performance, as it tends to diverge. KO has a better performance than DN and NS in many cases. DN generates lower risk than KO when $R^2$ is high but higher risk when $R^2$ is low. MAIV has the best performance for all $R^2$. 

Figure 4.1: KO Estimators with Different Restrictions on $w$
4.5. MONTE CARLO SIMULATION

Figure 4.2: Risk Against $R^2$ for $n = 25$

![Figure 4.2: Risk Against $R^2$ for $n = 25$](image)

I also plot the risk of the four estimators for various sample sizes ($n = 50, 100, 200, 400$) in Figure 4.3. The $R^2$ is presented on the horizontal axis and the risk is displayed on the vertical axis. The dashed line, dots, solid line and dot-solid line correspond to DN, KO, MAIV, and NS, respectively. All methods tend to generate lower risk as $n$ increases. NS diverges as $R^2$ increases despite the value of $n$. MAIV has the best performance in all cases. As $n$ gets larger, DN and KO tend to merge with MAIV as $R^2$ increases.
4.5.2 The Case of Multiple Endogenous Variables

In this part, I extend the simulation design in the previous subsection to consider multiple endogenous variables, where

\[ y = \beta_1 \tau + \beta_2 y_1 + \beta_2 y_2 + u \]
with \( y_1 \) and \( y_2 \) being the endogenous variables, which are generated as functions of the instruments in the \( n \times 7 \) matrix \( W = [\nu \ W_1 \ W_2] \) by

\[
\begin{align*}
y_1 &= W_1 \pi_1 + v_1, \\
y_2 &= W_2 \pi_2 + v_2.
\end{align*}
\]

Both \( W_1 \) and \( W_2 \) are \( n \times 3 \) matrices. All columns of \( W \), except the first column, are I.I.D. standard normal random variables. The elements of \( u, v_1, \) and \( v_2 \) are generated by

\[
\begin{bmatrix}
u_i \\
v_{i1} \\
v_{i2}
\end{bmatrix}
\sim N\left(
\begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix},
\begin{bmatrix}
s_{u} & \rho_1 & \rho_1 \\
\rho_1 & s_{v_1} & 0 \\
\rho_1 & 0 & s_{v_2}
\end{bmatrix}
\right).
\]

I set \( \beta = [1, 1/2, 1/3]^{\top} \), \( \rho_1 = \rho_2 = 0.5 \) and \( s_{u}^2 = s_{v_1}^2 = s_{v_2}^2 = 1 \) in all experiments. Let \( \pi_1 = \pi_2 = \pi \) to control the population \( R^2 = ||\pi||^2/(||\pi||^2 + \sigma_{v}^2) \) so as to vary on a grid between 0.01 and 0.80. In addition, the values of the \( \pi_j \) for \( j = 1,2,3 \) are constrained to be equal.

In total, there are 49 sets of auxiliary instruments. I compare DN, KO, NS, and MAIV for five sample sizes: \( n = 25, 50, 100, 200, \) and 400. I evaluate the performance of the above estimators using the risk function defined in (4.19). The arbitrary \( \lambda \) for DN and KO now matters, since there are two endogenous variables on the RHS.

In Figure 4.4, I demonstrate that DN and KO are sensitive to different values of \( \lambda \) when there are multiple endogenous variables on the RHS. Consider \( \lambda_1 = [1/3, 1/3, 1/3]^{\top} \) and \( \lambda_2 = [0, 0, 1]^{\top} \). Denote DN1 and KO1 as the DN and KO that are associated with \( \lambda_1 \). Similarly, DN2 and KO2 are associated with \( \lambda_2 \). I normalize the results with respect to DN1. DN1, KO1, DN2, and KO2 are represented by the
dashed line, dots, dash-x line, and dash-+ line, respectively. The results clearly show that DN and KO are sensitive to different values of $\lambda$. Notice that $\lambda_1$ generally provides better results for both DN and KO. In the rest of this subsection, I use $\lambda_1$ to compute DN and KO, henceforth DN1 and KO1.

In Figure 4.5, I show the risk of the four estimators against $R^2$ for $n = 25$. The $R^2$ is presented on the horizontal axis, and the risk is displayed on the vertical axis. The dashed line, dots, solid line, and dot-solid line correspond to DN1, KO1, MAIV, and NS, respectively. NS generates lower risk than DN1 and KO1 when $R^2$ is lower. KO1 beats DN1 and NS when $R^2$ is high, but it has the worst performance when $R^2$ is low. MAIV has the best performance for all values of $R^2$.

I also plot the risk against various sample sizes ($n = 50, 100, 200, 400$) in Figure
4.5. MONTE CARLO SIMULATION

Figure 4.5: Risk Against $R^2$ for $n = 25$ (Multiple Endogenous Variables)

4.6. The $R^2$ is presented on the horizontal axis, and the risk is displayed on the vertical axis. The dashed line, dots, solid line, and dot-solid line correspond to DN1, KO1, MAIV, and NS, respectively. The results are very similar to Figure 4.3. MAIV has the best performance in all cases. NS tends to diverge as $R^2$ increases for all values of $n$. KO1 and DN1 have better performance with larger values of $n$. KO1 and DN1 tend to merge with MAIV as $n$ increases.
4.6 Conclusion

In this chapter, we consider the situation in which there are many potential instruments for IV estimation. I proposed using the model averaging method to construct optimal instruments for IV estimation. The projection matrix for the optimal instruments is a weighted sum of the projection matrices for all potential instrument sets. The empirical weight on each potential instrument set is computed by minimizing
the model averaging IV (MAIV) criterion through convex optimization. To evaluate the performance of the constructed instruments, I proposed a new loss function. I proved that the instrument set by the MAIV estimator is asymptotically optimal in the sense of achieving the lowest possible value of the loss function as $n \to \infty$. Simulation results showed that the MAIV estimator generated significantly lower risk than existing methods in finite samples.
Chapter 5

Least Squares Forecast Combination by MAPC

5.1 Introduction

This chapter develops a new forecast combination method based on the model averaging prediction criterion (MAPC) introduced in Chapter 2. In this chapter, I extend the theoretical results of Chapter 2 by showing that, with stationary observations, the MAPC estimator is asymptotically optimal for forecast combination in the sense of achieving the lowest possible one-step-ahead second-order mean squared forecast error (MSFE) asymptotically. I also show that the model averaging prediction criterion is asymptotically equivalent to the in-sample mean squared error (MSE) and MSFE. Simulation results indicate that the point forecast by MAPC estimator has significant efficiency gains over existing forecast combination methods in finite samples.

The remainder of this chapter is organized as follows. Section 5.2 introduces the concept of forecast combination and reviews some existing methods. Section 5.3 reviews the original MAPC method and proposes the MAPC forecast combination. Section 5.4 discusses the asymptotic properties of the MAPC forecast combination. Section 5.5 presents a simulation experiment. Section 5.6 concludes this chapter.
5.2. FORECAST COMBINATION

Proofs in this chapter are presented in Appendix C.

5.2 Forecast Combination

Let \((y_t, x_t) : t = 1, \ldots, T\) be a random time series\(^1\) sample, where \(y_t\) and \(x_t = [x_{t1}, x_{t2}, \ldots]\) are real-valued. The data generating process is given by

\[
y_t = \mu_t + u_t, \quad (5.1)
\]

where \(\mu_t = \sum_{j=1}^{\infty} \beta_j x_{tj}, \mathbb{E}(u_t|x_t) = 0\) and \(\mathbb{E}(u_t^2|x_t) = \sigma^2\). Note that lags of \(y_t\) can be included in \(x_t\) as long as \(u_t\) is independent of lags of \(y_t\).

There exists a sequence of linear approximation models \(m = 1, 2, \ldots M\). The \(m\)th approximation model uses any \(k^{(m)}\) regressors belonging to \(x_t\) such that

\[
y_t = \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{tj}^{(m)} + u_t^{(m)}, \quad (5.2)
\]

where \(\beta_j^{(m)}\) is a coefficient in model \(m\) and \(x_{tj}^{(m)}\) is a regressor in model \(m\).

The DGP (5.1) and the approximation model (5.2) can be represented in matrix form:

\[
y = \mu + u,
\]

\[
y = X^{(m)} \beta^{(m)} + u^{(m)},
\]

where \(y\) is \(T \times 1\), \(\mu\) is \(T \times 1\), \(\beta^{(m)}\) is \(k^{(m)} \times 1\), \(X^{(m)}\) is \(T \times k^{(m)}\) with \(t_j\)th element

\(^1\)In this chapter, I use \(t\) and \(T\) instead of \(i\) and \(n\).
being $x_{ij}^{(m)}$, and $u^{(m)}$ is the error term. The least squares estimate of $\beta^{(m)}$ is simply

$$\hat{\beta}^{(m)} = \left( X^{(m)^\top} X^{(m)} \right)^{-1} X^{(m)^\top} y.$$  

The forecast combination of $y_{T+1}$ is described as follows. First, define the point forecast of $y_{T+1}$ as $f_{T+1}$. The estimated forecast of $y_{T+1}$ by model $m$ is then

$$\hat{f}_{T+1|T}^{(m)} = x_{T+1|T}^{(m)} \hat{\beta}^{(m)}, \quad (5.3)$$

where $x_{T+1|T}^{(m)}$ is the regressor vector for model $m$ in period $T + 1$ and $x_{T+1|T} \in F_T$, which means that $x_{T+1|T}$ is known in period $T$. We estimate the point forecasts of $y_{T+1}$ from $M$ approximation models using (5.3). Define a $M \times 1$ vector $\hat{f}_{T+1}$ that accumulates all the estimated forecasts of $y_{T+1}$ by $M$ approximation models, where

$$\hat{f}_{T+1} \equiv \left[ \hat{f}_{T+1}^{(1)}, \hat{f}_{T+1}^{(2)}, \ldots, \hat{f}_{T+1}^{(M)} \right]^\top.$$  

Let $w = [w^{(1)}, \ldots, w^{(M)}]^\top$ be a weight vector in the unit simplex in $\mathbb{R}^M$,

$$H_M \equiv \left\{ w \in [0, 1]^M : \sum_{m=1}^M w^{(m)} = 1 \right\}, \quad (5.4)$$

which is a continuous set. The forecast combination of $y_{T+1}$ is then

$$\hat{f}_{T+1}(w) \equiv w^\top \hat{f}_{T+1} = \sum_{m=1}^M w^{(m)} \hat{f}_{T+1}^{(m)}, \quad (5.5)$$
Similarly, the model average estimator of $\mu$ is defined as

$$
\mu(w) \equiv \sum_{m=1}^{M} w^{(m)} \hat{\mu}^{(m)} = \sum_{m=1}^{M} w^{(m)} X^{(m)} \hat{\beta}^{(m)} = P(w)y, \quad (5.6)
$$

where $P(w) = \sum_{m=1}^{M} w^{(m)} X^{(m)} \left( X^{(m)} \top X^{(m)} \right)^{-1} X^{(m)} \top$ is the averaged projection matrix.

There are many ways to compute the weight vector $w$ based on the existing literature. Following Buckland et al. (1997), the weight estimator for model averaging AIC (MA-AIC) is given by

$$
\hat{w}_{\text{AIC}}^{(m)} = \exp \left( -\frac{1}{2} \text{AIC}^{(m)} \right) / \sum_{m=1}^{M} \exp \left( -\frac{1}{2} \text{AIC}^{(m)} \right), \quad (5.7)
$$

where $\text{AIC}^{(m)} = T \log(\hat{\sigma}_m^2) + 2k^{(m)}$ and $\hat{\sigma}_m^2$ is the estimated $\sigma^2$ by model $m$.

