Queen's Economics Department Working Paper No. 1299

Least Squares Model Averaging by Prediction Criterion

Tian Xie

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

11-2012

Least Squares Model Averaging by Prediction Criterion (Job Market Paper)

Tian Xie *

November 8, 2012

Abstract

This paper 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. We propose computing weights by minimizing a model average prediction criterion (MAPC). We prove that the MAPC estimator is asymptotically optimal in the sense of achieving the lowest possible mean squared error. For statistical inference, we derive asymptotic tests for single hypotheses and joint hypotheses on the average coefficients for the "core" regressors. These regressors are of primary interest to us and are included in every approximation model. To improve the finite sample performance, we 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. We also show that the bootstrap tests have more reasonable rejection frequency than the asymptotic tests in small samples. As an empirical illustration, we apply the MAPC estimator to cross-country economic growth models.

JEL classification: C52, C53, O40

Keywords: Model Averaging, MAPC, Convex Optimization, Optimality, Statistical Inference

*Queen's University, Department of Economics, 94 University Ave., Kingston, Ontario, Canada K7L 3N6, xietian@econ.queensu.ca. I am extremely grateful to James G. MacKinnon and Morten Ørregaard Nielsen for their valuable supervision. I am also very grateful to Bruce E. Hansen, Jeffrey S. Racine, Joris Pinkse, Donald W. K. Andrews, and Joon Y. Park for their comments and suggestions. Helpful comments were received from participants at the Canadian Economics Association Conference held in Calgary (2012) and the Canadian Econometrics Study Group Annual Meeting in Kingston (2012).

1 Introduction

Economists formulate approximation models to capture the effects or factors supported by the empirical data. However, different approximation models usually yield different empirical results, which give rise to model uncertainty. There are two popular approaches for dealing with model uncertainty: model selection and model averaging.

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 them accordingly. One of the most widely used criterion functions is the Akaike information criterion (AIC) proposed by Akaike (1973). There are multiple versions of the AIC, the simplest of which is composed of a log-likelihood maximum and a penalty term. 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).

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. Leamer (1978) proposed the basic paradigm for Bayesian model averaging (BMA). Buckland, Burnham and Augustin (1997) suggested using exponential AIC estimates as model weights and proposed the model average AIC (MA-AIC). 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 (2011) proposed a jackknife model averaging (JMA) that considers heteroskedastic error settings. 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 continuous model weights with non-nested models.

There are alternatives to model selection and model averaging. Fan and Li (2001) studied penalized likelihood estimators. Knox, Stock and Watson (2004) proposed the empirical Bayes estimator. 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 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 (2005) provide an overview of, and selected bibliography regarding, GETS modeling.

In this paper, we propose a model average estimator with empirical weights computed through numerical minimization of a model average prediction criterion (MAPC). Our criterion can be seen as a model averaging version of the original prediction criterion proposed by Amemiya (1980). We 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 to non-nested approximation models. We 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, we derive asymptotic tests for single hypotheses and joint hypotheses on the core average coefficients. To improve the finite sample performance, we 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. We also show that the bootstrap tests have more reasonable rejection frequency than the asymptotic tests in small samples. As an empirical illustration, we apply the MAPC estimator to the cross-country economic growth models in Barro (1991).

This paper continues with an introduction of the framework of MAPC in Section 2. Section 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 4 presents simulation experiments. In Section 5, we provide an empirical application, in which the MAPC estimator is applied to the economic growth models in Barro (1991). Section 6 concludes the paper. Proofs are presented in Appendix A and a description of the data set used in Section 5 can be found in Appendix B.

2 Model Averaging Prediction Criterion

Let $(y_i, \boldsymbol{x_i}) : i = 1, ..., n$ be a random sample, where y_i and $\boldsymbol{x_i} = [x_{i1}, x_{i2}, ...]$ are real-valued. We let $\boldsymbol{x}_i$ be countably infinite. The same design can be seen in Hansen (2007) and Wan et al. (2010). We assume the data generating process is

$$y_i = \mu_i + u_i, \tag{1}$$

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

We consider a sequence of linear approximation models $m = 1, 2, ..., M$. The concept of "approximation model" can be vague. In this paper, an approximation model m uses $k^{(m)}$ regressors belonging to $\boldsymbol{x_i}$ such that

$$y_i = \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)} + u_i^{(m)} \quad \text{for } i = 1, ..., n, \tag{2}$$

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 paper. The approximation error for model m is defined as

$$b_i^{(m)} \equiv \mu_i - \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)} \quad \text{for } i = 1, ..., n. \tag{3}$$

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

Hansen (2007) assumed that the regressors $\boldsymbol{x_i}$ were an ordered set and an approximation model m contained the first $k^{(m)}$ regressors from $\boldsymbol{x_i}$. 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 (2011). In this paper, we place no such restrictions on the orders of $\boldsymbol{x_i}$ and the $k^{(m)}$. Approximation models in our paper can be either nested or non-nested, which makes our method more widely applicable.

The DGP (1) and approximation model (2) can be represented in the following matrix forms:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{u}$$

and

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

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

$$\boldsymbol{P}^{(m)} = \boldsymbol{X}^{(m)} \left(\boldsymbol{X}^{(m)\top} \boldsymbol{X}^{(m)} \right)^{-1} \boldsymbol{X}^{(m)\top}. \tag{4}$$

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

Let $\boldsymbol{w} = \left[w^{(1)}, ..., w^{(M)}\right]^\top$ be a weight vector in the unit simplex in $\mathbb{R}^M$ and define

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

where $\boldsymbol{H_M}$ is a continuous set. Note that Hansen (2007) and Hansen and Racine (2011) assumed a discrete set $\boldsymbol{H_M^*}$ for $\boldsymbol{w}$, in which

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

for some fixed integer N. We will discuss this further in Section 3.

We define the model average estimator of $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^{(m)} \hat{\boldsymbol{\mu}}^{(m)} = \sum_{m=1}^{M} w^{(m)} \boldsymbol{P}^{(m)} \boldsymbol{y}. \tag{5}$$

For the sake of simplicity, we define a weighted average projection matrix $\boldsymbol{P}(\boldsymbol{w})$ as

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

Accordingly, (5) can be simplified to

$$\boldsymbol{\mu}(\boldsymbol{w}) = \boldsymbol{P}(\boldsymbol{w}) \boldsymbol{y}. \tag{6}$$

The effective number of parameters $k(\boldsymbol{w})$ in model averaging estimation is defined as

$$k(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^{(m)} k^{(m)}, \tag{7}$$

which is a weighted sum of the $k^{(m)}$. Note that $k(\boldsymbol{w})$ is not necessarily an integer.

We propose the model average prediction criterion (MAPC):

$$\text{MAPC}_n(\boldsymbol{w}) = \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big)^\top \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big) \left(\frac{n + k(\boldsymbol{w})}{n - k(\boldsymbol{w})}\right), \tag{8}$$

where $\boldsymbol{\mu}(\boldsymbol{w})$ and $k(\boldsymbol{w})$ are defined in (6) and (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 follows the idea of parsimony and balances between the fitness and the size of a model. MAPC can be used to calculate the empirical

weight vector $\hat{\boldsymbol{w}}$, in which

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\boldsymbol{H_M}}{\arg\min}\, \mathrm{MAPC}_n(\boldsymbol{w}).$$

According to Hansen (2007), the Mallows' model average (MMA) criterion is

$$\mathrm{MMA}_n(\boldsymbol{w}) = \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big)^\top \big(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\big) + 2\sigma^2 k(\boldsymbol{w}). \tag{9}$$

The empirical weights $\hat{\boldsymbol{w}}$ can be selected by minimizing (9) subject to $\boldsymbol{w} \in \boldsymbol{H}_{\boldsymbol{M}}^*$. 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 σ^2 that must be replaced by a sample estimate.

For convenience in calculations, we rewrite both criteria. First, denote $\hat{\boldsymbol{u}}^{(m)}$ as an $n \times 1$ estimated residual vector from model m. Let $\hat{\boldsymbol{U}}$ be an $n \times M$ matrix consisting of these residuals such that $\hat{\boldsymbol{U}} \equiv [\hat{\boldsymbol{u}}^{(1)}, \hat{\boldsymbol{u}}^{(2)}, ..., \hat{\boldsymbol{u}}^{(M)}]$. Define an $M \times 1$ vector $\boldsymbol{k}$ which contains the number of parameters from each model such that $\boldsymbol{k} \equiv [k^{(1)}, k^{(2)}, ..., k^{(M)}]^\top$. Then, the MAPC in (8) can be written as

$$\mathrm{MAPC}_n(\boldsymbol{w}) = \boldsymbol{w}^\top \hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} \boldsymbol{w} \left(\frac{n + \boldsymbol{k}^\top \boldsymbol{w}}{n - \boldsymbol{k}^\top \boldsymbol{w}} \right). \tag{10}$$

Likewise, the MMA criterion in (9) becomes

$$\mathrm{MMA}_n(\boldsymbol{w}) = \boldsymbol{w}^\top \hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} \boldsymbol{w} + 2\sigma^2 \boldsymbol{k}^\top \boldsymbol{w}. \tag{11}$$

3 Asymptotic Properties

In this section, we first prove the asymptotic optimality of the MAPC estimator by showing that it achieves the lowest possible mean squared error. Then, we derive asymptotic tests for single hypotheses and joint hypotheses on the average coefficients. We also recommend bootstrap tests for improved inference in finite samples. Proofs for this section are presented in Appendix A.

3.1 Asymptotic Optimality

We start by listing some properties of $\boldsymbol{P}(\boldsymbol{w})$.

Lemma 1 *Define* $\boldsymbol{M}(\boldsymbol{w}) \equiv \boldsymbol{I} - \boldsymbol{P}(\boldsymbol{w})$. *We have*

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

(ii) $\lambda_{\max}\big(\boldsymbol{P}(\boldsymbol{w})\big) \leq 1$, *where* $\lambda_{\max}(\cdot)$ *returns the largest eigenvalue of its argument,*

(iii) $\left\|\boldsymbol{P}(\boldsymbol{w}^*)\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}\right\|^2 \leq \left\|\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}\right\|^2$ *for any* $\boldsymbol{w}^*, \boldsymbol{w} \in \boldsymbol{H_M}$,

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

The matrix $\boldsymbol{P}(\boldsymbol{w})$ is not a traditional projection matrix. For example, it is not idempotent. Therefore, we cannot further simplify Lemma 1 (iv).

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

Lemma 2 *We have*

(i) $R_n(\boldsymbol{w}) \geq \left\|\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}\right\|^2$,

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

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

Assumption 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

$$\mathbb{E}(|u_i|^{4(N+1)}|\boldsymbol{x}_i) \leq \kappa < \infty. \tag{12}$$

Note that (12) depends on model weights $\boldsymbol{w}$ through Hansen's assumption that $w^{(m)}$ is restricted to the set $\big\{0, \frac{1}{N}, \frac{2}{N}, ..., 1\big\}$ for some integer N.

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

Assumption 2 is the convergence condition. It is the same as the convergence condition in Wan et al. (2010). A necessary condition for Assumption 2 to hold is $\xi_n \to \infty$, which indicates that there is no finite approximating model for which the bias is zero. Moreover, we assume that as $n \to \infty$, ξ_n^{2G} goes to infinity at a faster rate than $M \sum_{m=1}^{M} \big(R_n(\boldsymbol{w}_m^0)\big)^G$. 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 $\boldsymbol{w}$; in contrast, our theorem is built on a more general set-up with non-nested models and continuous $\boldsymbol{H_M}$. In practice, Assumption 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 holds.

Assumption 3 *As* $n \to \infty$, $k^{(m)} \to \infty$ *and* $k^{(m)}/n \to 0$ *for all* m.

Assumption 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 3, we have

Lemma 3 *Let Assumption 3 hold. Then, $k(\boldsymbol{w})/n \to 0$ as $n \to \infty$.*

With the above lemmas and assumptions, we now show that the MAPC estimator is asymptotically optimal in the following theorem.

Theorem 1 *Let Assumptions 1, 2 and 3 hold. Then, as $n \to \infty$*

$$\frac{L_n(\hat{\boldsymbol{w}})}{L_n(\boldsymbol{w}_{opt})} \xrightarrow{p} 1,$$

where

$$\boldsymbol{w}_{opt} = \underset{\boldsymbol{w} \in \boldsymbol{H_M}}{\arg\inf} L_n(\boldsymbol{w}).$$

Theorem 1 states that by using the empirical weight vector $\hat{\boldsymbol{w}}$, 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 (5) where the weight vector belongs to the set $\boldsymbol{H_M}$.

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

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

By inserting $\hat{\sigma}^2(\boldsymbol{w})$ into (10) and rearranging the equation, we can rewrite MAPC as

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

which is similar to MMA in (11) with σ^2 being replaced by an average estimate $\hat{\sigma}^2(\boldsymbol{w})$.

