Distribution Theory of the Least Squares Averaging Estimator

Chu-An Liu

National University of Singapore

23. October 2013

Distribution Theory of the Least Squares Averaging Estimator[*]

Chu-An Liu[†]
National University of Singapore[‡]
ecslca@nus.edu.sg

First Draft: July 2011
This Draft: October 2013

Abstract

This paper derives the limiting distributions of least squares averaging estimators for linear regression models in a local asymptotic framework. We show that the averaging estimators with fixed weights are asymptotically normal and then develop a plug-in averaging estimator that minimizes the sample analog of the asymptotic mean squared error. We investigate the focused information criterion (Claeskens and Hjort, 2003), the plug-in averaging estimator, the Mallows model averaging estimator (Hansen, 2007), and the jackknife model averaging estimator (Hansen and Racine, 2012). We find that the asymptotic distributions of averaging estimators with data-dependent weights are nonstandard and cannot be approximated by simulation. To address this issue, we propose a simple procedure to construct valid confidence intervals with improved coverage probability. Monte Carlo simulations show that the plug-in averaging estimator generally has smaller expected squared error than other existing model averaging methods, and the coverage probability of proposed confidence intervals achieves the nominal level. As an empirical illustration, the proposed methodology is applied to cross-country growth regressions.

Keywords: Local asymptotic theory, Model averaging, Model selection, Plug-in estimators.

JEL Classification: C51, C52.

[*] A previous version was circulated under the title “A Plug-In Averaging Estimator for Regressions with Heteroskedastic Errors.”

[†] I am deeply indebted to Bruce Hansen and Jack Porter for guidance and encouragement. I thank the co-editor, the associate editor, and three referees for very constructive comments and suggestions. The paper has significantly benefited from them. I also thank Xiaoxia Shi, Biing-Shen Kuo, Yu-Chin Hsu, Alan T. K. Wan, and Xinyu Zhang for helpful discussions. Comments from the seminar participants of University of Wisconsin-Madison, National University of Singapore, National Chengchi University, Academia Sinica, and City University of Hong Kong also helped to shape the paper. All errors remain the author’s.

[‡] Department of Economics, National University of Singapore, AS2 Level 6, 1 Arts Link, 117570 Singapore.

1 Introduction

In recent years, interest has increased in model averaging from the frequentist perspective. Unlike model selection, which picks a single model among the candidate models, model averaging incorporates all available information by averaging over all potential models. Model averaging is more robust than model selection since the averaging estimator considers the uncertainty across different models as well as the model bias from each candidate model. The central questions of concern are how to optimally assign the weights for candidate models and how to make inference based on the averaging estimator. This paper investigates the averaging estimators in a local asymptotic framework to deal with these issues. The main contributions of the paper are the following: First, we characterize the optimal weights of the model averaging estimator and propose a plug-in estimator to estimate the infeasible optimal weights. Second, we investigate the focused information criterion (FIC; Claeskens and Hjort, 2003), the plug-in averaging estimator, the Mallows model averaging (MMA; Hansen, 2007), and the jackknife model averaging (JMA; Hansen and Racine, 2012). We show that the asymptotic distributions of averaging estimators with data-dependent weights are nonstandard and cannot be approximated by simulation. Third, we propose a simple procedure to construct valid confidence intervals to address the problem of inference post model selection and averaging.

In finite samples, adding more regressors reduces the model bias but causes a large variance. To yield a good approximation to the finite sample behavior, we follow Hjort and Claeskens (2003a) and Claeskens and Hjort (2008) and investigate the asymptotic distribution of averaging estimators in a local asymptotic framework where the regression coefficients are in a local $n^{-1/2}$ neighborhood of zero. This local asymptotic framework ensures the consistency of the averaging estimator while in general presents an asymptotic bias. Excluding some regressors with little information introduces the model bias but reduces the asymptotic variance. The trade-off between omitted variable bias and estimation variance remains in the asymptotic theory. Under drifting sequences of parameters, the asymptotic mean squared error (AMSE) remains finite and provides a good approximation to finite sample mean squared error. The $O(n^{-1/2})$ framework is canonical in the sense that both squared model biases and estimator variances have the same order $O(n^{-1})$. Therefore, the optimal model is the one that has the best trade-off between bias and variance in this context.

Under the local-to-zero assumption, we derive the asymptotic distributions of least squares averaging estimators with both fixed weights and data-dependent weights. We show that the submodel estimators are asymptotically normal and develop a model selection criterion, FIC, which is an unbiased estimator of the AMSE of the submodel estimator. The FIC chooses the model that achieves the minimum estimated AMSE. We extend the idea of FIC to the model averaging. We first derive the asymptotic distribution of the averaging estimator with fixed weights, which allows us to characterize the optimal weights under the quadratic loss function. The optimal weights are found by numerical minimization of the AMSE of the averaging estimator. We then propose a plug-in estimator of the infeasible optimal fixed weights, and use these estimated weights to construct a plug-in averaging estimator of the parameter of interest. Since the estimated weights depend on the covariance matrix, it is quite easy to model the heteroskedasticity.

Estimated weights are asymptotically random, and this must be taken into account in the asymptotic distribution of the plug-in averaging estimator. This is because the optimal weights depend on the local parameters, which cannot be estimated consistently. To address this issue, we first show the joint convergence in distribution of all candidate models and the data-dependent weights. We then show that the asymptotic distribution of the plug-in estimator is a nonlinear function of the normal random vector. Under the same local asymptotic framework, we show that both MMA and JMA estimators have nonstandard asymptotic distributions.