Recent literature has paid increasing attention to forecast combination by Bayesian model averaging (BMA). Following Hansen (2007, 2008), if the priors are fixed, the BMA weights are approximately equivalent to

$$
\hat{w}_{\text{BMA}}^{(m)} = \exp \left( -\frac{1}{2} \text{BIC}^{(m)} \right) / \sum_{m=1}^{M} \exp \left( -\frac{1}{2} \text{BIC}^{(m)} \right), \quad (5.8)
$$

where $\text{BIC}^{(m)} = T \log(\hat{\sigma}_m^2) + \log(T)k^{(m)}$.

Hansen (2008) proposed a forecast combination method based on the Mallows Model Averaging (MMA) criterion

$$
\text{MMA}_T(w) = (y - \mu(w)) \top (y - \mu(w)) + 2\sigma^2 k(w), \quad (5.9)
$$
where $\mu(w)$ is defined in (5.6) and $k(w)$ is the effective number of coefficients such that

$$k(w) \equiv \sum_{m=1}^{M} w^{(m)} k^{(m)}.$$  

This version of MMA is almost identical to the original MMA of Hansen (2007) in (2.9). Three limitations were inherited. First, Hansen (2008) assumed that the regressors $x_t$ were an ordered set and the approximation model $m$ contained the first $k^{(m)}$ regressors from $x_t$. As a result, models with fewer regressors would always nest within larger models. Second, Hansen (2008) assumed a discrete set $H_M^*$ for $w$, in which

$$H_M^*(N) \equiv \left\{ w^{(m)} \in \left[0, \frac{1}{N}, \frac{2}{N}, ..., 1 \right] : \sum_{m=1}^{M} w^{(m)} = 1 \right\}$$

for some fixed integer $N$. The empirical weights $\hat{w}$ are obtained by minimizing equation (5.9) with constraints $w \in H_M^*$ such that

$$\hat{w} = \arg \min_{w \in H_M^*} \text{MMA}_T(w).$$

Finally, there is an infeasible $\sigma^2$ in (5.9). Hansen (2008) recommended using a sample estimate $\hat{\sigma}_L^2$ instead, where

$$\hat{\sigma}_L^2 = \frac{(y - \hat{\mu}^{(L)})^\top (y - \hat{\mu}^{(L)})}{T - k^{(L)}}$$

is the estimated $\sigma^2$ from the largest approximation model $L$ that contains more regressors than all other models. This set-up is different from the one in Hansen (2007), in which $\hat{\sigma}^2$ is computed from any “large” approximation model.
5.3 Model Averaging Prediction Criterion

As we can see from (5.5), forecast combination estimates essentially rely on the weight vector $\mathbf{w}$. I propose estimating $\mathbf{w}$ by the model averaging prediction criterion (MAPC):

$$\text{MAPC}_T(\mathbf{w}) = (\mathbf{y} - \mathbf{\mu}(\mathbf{w}))^\top (\mathbf{y} - \mathbf{\mu}(\mathbf{w})) \left( \frac{T + k(\mathbf{w})}{T - k(\mathbf{w})} \right),$$

(5.12)

where $\mathbf{\mu}(\mathbf{w})$ and $k(\mathbf{w})$ are defined in (5.6) and (5.10) respectively. Similar to MMA, the empirical weights $\mathbf{w}$ can be obtained by minimizing (5.12):

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in H_M} \text{MAPC}_T(\mathbf{w}).$$

Note that $\mathbf{w}$ is restricted to lie in the unit simplex $H_M$, which is defined in (5.4). No restrictions are placed on the orders of $\mathbf{x}_t$ and the $k(m)$. Approximation models in this method can be either nested or non-nested.

For convenience in calculations, (5.12) can be transformed into the following form. First, denote $\hat{\mathbf{u}}^{(m)}$ as a $T \times 1$ estimated residual vector from model $m$. Let $\hat{\mathbf{U}}$ be a $T \times M$ matrix consisting of the residuals $\hat{\mathbf{U}} \equiv [\hat{\mathbf{u}}^{(1)}, \hat{\mathbf{u}}^{(2)}, ..., \hat{\mathbf{u}}^{(M)}]$. Then, define an $M \times 1$ vector $\mathbf{k}$ that contains the number of parameters from each model such that $\mathbf{k} \equiv [k^{(1)}, k^{(2)}, ..., k^{(M)}]^\top$.

Consequently, (5.12) becomes

$$\text{MAPC}_T(\mathbf{w}) = \mathbf{w}^\top \hat{\mathbf{U}}^\top \hat{\mathbf{U}} \mathbf{w} \left( \frac{T + \mathbf{k}^\top \mathbf{w}}{T - \mathbf{k}^\top \mathbf{w}} \right).$$

(5.13)
Likewise, the MMA criterion can be transformed into

\[ \text{MMA}_T(w) = w^\top \hat{U}^\top \hat{U}w + 2\hat{\sigma}_L^2 k^\top w, \] (5.14)

where \( \hat{\sigma}_L^2 \) is defined in (5.11).

If all regressors \( y \) have strong explanatory power, \( \hat{\sigma}_L^2 \) is a good approximation to the true value \( \sigma^2 \). However, \( \hat{\sigma}_L^2 \) may not be an efficient estimator of \( \sigma^2 \) if many regressors have no or little explanatory power. Define the average estimate of \( \sigma^2 \) as \( \hat{\sigma}^2(w) \) such that

\[ \hat{\sigma}^2(w) \equiv \frac{(y - \mu(w))^\top (y - \mu(w))}{T - k(w)} = \frac{w^\top \hat{U}^\top \hat{U}w}{T - k^\top w}. \] (5.15)

A model averaging estimator should assign large weight to the largest model asymptotically if there is evidence that all regressors have strong explanatory power. In this case, \( \hat{\sigma}^2(w) \) is very close to \( \hat{\sigma}_L^2 \). If many regressors are irrelevant, \( \hat{\sigma}_L^2 \) is highly inefficient, but a model averaging estimator should assign large weights to models that contain relevant regressors, which makes \( \hat{\sigma}^2(w) \) more efficient than \( \hat{\sigma}_L^2 \).

It is straightforward to show that MAPC can be rewritten as

\[ \text{MAPC}_T(w) = w^\top \hat{U}^\top \hat{U}w + 2\hat{\sigma}^2(w) k^\top w, \]

which is similar to (5.14) with \( \sigma^2 \) being replaced by a more efficient estimator \( \hat{\sigma}^2(w) \).

In fact, the feasible MMA criterion in (5.14) is a special case of MAPC, since \( \hat{\sigma}_L^2 \) in
(5.11) can be rewritten as a special average estimator such that

\[ \hat{\sigma}^2_L = \frac{(y - \hat{\mu}(L))^T (y - \hat{\mu}(L))}{T - k(L)} = \frac{(y - \mu(w^0_L))^T (y - \mu(w^0_L))}{T - k(w^0_L)}, \]

where \( w^0_L \) is a weight vector in which the \( L^{th} \) element is one and all of the other elements are zero. Simulation evidence in Section 5.5 shows that the MMA estimator with \( \hat{\sigma}^2_L \) generates a higher forecast risk than the MAPC estimator.

5.4 Asymptotic Properties of MAPC

Define in-sample mean squared error (MSE) as

\[ L_T(w) \equiv \frac{1}{T} \mathbb{E} [\mu(w) - \mu]^T [\mu(w) - \mu], \tag{5.16} \]

where \( \mu(w) \), defined in (5.6), is the in-sample averaged estimate for \( \mu \). The one-step-ahead out-of-sample forecast of \( y_{T+1} \) is defined by \( \hat{f}_{T+1}(w) \) in (5.5). Then, define one-step-ahead second-order mean squared forecast error (MSFE) as

\[ R_T(w) = \mathbb{E} \left[ y_{T+1} - \hat{f}_{T+1}(w) \right]^2 - \sigma^2. \tag{5.17} \]

Theorem 2.1 in Chapter 2 established the asymptotic optimality of MAPC in in-sample estimation by showing that MAPC achieves the lowest possible MSE asymptotically. For the sake of completeness, Theorem 2.1 and the necessary assumptions are reported here in the time series notations.

Assumption 5.1. As \( T \to \infty \), \( k^{(m)} \to \infty \) and \( k^{(m)}/T \to 0 \) for all \( m \).
5.4. ASYMPTOTIC PROPERTIES OF MAPC

Assumption 5.2. Assume that \( \mathbb{E}(|u_t|^4 G | x_t) \leq \kappa < \infty \) for some fixed integer \( 1 \leq G < \infty \).

Assumption 5.3. As \( T \to \infty \), \( \xi^2 G M \sum_{m=1}^M (\mathbb{E}(T L_T(w^0_m)))^G \to 0 \), where \( w^0_m \) is an \( M \times 1 \) vector of which the \( m^{th} \) element is one and the other elements are zero and \( \xi_T = \inf_{w \in H_M} \mathbb{E}(T L_T(w^0_m)) \).

Theorem 5.1 (Theorem 2.1 in time series notations). Let Assumptions 5.1, 5.2, and 5.3 hold. Then, as \( T \to \infty \)

\[
\frac{L_T(\hat{w})}{L_T(w_{opt})} \to_p 1,
\]

where

\[
w_{opt} = \arg\inf_{w \in H_M} L_T(w).
\]

Assumption 5.1 states that as \( T \) goes to infinity, \( k^{(m)} \) goes to infinity at a slower rate for \( m = 1, ..., M \). Assumption 5.2 is a boundary condition on the conditional moment. Assumption 5.3 is the crucial convergence condition. I assume that as \( T \to \infty \), \( \xi^2 G \to \infty \) at a faster rate than \( M \sum_{m=1}^M (\mathbb{E}(T L_T(w^0_m)))^G \to \infty \). See Chapter 2 for a detailed explanation of these assumptions. Theorem 5.1 delivers the optimality of MAPC with respect to MSE.

To show the optimality of MAPC with respect to MSFE, the following assumption needs to be imposed:

Assumption 5.4. Assume that \((x_t, u_t)\) is strictly stationary and ergodic. Also, \( \mathbb{E}(x_t^\top x_t) < \infty \).

Stationarity is a crucial assumption in this chapter. The extension to non-stationary cases is beyond the scope of this chapter and is left for future research.
The asymptotic property of $R_T(w)$ is investigated and presented in the following theorem:

**Theorem 5.2.** Let Assumption 5.4 hold. Then, $R_T(w) = L_T(w)$.

Theorem 5.2 shows that $R_T(w)$ and $L_T(w)$ are equivalent. Combining Theorem 5.1 with Theorem 5.2, we conclude that MAPC is also asymptotically optimal in forecast estimation, as MAPC estimator achieves the lowest possible one-step-ahead second-order MSFE asymptotically.

The criterion $\text{MAPC}_T(w)$ also has the classic property of the Mallows criterion.

**Theorem 5.3.** Let Assumptions 5.1, 5.2, 5.3, and 5.4 hold. Then, as $T \to \infty$, the result is

$$\frac{\text{MAPC}_T(w)}{T(L_T(w) + \sigma^2)} \to_p 1.$$

The criterion $\text{MAPC}_T(w)$ is asymptotically equivalent to the in-sample MSE (5.16). The additive and multiplicative terms do not matter when calculating the weights $w$. Combining Theorem 5.2 with Theorem 5.3, we can deduce that MAPC is asymptotically equivalent to MSE (5.16) and MSFE (5.17).