Theorem 2 *Let Assumption 3 hold. Then $\hat{\sigma}^2(\boldsymbol{w}) \xrightarrow{p} \sigma^2$ as $n \to \infty$.*

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

$$\hat{\sigma}^2(\hat{\boldsymbol{w}}) = \frac{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} \hat{\boldsymbol{w}}}{n - \boldsymbol{k}^\top \hat{\boldsymbol{w}}}. \tag{13}$$

The MMA criterion includes an infeasible σ^2. Therefore, this must be computed with a sample estimate in practice. Hansen (2007, 2008) recommended using $\hat{\sigma}_L^2$ to replace the unknown σ^2, where

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

is the estimated σ^2 from a large approximation model L.[1] As a result, the MMA criterion in practice becomes

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

We can rewrite $\hat{\sigma}_L^2$ as an average estimator such that

$$\hat{\sigma}_L^2 = \frac{\left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(L)}\right)^\top \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(L)}\right)}{n - k^{(L)}} = \frac{\left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w}_L^0)\right)^\top \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w}_L^0)\right)}{n - \boldsymbol{k}^\top \boldsymbol{w}_L^0},$$

where $\boldsymbol{w}_L^0$ is a weight vector in which the L^{th} element is one and the others are zeros. In practice, using $\hat{\sigma}_L^2$ as an approximation to the infeasible σ^2 can be inefficient at times. As we show in the next section, simulation evidence indicates that the MMA estimator with $\hat{\sigma}_L^2$ yields a higher mean squared error than the MAPC estimator in many cases.

The MMA estimator is a two-step estimator since a sample estimate of σ^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 $\hat{\boldsymbol{w}}$ from the MMA criterion with constraints is a classic quadratic programming problem, while estimating $\hat{\boldsymbol{w}}$ by the MAPC estimator is a convex optimization problem.[2]

3.2 Estimating the Variance-Covariance Matrix for $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})$

The average coefficient $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})$ is not easy to compute. The number of regressors, $k^{(m)}$, is usually not the same for different models; even if the $k^{(m)}$ are 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 $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})$ can be computed by

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}}) = \sum_{m=1}^{M} \hat{w}^{(m)} \left(\boldsymbol{\Gamma}^{(m)} \hat{\boldsymbol{\beta}}^{(m)}\right),$$

where $\boldsymbol{\Gamma}^{(m)}$ is $k^{(L)} \times k^{(m)}$ and plays the role of mapping the $k^{(m)} \times 1$ vector $\hat{\boldsymbol{\beta}}^{(m)}$ to $k^{(L)} \times 1$ by filling the extra parameters with 0. A convenient way to construct $\boldsymbol{\Gamma}^{(m)}$ is to use the following equation:

$$\boldsymbol{\Gamma}^{(m)} = \left(\boldsymbol{X}^{(L)\top} \boldsymbol{X}^{(L)}\right)^{-1} \boldsymbol{X}^{(L)\top} \boldsymbol{X}^{(m)}. \tag{14}$$

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

[1] In fact, Hansen (2007, 2008) used the largest approximation model in his simulation experiments.

[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.

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

$$\begin{aligned}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})\right) &= \mathrm{Var}\left(\sum_{m=1}^{M}\hat{w}^{(m)}\left(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}\right)\right)\\ &= \sum_{m=1}^{M}\sum_{s=1}^{M}\hat{w}^{(m)}\hat{w}^{(s)}\mathrm{Cov}\left(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)},\boldsymbol{\Gamma}^{(s)}\hat{\boldsymbol{\beta}}^{(s)}\right). \qquad (15)\end{aligned}$$

The right-hand-side of equation (15) is a linear combination of $\mathrm{Cov}(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)},\boldsymbol{\Gamma}^{(s)}\hat{\boldsymbol{\beta}}^{(s)})$. When s equals m, the covariance matrix becomes the variance-covariance matrix of $\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}$. Each $\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}$ includes possible model misspecification bias. We define the following $k^{(L)}\times 1$ vector

$$\boldsymbol{d}^{(m)} = \mathbb{E}\left(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}\right) - \boldsymbol{\beta}(\boldsymbol{w}) \qquad (16)$$

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

Lemma 4 *For any approximation models m and s, in which m and s can represent the same model, we have*

$$\begin{aligned}&\widehat{\mathrm{Cov}}\left(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)},\boldsymbol{\Gamma}^{(s)}\hat{\boldsymbol{\beta}}^{(s)}\right)\\ =\ &\hat{\sigma}^2(\hat{\boldsymbol{w}})\boldsymbol{\Gamma}^{(m)}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(s)}\left(\boldsymbol{X}^{(s)\top}\boldsymbol{X}^{(s)}\right)^{-1}\boldsymbol{\Gamma}^{(s)\top}+\hat{\boldsymbol{d}}^{(m)}\left(\hat{\boldsymbol{d}}^{(s)}\right)^{\top},\end{aligned}$$

where

$$\hat{\boldsymbol{d}}^{(m)} = \boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)} - \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}}). \qquad (17)$$

For a particular average coefficient, for example $\hat{\beta}_j(\hat{\boldsymbol{w}})$, the variance of $\hat{\beta}_j(\hat{\boldsymbol{w}})$ is

$$\mathrm{Var}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big) = \Big[\mathrm{Var}\big(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})\big)\Big]_{jj},$$

which is the j^{th} element on the diagonal of the variance-covariance matrix $\mathrm{Var}\big(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})\big)$. Buckland et al. (1997) proposed other estimators for $\mathrm{Var}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big)$. Some of the estimators are based on a restrictive assumption that there is perfect correlation between each $\boldsymbol{\beta}^{(m)}$. They also proposed computing $\mathrm{Var}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big)$ 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. We name these regressors the core regressors. Let the $n\times k_c$ matrix $\boldsymbol{X}_c$ represent the core regressors and let $\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}})$ be the

corresponding averaged core coefficients. We define $\mathbf{\Gamma}_c^{(m)}$ as a $k^{(m)} \times k_c$ matrix that plays the role of subtracting the $n \times k_c$ matrix $\boldsymbol{X}_c$ out of any $\boldsymbol{X}^{(m)}$ such that $\boldsymbol{X}^{(m)}\mathbf{\Gamma}_c^{(m)} = \boldsymbol{X}_c$. Similar to $\mathbf{\Gamma}^{(m)}$ defined in (14), we can construct $\mathbf{\Gamma}_c^{(m)}$ using $\mathbf{\Gamma}_c^{(m)} = \left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\boldsymbol{X}_c$. The variance-covariance matrix for $\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}})$ is then

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}})\right) = \mathbf{\Gamma}_c^{(m)\top}\text{Var}\left(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})\right)\mathbf{\Gamma}_c^{(m)}.$$

We name the regressors that are of marginal interest to us potential regressors. These potential regressors are not included in every approximation model. We let $\boldsymbol{\beta}_p$ be the corresponding coefficient. The coefficient $\boldsymbol{\beta}_p$ is a $k_p \times 1$ vector with $k_p \leq k^{(L)}$. We define the regressors that are not included in $\boldsymbol{X}^{(m)}$ as $\boldsymbol{X}^{(-m)}$ and the associated coefficient as $\boldsymbol{\beta}_{-m}$. By definition, the set of $\boldsymbol{\beta}_{-m}$ belongs to the set of $\boldsymbol{\beta}_p$ for all m. Similar to the definition of $\mathbf{\Gamma}^{(m)}$ in (14), we let $\mathbf{\Gamma}^{(-m)} = \left(\boldsymbol{X}^{(L)\top}\boldsymbol{X}^{(L)}\right)^{-1}\boldsymbol{X}^{(L)\top}\boldsymbol{X}^{(-m)}$ play the role of mapping the $\left(k^{(L)} - k^{(m)}\right) \times 1$ vector $\boldsymbol{\beta}_{-m}$ to $k^{(L)} \times 1$. How to distinguish potential regressors from core regressors is not the primary concern of this paper. We leave that for future research.

3.3 Asymptotic Inference and Bootstrap Based Inference

We start by listing more assumptions.

Assumption 4 *We have*

(i) $(\boldsymbol{x}_i, u_i)$ *is an iid sequence,*

(ii) $\mathbb{E}(\boldsymbol{x}_i u_i) = \mathbf{0}$, $\mathbb{E}|x_{ij}u_i|^2 < \infty$, $\text{Var}(n^{-1/2}\boldsymbol{X}^\top\boldsymbol{u})$ *is positive definite,*

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

Assumption 5 $\boldsymbol{\beta}_p = \boldsymbol{h}_p/\sqrt{n}$, *where* $\boldsymbol{h}_p$ *is a fixed vector.*

Assumption 6 *The average coefficient* $\boldsymbol{\beta}(\boldsymbol{w})$ *is a function of* $\boldsymbol{\beta}^{(m)}$ *and* $\boldsymbol{\beta}_{-m}$ *that can be written as* $\boldsymbol{\beta}(\boldsymbol{w}) = \boldsymbol{f}\left(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}_{-m}\right)$, *where* $\boldsymbol{\beta}(\boldsymbol{w}) = \boldsymbol{f}\left(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}_{-m}\right)$ *is twice differentiable in a neighborhood of* $\boldsymbol{\beta}_{-m}$ *and* $\boldsymbol{f}\left(\boldsymbol{\beta}^{(m)}, \mathbf{0}\right) = \mathbf{\Gamma}^m\boldsymbol{\beta}^{(m)}$.

Assumption 4 is a standard assumption about regressors and error terms. Assumption 5 follows the logic that the potential regressors are only of marginal interest to us as they are weakly correlated with $\boldsymbol{y}$. This idea of weak variables is similar to ideas put forth in the weak instruments literature. For example, the Assumption L_Π in Staiger and Stock (1997)

is almost identical to our Assumption 5. Assumption 6 is a standard assumption that allows us to use Taylor expansion on $\boldsymbol{\beta}(\boldsymbol{w})$.

Let $\hat{\boldsymbol{\beta}}_c^{(m)} = \boldsymbol{\Gamma}_c^{(m)\top}\hat{\boldsymbol{\beta}}^{(m)}$ be the core part of $\hat{\boldsymbol{\beta}}^{(m)}$. We investigate the asymptotic distribution of $\hat{\boldsymbol{\beta}}_c^{(m)}$ in the following lemma:

Lemma 5 *Let Assumptions 4, 5, and 6 hold. Then, as $n \to \infty$,*

(i) $\boldsymbol{\beta}_{-m} = \boldsymbol{h}_{-m}/\sqrt{n}\ \forall m$, *where $\boldsymbol{h}_{-m}$ is a fixed vector.*

(ii) $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_c^{(m)} - \boldsymbol{\beta}_c\right) \xrightarrow{d} N\left(\boldsymbol{\delta}_1^{(m)}, \sigma^2\boldsymbol{\Gamma}_c^{(m)\top}\left(\boldsymbol{S}^{(m)}\right)^{-1}\boldsymbol{\Gamma}_c^{(m)}\right)$, *where*

$$\boldsymbol{\delta}_1^{(m)} \equiv \boldsymbol{\Gamma}_c^{(m)\top}\left(\boldsymbol{S}^{(m)}\right)^{-1}\boldsymbol{S}^{(m,-m)}\boldsymbol{h}_{-m}.$$

(iii) *Define* $\boldsymbol{F}_{\boldsymbol{\beta}_{-m}} \equiv \partial \boldsymbol{f}\left(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}_{-m}\right)^\top \Big/ \partial \boldsymbol{\beta}_{-m}$, *then*

$$\sqrt{n}\left(\boldsymbol{\beta}_c - \boldsymbol{\beta}_c(\boldsymbol{w})\right) \xrightarrow{p} \boldsymbol{\delta}_2^{(m)},$$

where $\boldsymbol{\delta}_2^{(m)} \equiv -\boldsymbol{\Gamma}_c^{(L)\top}\left(\boldsymbol{F}_{\boldsymbol{\beta}_{-m}}\big|_{\boldsymbol{\beta}_{-m}=\mathbf{0}}\right)\boldsymbol{h}_{-m}$.

(iv) $\sqrt{n}\boldsymbol{d}^{(m)} \xrightarrow{p} \boldsymbol{\delta}^{(m)}$, *where* $\boldsymbol{\delta}^{(m)} \equiv \boldsymbol{\Gamma}^{(m)\top}\left(\boldsymbol{S}^{(m)}\right)^{-1}\boldsymbol{S}^{(m,-m)}\boldsymbol{h}_{-m} - \left(\boldsymbol{F}_{\boldsymbol{\beta}_{-m}}\big|_{\boldsymbol{\beta}_{-m}=\mathbf{0}}\right)\boldsymbol{h}_{-m}$.

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

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_c^{(m)} - \boldsymbol{\beta}_c(\boldsymbol{w})\right) \to_d \boldsymbol{\Lambda}_c^{(m)} \sim \mathrm{N}\left(\boldsymbol{\delta}_c^{(m)}, \boldsymbol{V}_c^{(m)}\right), \tag{18}$$

where $\boldsymbol{\delta}_c^{(m)} = \boldsymbol{\delta}_1^{(m)} + \boldsymbol{\delta}_2^{(m)}$ *and* $\boldsymbol{V}_c^{(m)} \equiv \sigma^2\boldsymbol{\Gamma}_c^{(m)\top}\left(\boldsymbol{S}^{(m)}\right)^{-1}\boldsymbol{\Gamma}_c^{(m)}$.

Based on Lemma 5, we derive the asymptotic distribution of the core average coefficient $\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}})$ conditional on $\hat{\boldsymbol{w}}$ in the following theorem:

Theorem 3 *Let Assumptions 4, 5, and 6 hold. Then, as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}_c(\boldsymbol{w})\right)\Big|\hat{\boldsymbol{w}} \to_d \boldsymbol{\Lambda}_c = \sum_{m=1}^{M}\hat{w}^{(m)}\boldsymbol{\Lambda}_c^{(m)} \sim \mathrm{N}\Big(\mathbf{0}, \boldsymbol{\Gamma}_c^{(m)\top}\boldsymbol{V}(\hat{\boldsymbol{w}})\boldsymbol{\Gamma}_c^{(m)}\Big),$$

where $\boldsymbol{\Lambda}_c^{(m)}$ *is defined in* (18) *and*

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

The terms $\boldsymbol{S}^{(m)}$, $\boldsymbol{S}^{(m,s)}$, and $\boldsymbol{\delta}^{(m)}$ are defined in Assumption 4 and Lemma 5.