### 5.5 Monte Carlo Simulation

In this section, I investigate the finite sample performance of MAPC in forecast combination using the Monte Carlo Simulation. Consider a regression model with autoregressive regressors

$$y = X\beta + u,$$

where $X = [x_1, x_2, x_3, x_4, x_5, x_6]$ is $T \times 6$, $\beta$ is $6 \times 1$, and $u$ is $T \times 1$. The coefficient vector $\beta$ is determined by $\beta_j = j^{-1}$ for $j = 1, 2, ..., 6$. The error term $u$ is distributed
5.5. MONTE CARLO SIMULATION

as \( \text{N}(0, \sigma_u^2) \) and is independent of \( \mathbf{X} \). The parameter \( \sigma_u \) controls the population \( R^2 = 1/(1 + \sigma_u^2) \) so as to vary on a grid between 0.01 and 0.99.

The first column of \( \mathbf{X} \), \( \mathbf{x}_1 \), is set to be the intercept; the remaining \( \mathbf{x}_i \) follow an AR(1) process

\[
x_{it} = \rho x_{i,t-1} + \epsilon_{it} \quad \text{for } i = 2, \ldots, 6,
\]

where \( |\rho| < 1, x_{i0} = 0 \) and \( \epsilon_{it} \) is independent and identically distributed as \( \text{N}(0, \sigma_e^2) \), where the variance \( \sigma_e \) equals 1. I consider six sample sizes \( n = 25, 50, 100, 200, 400 \) and 800.

All submodels, which are nested in the general unrestricted model (5.18), are treated as approximation models. The approximation models are clearly non-nested in our experiment. The total number of approximation models, \( M \), is equal to the total number of combinations made by \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_6 \) (except the empty set) such that

\[
M = \sum_{k=1}^{6} \binom{6}{k} = 2^6 - 1 = 63.
\]

I compare the finite sample performance of MA-AIC, BMA, and MMA with MAPC by computing their risk

\[
\text{Risk} = \left[ y_{T+1} - \hat{f}_{T+1}(\mathbf{w}) \right]^2 - \sigma^2,
\]

which is the MSFE defined in (5.17) with no expectation sign. The weighting vectors \( \mathbf{w} \) for MA-AIC and BMA are computed following equations (5.7) and (5.8), while \( \mathbf{w} \) for MAPC and MMA are estimated through numerical minimization of (5.13) and (5.14) under their respective constraints. For each parameterization, the result is normalized by dividing by the risk of the MAPC estimator and averaging across
100,000 simulation draws.

Figure 5.1 shows the simulation results for forecasting $y_{T+1}$ with $T = 25$. I set $ho = 0.5$. Although not included here, I also tried other values of $\rho$. The ranking of the performance of each method is not sensitive to $\rho$. The $R^2$ is presented on the horizontal axis and the average MSFE is presented on the vertical axis. The dots, dash-dotted line, solid line, and dash line correspond to MA-AIC, BMA, MAPC, and MMA, respectively. Note that when $R^2 < 0.1$, BMA has the best performance and MA-AIC has a better performance than MMA and MAPC. Also, MAPC is slightly outperformed by MA-AIC when $R^2 > 0.9$. However, for other values of $R^2$, which are more common in application, MAPC and MMA generate significantly lower MSFE than MA-AIC and BMA. Moreover, MAPC generates significantly smaller MSFE than MMA for most values of $R^2$.

Figure 5.2 represents the simulation results for forecasting $y_{T+2}$ and $y_{T+3}$. The risk is modified to

$$\text{Risk} = \left[ y_{T+j} - \hat{f}_{T+j}(w) \right]^2 - \sigma_u^2$$

(5.19)

for $j = 1, \ldots, \infty$. Figure 5.2(a) and Figure 5.2(b) show the normalized results for forecasting $y_{T+2}$ and $y_{T+3}$ with $T = 25$ and $\rho = 0.5$. The normalized results are very similar to Figure 5.1. For most values of $R^2$, MAPC and MMA generate significantly lower risk than MA-AIC and BMA. Also, MAPC generates significantly smaller MSFE than MMA for most values of $R^2$. Figure 5.2(c) and Figure 5.2(d) show the simulation results for forecasting $y_{T+2}$ and $y_{T+3}$ with no normalization. Although the figures are cluttered (the major reason for normalization), we find that the risk for forecasting $y_{T+2}$ is smaller than that for $y_{T+3}$.

I also examine the estimators for various sample sizes in which $T = 25, 50, 100,$
In this simulation, we consider one-step-ahead forecasting. The results are shown in Figure 5.3. To keep our figure uncluttered, only MMA and MAPC are displayed and are represented by the dashed line and the solid line, respectively. MAPC generates smaller MSFE than MMA for most values of $R^2$ in all panels. As the sample size increases, the gap between MMA and MAPC shrinks and MMA merges with MAPC. Although not shown in the figure, the unnormalized risk is also examined. As $T$ increases, the unnormalized risks for both estimators decrease.
Figure 5.2: Simulation Results for Forecasting $y_{T+2}$ and $y_{T+3}$
Figure 5.3: Simulation Results for Various Sample Sizes
5.6 Conclusion

The new forecast combination method in this chapter was developed based on the MAPC method introduced in Chapter 2. The empirical weights on each potential forecast model are obtained through a convex optimization of MAPC. With stationary observations, I showed that the MAPC estimator is asymptotically optimal for forecast combination in that it achieves the lowest possible one-step-ahead second-order mean squared forecast error (MSFE) as $n \to \infty$. The simulation results demonstrated that the MAPC estimator generates significantly smaller risks than others for most values of $R^2$, which are more reasonable in practice.
Chapter 6

Conclusion and Future Work

This thesis contributes to the literature on least squares model averaging by studying and extending current least squares model averaging techniques. I proposed a new estimator for least squares model averaging in Chapter 2. I was largely inspired by Bruce Hansen’s 2007 *Econometrica* paper, in which Hansen (2007) proposed the Mallows model averaging (MMA) method which is based on the Mallows’ $C_p$ model selection method. My MAPC estimator can be seen as the model averaging version of Amemiya’s (1980) prediction criterion. Amemiya (1980) demonstrated that his prediction criterion is more finite-sample robust than Mallows’ $C_p$.

Although it is the pioneer work in least squares model averaging, Hansen’s (2007) paper has several limitations: nested models, discontinuous weighting vector, preliminary $\hat{\sigma}^2$, and the lack of a testing method. The MAPC estimator I proposed in Chapter 2 eliminates all these limitations. The results of empirical applications in Chapter 3 suggested that the asymptotic test statistics derived in Section 2.3 work well in large sample sizes and that bootstrap tests are highly recommended for small sample sizes.

I extended the MAPC method to IV estimation in Chapter 4. It is difficult to
obtain a better prediction using IV estimation when there are multiple instrument sets. The potential instrument sets are usually not the same size. Therefore, we cannot take averages across all the instrument sets. Kuersteiner and Okui (2010) bypassed this issue by taking an average across all the projection matrices for all the instrument sets. I followed this idea and proposed the model averaging IV (MAIV) method. One merit of the MAIV method is its simplicity. I also prepared an algorithm for convenience in calculations. Simulation results showed that the MAIV estimator generates significantly lower risks than existing methods.

In Chapter 5, I extended the MAPC method to forecast combination. I showed that the MAPC estimator can be used in forecasting. One crucial assumption is that the observations must be stationary. This limits the application of this method and demonstrates the need for further extension.

Besides the stationarity restriction, the MAPC method (and the MMA method) also suffers from a curse of dimensionality. As the total number of potential variables increases, the total number of approximation models increases exponentially. The MAPC method relies on convex optimization, which means that the computation cost can be huge for a large number of approximation models. On the other hand, Bayesian model averaging does not have this problem as it uses the MC³ method. Further research to solve this issue is required.

Also, models in this thesis are restricted to be linear. Extension to nonlinear models are also left for future research.
Bibliography


Appendix A

Appendix for Chapter 2

Proof of Lemma 2.1. To prove (i), we make use of the fact that \( \text{Tr}(P^{(m)}) = k^{(m)} \).

Then,

\[
\text{Tr}(P(w)) = \text{Tr} \left( \sum_{m=1}^{M} w^{(m)} P^{(m)} \right) = \sum_{m=1}^{M} w^{(m)} \text{Tr} \left( P^{(m)} \right) = \sum_{m=1}^{M} w^{(m)} k^{(m)} = k(w).
\]

To prove (ii), we first note that the projection matrix \( P^{(m)} \) is idempotent. Then for an \( n \times 1 \) vector \( \eta \)

\[
\max_{\eta} \frac{\eta^\top P^{(m)} \eta}{\eta^\top \eta} = 1 \quad \text{for all } m.
\]

By the definition of eigenvalue, we have

\[
\lambda_{\text{max}}(P(w)) = \max_{\eta} \frac{\eta^\top P(w) \eta}{\eta^\top \eta} \leq \sum_{m=1}^{M} w^{(m)} \left( \max_{\eta} \frac{\eta^\top P^{(m)} \eta}{\eta^\top \eta} \right) = 1.
\]

Parts (iii) and (iv) can be obtained using result (ii). We have

\[
\|P(w^*) M(w) \mu\|^2 \leq \lambda_{\text{max}}^2(P(w^*)) \|M(w) \mu\|^2 \leq \|M(w) \mu\|^2.
\]
and

\[
\text{Tr}
\left[
(P(w^*)P(w^*))P(w)P(w)
\right] \leq \lambda_{\max}^2(P(w^*)) \text{Tr}
\left[
P(w)P(w)
\right] \leq \text{Tr}
\left[
P(w)P(w)
\right].
\]

\(\square\)

**Proof of Lemma 2.2.** \(R_n(w)\) can be written as

\[
R_n(w) = \mathbb{E}
\left[
(P(w)\mu + P(w)u - \mu)^\top(P(w)\mu + P(w)u - \mu)|X\right]
\]

\[
= \mu^\top M(w)M(w)\mu + \sigma^2 \text{Tr}(P(w)P(w))
\]

\[
= \|M(w)\mu\|^2 + \sigma^2 \text{Tr}(P(w)P(w)).
\] (A.1)

Both terms in (A.1) are non-negative, which implies Lemma 2.2. \(\square\)

**Proof of Lemma 2.3.** The proof is straightforward. If \(k^{(m)}/n \rightarrow 0\) for all \(m\), then the linear combination of \(k^{(m)}, k(w) = \sum_{m=1}^{M} w^{(m)} (k^{(m)}/n) \rightarrow 0\). \(\square\)

**Proof of Theorem 2.1.** Our proof follows the techniques derived in Li (1987) and Wan et al. (2010). Rewrite MAPC\(_n(w)\) to contain \(L_n(w)\):

\[
\text{MAPC}_n(w) = (\mu - \mu(w) + u)^\top(\mu - \mu(w) + u) \left(\frac{n + k(w)}{n - k(w)}\right)
\]

\[
= L_n(w) + 2\mu^\top M(w)u + 2L_n(w) \left(\frac{k(w)}{n - k(w)}\right)
\]

\[
+ 4\mu^\top M(w)u \left(\frac{k(w)}{n - k(w)}\right) + u^\top u \left(\frac{n + k(w)}{n - k(w)}\right).
\] (A.2)

As \(n \rightarrow \infty\), \((n + k(w))/(n - k(w)) \rightarrow 1\) by Lemma 2.3, which implies that the last
term in (A.2) is independent of $w$. Therefore, $\hat{w}$ minimizes

$$L_n(w) + 2\mu^T M(w)u + 2L_n(w)\left(\frac{k(w)}{n-k(w)}\right) + 4\mu^T M(w)u \left(\frac{k(w)}{n-k(w)}\right). \quad (A.3)$$

If we can show that as $n \to \infty$, all terms in (A.3) except $L_n(w)$ are negligible compared with $L_n(w)$, for any $w \in H_M$, then the asymptotic optimality of $\hat{w}$ is established.