There is joint convergence of $\sqrt{n}(\hat{\boldsymbol{\beta}}_c^{(m)} - \boldsymbol{\beta}_c(\boldsymbol{w}))$ and the stochastic weights $\hat{\boldsymbol{w}}$. The conditional asymptotic distribution $\boldsymbol{\Lambda}$ is normal, which implies that the unconditional asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}_c(\boldsymbol{w}))$ is a mixed normal distribution. There is a large literature in time series that studies the mixed normal distribution and its inference (see Johansen (1995, pp.177–178) for a detailed explanation).

To test a single restriction, for example $\beta_j(\boldsymbol{w}) = \beta_{j0}$, we derive the t-statistic for the average core coefficient $\hat{\beta}_j(\hat{\boldsymbol{w}})$

$$t_{\hat{\beta}_j(\hat{\boldsymbol{w}})} \equiv \frac{\hat{\beta}_j(\hat{\boldsymbol{w}}) - \beta_{j0}}{\widehat{\mathrm{Var}}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big)^{1/2}},$$

where the estimated variance $\widehat{\mathrm{Var}}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big)$ is the j^{th} element on the diagonal of $\widehat{\mathrm{Var}}\big(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})\big)$:

$$\begin{aligned}\widehat{\mathrm{Var}}\big(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}})\big) \;=\; & \sum_{m=1}^{M}\sum_{s=1}^{M} \hat{w}^{(m)}\hat{w}^{(s)} \Bigg(\hat{\boldsymbol{d}}^{(m)} \left(\hat{\boldsymbol{d}}^{(s)}\right)^{\top} \\ & + \hat{\sigma}^2(\hat{\boldsymbol{w}}) \boldsymbol{\Gamma}^{(m)} \left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1} \boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(s)} \left(\boldsymbol{X}^{(s)\top}\boldsymbol{X}^{(s)}\right)^{-1} \boldsymbol{\Gamma}^{(s)\top} \Bigg),\end{aligned}$$

where $\hat{\boldsymbol{d}}^{(m)}$ and $\hat{\boldsymbol{d}}^{(s)}$ are defined in (17) and $\hat{\sigma}^2(\hat{\boldsymbol{w}})$ is defined in (13). To test the joint null hypothesis that $\boldsymbol{R}\boldsymbol{\beta}_c(\boldsymbol{w}) = \boldsymbol{r}$, where $\boldsymbol{r}$ is $k_r \times 1$, we use the Wald statistic such that

$$W_{\boldsymbol{r}} \equiv \left(\boldsymbol{R}\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{r}\right)^{\top} \left(\boldsymbol{R}\widehat{\mathrm{Var}}\big(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}})\big)\boldsymbol{R}^{\top}\right)^{-1} \left(\boldsymbol{R}\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{r}\right).$$

We derive the asymptotic distribution of the t-statistic and the Wald statistic in the following theorem.

Theorem 4 *Let Assumptions 4, 5, and 6 hold. Then*

$$t_{\hat{\beta}_j(\hat{\boldsymbol{w}})} \longrightarrow_d \mathrm{N}(0,1) \qquad \textit{and} \qquad W_{\boldsymbol{r}} \longrightarrow_d \chi^2(k_r).$$

We can construct the $1-\alpha$ confidence interval for $\hat{\beta}_j(\hat{\boldsymbol{w}})$ in the classical way:

$$\Big[\beta_{\text{lower}}, \quad \beta_{\text{upper}}\Big] = \Big[\hat{\beta}_j(\hat{\boldsymbol{w}}) - \hat{s}_j z_{1-\alpha/2}, \quad \hat{\beta}_j(\hat{\boldsymbol{w}}) + \hat{s}_j z_{1-\alpha/2}\Big],$$

where $\hat{s}_j$ is the standard error for $\hat{\beta}_j(\hat{\boldsymbol{w}})$ and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution.

Since both test statistics are asymptotically pivotal, we can use a semiparametric bootstrap test to provide improved statistical inference in finite samples. We denote $\hat{\tau}$ as an estimated test statistic (t or Wald). We first compute the estimated residuals $\tilde{\boldsymbol{u}}$ by plugging in the estimated average coefficients $\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{w}})$ under the null hypothesis. We then resample $\tilde{\boldsymbol{u}}$

B times and obtain B bootstrap samples $\boldsymbol{y}_l^*$ for $l = 1, ..., B$. The value of B should satisfy the condition that $\alpha(1 + B)$ is an integer,[4] where α is the desired level of significance. For each bootstrap sample $\boldsymbol{y}_l^*$, 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:

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

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 τ is symmetrically distributed around zero, as is the case in the t test, we can use the symmetric bootstrap:

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

If we are not willing to make the assumption that τ 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 4.2 as an example).

To construct symmetric bootstrap confidence intervals, we just need to invert the symmetric bootstrap test we described above. We can also use the bootstrap-t method to construct asymmetric bootstrap confidence intervals. We first sort the $\hat{\tau}_l^*$ in ascending order. We let the value of $c_{\alpha/2}^*$ be the value of number $\alpha(1 + B)/2$ bootstrap t-statistic in the sorted list. Similarly, the value of $c_{1-\alpha/2}^*$ is the value of number $(1 - \alpha/2)(1 + B)$ bootstrap t-statistic. Both $\alpha(1 + B)/2$ and $(1 - \alpha/2)(1 + B)$ should be integers. The bootstrap-t, or asymmetric equal-tail bootstrap confidence interval, is then

$$\Big[\beta_{\text{lower}}, \quad \beta_{\text{upper}}\Big] = \Big[\hat{\beta}_j(\hat{\boldsymbol{w}}) - \hat{s}_j c_{1-\alpha/2}^*, \quad \hat{\beta}_j(\hat{\boldsymbol{w}}) + \hat{s}_j c_{\alpha/2}^*\Big].$$

There are many ways to compute bootstrap confidence intervals. For a literature review, see DiCiccio and Efron (1996).

Computational cost can be a concern for model averaging estimation when the total number of approximation models is large. In this situation, we do not want to assign a huge

[4] If we are using a two-tailed test, it is helpful to make $\alpha(1+B)/2$ an integer for the purpose of constructing confidence intervals.

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).

4 Finite Sample Performance

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

4.1 The MAPC Estimator Versus Other Estimators

The general unrestricted model is the simple regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{19}$$

where $\boldsymbol{X}$ is $n \times k$, $\boldsymbol{\beta}$ is $k \times 1$ and $\boldsymbol{u}$ is $n \times 1$. The number of regressors k increases as n increases but at a slower rate, in which $k = \text{round}(3+n^{1/5})$. We set the first column of $\boldsymbol{X}$, $\boldsymbol{x}_1$, to be the intercept; the remaining $\boldsymbol{x_i}$ are assumed to be correlated with each other and are generated by $\text{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a $(k-1) \times (k-1)$ symmetric matrix with all diagonal terms equal to 1 and all off-diagonal terms equal to ρ. The $k \times 1$ coefficients $\boldsymbol{\beta}$ are determined by $\boldsymbol{\beta} = [1/5, 1/5, 5/\sqrt{n}, 4/\sqrt{n}, ..., (8-k)/\sqrt{n}]^\top$, where the first two regressors are core regressors and the remaining regressors are potential regressors. The error term $\boldsymbol{u}$ is independent of the regressors $\boldsymbol{X}$ and is distributed as $\text{N}(0, \sigma_u^2 \boldsymbol{I})$. The parameter σ_u controls the population $R^2 = \boldsymbol{\beta}_2^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_2 / \left(\boldsymbol{\beta}_2^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_2 + \sigma_u^2\right)$ so as to vary on a grid between 0.01 and 0.99, where $\boldsymbol{\beta}_2 = [\beta_2, ..., \beta_k]^\top$. We consider six different sample sizes, in which $n = 25, 50, 100, 200, 400$ and 800. Other simulation results, which are not reported here, demonstrate that the findings are not sensitive to alternative distributions. We also find that the results are not sensitive to different values of ρ, which we will show later.

All submodels that are nested in the general unrestricted model (19) are treated as approximation models. Therefore, 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 all the potential regressors.

We study seven methods: (1) general-to-specific approach (GETS); (2) 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 our experiment, we adopt a simple

GETS approach from Hendry and Nielsen (2007). We first estimate model (19). Then, regressors with the absolute value of the t-statistics smaller than $c_\alpha = 2$ are eliminated. If multiple t-statistics are smaller than 2, we eliminate the smallest one. 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

$$\text{AIC}^{(m)} = n\log(\hat{\sigma}_m^2) + 2k^{(m)}.$$

The model that achieves the lowest value among all of the estimated $\text{AIC}^{(m)}$ is selected. The model average AIC (MA-AIC) makes use of the estimated $\text{AIC}^{(m)}$ to compute the empirical weights, where

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

MA-BIC computes the empirical weights for its associated average estimator according to

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

where

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

Jackknife model averaging (JMA) (Hansen and Racine 2011) is also known as leave-one-out cross-validation. As its name indicates, JMA requires the jackknife residuals for the average estimator. The jackknife residual vector for model m can be conveniently written as $\hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(m)} = \boldsymbol{D}^{(m)}\hat{\boldsymbol{u}}^{(m)}$, where $\hat{\boldsymbol{u}}^{(m)}$ is the least squares residual vector and $\boldsymbol{D}^{(m)}$ 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 $\boldsymbol{P}^{(m)}$ defined in (4). Define an $n \times M$ matrix that collects all the jackknife residuals, in which $\hat{\boldsymbol{U}}_{\boldsymbol{J}} = \left[\hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(1)}, ..., \hat{\boldsymbol{u}}_{\boldsymbol{J}}^{(M)}\right]$. The least squares cross-validation criterion for JMA is simply

$$\text{CV}_n(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{w}^\top \hat{\boldsymbol{U}}_{\boldsymbol{J}}^\top \hat{\boldsymbol{U}}_{\boldsymbol{J}}\boldsymbol{w} \quad \text{with} \quad \hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\boldsymbol{H}_{\boldsymbol{M}}^*}{\text{argmin}}\, \text{CV}_n(\boldsymbol{w}).$$

The MAPC estimator and the MMA estimator are presented in previous sections. The infeasible σ^2 in (11) is replaced by a sample estimate, $\hat{\sigma}_L^2$, from the largest approximation model (19).

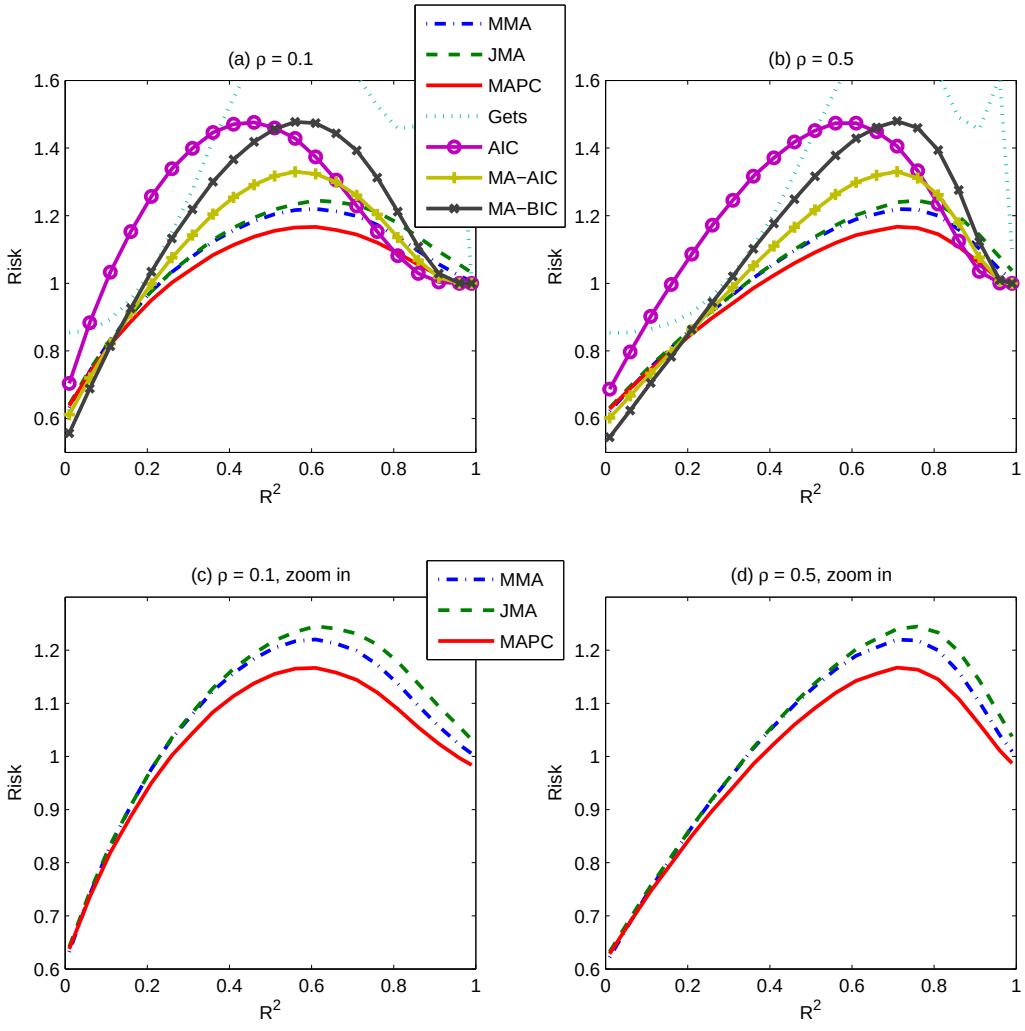
We define the risk of an estimator as the mean squared error such that

$$\text{Risk} \equiv \frac{1}{n}\big(\boldsymbol{\mu}(\hat{\boldsymbol{w}}) - \boldsymbol{\mu}_0\big)^\top \big(\boldsymbol{\mu}(\hat{\boldsymbol{w}}) - \boldsymbol{\mu}_0\big)$$

where $\boldsymbol{\mu}(\hat{\boldsymbol{w}})$ is the averaged $\boldsymbol{\mu}$ by the estimated $\hat{\boldsymbol{w}}$ and $\boldsymbol{\mu}_0$ is the true value of $\boldsymbol{\mu}$ (feasible in simulation). We compute risks for all seven estimators and average across 100,000 simulation draws. We normalize risks by dividing by the risk of the infeasible optimal least squares estimator (the OLS estimator of model (19)). We present the risk calculations for $n = 25$ with two different values of ρ (0.1 and 0.5) in Figures 1(a) and 1(b). The R^2 is presented on the x-axis and the risk is displayed on the y-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 1(a) and 1(b) show similar results. As $R^2 \to 1$, risks of all seven methods tend to converge to 1, which suggests that all corresponding estimators are converging to the infeasible optimal least squares estimator. In many cases, the GETS method shows a much