Theorem 2.1 is valid, if, as $n \to \infty$

$$\sup_{w \in H_M} \left| \frac{\mu^T M(w)u}{R_n(w)} \right| \to_p 0, \quad (A.4)$$

$$\sup_{w \in H_M} \left| \frac{L_n(w)}{R_n(w)} - 1 \right| \to_p 0, \quad (A.5)$$

$$\sup_{w \in H_M} \left| \frac{\mu^T M(w)u}{R_n(w)} \left(\frac{k(w)}{n-k(w)}\right) \right| \to_p 0, \quad (A.6)$$

$$\sup_{w \in H_M} \left| \frac{L_n(w)}{R_n(w)} \left(\frac{k(w)}{n-k(w)}\right) \right| \to_p 0. \quad (A.7)$$

We shall prove (A.4) first. Given any $\delta > 0$, by triangular inequality and Bonferroni’s inequality we have

$$P \left\{ \sup_{w \in H_M} \left| \frac{\mu^T M(w)u}{R_n(w)} \right| > \delta \right\} \leq P \left\{ \sup_{w \in H_M} \sum_{m=1}^M w^{(m)} \left| \mu^T M^{(m)} u \right| > \delta \xi_n \right\}$$

$$= P \left\{ \max_m \left| \mu^T M^{(m)} u \right| > \delta \xi_n \right\}$$

$$\leq \sum_{m=1}^M P \left\{ \left| \mu^T M(w_m^0) u \right| > \delta \xi_n \right\},$$
which, by Chebyshev’s inequality, is no greater than
\[
\sum_{m=1}^{M} \mathbb{E} \left\{ \frac{(\mu^\top M(w_0^m)u)^{2G}}{\delta^{2G} \varepsilon_n^{2G}} \right\}.
\]

By Theorem 2 of Whittle (1960) and our Lemma 2.2(i), we observe that
\[
\sum_{m=1}^{M} \mathbb{E} \left\{ \frac{(\mu^\top M(w_0^m)u)^{2G}}{\delta^{2G} \varepsilon_n^{2G}} \right\} \leq C_1 \delta^{-2G} \varepsilon_n^{-2G} \sum_{m=1}^{M} \| M(w_0^m)\mu \|^{2G}
\]
\[
\leq C_1 \delta^{-2G} \varepsilon_n^{-2G} \sum_{m=1}^{M} (R_n(w_0^m))^G
\]
for some constant $C_1$. The last term above goes to zero by Assumption 2.2. Thus, (A.4) is proved.

To prove (A.5), we first see
\[
\sup_{w \in H_M} \left\| \frac{L_n(w)}{R_n(w)} - 1 \right\|_p \to 0
\]
\[
\iff \sup_{w \in H_M} \left\| \frac{u^\top P(w)P(w)u - \sigma^2 \text{Tr}(P(w)P(w)) - 2 \mu^\top M(w)P(w)u}{R_n(w)} \right\|_p \to 0.
\]

Then, it suffices to prove
\[
\sup_{w \in H_M} \left| \frac{\mu^\top M(w)P(w)u}{R_n(w)} \right| \to_p 0 \quad \text{(A.8)}
\]
and
\[
\sup_{w \in H_M} \left| \frac{u^\top P(w)P(w)u - \sigma^2 \text{Tr}(P(w)P(w))}{R_n(w)} \right| \to_p 0. \quad \text{(A.9)}
\]
By Chebyshev’s inequality and Theorem 2 of Whittle (1960), given any $\delta > 0$,

$$
\begin{align*}
P \left\{ \sup_{w \in H_M} \left| \frac{\mu^\top M(w)P(w)u}{R_n(w)} \right| > \delta \right\} &
\leq \sum_{m=1}^M \sum_{l=1}^M \mathbb{E} \left\{ \left( \frac{\mu^\top M(w_0^m)P(w_0^l)u}{\delta^{2G} \xi_n^{2G}} \right)^{2G} \right\} \\
&\leq C_2 \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^M \sum_{l=1}^M \left( \mu^\top M(w_0^m)P(w_0^l)M(w_0^m)\mu \right)^G \\
&= C_2 \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^M \sum_{l=1}^M \|P(w_0^l)M(w_0^m)\mu\|^{2G},
\end{align*}
$$

where $C_2$ is a constant. By Lemma 2.1 (iii), Lemma 2.2 (i) and Assumption 2.2, we obtain

$$
C_2 \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^M \sum_{l=1}^M \left\|P(w_0^l)M(w_0^m)\mu\right\|^{2G} \leq C_2 \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^M \sum_{l=1}^M \left\|M(w_0^m)\mu\right\|^{2G}
$$

$$
\leq C_2 \delta^{-2G} \xi_n^{-2G} M \sum_{m=1}^M \left( R_n(w_0^m) \right)^G \to 0.
$$

Likewise, by Chebyshev’s inequality, Theorem 2 of Whittle (1960), Lemma 2.1 (iii), Lemma 2.2 (ii) and Assumption 2.2, given any $\delta > 0$, we observe that, for a constant $C_3$,

$$
\begin{align*}
P \left\{ \sup_{w \in H_M} \left| \frac{u^\top P(w)P(w)u - \sigma^2 \text{Tr} (P(w)P(w))}{R_n(w)} \right| > \delta \right\} &
\leq \sum_{m=1}^M \sum_{l=1}^M \mathbb{E} \left\{ \left( \frac{u^\top P(w_0^l)P(w_0^l)u - \sigma^2 \text{Tr} (P(w_0^l)P(w_0^l))}{\delta^{2G} \xi_n^{2G}} \right)^{2G} \right\} \\
&\leq C_3 \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^M \sum_{l=1}^M \left[ \text{Tr} (P(w_0^l)P(w_0^l)P(w_0^m)) \right]^G
\end{align*}
$$
\[ \leq C_3 \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^{M} \sum_{t=1}^{M} \left[ \text{Tr} \left( P(w_m^0) P(w_m^0) \right) \right]^G \]

\[ \leq \frac{C_3}{\sigma^2} \delta^{-2G} \xi_n^{-2G} M \sum_{m=1}^{M} [R_n(w_m^0)]^G \to 0. \]

Proving (A.6) and (A.7) becomes straightforward once we validate (A.4) and (A.5). We obtain, as \( n \to \infty \), that

\[ \sup_{w \in H_M} \left\| \mu^\top M(w) u \right\| \frac{k(w)}{n - k(w)} \leq \sup_{w \in H_M} \left\| \mu^\top M(w) u \right\| \cdot \sup_{w \in H_M} \left\| \frac{k(w)}{n - k(w)} \right\| \to_p 0 \]

and

\[ \sup_{w \in H_M} \left\| L_n(w) \cdot \frac{k(w)}{n - k(w)} \right\| \leq \sup_{w \in H_M} \left\| L_n(w) \right\| \cdot \sup_{w \in H_M} \left\| \frac{k(w)}{n - k(w)} \right\| \to_p 0. \]

This completes the proof of Theorem 2.1.

**Proof of Theorem 2.2.** Since \( \hat{u}^{(m)} = y - \hat{\mu}^{(m)} = y - P^{(m)} y = M^{(m)} u + M^{(m)} \mu \) for model \( m \), the average estimate of \( u \) becomes

\[ u(w) = \sum_{m=1}^{M} w^{(m)} (y - \hat{\mu}^{(m)}) = \sum_{m=1}^{M} w^{(m)} M^{(m)} (\mu + u) = M(w)(\mu + u). \]

Then, the average estimate for \( \sigma^2 \) becomes

\[ \sigma^2(w) = \frac{u(w)^\top u(w)}{n - k(w)} = \frac{u^\top M(w) M(w)(\mu + u)}{n - k(w)} \]

\[ = \frac{u^\top M(w) M(w) u}{n - k(w)} + \frac{\mu^\top M(w) M(w) \mu}{n - k(w)} + \frac{2 \mu^\top M(w) M(w) u}{n - k(w)}. \]
Therefore, Theorem 2.2 is valid if the following hold: as $n \to \infty$,

\[
\sup_{w \in H_M} \left| \frac{u^\top M(w)M(w)u}{n - k(w)} \right| \to_p \sigma^2, \quad (A.11)
\]

\[
\sup_{w \in H_M} \left| \frac{\mu^\top M(w)M(w)\mu}{n - k(w)} \right| \to_p 0, \quad (A.12)
\]

\[
\sup_{w \in H_M} \left| \frac{\mu^\top M(w)M(w)u}{n - k(w)} \right| \to_p 0. \quad (A.13)
\]

Equation (A.11) is equivalent to

\[
\sup_{w \in H_M} \left| \frac{u^\top M(w)u}{n - k(w)} - \frac{u^\top P(w)M(w)u}{n - k(w)} \right| \to_p \sigma^2. \quad (A.14)
\]

To prove (A.11), it suffices to show that as $n \to \infty$,

\[
\sup_{w \in H_M} \left| \frac{u^\top M(w)u}{n - k(w)} \right| \to_p \sigma^2 \quad (A.15)
\]

and

\[
\sup_{w \in H_M} \left| \frac{u^\top P(w)M(w)u}{n - k(w)} \right| \to_p 0. \quad (A.16)
\]

First, because $\mathbb{E} \left( u^\top M(w)u \right) = \sigma^2(n - k(w))$, by Theorem 2 of Whittle (1960),

\[
\mathbb{E} \left| u^\top M(w)u - \sigma^2(n - k(w)) \right|^2 \leq C_4^2 \text{Tr} \left( M(w)M(w) \right) \\
\leq C_4^2 \text{Tr} \left( M(w) \right) \\
= C_4^2(n - k(w)),
\]

where $C_4$ is some constant. Thus, for any $\delta > 0$, by Markov’s inequality and Lemma
2.3, we have

\[
P \left\{ \sup_{w \in H_M} \left| \frac{u^\top M(w)u - \sigma^2(n - k(w))}{n - k(w)} \right| > \delta \right\} \leq \frac{\mathbb{E} \left| u^\top M(w)u - \sigma^2(n - k(w)) \right|^2}{\delta^2 (n - k(w))^2} \leq \frac{C_4}{\delta^2 (n - k(w))} \to 0.
\]

Second,

\[
P \left\{ \sup_{w \in H_M} \left| \frac{u^\top P(w)M(w)u}{n - k(w)} \right| > \delta \right\} \leq P \left\{ \sup_{w \in H_M} \left| \frac{u^\top P(w)u}{n - k(w)} \right| > \delta \right\} = P \left\{ \sup_{w \in H_M} \sum_{m=1}^{M} w^{(m)} \left| \frac{u^\top P^{(m)}u}{n - k(w)} \right| > \delta \right\} \leq P \left\{ \sup_{w \in H_M} \max_{m} \left| \frac{u^\top P^{(m)}u}{n - k(w)} \right| > \delta \right\} \to 0
\]

since \((n - k(w))^{-1} \left( u^\top P^{(m)}u \right) \to_p 0\) as \(n \to \infty\) for all \(m\). (A.11) is obtained.