Figure 1: Simulation Results for $n = 25$



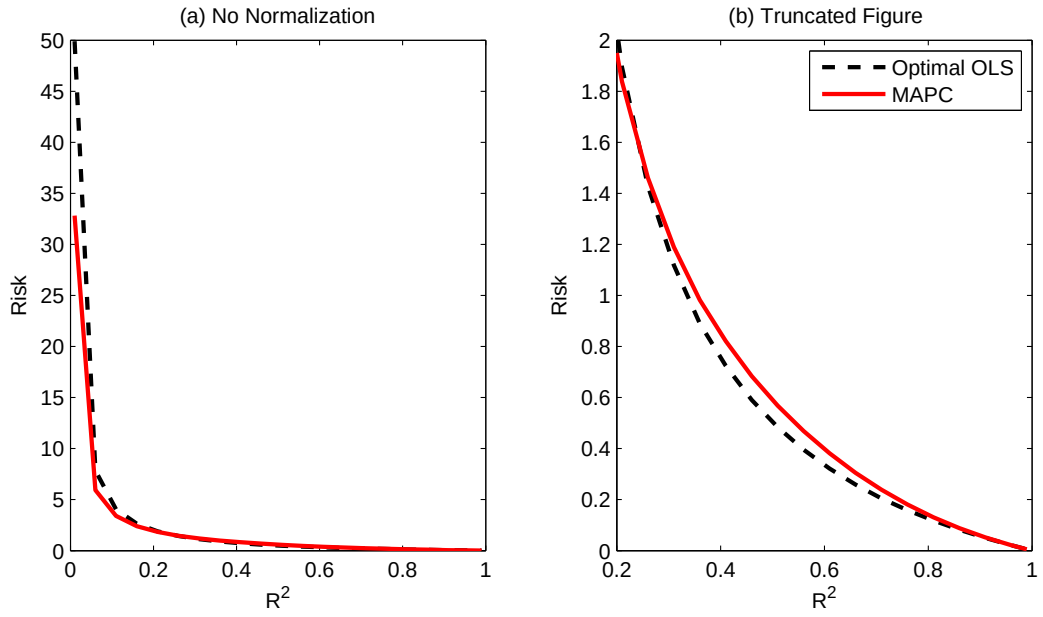
higher risk relative to other methods. We notice that when R^2 is close to 1, AIC yields the lowest risk. This means that the true model (19) is frequently selected by the AIC method, when there is sufficient information. However, AIC has a relatively poor performance for other values of R^2. The MA-AIC 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 yield lower risks than MMA, JMA, and MAPC. However, they are outperformed by MMA, JMA, and MAPC in most other cases that are more reasonable in practice.

It is instructive to compare the performance of MMA, JMA, with MAPC. In most cases, these three methods have better performance when compared to other methods. We provide Figures 1(c) and 1(d) to show only these three methods. For all values of R^2, the MAPC estimator achieves lower risks than the MMA estimator and the JMA estimator. This suggests that, in finite samples, MAPC is more efficient than MMA and JMA.

We notice that all the risk curves are hump-shaped. As we 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, we use the MAPC estimator as an example and present its estimated risk without normalization (denoted by solid lines) in Figure 2, along with the estimated risk of the optimal OLS estimator (denoted by dashed lines).

In Figure 2, part (a) represents the estimated risks for the MAPC estimator along with the OLS estimator. When R^2 is low, both estimators yield high risks. A low R^2 means

Figure 2: The Estimated Risks without Normalization



that the model does not fit the data very well. As a consequence, the estimated risks for the OLS estimator are high. The MAPC estimator, however, considers other approximation models as opposed to limiting itself to the unrestricted model (19). Therefore, as we see from the figure, the MAPC estimator yields 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 is the truncated version of part (a) with y-axis being 0 to 2 and x-axis being 0.2 to 1. As R^2 increases, the estimated risk for the OLS estimator decreases and eventually becomes smaller than the MAPC risks. Therefore, the corresponding normalized risk for MAPC is greater than 1. As we can see from the figure, the gap between the two curves first expands and then shrinks. The two methods yield almost identical risks 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 $R^2 = 0.99$.

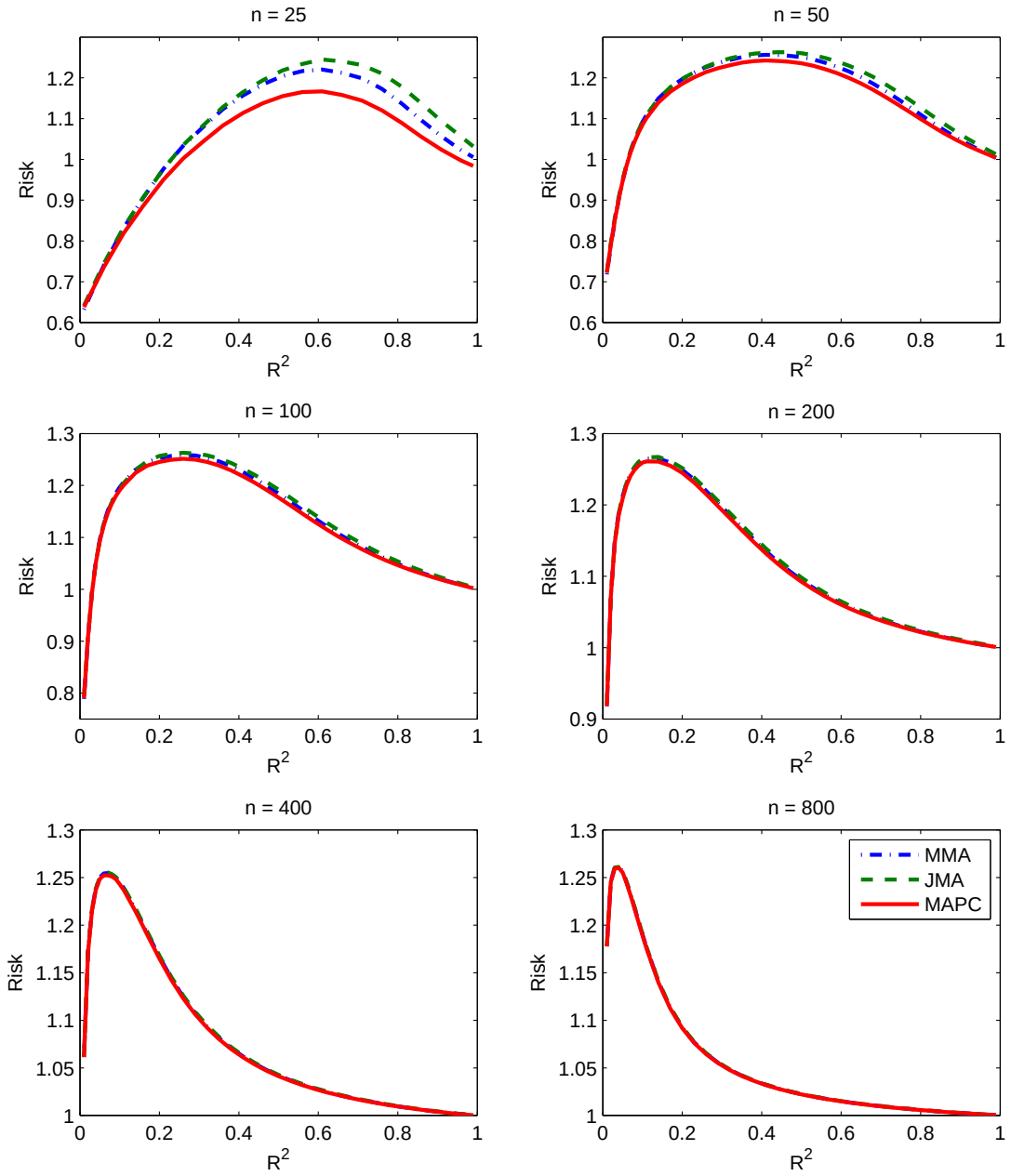
Risk calculations for various sample sizes ($n = 25, 50, 100, 200, 400$, and 800) are displayed in Figure 3. In this case, we set $\rho = 0.1$. To keep the figure uncluttered, only MMA, JMA and MAPC are displayed and represented by the dash-dotted line, dashed, and solid lines, respectively. In each panel, the MAPC estimator has a better performance than the MMA and JMA estimator by yielding lower risks. As n increases, the index on the y-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. We also 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. Therefore, the values of the normalized risks increases when R^2 is low.

4.2 Evaluate Tests via Rejection Frequency

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 data matrix 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 easily derive asymptotic tests (t and Wald) using the pairs bootstrap sample variance, henceforth pairs tests.

In Section 3.3, we derived the asymptotic tests (t and Wald) and the bootstrap tests for the average core coefficient. In this section, we compare the finite sample performance of these tests with pairs tests via the rejection frequency. We use the same simulation design we proposed in Section 4.1. We concentrate on the second coefficient, β_2 and the null hypothesis is that $\beta_2 = 1/5$, which is the true value of β_2. We generate 10,000 simulation draws for 15 sample sizes: $n = 25, 30, 35, ..., 95$. For each simulation draw, we compute the P values from the asymptotic tests, the semiparametric bootstrap tests and pairs tests for each sample size. In order to reduce the costs of doing the Monte Carlo experiments, the total number of

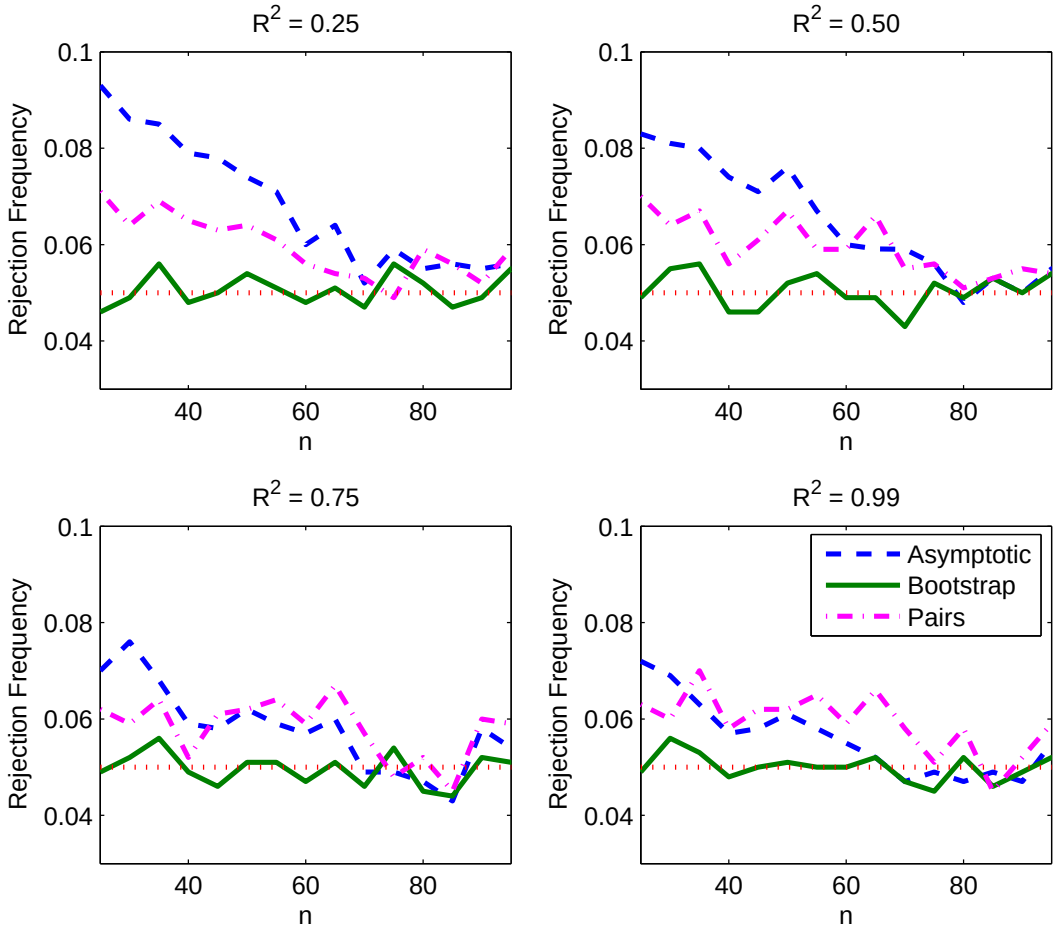
Figure 3: Simulation Results for Various Sample Sizes



bootstraps for each bootstrap test and pairs test is $B = 199$. We then record the frequencies with which tests based on the three p values reject at the 0.05 level. We consider four different R^2: 0.25, 0.5, 0.75, and 0.99. The results of this simulation are presented in the following figure.

In Figure 4, asymptotic t tests, bootstrap tests, and pairs tests are represented by dashed lines, solid lines, and dash-dotted lines, respectively. Although not shown for the sake of brevity, we also compare the three tests using the Wald statistic. Simulation results are very similar to what is shown here. The results suggest that both asymptotic test and pairs test tend to overreject severely when n is small, although their performance improves quickly as n increases. Note that the pairs test becomes more computer intensive when the bootstrap method is used. In contrast, the bootstrap tests overreject only very slightly on average for all values of R^2. The rejection frequencies are always very close to 0.05. Overall, the bootstrap tests outperform the asymptotic tests and pairs tests, especially when the sample

Figure 4: Rejection Frequency for Three Tests



size is small. One drawback of the bootstrap tests is that they require more computation time than the asymptotic tests.

Simulation results also suggest that asymptotic tests and pairs tests have similar performance. The performance of both tests are sensitive to different values of R^2. When $R^2 = 0.25$, rejection frequencies yielded 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 yield similar rejection frequencies for most values of n. In the extreme case of $R^2 = 0.99$, we see that the asymptotic test has better performance than the pairs test by yielding rejection frequencies closer to 0.05 for most values of n.

5 Empirical Application

In this section, we apply our 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 we consider. The main results in Barro (1991) are summarized in Table 1.

We create our data set based on the Barro and Lee (1994) updated growth data for 98 countries. Variables that are used in our estimation are GR6085, GDP60, SEC60, PRIM60,

Table 1: A Summary of Barro's (1991) Arguments

- The growth rate of real per capital 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.

g^c/y, REV, AS, PPI60DEV, GDP60SQ, RPRI, RSEC, AFRICA, LAT.AMER, i/y, and FERTNET. Means and standard deviations are reported in Appendix B, along with detailed definitions for all of the variables. There are eight variables that are of primary interest to us (core variables, including a constant term) and the other seven variables are only of marginal interest to us(potential variables).

We first revisit Barro's (1991) approach using our updated data set. The average growth rate between 1960–1985 (GR6085) is the common dependent variable. There are eight models that are summarized from Barro (1991). Table 2 shows regression results of these models, which are labeled from [1] to [8]. The smallest model is model [1], which only contains the

Table 2: Regressions for Per Capita Growth

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Const.	0.0287	0.0304	0.0368	0.0355	0.0332	0.0218	0.0449	0.0438
	*(0.0076)	(0.0076)	(0.0103)	(0.0103)	(0.0069)	(0.0071)	(0.0119)	(0.0111)
GDP60	-0.0473	-0.0826	-0.0495	-0.0490	-0.0435	-0.0489	-0.0554	-0.0501
	(0.0091)	(0.0284)	(0.0093)	(0.0093)	(0.0080)	(0.0083)	(0.0085)	(0.0080)
SEC60	0.0331	0.0352	0.0321	0.0306	0.0107	0.0213	0.0127	0.0013
	(0.0108)	(0.0109)	(0.0108)	(0.0109)	(0.0103)	(0.0102)	(0.0106)	(0.0102)
PRIM60	0.0202	0.0237	0.0196	0.0187	0.0249	0.0071	0.0037	0.0110
	(0.0071)	(0.0076)	(0.0072)	(0.0072)	(0.0069)	(0.0072)	(0.0071)	(0.0075)
g^c/y	-0.1071	-0.1102	-0.1060	-0.1047	-0.0800	-0.1104	-0.1071	-0.0853
	(0.0286)	(0.0286)	(0.0286)	(0.0286)	(0.0260)	(0.0261)	(0.0255)	(0.0244)
REV	-0.0185	-0.0183	-0.0197	-0.0202	-0.0144	-0.0129	-0.0123	-0.0111
	(0.0067)	(0.0067)	(0.0068)	(0.0068)	(0.0060)	(0.0062)	(0.0061)	(0.0057)
AS	-0.0441	-0.0430	-0.0440	-0.0428	-0.0242	-0.0401	-0.0375	-0.0254
	(0.0184)	(0.0183)	(0.0183)	(0.0183)	(0.0170)	(0.0168)	(0.0164)	(0.0160)
PPI60DEV	-0.0074	-0.0080	-0.0071	-0.0073	-0.0082	-0.0045	-0.0051	-0.0061
	(0.0036)	(0.0037)	(0.0036)	(0.0037)	(0.0033)	(0.0034)	(0.0033)	(0.0032)
GDP60SQ	-	0.0377	-	-	-	-	-	-
	-	(0.0287)	-	-	-	-	-	-
RPRI	-	-	-0.1929	-0.2637	-	-	-	-
	-	-	(0.1646)	(0.1770)	-	-	-	-
RSEC	-	-	-	0.2452	-	-	-	-
	-	-	-	(0.2267)	-	-	-	-
AFRICA	-	-	-	-	-0.0141	-	-	-0.0114
	-	-	-	-	(0.0037)	-	-	(0.0035)
LAT.AMER	-	-	-	-	-0.0171	-	-	-0.0126
	-	-	-	-	(0.0035)	-	-	(0.0035)
i/y	-	-	-	-	-	0.0955	0.0866	0.0640
	-	-	-	-	-	(0.0218)	(0.0216)	(0.0208)
FERTNET	-	-	-	-	-	-	-0.0035	-0.0025
	-	-	-	-	-	-	(0.0015)	(0.0014)
R_c^2	0.4679	0.4780	0.4760	0.4829	0.6015	0.5624	0.5888	0.6578

* Values in parentheses are standard errors.

eight core variables, while the largest is model [8], which contains twelve variables (eight cores and four potentials). We also report 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 (coefficient and P value) for the core variable across different models using different potential variables. Take SEC60 for 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), we may end up with contradictory results.

To solve this model uncertainty problem, we conduct model average estimation using the MAPC estimator. Following Barro (1991), we 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, we include 128 approximation models in our estimation.

In Table 3, we report the estimated average coefficients for the core variables and their standard errors. We also estimate the P value for each coefficient equal to zero using the t-test we derived in Section 3. We also calculate the bootstrap P value for each average coefficient. We set $B = 9999$ for all variables. We also show the top eight models with associated weights in Table 4.

The result of the MAPC estimation provides significant evidence to support Barro's (1991) argument 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 potential variables. The estimated coefficient for GDP60 is -0.0713, which indicates a negative relationship between growth rate and the initial per capita product. We notice 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

Table 3: Results for the MAPC Estimation

	Estimates	s.e.	Asymptotic	Bootstrap
Const.	0.0438	0.0089	0.0000	0.0002
GDP60	-0.0713	0.0148	0.0000	0.0100
SEC60	0.0070	0.0097	0.4715	0.5445
PRIM60	0.0136	0.0069	0.0483	0.0911
g^c/y	-0.0902	0.0240	0.0002	0.0020
REV	-0.0122	0.0056	0.0289	0.0361
AS	-0.0281	0.0156	0.0716	0.0861
PPI60DEV	-0.0064	0.0031	0.0376	0.0501

Table 4: Top Eight Models in MAPC Estimation

	Assigned Weights							
	0.0814	0.0795	0.0792	0.0626	0.0547	0.0522	0.0518	0.0504
Const.	+	+	+	+	+	+	+	+
GDP60	+	+	+	+	+	+	+	+
SEC60	+	+	+	+	+	+	+	+
PRIM60	+	+	+	+	+	+	+	+
g^c/y	+	+	+	+	+	+	+	+
REV	+	+	+	+	+	+	+	+
AS	+	+	+	+	+	+	+	+
PPI60DEV	+	+	+	+	+	+	+	+
GDP60SQ	-	+	-	+	+	+	+	-
RPRI	-	-	-	-	-	-	+	-
RSEC	-	-	-	-	-	-	-	-
AFRICA	+	+	+	+	+	-	+	+
LAT.AMER	+	+	+	+	-	-	+	+
i/y	+	+	+	+	+	+	+	-
FERTNET	-	+	+	-	+	+	+	+

The "+" indicates that the corresponding parameters are included in the model and the "-" indicates the opposite.

growth rate and the share of government consumption in GDP. Political instability, which is proxied by REV and AS, reflects negative effects on the growth rate.

We also compare the relative out-of-sample predictive efficiency of the MAPC estimator with other estimators. For the original data sample $n = 98$, we shuffle the sample into a training set of n_1 and an evaluation set of size $n_2 = n - n_1$. We apply the seven methods (GETS, AIC, MA-AIC, MA-BIC, MMA, JMA, and MAPC) in Section 4 to the training set. We evaluate the selected models and computed estimates via mean squared prediction error (MSPE):

$$\text{MSPE} = \frac{1}{n_2}(\boldsymbol{y}_2 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{y}_2 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_1),$$

where $(\boldsymbol{y}_2, \boldsymbol{X}_2)$ is the evaluation set, n_2 is the number of observations of the evaluation set, and $\hat{\boldsymbol{\beta}}_1$ is the estimated coefficients by a particular method based on the training set. We normalize the MSPEs by the MSPE by the MAPC estimator. We repeat the entire procedure 10001 times and report the median MSPE. We vary n_1 and consider $n_1 = 20, 30, ..., 80$. Table 5 reports the relative out-of-sample predictive efficiency. Entries larger than one indicate inferior performance relative to the MAPC estimator.

Table 5 suggests that the MAPC estimator delivers models that have better out-of-sample predictive efficiency than those by the six existing methods. As n_1 increases, the relative performance for all six methods improves. We notice that the MA-AIC method works better than MMA. As mentioned in Hansen and Racine (2012), the MMA method is somewhat sensitive to the preliminary estimate of σ^2 needed for its computation, and relying on a "large" approximating model may not be sufficient to deliver optimal results. In contrast,

Table 5: Relative Out-of-sample Predictive Efficiency

n_1	GETS	AIC	MA-AIC	MA-BIC	MMA	JMA
20	1.1808	1.0168	1.0052	1.0290	1.0111	1.0100
30	1.1801	1.0154	1.0049	1.0284	1.0101	1.0098
40	1.1680	1.0148	1.0046	1.0283	1.0100	1.0096
50	1.1665	1.0139	1.0046	1.0281	1.0099	1.0095
60	1.1659	1.0137	1.0032	1.0281	1.0097	1.0094
70	1.1642	1.0131	1.0024	1.0274	1.0096	1.0094
80	1.1630	1.0109	1.0023	1.0271	1.0093	1.0070

our MAPC estimator does not require a preliminary estimate, as it is a continuous-updating estimator.

6 Conclusion

Our MAPC estimator computes the weighted average across all approximation models; in doing so, it reduces the chance of a poor model being selected. The eight models selected from Barro (1991) are a small subset of the 128 models in the MAPC estimation. The idea of considering every possible approximation model helps avoid the risk of omitting important variables or retaining irrelevant variables. Both simulation and application suggest that as a continuous-updating estimator, our MAPC estimator yields better efficiency than the two-steps estimator MMA, especially when sample sizes are small. The asymptotic test statistics we derived in Section 3 work well in large sample sizes and bootstrap tests are highly recommended for small sample sizes.