To prove (A.12), we see that for any approximation model \(m\),

\[
\mathbb{E} \left[ \frac{(b^{(m)})^\top M^{(m)}b^{(m)}}{n - k(w)} \right] \leq \mathbb{E} \left[ \frac{(b^{(m)})^\top b^{(m)}}{n - k(w)} \right] \leq \frac{n}{n - k(w)} \mathbb{E} \left( b^{(m)}_i \right)^2 \to 0
\]

since \(k^{(m)} \to \infty\) as \(n \to \infty\) and the square integrability of \(\mu_i^{(m)}\) implies \(\mathbb{E} \left( b^{(m)}_i \right)^2 \to 0\) as \(k^{(m)} \to \infty\). This implies

\[
\frac{(b^{(m)})^\top M^{(m)}b^{(m)}}{n - k(w)} \to_p 0 \quad \text{(A.17)}
\]
for all \( m \). Therefore, by (A.17), we observe that, for any \( \delta > 0 \),

\[
P \left\{ \sup_{w \in H_M} \left| \frac{\mu^\top M(w) M(w) \mu}{n - k(w)} \right| > \delta \right\} \\
\leq P \left\{ \sup_{w \in H_M} \left| \frac{\mu^\top M(w) \mu}{n - k(w)} \right| > \delta \right\} \\
= P \left\{ \sup_{w \in H_M} \sum_{m=1}^{M} w^{(m)} \left| \frac{\mu^\top M^{(m)} \mu}{n - k(w)} \right| > \delta \right\} \\
\leq P \left\{ \sup_{w \in H_M} \max_{m} \left| \frac{\mu^\top M^{(m)} \mu}{n - k(w)} \right| > \delta \right\} \\
= P \left\{ \sup_{w \in H_M} \max_{m} \left| \frac{(b^{(m)})^\top M^{(m)} b^{(m)}}{n - k(w)} \right| > \delta \right\} \to 0.
\]

Finally, since \( (n - k(w))^{-1} \left( \mu^\top M^{(m)} u \right) \to_p 0 \) for all model \( m \), we obtain

\[
P \left\{ \sup_{w \in H_M} \left| \frac{\mu^\top M(w) M(w) u}{n - k(w)} \right| > \delta \right\} \leq P \left\{ \sup_{w \in H_M} \left| \frac{\mu^\top M(w) u}{n - k(w)} \right| > \delta \right\} \\
\leq P \left\{ \sup_{w \in H_M} \max_{m} \left| \frac{\mu^\top M^{(m)} u}{n - k(w)} \right| > \delta \right\} \to_p 0.
\]

Therefore, we conclude that \( \sigma^2(w) \to_p \sigma^2 \).

\( \square \)

**Proof of Lemma 2.4.** Since the covariance matrix is conditional on \( X^{(L)} \), we can assume that \( X^{(m)} \) is exogenous for all \( m \). The covariance matrix can be written as

\[
\text{Cov} \left( \Gamma^{(m)} \hat{\beta}^{(m)}, \Gamma^{(s)} \hat{\beta}^{(s)} \right) \\
= \mathbb{E} \left[ \left( \Gamma^{(m)} \beta^{(m)} - \beta(w) \right) \left( \Gamma^{(s)} \beta^{(s)} - \beta(w) \right)^\top \right] \\
= \mathbb{E} \left( \left( \Gamma^{(m)} (\hat{\beta}^{(m)} - \beta^{(m)}) + d_1^{(m)} \right) \left( \Gamma^{(s)} (\hat{\beta}^{(s)} - \beta^{(s)}) + d_1^{(s)} \right)^\top \right), \quad (A.18)
\]
where $d_1^{(m)} = \Gamma^{(m)}\beta^{(m)} - \beta(w)$ is non-random. First, we have

$$
\Gamma^{(m)}\left(\beta^{(m)} - \beta^{(m)}\right)
= \Gamma^{(m)}\left(\left(X^{(m)^\top}X^{(m)}\right)^{-1}X^{(m)^\top}\left(X^{(m)}\beta^{(m)} + X^{(-m)}\beta_{-m} + u\right) - \beta^{(m)}\right)
= \Gamma^{(m)}\left(X^{(m)^\top}X^{(m)}\right)^{-1}X^{(m)^\top}X^{(-m)}\beta_{-m} + \Gamma^{(m)}\left(X^{(m)^\top}X^{(m)}\right)^{-1}\Gamma^{(m)^\top}X^{\top}u.
$$

Define

$$
A^{(m)} \equiv \Gamma^{(m)}\left(X^{(m)^\top}X^{(m)}\right)^{-1}X^{(m)^\top}X^{(-m)}\beta_{-m},
B^{(m)} \equiv \Gamma^{(m)}\left(X^{(m)^\top}X^{(m)}\right)^{-1}\Gamma^{(m)^\top}X^{\top}u.
$$

We expand the expectation in (A.18) and obtain

$$
\mathbb{E}\left(A^{(m)}A^{(s)^\top} + A^{(m)}B^{(s)^\top} + A^{(m)}d_1^{(s)^\top} + B^{(m)}A^{(s)^\top} + B^{(m)}B^{(s)^\top} + \Gamma^{(m)}\Gamma^{(s)^\top}X^{(s)^\top}X^{(s)}\right) = \sigma^2 \Gamma^{(m)}\left(X^{(m)^\top}X^{(m)}\right)^{-1}X^{(m)^\top}X^{(s)}\left(X^{(s)^\top}X^{(s)}\right)^{-1}\Gamma^{(s)^\top}.
$$

Note that $\mathbb{E}(B^{(m)}) = 0$ since $\mathbb{E}(u|X) = 0$. This implies that $\mathbb{E}(A^{(m)}B^{(s)^\top}) = 0$, $\mathbb{E}(B^{(m)}A^{(s)^\top}) = 0$, $\mathbb{E}(B^{(m)}d_1^{(s)^\top}) = 0$, and $\mathbb{E}(d_1^{(m)}B^{(s)^\top}) = 0$. Also,

$$
\mathbb{E}\left(B^{(m)}B^{(s)^\top}\right) = \sigma^2 \Gamma^{(m)}\left(X^{(m)^\top}X^{(m)}\right)^{-1}X^{(m)^\top}X^{(s)}\left(X^{(s)^\top}X^{(s)}\right)^{-1}\Gamma^{(s)^\top}.
$$

Finally, (A.19) is equivalent to

$$
\sigma^2 \Gamma^{(m)}\left(X^{(m)^\top}X^{(m)}\right)^{-1}X^{(m)^\top}X^{(s)}\left(X^{(s)^\top}X^{(s)}\right)^{-1}\Gamma^{(s)^\top} + \left(A^{(m)} + d_1^{(m)}\right)\left(A^{(s)} + d_1^{(s)}\right)^{\top}.
$$
Note that $A^{(m)} + d_1^{(m)} = \mathbb{E} \left( \Gamma^{(m)} \hat{\beta}^{(m)} \right) - \beta(w) = d^{(m)}$. In practice, we replace the infeasible $\sigma^2$ and $d^{(m)}$ with their sample estimates. This completes the proof.

Proof of Lemma 2.5. Part (i) follows the fact that the set of $\beta_m$ belongs to the set of $\beta_p$ for all $m$. Therefore, $\sqrt{n}\beta_m \rightarrow h_m$ by Assumption 2.5.

To prove part (ii), we first expand the estimator $\hat{\beta}^{(m)}$ around $\beta^{(m)}$ and obtain

$$\sqrt{n} \left( \hat{\beta}^{(m)} - \beta^{(m)} \right) = \sqrt{n} \left( X^{(m)}^T X^{(m)} \right)^{-1} X^{(m)}^T X^{(-m)} \beta_m + \sqrt{n} \left( X^{(m)}^T X^{(m)} \right)^{-1} \Gamma^{(m)}^T X^T u. $$

By Assumption 2.5 and Lemma 2.5 (i),

$$\sqrt{n} \left( X^{(m)}^T X^{(m)} \right)^{-1} X^{(m)}^T X^{(-m)} \beta_m \overset{p}{\rightarrow} \left( S^{(m)} \right)^{-1} S^{(m-m)} h_m \equiv \delta^{(m)}.$$ 

By Assumption 2.4 and the central limit theorem,

$$\sqrt{n} \left( X^{(m)}^T X^{(m)} \right)^{-1} \Gamma^{(m)}^T X^T u \overset{d}{\rightarrow} N \left( 0, \sigma^2 \left( S^{(m)} \right)^{-1} \right).$$

This implies $\sqrt{n} \left( \hat{\beta}_c^{(m)} - \beta_c \right) = \sqrt{n} \mathbb{I}^{(m)} ( \hat{\beta}^{(m)} - \beta^{(m)} ) \overset{d}{\rightarrow} N \left( \delta^{(m)}_c, \sigma^2 \mathbb{I}^{(m)} \right).$

To prove part (iii), we first have $\beta_m = O(n^{-1/2})$ by Assumption 2.5. Under Assumption 2.6, we apply Taylor expansion to $\beta(w)$ and obtain

$$\beta(w) = f(\beta^{(m)}, \beta_m) = f(\beta^{(m)}, 0) + \left( F_{\beta_m} \big|_{\beta_m = 0} \right) \beta_m + \frac{1}{2} \beta_m^T \bar{H} \beta_m, \quad (A.20)$$

where $\bar{H}$ is the Hessian matrix $H \equiv \partial^2 f(\cdot)/\partial \beta_m \partial \beta_m^T$ evaluated at an intermediate point between $0$ and $\beta_m$. Since $\beta_m = O(n^{-1/2})$, the last term in (A.20) is $O(n^{-1})$. 
With \( f(\beta^{(m)}, 0) = \Gamma^{(m)}\beta^{(m)} \), we have

\[
\sqrt{n} \left( \Gamma^{(m)}\beta^{(m)} - \beta(w) \right) = \sqrt{n} \left( - \left( F_{\beta_{-m}|_{\beta_{-m}=0}} \right) \beta_{-m} - O(n^{-1}) \right) \\
\to - \left( F_{\beta_{-m}|_{\beta_{-m}=0}} \right) h_{-m}.
\]

Therefore, \( \sqrt{n} (\beta_c - \beta_c(w)) = \sqrt{n} \Gamma^{(m)}_c \left( \Gamma^{(m)}\beta^{(m)} - \beta(w) \right) \to \delta_2. \)

Part (iv) follows Lemma 2.5 (ii) and (iii). We have

\[
\sqrt{n} d^{(m)} = \sqrt{n} \left( \mathbb{E}(\Gamma^{(m)}\hat{\beta}^{(m)}) - \beta(w) \right) \\
\overset{p}{\to} \Gamma^{(m)} \left( S^{(m,m)} \right)^{-1} S^{(m,m)} h_{-m} - \left( F_{\beta_{-m}|_{\beta_{-m}=0}} \right) h_{-m} = \delta^{(m)}.
\]

Part (v) is the main result of Lemma 2.5.