A Proof

Proof of Lemma 1. To prove (i), we make use of $\mathrm{Tr}\Big(\boldsymbol{P}^{(m)}\Big) = k^{(m)}$, then

$$\mathrm{Tr}\big(\boldsymbol{P}(\boldsymbol{w})\big) = \mathrm{Tr}\left(\sum_{m=1}^{M} w^{(m)}\boldsymbol{P}^{(m)}\right) = \sum_{m=1}^{M} w^{(m)}\mathrm{Tr}\left(\boldsymbol{P}^{(m)}\right) = \sum_{m=1}^{M} w^{(m)}k^{(m)} = k(\boldsymbol{w}).$$

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

$$\max_{\boldsymbol{\eta}} \frac{\boldsymbol{\eta}^\top \boldsymbol{P}^{(m)}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top\boldsymbol{\eta}} = 1 \qquad \text{for all } m.$$

By the definition of eigenvalue, we have

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

Parts (iii) and (iv) can be obtained using result (ii) such that

$$\big\|\boldsymbol{P}(\boldsymbol{w}^*)\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}\big\|^2 \leq \lambda^2_{\max}\big(\boldsymbol{P}(\boldsymbol{w}^*)\big)\big\|\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}\big\|^2 \leq \big\|\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}\big\|^2$$

and

$$\mathrm{Tr}\Big[\big(\boldsymbol{P}(\boldsymbol{w}^*)\boldsymbol{P}(\boldsymbol{w}^*)\big)\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\Big] \leq \lambda^2_{\max}\big(\boldsymbol{P}(\boldsymbol{w}^*)\big)\mathrm{Tr}\big[\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\big] \leq \mathrm{Tr}\big[\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\big].$$

∎

Proof of Lemma 2. $R_n(\boldsymbol{w})$ can be written as

$$\begin{aligned} R_n(\boldsymbol{w}) &= \mathbb{E}\left[(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu} + \boldsymbol{P}(\boldsymbol{w})\boldsymbol{u} - \boldsymbol{\mu})^\top(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu} + \boldsymbol{P}(\boldsymbol{w})\boldsymbol{u} - \boldsymbol{\mu})|\boldsymbol{X}\right] \\ &= \boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu} + \sigma^2\mathrm{Tr}(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})) \\ &= \|\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}\|^2 + \sigma^2\mathrm{Tr}\big(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\big). \end{aligned} \tag{20}$$

Both terms in (20) are non-negative, which implies Lemma 2. ∎

Proof of Lemma 3. The proof is straightforward. If $k^{(m)}/n \to 0$ for all m, then the linear combination of $k^{(m)}$, $k(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)}\left(k^{(m)}/n\right) \to 0$. ∎

Proof of Theorem 1. Our proof follows the techniques derived in Li (1987) and Wan et

al. (2010). Rewrite $\mathrm{MAPC}_n(\boldsymbol{w})$ to contain $L_n(\boldsymbol{w})$,

$$\begin{aligned}\mathrm{MAPC}_n(\boldsymbol{w}) &= (\boldsymbol{\mu}-\boldsymbol{\mu}(\boldsymbol{w})+\boldsymbol{u})^\top(\boldsymbol{\mu}-\boldsymbol{\mu}(\boldsymbol{w})+\boldsymbol{u})\left(\frac{n+k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right)\\ &= L_n(\boldsymbol{w})+2\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}+2L_n(\boldsymbol{w})\left(\frac{k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right)\\ &\quad+4\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}\left(\frac{k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right)+\boldsymbol{u}^\top\boldsymbol{u}\left(\frac{n+k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right). \qquad (21)\end{aligned}$$

As $n\to\infty$, $\big(n+k(\boldsymbol{w})\big)/\big(n-k(\boldsymbol{w})\big)\to 1$ by Lemma 3, which implies that the last term in (21) is independent of $\boldsymbol{w}$. Therefore, $\hat{\boldsymbol{w}}$ minimizes

$$L_n(\boldsymbol{w})+2\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}+2L_n(\boldsymbol{w})\left(\frac{k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right)+4\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}\left(\frac{k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right). \qquad (22)$$

If we can show that as $n\to\infty$, all terms in (22) except $L_n(\boldsymbol{w})$ are negligible compared with $L_n(\boldsymbol{w})$, for any $\boldsymbol{w}\in\boldsymbol{H_M}$, then the asymptotic optimality of $\hat{\boldsymbol{w}}$ is established.

Theorem 1 is valid, if, as $n\to\infty$

$$\begin{aligned}\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{R_n(\boldsymbol{w})}\right| &\to_p 0, \qquad (23)\\ \sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{L_n(\boldsymbol{w})}{R_n(\boldsymbol{w})}-1\right| &\to_p 0, \qquad (24)\\ \sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{R_n(\boldsymbol{w})}\left(\frac{k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right)\right| &\to_p 0, \qquad (25)\\ \sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{L_n(\boldsymbol{w})}{R_n(\boldsymbol{w})}\left(\frac{k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right)\right| &\to_p 0. \qquad (26)\end{aligned}$$

We shall prove (23) first. Given any $\delta>0$, by triangular inequality, Bonferroni's inequality we have

$$\begin{aligned}\mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{R_n(\boldsymbol{w})}\right|>\delta\right\} &\le \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\sum_{m=1}^{M}w^{(m)}\left|\boldsymbol{\mu}^\top\boldsymbol{M}^{(m)}\boldsymbol{u}\right|>\delta\xi_n\right\}\\ &= \mathrm{P}\left\{\max_m\left|\boldsymbol{\mu}^\top\boldsymbol{M}^{(m)}\boldsymbol{u}\right|>\delta\xi_n\right\}\\ &\le \sum_{m=1}^{M}\mathrm{P}\left\{\left|\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{u}\right|>\delta\xi_n\right\},\end{aligned}$$

which, by Chebyshev's inequality, is no greater than

$$\sum_{m=1}^{M} \mathbb{E}\left\{ \frac{\left(\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{u}\right)^{2G}}{\delta^{2G}\xi_n^{2G}} \right\}.$$

By Theorem 2 of Whittle (1960) and our Lemma 2(i), we observe that

$$\begin{aligned}
\sum_{m=1}^{M} \mathbb{E}\left\{ \frac{\left(\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{u}\right)^{2G}}{\delta^{2G}\xi_n^{2G}} \right\} &\leq C_1 \delta^{-2G}\xi_n^{-2G} \sum_{m=1}^{M} \left\| \boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{\mu} \right\|^{2G} \\
&\leq C_1 \delta^{-2G}\xi_n^{-2G} \sum_{m=1}^{M} \left(R_n(\boldsymbol{w}_m^0)\right)^G
\end{aligned}$$

for some constant C_1. The last term above goes to zero by Assumption 2. Thus, (23) is proved.

To prove (24), we first see

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\boldsymbol{H_M}} \left| \frac{L_n(\boldsymbol{w})}{R_n(\boldsymbol{w})} - 1 \right| \to_p 0 \\
\Leftrightarrow \quad &\sup_{\boldsymbol{w}\in\boldsymbol{H_M}} \left| \frac{\boldsymbol{u}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{u} - \sigma^2 \mathrm{Tr}\left(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\right) - 2\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{u}}{R_n(\boldsymbol{w})} \right| \to_p 0.
\end{aligned}$$

Then, it suffices to prove

$$\sup_{\boldsymbol{w}\in\boldsymbol{H_M}} \left| \frac{\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{u}}{R_n(\boldsymbol{w})} \right| \to_p 0 \tag{27}$$

and

$$\sup_{\boldsymbol{w}\in\boldsymbol{H_M}} \left| \frac{\boldsymbol{u}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{u} - \sigma^2 \mathrm{Tr}\left(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\right)}{R_n(\boldsymbol{w})} \right| \to_p 0. \tag{28}$$

By Chebyshev's inequality and Theorem 2 of Whittle (1960), given any $\delta > 0$,

$$\begin{aligned}
\mathrm{P}\left\{ \sup_{\boldsymbol{w}\in\boldsymbol{H_M}} \left| \frac{\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{u}}{R_n(\boldsymbol{w})} \right| > \delta \right\} &\leq \sum_{m=1}^{M}\sum_{l=1}^{M} \mathbb{E}\left\{ \frac{\left(\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{P}(\boldsymbol{w}_l^0)\boldsymbol{u}\right)^{2G}}{\delta^{2G}\xi_n^{2G}} \right\} \\
&\leq C_2 \delta^{-2G}\xi_n^{-2G} \sum_{m=1}^{M}\sum_{l=1}^{M} \left(\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{P}(\boldsymbol{w}_l^0)\boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{\mu}\right)^G \\
&= C_2 \delta^{-2G}\xi_n^{-2G} \sum_{m=1}^{M}\sum_{l=1}^{M} \left\| \boldsymbol{P}(\boldsymbol{w}_l^0)\boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{\mu} \right\|^{2G},
\end{aligned}$$

where C_2 is a constant. By Lemma 1 (iii), Lemma 2 (i) and Assumption 2, we obtain

$$\begin{aligned} C_2\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^{M}\sum_{l=1}^{M}\left\|\boldsymbol{P}(\boldsymbol{w}_l^0)\boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{\mu}\right\|^{2G} &\leq C_2\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^{M}\sum_{l=1}^{M}\left\|\boldsymbol{M}(\boldsymbol{w}_m^0)\boldsymbol{\mu}\right\|^{2G} \\ &\leq C_2\delta^{-2G}\xi_n^{-2G}M\sum_{m=1}^{M}\left(R_n(\boldsymbol{w}_m^0)\right)^G \to 0. \end{aligned}$$

Likewise, by Chebyshev's inequality, Theorem 2 of Whittle (1960), Lemma 1 (iii), Lemma 2 (ii) and Assumption 2, given any $\delta > 0$, we observe that, for a constant C_3,

$$\begin{aligned} &\mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{u}^\top\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{u}-\sigma^2\mathrm{Tr}\left(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\right)}{R_n(\boldsymbol{w})}\right|>\delta\right\} \\ &\quad\leq\sum_{m=1}^{M}\sum_{l=1}^{M}\mathbb{E}\left\{\frac{\left[\boldsymbol{u}^\top\boldsymbol{P}(\boldsymbol{w}_l^0)\boldsymbol{P}(\boldsymbol{w}_m^0)\boldsymbol{u}-\sigma^2\mathrm{Tr}\left(\boldsymbol{P}(\boldsymbol{w}_l^0)\boldsymbol{P}(\boldsymbol{w}_m^0)\right)\right]^{2G}}{\delta^{2G}\xi_n^{2G}}\right\} \\ &\quad\leq C_3\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^{M}\sum_{l=1}^{M}\left[\mathrm{Tr}\left(\boldsymbol{P}(\boldsymbol{w}_m^0)\boldsymbol{P}(\boldsymbol{w}_l^0)\boldsymbol{P}(\boldsymbol{w}_m^0)\right)\right]^G \\ &\quad\leq C_3\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^{M}\sum_{l=1}^{M}\left[\mathrm{Tr}\left(\boldsymbol{P}(\boldsymbol{w}_m^0)\boldsymbol{P}(\boldsymbol{w}_m^0)\right)\right]^G \\ &\quad\leq\frac{C_3}{\sigma^2}\delta^{-2G}\xi_n^{-2G}M\sum_{m=1}^{M}[R_n(\boldsymbol{w}_m^0)]^G\to 0. \end{aligned}$$

Proving (25) and (26) becomes straightforward once we validate (23) and (24). We obtain, as $n\to\infty$, that

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

and

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

This completes the proof of Theorem 1. ∎

Proof of Theorem 2. Since $\hat{\boldsymbol{u}}^{(m)}=\boldsymbol{y}-\hat{\boldsymbol{\mu}}^{(m)}=\boldsymbol{y}-\boldsymbol{P}^{(m)}\boldsymbol{y}=\boldsymbol{M}^{(m)}\boldsymbol{u}+\boldsymbol{M}^{(m)}\boldsymbol{\mu}$ for model m, the average estimate of $\boldsymbol{u}$ becomes

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

Then, the average estimate for σ^2 becomes

$$\begin{aligned}\sigma^2(\boldsymbol{w}) &= \frac{\boldsymbol{u}(\boldsymbol{w})^\top \boldsymbol{u}(\boldsymbol{w})}{n-k(\boldsymbol{w})} = \frac{(\boldsymbol{\mu}+\boldsymbol{u})^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})(\boldsymbol{\mu}+\boldsymbol{u})}{n-k(\boldsymbol{w})} \\ &= \frac{\boldsymbol{u}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} + \frac{\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}}{n-k(\boldsymbol{w})} + \frac{2\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})}. \end{aligned} \tag{29}$$

Therefore, Theorem 2 is valid if the following hold: as $n \to \infty$,

$$\sup_{\boldsymbol{w}\in\boldsymbol{H}_M} \left| \frac{\boldsymbol{u}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} \right| \to_p \sigma^2, \tag{30}$$

$$\sup_{\boldsymbol{w}\in\boldsymbol{H}_M} \left| \frac{\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}}{n-k(\boldsymbol{w})} \right| \to_p 0, \tag{31}$$

$$\sup_{\boldsymbol{w}\in\boldsymbol{H}_M} \left| \frac{\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} \right| \to_p 0. \tag{32}$$

Equation (30) is equivalent to

$$\sup_{\boldsymbol{w}\in\boldsymbol{H}_M} \left| \frac{\boldsymbol{u}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} - \frac{\boldsymbol{u}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} \right| \to_p \sigma^2. \tag{33}$$

To prove (30), it suffices to show, as $n \to \infty$, that

$$\sup_{\boldsymbol{w}\in\boldsymbol{H}_M} \left| \frac{\boldsymbol{u}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} \right| \to_p \sigma^2 \tag{34}$$

and

$$\sup_{\boldsymbol{w}\in\boldsymbol{H}_M} \left| \frac{\boldsymbol{u}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} \right| \to_p 0. \tag{35}$$

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

$$\begin{aligned}\mathbb{E}\left|\boldsymbol{u}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{u} - \sigma^2\left(n-k(\boldsymbol{w})\right)\right|^2 &\leq C_4 \mathrm{Tr}\left(\boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\right) \\ &\leq C_4 \mathrm{Tr}\left(\boldsymbol{M}(\boldsymbol{w})\right) \\ &= C_4\left(n-k(\boldsymbol{w})\right), \end{aligned}$$

where C_4 is some constant. Thus, for any $\delta > 0$, by Markov's inequality and Lemma 3,

$$\begin{aligned}\mathrm{P}\left\{ \sup_{\boldsymbol{w}\in\boldsymbol{H}_M} \left| \frac{\boldsymbol{u}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})} - \sigma^2 \right| > \delta \right\} &\leq \frac{\mathbb{E}\left|\boldsymbol{u}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{u} - \sigma^2\left(n-k(\boldsymbol{w})\right)\right|^2}{\delta^2\left(n-k(\boldsymbol{w})\right)^2} \\ &\leq \frac{C_4}{\delta^2\left(n-k(\boldsymbol{w})\right)} \to 0. \end{aligned}$$

Second,

$$\begin{aligned}
\mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{u}^\top\boldsymbol{P}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})}\right|>\delta\right\} &\leq \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{u}^\top\boldsymbol{P}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})}\right|>\delta\right\} \\
&= \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\sum_{m=1}^{M}w^{(m)}\left|\frac{\boldsymbol{u}^\top\boldsymbol{P}^{(m)}\boldsymbol{u}}{n-k(\boldsymbol{w})}\right|>\delta\right\} \\
&\leq \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\max_{m}\left|\frac{\boldsymbol{u}^\top\boldsymbol{P}^{(m)}\boldsymbol{u}}{n-k(\boldsymbol{w})}\right|>\delta\right\}\to 0
\end{aligned}$$

since $(n-k(\boldsymbol{w}))^{-1}\left(\boldsymbol{u}^\top\boldsymbol{P}^{(m)}\boldsymbol{u}\right)\to_p 0$ as $n\to\infty$ for all m. (30) is obtained.