\[
\sqrt{n} \left( \hat{\beta}_c^{(m)} - \beta_c(w) \right) = \sqrt{n} \left( \hat{\beta}_c^{(m)} - \beta_c \right) + \sqrt{n} (\beta_c - \beta_c(w)) \\
\overset{d}{\to} N \left( \delta_c^{(m)}, \sigma^2 \Gamma^{(m)}_{\beta} \left( S^{(m)} \right)^{-1} \Gamma^{(m)}_{\beta} \right) \sim \Lambda_c^{(m)}.
\]

\[\Box\]

**Proof of Theorem 2.3.** It is straightforward to show that the asymptotic distribution of \( \Lambda_c \) is a normal distribution, since

\[
\sqrt{n} \left( \hat{\beta}_c(\hat{w}) - \beta_c(w) \right) | \hat{w} = \sum_{m=1}^{M} \hat{w}(m) \sqrt{n} \left( \hat{\beta}_c^{(m)} - \beta_c(w) \right) | \hat{w} \overset{d}{\to} \sum_{m=1}^{M} \hat{w}(m) \Lambda_c^{(m)}
\]

is a linear combination of normal distribution \( \Lambda_c^{(m)} \) for all \( m \). Given \( \mathbb{E} \left( \hat{\beta}_c(\hat{w}) - \beta_c(w) \right) = 0 \), the mean of \( \Lambda_c \) is 0. For the variance of \( \Lambda_c \), we have \( V_c(\hat{w}) = \Gamma^{(m)}_{\beta} V(\hat{w}) \Gamma^{(m)}_{\beta} \).
where

\[
V(\hat{w}) = \lim_{n \to \infty} \text{Var} \left( \sqrt{n} \left( \beta(\hat{w}) - \beta(w) \right) \bigg| \hat{w} \right) = \lim_{n \to \infty} n \sum_{m=1}^{M} \sum_{s=1}^{M} \hat{w}(m) \hat{w}(s) \left( \hat{\beta}(m) \hat{\beta}(s) \right)^T + \sigma^2 \Gamma(m) \left( X(m)^T X(m) \right)^{-1} X(m)^T X(s) \left( X(s)^T X(s) \right)^{-1} \Gamma(s)^T \bigg| \hat{w} \right)
\]

\[
= \lim_{n \to \infty} \sum_{m=1}^{M} \sum_{s=1}^{M} \hat{w}(m) \hat{w}(s) \left( \sqrt{n}d(m) \sqrt{n}d(s)^T + \sigma^2 \Gamma(m) \left( n^{-1} X(m)^T X(m) \right)^{-1} n^{-1} X(m)^T X(s) \left( n^{-1} X(s)^T X(s) \right)^{-1} \Gamma(s)^T \right) .
\]

We have proved in Lemma 2.5 that \( \sqrt{n}d(m) \overset{P}{\to} \delta(m) \). Therefore,

\[
V(\hat{w}) = \sum_{m=1}^{M} \sum_{s=1}^{M} \hat{w}(m) \hat{w}(s) \left( \sigma^2 \Gamma(m) \left( S(m) \right)^{-1} S(m,s) \left( S(s) \right)^{-1} \Gamma(s)^T + \delta(m) \delta(s)^T \right) ,
\]

where \( S(m) \), \( S(m,s) \), and \( \delta(m) \) are defined in Lemma 2.5.

**Proof of Theorem 2.4.** Since \( \sqrt{n}(\hat{\beta}(w) - \beta(w))|\hat{w} \to_d N(0, V_c(\hat{w})) \) and \( V_c(\hat{w}) \) is a linear function of \( \hat{w} \), it is equivalent to have \( \sqrt{n}(\hat{\beta}(w) - \beta(w))|V_c(\hat{w}) \to_d N(0, V_c(\hat{w})) \).

We can rewrite the \( t \)-statistic as

\[
t_{\beta_j(w)} = \frac{\hat{\beta}_j(\hat{w}) - \beta_j(0)}{\sqrt[1/2]{\text{Var}(\hat{\beta}_j(\hat{w}))}} = \frac{\sqrt{n}(\hat{\beta}_j(\hat{w}) - \beta_j(0))}{n\sqrt[1/2]{\text{Var}(\hat{\beta}_j(\hat{w}))}} .
\]
Following Theorem 2.3, for fixed $\mathbf{V}_c(\hat{w})$, we have

$$\sqrt{n} \left( \hat{\beta}_j(\hat{w}) - \beta_{j0} \right) = \sqrt{n} \left( \hat{\beta}_j(\hat{w}) - \beta_{j0} \right) \bigg| \mathbf{V}_c(\hat{w}) \to_d \Lambda_j$$

with mean 0 and variance $V_j(\hat{w})$, which is the $j^{th}$ element on the diagonal of $\mathbf{V}_c(\hat{w})$, and

$$n \overline{\text{Var}}(\hat{\beta}_j(\hat{w})) = \overline{\text{Var}} \left( \sqrt{n} (\hat{\beta}_j(\hat{w}) - \beta_{j0}) \bigg| \mathbf{V}_c(\hat{w}) \right) \to_p V_j(\hat{w}) \text{.}$$

Therefore, for fixed $\mathbf{V}_c(\hat{w})$, the conditional asymptotic distribution of $t_{\beta_j(w)}$ is clearly $N(0, 1)$. Since this conditional distribution does not depend on $\mathbf{V}_c(\hat{w})$, it also holds marginally.

Similarly, for the Wald statistic, we have

$$W_r = \left( \mathbf{R} \hat{\beta}_c(\hat{w}) - r \right)^\top \left( \mathbf{R} \overline{\text{Var}}(\hat{\beta}_c(\hat{w})) \mathbf{R}^\top \right)^{-1} \left( \mathbf{R} \hat{\beta}_c(\hat{w}) - r \right) \to_d \chi^2(k_r)$$

for fixed $\mathbf{V}_c(\hat{w})$. Since this conditional distribution does not depend on $\mathbf{V}_c(\hat{w})$, it also holds marginally. \qed
Appendix B

Appendix for Chapter 4

Proof of Lemma 4.1. See Appendix A for proof.

Proof of Lemma 4.2. The proof is straightforward. If $l^{(q)}/n \to 0$ for all $m$, then the linear combination of $l^{(q)}$, $l(w) = \sum_{q=1}^{Q} w^{(q)} (l^{(q)}/n) \to 0$. Therefore,

$$
\frac{n + l(w)}{n - l(w)} = 1 + \frac{2l(w)/n}{1 - l(w)/n} = 1 + o(1).
$$

Proof of Lemma 4.3. $R_n(w)$ can be written as

$$
R_n(w) = \mathbb{E} \left\| P_W X \beta - P(w) X \hat{\beta}(w) \right\| W \right\|^2 \\
= \mathbb{E} \left\| P_W X \beta - P(w) X (X^\top P(w) P(w) X)^{-1} X^\top P(w) P(w) y \right\| W \right\|^2 \\
= \mathbb{E} \left\| P_W X \beta - P(w) X \beta - P_{P(w)X} P(w) u \right\| W \right\|^2 \\
= \mathbb{E} \left\| (P_W - P(w)) X \beta \right\|^2 + \sigma^2 \text{Tr} \left( P(w) P_{P(w)X} P(w) \right)
$$
\[-2\mathbb{E}(\mathbf{u}^\top \mathbf{P}(\mathbf{w}) \mathbf{P}(\mathbf{w})^\top \mathbf{X} (\mathbf{P}_W - \mathbf{P}(\mathbf{w})) \mathbf{X} \beta | \mathbf{W}) \, . \tag{B.1} \]

By the assumption that \( \mathbb{E}(\mathbf{u} | \mathbf{W}) = \mathbb{E}(\mathbf{u}^\top \mathbf{W}) = \mathbb{E}(\mathbf{u}^\top \mathbf{P}(\mathbf{w})) = 0 \), the last term of (B.1) is zero. Also,

\[ \sigma^2 \text{Tr} (\mathbf{P}(\mathbf{w}) \mathbf{P}(\mathbf{w})^\top \mathbf{X} \mathbf{P}(\mathbf{w})) = \sigma^2 \text{Tr} (\mathbf{P}(\mathbf{w}) \mathbf{P}(\mathbf{w})). \tag{B.2} \]

Therefore,

\[ R_n(\mathbf{w}) = \mathbb{E} \left\| (\mathbf{P}_W - \mathbf{P}(\mathbf{w})) \mathbf{X} \beta \right\|^2 + \sigma^2 \text{Tr} (\mathbf{P}(\mathbf{w}) \mathbf{P}(\mathbf{w})). \tag{B.3} \]

Both terms in (B.3) are non-negative, which implies Lemma 3. \qed

Proof of Theorem 4.1. The proof of Theorem 4.1 follows the techniques derived in Appendix A. By definition, we have

\[ \tilde{\beta} = (\mathbf{X}^\top \mathbf{P}_{\mathbf{W}(\mathbf{w})} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_{\mathbf{W}(\mathbf{w})} \mathbf{y} = \beta + (\mathbf{X}^\top \mathbf{P}_{\mathbf{W}(\mathbf{w})} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_{\mathbf{W}(\mathbf{w})} \mathbf{u} \]

and

\[ \mathbf{X} - \mathbf{F}(\mathbf{w}) = \mathbf{X} - \mathbf{P}(\mathbf{w}) \mathbf{X} = \mathbf{M}(\mathbf{w}) \mathbf{X}. \]

Then, the MAIV criterion (4.13) can be written as

\[
\text{MAIV}_n(\mathbf{w}) \\
= \left\| (\mathbf{X} - \mathbf{F}(\mathbf{w})) \tilde{\beta} \right\|^2 \left( \frac{n + l(\mathbf{w})}{n - l(\mathbf{w})} \right)
\]
\[
\begin{align*}
&= \left\| (X - F(w)) \left( \beta + (X^\top P_{W(l)} X)^{-1} X^\top P_{W(l)} u \right) \right\|^2 (1 + o(1)) \\
&= \left\| X\beta - F(w)\beta + M(w)X (X^\top P_{W(l)} X)^{-1} X^\top P_{W(l)} u \right\|^2 (1 + o(1)) \\
&= \left\| \left( P_{W} X\beta - P(w)X\hat{\beta}(w) \right) + M_{W} X\beta \right\|^2 (1 + o(1)).
\end{align*}
\]

Define

\[
A \equiv P_{W} X\beta - P(w)X\hat{\beta}(w),
\]

\[
B \equiv M_{W} X\beta,
\]

\[
C \equiv \left( P_{P(w)} X P(w) + M(w)X (X^\top P_{W(l)} X)^{-1} X^\top P_{W(l)} \right) u.
\]

We have

\[
\text{MAIV}_n(w) = \left( \|A\|^2 + \|B\|^2 + \|C\|^2 + 2(B^\top A + B^\top C + C^\top A) \right) (1 + o(1)).
\]

We have \(\|A\|^2 = L_n(w)\) and \(\|B\|^2\) is independent of \(w\). Expand \(\|C\|^2\):

\[
\|C\|^2 = u^\top P(w) P_{P(w)} X P(w) u + \left\| M(w)X (X^\top P_{W(l)} X)^{-1} X^\top P_{W(l)} u \right\|^2 \\
+ 2u^\top P(w) P_{P(w)} X M(w)X (X^\top P_{W(l)} X)^{-1} X^\top P_{W(l)} u.
\]  

(B.4)