To prove (31), we see that for any approximation model m,

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

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

$$\frac{\left(\boldsymbol{b}^{(m)}\right)^\top\boldsymbol{M}^{(m)}\boldsymbol{b}^{(m)}}{n-k(\boldsymbol{w})}\to_p 0 \tag{36}$$

for all m. Therefore, by (36), we observe, for any $\delta>0$, that

$$\begin{aligned}
\mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}}{n-k(\boldsymbol{w})}\right|>\delta\right\} &\leq \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{\mu}^\top\boldsymbol{M}(\boldsymbol{w})\boldsymbol{\mu}}{n-k(\boldsymbol{w})}\right|>\delta\right\} \\
&= \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\sum_{m=1}^{M}w^{(m)}\left|\frac{\boldsymbol{\mu}^\top\boldsymbol{M}^{(m)}\boldsymbol{\mu}}{n-k(\boldsymbol{w})}\right|>\delta\right\} \\
&\leq \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\max_{m}\left|\frac{\boldsymbol{\mu}^\top\boldsymbol{M}^{(m)}\boldsymbol{\mu}}{n-k(\boldsymbol{w})}\right|>\delta\right\} \\
&= \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\max_{m}\left|\frac{\left(\boldsymbol{b}^{(m)}\right)^\top\boldsymbol{M}^{(m)}\left(\boldsymbol{b}^{(m)}\right)}{n-k(\boldsymbol{w})}\right|>\delta\right\}\to 0.
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})}\right|>\delta\right\} &\leq \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\left|\frac{\boldsymbol{\mu}^\top \boldsymbol{M}(\boldsymbol{w})\boldsymbol{u}}{n-k(\boldsymbol{w})}\right|>\delta\right\} \\
&\leq \mathrm{P}\left\{\sup_{\boldsymbol{w}\in\boldsymbol{H_M}}\max_m\left|\frac{\boldsymbol{\mu}^\top \boldsymbol{M}^{(m)}\boldsymbol{u}}{n-k(\boldsymbol{w})}\right|>\delta\right\} \to_p 0.
\end{aligned}$$

Therefore, we conclude that $\sigma^2(\boldsymbol{w}) \to_p \sigma^2$. ∎

Proof of Lemma 4. Since the covariance matrix is conditional on $\boldsymbol{X}^{(L)}$, we can assume that $\boldsymbol{X}^{(m)}$ is exogenous for all m. The covariance matrix can be written as

$$\begin{aligned}
&\mathrm{Cov}\left(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\Gamma}^{(s)}\hat{\boldsymbol{\beta}}^{(s)}\right) \\
=\ &\mathbb{E}\left[\left(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}-\boldsymbol{\beta}(\boldsymbol{w})\right)\left(\boldsymbol{\Gamma}^{(s)}\hat{\boldsymbol{\beta}}^{(s)}-\boldsymbol{\beta}(\boldsymbol{w})\right)^\top\right] \\
=\ &\mathbb{E}\left(\left(\boldsymbol{\Gamma}^{(m)}\left(\hat{\boldsymbol{\beta}}^{(m)}-\boldsymbol{\beta}^{(m)}\right)+\boldsymbol{d}_1^{(m)}\right)\left(\boldsymbol{\Gamma}^{(s)}\left(\hat{\boldsymbol{\beta}}^{(s)}-\boldsymbol{\beta}^{(s)}\right)+\boldsymbol{d}_1^{(s)}\right)^\top\right), \qquad (37)
\end{aligned}$$

where $\boldsymbol{d}_1^{(m)} = \boldsymbol{\Gamma}^{(m)}\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}(\boldsymbol{w})$ is non-random. First, we have

$$\begin{aligned}
&\boldsymbol{\Gamma}^{(m)}\left(\hat{\boldsymbol{\beta}}^{(m)}-\boldsymbol{\beta}^{(m)}\right) \\
=\ &\boldsymbol{\Gamma}^{(m)}\left(\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\left(\boldsymbol{X}^{(m)}\boldsymbol{\beta}^{(m)}+\boldsymbol{X}^{(-m)}\boldsymbol{\beta}^{(-m)}+\boldsymbol{u}\right)-\boldsymbol{\beta}^{(m)}\right) \\
=\ &\boldsymbol{\Gamma}^{(m)}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(-m)}\boldsymbol{\beta}^{(-m)}+\boldsymbol{\Gamma}^{(m)}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{\Gamma}^{(m)\top}\boldsymbol{X}^\top\boldsymbol{u}.
\end{aligned}$$

Define

$$\begin{aligned}
\boldsymbol{A}^{(m)} &\equiv \boldsymbol{\Gamma}^{(m)}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(-m)}\boldsymbol{\beta}^{(-m)}, \\
\boldsymbol{B}^{(m)} &\equiv \boldsymbol{\Gamma}^{(m)}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{\Gamma}^{(m)\top}\boldsymbol{X}^\top\boldsymbol{u}.
\end{aligned}$$

We expand the expectation in (37) and obtain

$$\begin{aligned}
\mathbb{E}\Big(&\boldsymbol{A}^{(m)}\boldsymbol{A}^{(s)\top}+\boldsymbol{A}^{(m)}\boldsymbol{B}^{(s)\top}+\boldsymbol{A}^{(m)}\boldsymbol{d}_1^{(s)\top}+\boldsymbol{B}^{(m)}\boldsymbol{A}^{(s)\top}+\boldsymbol{B}^{(m)}\boldsymbol{B}^{(s)\top} \\
&+\boldsymbol{B}^{(m)}\boldsymbol{d}_1^{(s)\top}+\boldsymbol{d}_1^{(m)}\boldsymbol{A}^{(s)\top}+\boldsymbol{d}_1^{(m)}\boldsymbol{B}^{(s)\top}+\boldsymbol{d}_1^{(m)}\boldsymbol{d}_1^{(s)\top}\Big). \qquad (38)
\end{aligned}$$

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

$\mathbf{0}$, $\mathbb{E}(\boldsymbol{B}^{(m)}\boldsymbol{d}_1^{(s)\top}) = \mathbf{0}$, and $\mathbb{E}(\boldsymbol{d}_1^{(m)}\boldsymbol{B}^{(s)\top}) = \mathbf{0}$. Also,

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

Finally, (38) is equivalent to

$$\sigma^2\boldsymbol{\Gamma}^{(m)}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(s)}\left(\boldsymbol{X}^{(s)\top}\boldsymbol{X}^{(s)}\right)^{-1}\boldsymbol{\Gamma}^{(s)\top} + \left(\boldsymbol{A}^{(m)} + \boldsymbol{d}_1^{(m)}\right)\left(\boldsymbol{A}^{(s)} + \boldsymbol{d}_1^{(s)}\right)^{\top}.$$

Note that $\boldsymbol{A}^{(m)} + \boldsymbol{d}_1^{(m)} = \mathbb{E}\left(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}\right) - \boldsymbol{\beta}(\boldsymbol{w}) = \boldsymbol{d}^{(m)}$. In practice, we replace the infeasible σ^2 and $\boldsymbol{d}^{(m)}$ with their sample estimates. This completes the proof. ∎

Proof of Lemma 5. Part (i) follows the fact that the set of $\boldsymbol{\beta}_{-m}$ belongs to the set of $\boldsymbol{\beta}_p$ for all m. Therefore, $\sqrt{n}\boldsymbol{\beta}_{-m} \to \boldsymbol{h}_{-m}$ by Assumption 5.

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

$$\begin{aligned}&\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}^{(m)}\right)\\ =\ &\sqrt{n}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(-m)}\boldsymbol{\beta}^{(-m)} + \sqrt{n}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{\Gamma}^{(m)\top}\boldsymbol{X}^{\top}\boldsymbol{u}.\end{aligned}$$

By Assumption 5 and Lemma 5(i),

$$\sqrt{n}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(-m)}\boldsymbol{\beta}^{(-m)} \xrightarrow{p} \left(\boldsymbol{S}^{(m)}\right)^{-1}\boldsymbol{S}^{(m,-m)}\boldsymbol{h}_{-m} \equiv \boldsymbol{\delta}_1^{(m)}.$$

By Assumption 4 and central limit theorem,

$$\sqrt{n}\left(\boldsymbol{X}^{(m)\top}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{\Gamma}^{(m)\top}\boldsymbol{X}^{\top}\boldsymbol{u} \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \sigma^2\left(\boldsymbol{S}^{(m)}\right)^{-1}\right).$$

This implies $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_c^{(m)} - \boldsymbol{\beta}_c\big) = \sqrt{n}\boldsymbol{\Gamma}_c^{(m)\top}\big(\hat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}^{(m)}\big) \xrightarrow{d} \mathrm{N}\big(\boldsymbol{\delta}_1^{(m)}, \sigma^2\boldsymbol{\Gamma}_c^{(m)\top}(\boldsymbol{S}^{(m)})^{-1}\boldsymbol{\Gamma}_c^{(m)}\big)$.

To prove part (iii), we first have $\boldsymbol{\beta}_{-m} = O(n^{-1/2})$ by Assumption 5. Under Assumption 6, we apply Taylor expansion to $\boldsymbol{\beta}(\boldsymbol{w})$ and obtain

$$\boldsymbol{\beta}(\boldsymbol{w}) = \boldsymbol{f}\big(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}_{-m}\big) = \boldsymbol{f}\big(\boldsymbol{\beta}^{(m)}, \mathbf{0}\big) + \left(\boldsymbol{F}_{\boldsymbol{\beta}_{-m}}\big|_{\boldsymbol{\beta}_{-m}=\mathbf{0}}\right)\boldsymbol{\beta}_{-m} + \frac{1}{2}\boldsymbol{\beta}_{-m}^{\top}\bar{\boldsymbol{H}}\boldsymbol{\beta}_{-m}, \tag{39}$$

where $\bar{\boldsymbol{H}}$ is the Hessian matrix $\boldsymbol{H} \equiv \partial^2\boldsymbol{f}(\cdot)/\partial\boldsymbol{\beta}_{-m}\partial\boldsymbol{\beta}_{-m}^{\top}$ evaluated at an intermediate point between $\mathbf{0}$ and $\boldsymbol{\beta}_{-m}$. Since $\boldsymbol{\beta}_{-m} = O(n^{-1/2})$, the last term in (39) is $O(n^{-1})$. With

$\boldsymbol{f}\left(\boldsymbol{\beta}^{(m)}, \mathbf{0}\right) = \boldsymbol{\Gamma}^{(m)}\boldsymbol{\beta}^{(m)}$, we have

$$\begin{aligned}\sqrt{n}\left(\boldsymbol{\Gamma}^{(m)}\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}(\boldsymbol{w})\right) &= \sqrt{n}\left(-\left(\boldsymbol{F}_{\boldsymbol{\beta}_{-m}}\big|_{\boldsymbol{\beta}_{-m}=\mathbf{0}}\right)\boldsymbol{\beta}_{-m} - O(n^{-1})\right) \\ &\to -\left(\boldsymbol{F}_{\boldsymbol{\beta}_{-m}}\big|_{\boldsymbol{\beta}_{-m}=\mathbf{0}}\right)\boldsymbol{h}_{-m}.\end{aligned}$$

Therefore, $\sqrt{n}\left(\boldsymbol{\beta}_c - \boldsymbol{\beta}_c(\boldsymbol{w})\right) = \sqrt{n}{\boldsymbol{\Gamma}_c^{(m)}}^\top\left(\boldsymbol{\Gamma}^{(m)}\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}(\boldsymbol{w})\right) \to \boldsymbol{\delta}_2$.

Part (iv) follows Lemma 5 (ii) and (iii). We have

$$\begin{aligned}\sqrt{n}\boldsymbol{d}^{(m)} &= \sqrt{n}\left(\mathbb{E}(\boldsymbol{\Gamma}^{(m)}\hat{\boldsymbol{\beta}}^{(m)}) - \boldsymbol{\beta}(\boldsymbol{w})\right) \\ &\xrightarrow{p} \boldsymbol{\Gamma}^{(m)}\left(\boldsymbol{S}^{(m)}\right)^{-1}\boldsymbol{S}^{(m,-m)}\boldsymbol{h}_{-m} - \left(\boldsymbol{F}_{\boldsymbol{\beta}_{-m}}\big|_{\boldsymbol{\beta}_{-m}=\mathbf{0}}\right)\boldsymbol{h}_{-m} = \boldsymbol{\delta}^{(m)}.\end{aligned}$$

Part (v) is the main result of Lemma 5.