For \(B^\top A + B^\top C + C^\top A\), we have

\[
\begin{align*}
B^\top A + (A + B)^\top C \\
&= -\beta^\top X^\top M_{W} P(w)X\hat{\beta}(w) + \left( X\beta - P(w)X\hat{\beta}(w) \right)^\top P_{P(w)} X P(w) u \\
&\quad + \left( X\beta - P(w)X\hat{\beta}(w) \right)^\top M(w)X (X^\top P_{W(l)} X)^{-1} X^\top P_{W(l)} u.
\end{align*}
\]
To simplify the above equation, we make use of the following equations:

\[
P(w)X\hat{\beta}(w) = P_{P(w)}X P(w)y,
\]
\[
X^\top P(w)P_{P(w)}X = P(w)X,
\]
\[
-MWP(w) = MW(P_W - P(w)).
\]

Then, we have

\[
2 \left( B^\top A + B^\top C + C^\top A \right)
\]
\[
= 2 \left( \beta^\top X^\top (P_W - P(w)) P_{P(w)}X P(w)u + \beta^\top X^\top M_W (P_W - P(w)) X \beta 
+ \beta^\top X^\top (P_W - P(w)) M(w)X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u 
+ \beta^\top X^\top M_W (P_W - P(w)) X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u 
+ \beta^\top X^\top M_W X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u - u^\top P(w)P_{P(w)}X P(w)u 
- u^\top P(w)P_{P(w)}X M(w)X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u \right). \quad (B.5)
\]

Combining (B.5) with (B.4), we have

\[
2 \left( B^\top A + B^\top C + C^\top A \right) + \|C\|^2
\]
\[
= 2\beta^\top X^\top (P_W - P(w)) P_{P(w)}X P(w)u + 2\beta^\top X^\top M_W (P_W - P(w)) X \beta 
+ 2\beta^\top X^\top (P_W - P(w)) M(w)X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u 
+ 2\beta^\top X^\top M_W (P_W - P(w)) X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u 
+ 2\beta^\top X^\top M_W X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u 
+ \left\| M(w)X \left( X^\top P_{P_W(l)}X \right)^{-1} X^\top P_{P_W(l)}u \right\|^2 - u^\top P(w)P_{P(w)}X P(w)u.
\]
Note that the term $\beta^\top X^\top M_W X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u$ is independent of $w$. Also,

$$
\left\| M(w) X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u \right\|^2 - u^\top P(w) P_{P(w)} X P(w) u \\
= \left\| P(w) X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u \right\|^2 - u^\top P(w) P_{P(w)} X P(w) u \\
+ \left\| X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u \right\|^2 \\
- 2u^\top P_{W^{(L)}} X (X^\top P_{W^{(L)}} X)^{-1} X^\top P(w) X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u
$$

The term $\left\| X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u \right\|^2$ is independent of $w$. Define

$$
D_1 \equiv -u^\top P(w) P_{P(w)} X P(w) u, \\
D_2 \equiv \left\| P(w) X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u \right\|^2, \\
D_3 \equiv -2u^\top P_{W^{(L)}} X (X^\top P_{W^{(L)}} X)^{-1} X^\top P(w) X (X^\top P_{W^{(L)}} X)^{-1} X^\top P_{W^{(L)}} u
$$

and $D \equiv D_1 + D_2 + D_3 = (D_1 + \sigma^2 \text{Tr}(P(w) P(w))) + (D_2 - \sigma^2 \text{Tr}(P(w) P(w))) + D_3$. To prove that as $n \to \infty$

$$
\sup_{w \in H_Q^{(p)}} \left| \frac{D}{R_n(w)} \right| \to_p 0,
$$

it suffices to show that

$$
\sup_{w \in H_Q^{(p)}} \left| \left( D_1 + \sigma^2 \text{Tr}(P(w) P(w)) \right) / R_n(w) \right| \to_p 0, \quad (B.6) \\
\sup_{w \in H_Q^{(p)}} \left| \left( D_2 - \sigma^2 \text{Tr}(P(w) P(w)) \right) / R_n(w) \right| \to_p 0, \quad (B.7) \\
\sup_{w \in H_Q^{(p)}} \left| D_3 / R_n(w) \right| \to_p 0. \quad (B.8)
$$
To prove (B.6), we use the triangle inequality, Bonferroni’s inequality, Chebyshev’s inequality, Theorem 2 of Whittle (1960), Lemma 4.1 and Lemma 4.3(ii). Following condition (B.2) and the fact that $P_{P(w)X}$ is a projection matrix, we observe, for any $\delta > 0$, that

$$P \left\{ \sup_{w \in H_p^Q} \left| \left\| P_{P(w)X}P(w)u \right\|^2 - \sigma^2 \text{Tr} \left( P(w)P(w) \right) \right| \geq R_n(w) \right\} \leq P \left\{ \sup_{w \in H_p^Q} \sum_{q=1}^{Q} \sum_{s=1}^{Q} u^T P_{W(q)}P_{P(w)X}P_{W(s)}u - \sigma^2 \text{Tr} \left( P_{W(q)}P_{P(w)X}P_{W(s)} \right) \right| \geq \delta \xi_n \right\} \leq \sum_{q=1}^{Q} \sum_{s=1}^{Q} \left( \text{Tr} \left( P_{W(q)}P_{P(w)X}P_{W(s)} \right) \right) \frac{2G}{\delta^{2G}\xi_n^{2G}} \leq C_1 \delta^{-2G}\xi_n^{-2G} \sum_{q=1}^{Q} \sum_{s=1}^{Q} \left( \text{Tr} \left( P_{W(q)}P_{P(w)X}P_{W(s)} \right) \right)^G \leq C_1' \delta^{-2G}\xi_n^{-2G} Q \sum_{q=1}^{Q} \left( R_n(w_q^0) \right)^G,$$

for some fixed constants $C_1$ and $C_1'$. By Assumption 4.3, we obtain (B.6).

To prove (B.7), we see that since $W^{(L)}$ is the largest instruments in $W^{(q)}$ for all $q$, then

$$P(w)P_{W^{(L)}} = \sum_{q=1}^{Q} w^{(q)} P_{W^{(q)}} P_{W^{(L)}} = P(w). \quad (B.9)$$
Hence

\[ D_2 = \left\| P(w)P_{W(L)}X (X^\top P_{W(L)}X)^{-1} X^\top P_{W(L)}u \right\|^2 = \left\| P(w)P_{P_{W(L)}}xu \right\|^2. \]

Using the same technique we derived to prove (B.6), for any \( \delta > 0 \), we have

\[
P \left\{ \sup_{w \in H_Q^p} \frac{\left\| P(w)P_{P_{W(L)}}xu \right\|^2 - \sigma^2 \text{Tr} \left( P(w)P(w) \right) }{R_n(w) } > \delta \right\}
\]

\[ \leq C_2 \delta^{-2G} \xi_n^{-2G} \sum_{q=1}^Q \sum_{s=1}^Q \left( \text{Tr} \left( P_{P_{W(L)}}xP(w_q^0)P(w_s^0)P_{P_{W(L)}}x \right) \right)^G \]

\[ \leq C_2' \delta^{-2G} \xi_n^{-2G} Q \sum_{q=1}^Q \left( R_n(w_q^0) \right)^G \to 0, \]

where \( C_2 \) and \( C'_2 \) are some constants.

To prove (B.8), we make use of condition (B.9). Then,

\[ D_3 = -2u^\top P_{P_{W(L)}}xP(w)P_{W(L)}xu = -2 \sum_{q=1}^Q u^\top P_{P_{W(L)}}xP_{W(q)}P_{P_{W(L)}}xu. \]

Since we have

\[ u^\top P_{P_{W(L)}}xP_{W(q)}P_{P_{W(L)}}xu \to_p \sigma^2 [q] \]
for each \( q \). Then, \( D_3 \to_p \sigma^2 l(w) \), which is a constant. Therefore, (B.8) is obtained, since \( R_n(w) \to \infty \) as \( n \to \infty \).

To prove that Theorem 4.1 is valid, it remains to show that as \( n \to \infty \),

\[
\sup_{w \in H^p_Q} \left| \frac{\beta^T X^T (P_W - P(w)) P_{P(w)} X P(w) u}{R_n(w)} \right| \to_p 0, \quad (B.10)
\]

\[
\sup_{w \in H^p_Q} \left| \frac{\beta^T X^T M_W (P_W - P(w)) X \beta}{R_n(w)} \right| \to_p 0, \quad (B.11)
\]

\[
\sup_{w \in H^p_Q} \left| \frac{\beta^T X^T (P_W - P(w)) M(w) X (X^T P_{W_l} X)^{-1} X^T P_{W_l} u}{R_n(w)} \right| \to_p 0, \quad (B.12)
\]

\[
\sup_{w \in H^p_Q} \left| \frac{L_n(w) - 1}{R_n(w)} \right| \to_p 0. \quad (B.13)
\]

\[
\sup_{w \in H^p_Q} \left| \frac{\beta^T X^T (P_W - P(w)) P_{P(w)} X P(w) u}{R_n(w)} \right| \to_p 0. \quad (B.14)
\]

Using the triangle inequality, Bonferroni’s inequality, Chebyshev’s inequality, Theorem 2 of Whittle (1960), Lemma 4.1, Lemma 4.3(i) and Assumption 4.3, we observe that, for any \( \delta > 0 \),

\[
P \left\{ \sup_{w \in H^p_Q} \left| \frac{\beta^T X^T (P_W - P(w)) P_{P(w)} X P(w) u}{R_n(w)} \right| > \delta \right\} \leq P \left\{ \sup_{w \in H^p_Q} \left| \beta^T X^T (P_W - P(w)) P_{P(w)} u \right| > \delta \xi_n \right\}
\]

\[
\leq P \left\{ \sum_{q=1}^Q \sum_{s=1}^Q \left| \beta^T X^T (P_W - P_{w_q}) P_{w_s} u \right| > \delta \xi_n \right\}
\]

\[
\leq \sum_{q=1}^Q \sum_{s=1}^Q P \left\{ \left| \beta^T X^T (P_W - P(w_{q,s})) P_{w_{q,s}} u \right| > \delta \xi_n \right\}
\]
\[
\begin{align*}
&\leq \sum_{q=1}^{Q} \sum_{s=1}^{Q} \mathbb{E} \left[ \frac{(\beta^\top X^\top(P_W - P(w^0_q))P(w^0_s)u)_{2G}}{\delta^{2G} \xi^{2G}_n} \right] \\
&\leq C_3 \delta^{-2G} \xi^{-2G}_n \sum_{q=1}^{Q} \sum_{s=1}^{Q} \|P(w^0_s)(P_W - P(w^0_q))X\beta\|^{2G} \\
&\leq C_3 \delta^{-2G} \xi^{-2G}_n \sum_{q=1}^{Q} \sum_{s=1}^{Q} \|X\beta\|^{2G} \\
&\leq C_3 \delta^{-2G} \xi^{-2G}_n Q \sum_{q=1}^{Q} (R_n(w^0_q))^{G} \to 0
\end{align*}
\]

for a fixed constant \(C_3\).

Similarly, for (B.11), we observe that, for any \(\delta\),

\[
\begin{align*}
P \left\{ \sup_{w \in H^P_Q} \left| \frac{\beta^\top X^\top M_W(P_W - P(w))X\beta}{R_n(w)} \right| > \delta \right\} \\
&\leq P \left\{ \sup_{w \in H^P_Q} \left| \beta^\top X^\top (P_W - P(w))X\beta \right| > \delta \xi_n \right\} \\
&\leq P \left\{ \sup_{w \in H^P_Q} \sum_{q=1}^{Q} w^{(q)} \left| \beta^\top X^\top (P_W - P_{w^{(q)}})X\beta \right| > \delta \xi_n \right\}
\end{align*}
\]

Since the set of \(W^{(q)}\) is always nested in the set of \(W\) for all \(q\), we have

\[
P_W - P_{W^{(q)}} = (P_W - P_{W^{(q)}})(P_W - P_{W^{(q)}}).
\]

Therefore,

\[
\begin{align*}
P \left\{ \sup_{w \in H^P_Q} \left| \frac{\beta^\top X^\top M_W(P_W - P(w))X\beta}{R_n(w)} \right| > \delta \right\} \\
&\leq P \left\{ \sup_{w \in H^P_Q} \sum_{q=1}^{Q} w^{(q)} \left| \beta^\top X^\top (P_W - P_{W^{(q)}})X\beta \right| > \delta \xi_n \right\}
\end{align*}
\]
\[
\begin{align*}
&\leq P \left\{ \sup_{w \in H_Q^P} \sum_{q=1}^Q w(q) \left| \beta^\top X^\top (P_W - P_{W(q)}) (P_W - P_{W(0)}) X \beta \right| > \delta \xi_n \right\} \\
&\leq \sum_{q=1}^Q \mathbb{E} \left[ \left( \frac{\beta^\top X^\top (P_W - P_{W(q)}) (P_W - P_{W(0)}) X \beta}{\delta^2 \xi_n} \right)^{2G} \right] \\
&\leq C_4 \delta^{-2G} \xi_n^{-2G} \sum_{q=1}^Q \left( R_n(w_q^0) \right)^G \to 0,
\end{align*}
\]

where \( C_4 \) is a constant.

The same technique can be used to prove (B.12). For any \( \delta \), we have

\[
\begin{align*}
&\leq P \left\{ \sup_{w \in H_Q^P} \left| \beta^\top X^\top (P_W - P_{W(0)}) \right| \frac{M(w) X (X^\top P_{W(0)} X)^{-1} X^\top P_{W(0)} u}{R_n(w)} > \delta \xi_n \right\} \\
&\leq P \left\{ \sup_{w \in H_Q^P} \left| \beta^\top X^\top (P_W - P_{W(0)}) \right| M(w) X (X^\top P_{W(0)} X)^{-1} X^\top P_{W(0)} u > \delta \xi_n \right\} \\
&\leq P \left\{ \sup_{w \in H_Q^P} \sum_{q=1}^Q \sum_{s=1}^Q u(q) w(s) \left| \beta^\top X^\top (P_W - P_{W(0)}) \right| M_{W(s)} X (X^\top P_{W(0)} X)^{-1} X^\top P_{W(0)} u > \delta \xi_n \right\} \\
&\leq \sum_{q=1}^Q \sum_{s=1}^Q \mathbb{E} \left[ \left( \frac{\beta^\top X^\top (P_W - P_{W(q)}) M_{W(s)} X (X^\top P_{W(0)} X)^{-1} X^\top P_{W(0)} u}{\delta^2 \xi_n} \right)^{2G} \right] \\
&\leq \sum_{q=1}^Q \sum_{s=1}^Q \mathbb{E} \left[ \left( \frac{M_{W(s)} (P_W - P_{W(0)}) X \beta, X (X^\top P_{W(0)} X)^{-1} X^\top P_{W(0)} u}{\delta^2 \xi_n} \right)^{2G} \right].
\end{align*}
\]
Note that

\[ \mathbb{E} \left\| X \left( X^\top P_{W(L)} X \right)^{-1} X^\top P_{W(L)} u \right\|^2 \]
\[ = \sigma^2 \text{Tr} \left( P_{W(L)} X \left( X^\top P_{W(L)} X \right)^{-1} X^\top X \left( X^\top P_{W(L)} X \right)^{-1} X^\top P_{W(L)} \right) \]
\[ = \sigma^2 \text{Tr}(I_k) = \sigma^2 k. \]

By Theorem 2 of Whittle (1960), we have

\[
P \left\{ \sup_{w \in H_Q} \frac{\left| \beta^\top X^\top (P_W - P(w)) M(w) X \left( X^\top P_{W(L)} X \right)^{-1} X^\top P_{W(L)} u \right|}{R_n(w)} > \delta \right\} \]
\[ \leq C_5 \delta^{-2G} \xi_n^{-2G} \sum_{q=1}^{Q} \sum_{s=1}^{Q} \left\| M_{W(q)}(P_W - P_{W(q)}) X \beta \right\|^{2G} \]
\[ \leq C_5 \delta^{-2G} \xi_n^{-2G} \sum_{q=1}^{Q} \sum_{s=1}^{Q} \left( R_n(w_{q_s}) \right)^G \to 0 \]

for a fixed constant \( C_5 \).

For (B.13), we have, for any \( \delta \)

\[
P \left\{ \sup_{w \in H_Q} \frac{\left| \beta^\top X^\top M_W(P_W - P(w)) X \left( X^\top P_{W(L)} X \right)^{-1} X^\top P_{W(L)} u \right|}{R_n(w)} > \delta \right\} \]
\[ = P \left\{ \sup_{w \in H_Q} \frac{\left| \beta^\top X^\top M_W(P_W - P(w)) P_{W(L)} X \left( X^\top P_{W(L)} X \right)^{-1} X^\top P_{W(L)} u \right|}{R_n(w)} > \delta \right\} \]
\[ = P \left\{ \sup_{w \in H_Q} \frac{\left| \beta^\top X^\top M_W(P_W - P(w)) P_{W(L)} X \right| u}{R_n(w)} > \delta \xi_n \right\} \]
\[ \leq P \left\{ \sup_{w \in H_Q} \left| \beta^\top (P_W - P(w)) u \right| > \delta \xi_n \right\} \]
\[
\leq \sum_{q=1}^{Q} \mathbb{E} \left[ \frac{(\beta^T X^T (P_W - P(w_0^q)) u)^{2G}}{\delta^{2G} \xi^2 \zeta_n} \right]
\]
\[
\leq C_6 \delta^{2G} \xi_n^{-2G} \sum_{q=1}^{Q} \| (P_W - P(w_0^q)) X \beta \|^2 G
\]
\[
\leq C_6 \delta^{2G} \xi_n^{-2G} \sum_{q=1}^{Q} (R_n(w_0^q))^G \rightarrow 0,
\]

where \( C_6 \) is a constant.

For (B.14), it is readily seen that
\[
\sup_{w \in H^p_Q} \left| \frac{L_n(w)}{R_n(w)} - 1 \right| \rightarrow_p 0
\]
\[
\iff \sup_{w \in H^p_Q} \left| \left( \left\| P_{P(w)} X P(w) u \right\|^2 - \sigma^2 \text{Tr}(P(w)P(w)) \right) - 2\beta^T X^T (P_W - P(w)) P_{P(w)} X P(w) u \right| / R_n(w) \rightarrow_p 0.
\]

Then, it suffices to show that, as \( n \rightarrow \infty \),
\[
\sup_{w \in H^p_Q} \left| \frac{\beta^T X^T (P_W - P(w)) P_{P(w)} X P(w) u}{R_n(w)} \right| \rightarrow_p 0,
\]
\[
\sup_{w \in H^p_Q} \left| \left\| P_{P(w)} X P(w) u \right\|^2 - \sigma^2 \text{Tr}(P(w)P(w)) \right| / R_n(w) \rightarrow_p 0,
\]

which are proven in (B.10) and (B.6), respectively. This completes the proof of Theorem 1.
Proof of Theorem 5.2. We expand (5.17)

\[ R_T(w) = \mathbb{E} \left[ y_{T+1} - \hat{f}_{T+1}(w) \right]^2 - \sigma^2 \]

\[ = \mathbb{E} \left[ u_{T+1} + x_{T+1|T} \left( \beta - \hat{\beta}(w) \right) \right]^2 - \sigma^2 \]

\[ = \mathbb{E} \left[ x_{T+1|T} \left( \beta - \hat{\beta}(w) \right) \right]^2 + 2\mathbb{E} \left[ u_{T+1} x_{T+1|T} \left( \beta - \hat{\beta}(w) \right) \right] + \mathbb{E}(u_{T+1}^2 - \sigma^2) \]

\[ = \mathbb{E} \left[ \mu_{T+1} - \mu_{T+1}(w) \right]^2 + 2\mathbb{E} \left[ u_{T+1} x_{T+1|T} (\beta - \hat{\beta}(w)) \right] + \mathbb{E}(u_{T+1}^2 - \sigma^2) \quad (C.1) \]

The last two terms in (C.1) equal 0 since \( \mathbb{E}(u_t|x_t) = 0 \) and \( \mathbb{E}(u_t^2|x_t) = \sigma^2 \) for all \( t \).

With the assumption that \((x_t, u_t)\) is strictly stationary and ergodic, \( \mathbb{E} x_t^\top x_t < \infty \), the first term of (C.1) converges to its stationary value such that

\[ \mathbb{E} \left[ \mu_{T+1} - \mu_{T+1}(w) \right]^2 = \mathbb{E} \left[ \mu_T - \mu_T(w) \right]^2 \]

\[ = \frac{1}{T} \left( \mathbb{E} \left[ \mu_1 - \mu_1(w) \right]^2 + \ldots + \mathbb{E} \left[ \mu_T - \mu_T(w) \right]^2 \right) \]

\[ = \frac{1}{T} \mathbb{E} (\mu(w) - \mu)^\top (\mu(w) - \mu) \]

\[ = L_T(w) \]
Finally, we obtain
\[ R_T(w) = L_T(w). \]

\[ \square \]

Proof of Theorem 5.3. First, define the average projection matrix
\[ M(w) \equiv I - P(w). \]

We expand MAPC \( T(w) \), as follows
\[
\begin{align*}
\text{MAPC}_T(w) &= (y - \mu(w))^\top(y - \mu(w)) \left( \frac{T + k(w)}{T - k(w)} \right) \\
&= (TL_T(w) + u^\top u) \left( \frac{T + k(w)}{T - k(w)} \right) + 2(\mu^\top A(w)u) \left( \frac{T + k(w)}{T - k(w)} \right).
\end{align*}
\]

Therefore,
\[
\frac{\text{MAPC}_T(w)}{T(L_T(w) + \sigma^2)} = \frac{T + k(w)}{T - k(w)} + \frac{2(\mu^\top M(w)u) \left( \frac{T + k(w)}{T - k(w)} \right)}{T(L_T(w) + \sigma^2) \left( \frac{T + k(w)}{T - k(w)} \right)}.
\] (C.2)

Since \( k^{(m)}/T \rightarrow 0 \) as \( n \rightarrow \infty \), as demonstrated by Lemma 2.3 in Chapter 2, \( k(w)/T \rightarrow 0 \). Therefore,
\[
\frac{T + k(w)}{T - k(w)} \rightarrow 1.
\]

As \( T \rightarrow \infty \), following the proof of Theorem 2.1 in Chapter 2, the last term in (C.2)
converge in probability to zero. Finally, we have

\[
\frac{\text{MAPC}_T(w)}{T(L_T(w) + \sigma^2)} \to_p 1.
\]