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_c^{(m)} - \boldsymbol{\beta}_c(\boldsymbol{w})\right) &= \sqrt{n}\left(\hat{\boldsymbol{\beta}}_c^{(m)} - \boldsymbol{\beta}_c\right) + \sqrt{n}\left(\boldsymbol{\beta}_c - \boldsymbol{\beta}_c(\boldsymbol{w})\right) \\ &\xrightarrow{d} \mathrm{N}\left(\boldsymbol{\delta}_c^{(m)}, \sigma^2{\boldsymbol{\Gamma}_c^{(m)}}^\top\left(\boldsymbol{S}^{(m)}\right)^{-1}\boldsymbol{\Gamma}_c^{(m)}\right) \sim \boldsymbol{\Lambda}_c^{(m)}. \qquad \blacksquare\end{aligned}$$

Proof of Theorem 3. It is straightforward to show that the asymptotic distribution of $\boldsymbol{\Lambda}_c$ is a normal distribution, since

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}_c(\boldsymbol{w})\right)\Big|\hat{\boldsymbol{w}} = \sum_{m=1}^{M}\hat{w}^{(m)}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_c^{(m)} - \boldsymbol{\beta}_c(\boldsymbol{w})\right)\Big|\hat{\boldsymbol{w}} \to_d \sum_{m=1}^{M}\hat{w}^{(m)}\boldsymbol{\Lambda}_c^{(m)}$$

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

$$\begin{aligned}\boldsymbol{V}(\hat{\boldsymbol{w}}) &= \operatorname*{plim}_{n\to\infty}\mathrm{Var}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}(\boldsymbol{w})\right)\Big|\hat{\boldsymbol{w}}\right) \\ &= \operatorname*{plim}_{n\to\infty} n\sum_{m=1}^{M}\sum_{s=1}^{M}\hat{w}^{(m)}\hat{w}^{(s)}\Big(\boldsymbol{d}^{(m)}{\boldsymbol{d}^{(s)}}^\top \\ &\quad + \sigma^2\boldsymbol{\Gamma}^{(m)}\left({\boldsymbol{X}^{(m)}}^\top\boldsymbol{X}^{(m)}\right)^{-1}{\boldsymbol{X}^{(m)}}^\top\boldsymbol{X}^{(s)}\left({\boldsymbol{X}^{(s)}}^\top\boldsymbol{X}^{(s)}\right)^{-1}{\boldsymbol{\Gamma}^{(s)}}^\top\Big|\hat{\boldsymbol{w}}\Big) \\ &= \operatorname*{plim}_{n\to\infty}\sum_{m=1}^{M}\sum_{s=1}^{M}\hat{w}^{(m)}\hat{w}^{(s)}\Big(\sqrt{n}\boldsymbol{d}^{(m)}\sqrt{n}{\boldsymbol{d}^{(s)}}^\top \\ &\quad + \sigma^2\boldsymbol{\Gamma}^{(m)}\left(n^{-1}{\boldsymbol{X}^{(m)}}^\top\boldsymbol{X}^{(m)}\right)^{-1}n^{-1}{\boldsymbol{X}^{(m)}}^\top\boldsymbol{X}^{(s)}\left(n^{-1}{\boldsymbol{X}^{(s)}}^\top\boldsymbol{X}^{(s)}\right)^{-1}{\boldsymbol{\Gamma}^{(s)}}^\top\Big).\end{aligned}$$

We have proved in Lemma 5 that $\sqrt{n}\boldsymbol{d}^{(m)} \xrightarrow{p} \boldsymbol{\delta}^{(m)}$. Therefore,

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

where $\boldsymbol{S}^{(m)}$, $\boldsymbol{S}^{(m,s)}$, and $\boldsymbol{\delta}^{(m)}$ are defined in Lemma 5. ∎

Proof of Theorem 4. Since $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}_c(\boldsymbol{w})\big)\big|\hat{\boldsymbol{w}} \to_d \mathrm{N}\big(\boldsymbol{0}, \boldsymbol{V}_c(\hat{\boldsymbol{w}})\big)$ and $\boldsymbol{V}_c(\hat{\boldsymbol{w}})$ is a linear function of $\hat{\boldsymbol{w}}$, it is equivalent to have $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}_c(\boldsymbol{w})\big)|\boldsymbol{V}_c(\hat{\boldsymbol{w}}) \to_d \mathrm{N}\big(\boldsymbol{0}, \boldsymbol{V}_c(\hat{\boldsymbol{w}})\big)$.

We can rewrite the t-statistic as

$$t_{\beta_j(\boldsymbol{w})} \equiv \frac{\hat{\beta}_j(\hat{\boldsymbol{w}}) - \beta_{j0}}{\widehat{\mathrm{Var}}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big)^{1/2}} = \frac{\sqrt{n}\left(\hat{\beta}_j(\hat{\boldsymbol{w}}) - \beta_{j0}\right)}{\left(n\widehat{\mathrm{Var}}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big)\right)^{1/2}}.$$

Following Theorem 3, for fixed $\boldsymbol{V}_c(\hat{\boldsymbol{w}})$, we have

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

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

$$n\widehat{\mathrm{Var}}\big(\hat{\beta}_j(\hat{\boldsymbol{w}})\big) = \widehat{\mathrm{Var}}\left(\sqrt{n}\big(\hat{\beta}_j(\hat{\boldsymbol{w}}) - \beta_{j0}\big)\Big|\boldsymbol{V}_c(\hat{\boldsymbol{w}})\right) \to_p V_j(\hat{\boldsymbol{w}}).$$

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

Similarly, for the Wald statistic, we have

$$W_{\boldsymbol{r}} = \left(\boldsymbol{R}\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{r}\right)^{\top} \left(\boldsymbol{R}\widehat{\mathrm{Var}}\big(\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}})\big)\boldsymbol{R}^{\top}\right)^{-1} \left(\boldsymbol{R}\hat{\boldsymbol{\beta}}_c(\hat{\boldsymbol{w}}) - \boldsymbol{r}\right) \to_d \chi^2(k_r)$$

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

B Description of Data Set

	Variable	Mean	Std.dev
1.	GR6085	0.0219	0.0198
2.	GDP60	1.9146	1.8195
3.	SEC60	0.2314	0.2117
4.	PRIM60	0.7562	0.2710
5.	g^c/y	0.1080	0.0555
6.	REV	0.1806	0.2355
7.	AS	0.0279	0.0764
8.	PPPI60	0.8373	0.4716
9.	PPI60DEV	0.2884	0.3778
10.	GDP60SQ	6.9427	11.4995
11.	RPRI	35.7462	9.1282
12.	RSEC	19.3646	6.4440
13.	AFRICA	0.2653	0.4438
14.	LAT.AMER	0.2347	0.4260
15.	i/y	0.1912	0.0796
16.	FERT	4.8704	1.7520
17.	MORT01	0.0810	0.0503
18.	FERTNET	4.3946	1.4383

Definitions:

1. GR6085: Growth rate of real per capita GDP from 1960 to 1985.
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 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. PPPI60: 1960 PPP value for the investment deflator (U.S. = 1.0).
9. PPI60DEV: Magnitude of the deviation of PPI60 from the sample mean.
10. GDP60SQ: Square of GDP60.
11. RPRI: Student-teacher ratio in primary schools in 1960.

12. RSEC: Student-teacher ratio in secondary schools in 1960.
13. AFRICA: Dummy variable for sub-Saharan Africa.
14. LAT.AMER: Dummy variable for LATIN America.
15. i/y: Average from 1960-1985 of the ratio of real domestic investment to real GDP.
16. FERT: Total fertility rate, average of 1965 and 1985.
17. MORT01: Mortality rate for age 0 through 1, average of 1965 and 1985.
18. FERTNET: FERT$\times$(1-MORT01).

References

Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle," *Second International Symposium on Information Theory*, pp. 267–281.

Amemiya, T. (1980): "Selection of Regressors," *International Economic Review*, 21(2), 331–354.

Barnard, G. A. (1963): "New methods of quality control," *Journal of Royal Statistical Society Series: A*, 126, 255.

Barro, R. J. (1991): "Economic Growth in a Cross Section of Countries," *The Quarterly Journal of Economics*, 106(2), 407–443.

Barro, R. J., and J.-W. Lee (1994): "Sources of economic growth," *Carnegie-Rochester Conference Series on Public Policy*, 40(1), 1–46.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53(2), pp. 603–618.

Burnham, K. P., and D. R. Anderson (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* chap. 4, pp. 153–169. Springer.

Claeskens, G., and N. L. Hjort (2003): "The focused information criterion," *Journal of the American Statistical Association*, 98, 900–916.

Clyde, M., and E. I. George (2004): "Model uncertainty," *Statistical Science*, 19, 81–94.

Davidson, R., and J. MacKinnon (2000): "Bootstrap tests: how many bootstraps?," *Econometric Reviews*, 19(1), 55–68.

DiCiccio, T. J., and B. Efron (1996): "Bootstrap Confidence Intervals," *Statistical Science*, 11(3), pp. 189–212.

Draper, D. (1995): "Assessment and propagation of model uncertainty," *Journal of Royal Statistical Society Series: B*, 57, 45–97.

Fan, J., and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Hansen, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75(4), 1175–1189.

——— (2008): "Least-squares forecast averaging," *Journal of Econometrics*, 146, 342–350.

——— (2009): "Averaging estimators for regressions with a possible structural break," *Econometric Theory*, 25, 1498–1514.

——— (2010): "Averaging estimators for autoregressions with a near unit root," *Journal of Econometrics*, 158, 142–155.

Hansen, B. E., and J. S. Racine (2012): "Jackknife model averaging," *Journal of Econometrics*, 167(1), 38 – 46.

Hendry, D. F. (1976a): "The structure of simultaneous equations estimators," *Journal of Econometrics*, 4(1), 51–88.

——— (1976b): "The Structure of Simultaneous Equations Estimators," *Journal of Econometrics*, 4(1), 51–88.

——— (1980): "Econometrics – Alchemy or Science?," *Economica*, 47(188), 387–406.

——— (1983): "Econometric Modeling: The 'Consumption Function' in Retrospect," *Scottish Journal of Political Economy*, 30(3), 193–220.

Hendry, D. F., and H.-M. Krolzig (1999): "Improving on 'Data Mining Reconsidered' by K.D. Hoover and S.J. Perez," *Econometrics Journal*, 2, 41–58.

Hendry, D. F., and B. Nielsen (2007): *Econometric Modeling: A Likelihood Approach,* chap. 19, pp. 286–301. Princeton University Press.

Hjort, N. L., and G. Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98(464), 879–899.

Hoeting, J. A., D. Madian, A. E. Raftery, and C. T. Volinsky (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14(4), 382–417.

Hoover, K. D., and S. J. Perez (1999): "Data mining reconsidered: encompassing and the general-to-specific approach to specification search," *Econometrics Journal*, 2(2), 167–191.

Johansen, S. (1995): "13. Asymptotic Properties of the Estimators," *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, pp. 177–178.

Kass, R. E., and A. E. Raftery (1995): "Bayes Factors," *Journal of the American Statistical Association*, 90(430), 773–794.

Knox, T., J. H. Stock, and M. W. Watson (2004): "Empirical Bayes Regression With Many Regressors," *working paper*.

Krolzig, H.-M., and D. F. Hendry (2001): "Computer automation of general-to-specific model selection procedures," *Journal of Economic Dynamics & Control*, 25, 831–866.

Leamer, E. (1978): *Specification Searches*. Wiley, New York.

Ley, E., and M. F. Steel (2009): "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression," *Journal of Applied Econometrics*, 24(4), 651–674.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set," *The Annals of Statistics*, 15(3), 958–975.

Liu, C., and J. M. Maheu (2009): "Forecasting realized volatility: a Bayesian model-averaging approach," *Journal of Applied Econometrics*, 24(5), 709–733.

MacKinnon, J. G. (2009): "Bootstrap Hypothesis Testing," in *Handbook of Computational Econometrics*, ed. by D. A. Belsley, and J. Kontoghiorghes, chap. 6, pp. 183–213. Wiley.

Mallows, C. L. (1973): "Some Comments on C_p," *Technometrics*, 15(4), 661–675.

Pagan, A. R. (1987): "Three Econometric Methodologies: A Critical Appraisal," *Journal of Economic Surveys*, 1(1), 3–24.

Raftery, A. E., D. Madigan, and C. T. Volinsky (1997): "Bayesian model averaging for linear regression models," *Journal of American statistical association*, 92, 179–191.

Sala-i-Martin, X., G. Doppelhofer, and R. Miller (2004): "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94, 813–835.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shibata, R. (1981): "An Optimal Selection of Regression Variables," *Biometrika*, 68(1), 45–54.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Wan, A. T., X. Zhang, and G. Zou (2010): "Least Squares model averaging by Mallows criterion," *Journal of Econometrics*, 156, 277–283.

Welch, B. L. (1938): "The Significance of the Difference Between Two Means when the Population Variances are Unequal," *Biometrika*, 29(3/4), pp. 350–362.

Whittle, P. R. (1960): "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and Its Applications*, 5, 302–305.

Wright, J. H. (2009): "Forecasting US inflation by Bayesian model averaging," *Journal of Forecasting*, 28, 131–144.