The limiting distributions of averaging estimators can be used to address the important problem of inference after model selection and averaging. We first show that the asymptotic distribution of the model averaging t-statistic is nonstandard and not asymptotically pivotal. Thus, the traditional confidence intervals constructed by inverting the model averaging t-statistic lead to distorted inference. To address this issue, we propose a simple procedure for constructing valid confidence intervals. Simulations show that the coverage probability of traditional confidence intervals is generally too low, while the coverage probability of proposed confidence intervals achieves the nominal level.

In simulations, we compare the finite sample performance of the plug-in averaging estimator with other existing model averaging methods. Simulation studies show that the plug-in averaging estimator generally produces lower expected squared error than other data-driven averaging estimators. As an empirical illustration, we apply the least squares averaging estimators to cross-country growth regressions. Our estimator has the smaller variance of the log GDP per capita in 1960, though our regression coefficient of the log GDP per capita in 1960 is close to those of other estimators. Our results also find little evidence of the new fundamental growth theory.

The model setup in this paper is similar to that of Hansen (2007) and Hansen and Racine (2012). The main difference is that we consider a finite-order regression model instead of an infinite-order regression model. Hansen (2007) and Hansen and Racine (2012) propose the MMA and JMA estimators and demonstrate the asymptotic optimality in homoskedastic and heteroskedastic settings, respectively. However, it is difficult to make inference based on their estimators since there is no asymptotic distribution available in both papers. By considering a finite-order regression model, we are able to derive the asymptotic distributions of the MMA and JMA estimators in a local asymptotic framework.

The idea of using the local asymptotic framework to investigate the limiting distributions of model averaging estimators is developed by Hjort and Claeskens (2003a) and Claeskens and Hjort (2008). Like them, we employ a drifting asymptotic framework and use the AMSE to approximate the finite sample MSE. We, however, consider a linear regression model instead of the likelihood-based model, and allow for heteroskedastic error settings. Furthermore, we characterize the optimal weights of the averaging estimator in a general setting and propose a plug-in estimator to estimate the infeasible optimal weights.

Other work on the asymptotic properties of averaging estimators includes Leung and Barron (2006), Pötscher (2006), and Hansen (2009, 2010, 2013b). Leung and Barron (2006) study the risk bound of the averaging estimator under a normal error assumption. Pötscher (2006) analyzes the finite sample and asymptotic distributions of the averaging estimator for the two-model case. Hansen (2009) evaluates the AMSE of averaging estimators for the linear regression model with a possible structural break. Hansen (2010) examines the AMSE and forecast expected squared error of averaging estimators in an autoregressive model with a near unit root in a local-to-unity framework. Hansen (2013b) studies the asymptotic risk of least squares averaging estimator in a nested model framework. Most of these studies, however, are limited to the two-model case and the homoskedastic framework.

There is a growing body of literature on frequentist model averaging. Buckland, Burnham, and Augustin (1997) suggest selecting the weights using the exponential AIC. Yang (2000), Yang (2001), and Yuan and Yang (2005) propose an adaptive regression by mixing models. Hansen (2007) introduces the Mallows model averaging estimator for nested and homoskedastic models where the weights are selected by minimizing the Mallows criterion. Wan, Zhang, and Zou (2010) extend the asymptotic optimality of the Mallows model averaging estimator for continuous weights and a non-nested setup. Liang, Zou, Wan, and Zhang (2011) suggest selecting the weights by minimizing the trace of an unbiased estimator of mean squared error. Zhang and Liang (2011) propose an FIC and a smoothed FIC averaging estimator for generalized additive partial linear models. Hansen and Racine (2012) propose the jackknife model averaging estimator for non-nested and heteroskedastic models where the weights are chosen by minimizing a leave-one-out cross-validation criterion. DiTraglia (2013) proposes a moment selection criterion and a moment averaging estimator for the GMM framework. In contrast to frequentist model averaging, there is a large body of literature on Bayesian model averaging, see Hoeting, Madigan, Raftery, and Volinsky (1999) and Moral-Benito (2013) for a literature review.

There is a large body of literature on inference after model selection, including Pötscher (1991), Kabaila (1995, 1998), and Leeb and Pötscher (2003, 2005, 2006, 2008, 2012). These papers point out that the coverage probability of the confidence interval based on the model selection estimator is lower than the nominal level. They also argue that the conditional and unconditional distribution of post model selection estimators cannot be uniformly consistently estimated. In the model averaging literature, Hjort and Claeskens (2003a) and Claeskens and Hjort (2008) show that the traditional confidence interval based on normal approximations leads to distorted inference. Pötscher (2006) argues that the finite-sample distribution of the averaging estimator cannot be uniformly consistently estimated.

There are also alternatives to model selection and model averaging. Tibshirani (1996) introduces the LASSO estimator, a method for simultaneous estimation and variable selection. Zou (2006) proposes the adaptive LASSO approach and presents its oracle properties. Hansen, Lunde, and Nason (2011) propose the model confidence set, which is constructed based on an equivalence test. White and Lu (2014) propose a new Hausman (1978) type test of robustness for the core regression coefficients. They also provide a feasible optimally combined GLS estimator.

The outline of the paper is as follows. Section 2 presents the regression model, the submodel, and the averaging estimator. Section 3 presents the asymptotic framework and assumptions. Section 4 introduces the FIC and the plug-in averaging estimator. Section 5 derives the distribution theory of FIC, plug-in, MMA, and JMA estimators, and proposes a procedure to construct valid confidence intervals for averaging estimators. Section 6 examines the finite sample properties of averaging estimators. Section 7 presents the empirical application and Section 8 concludes the paper. Proofs are included in the Appendix.

2 The Model and the Averaging Estimator

Consider a linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + e_i, \tag{2.1}$$

$$\mathrm{E}(e_i|\mathbf{x}_i, \mathbf{z}_i) = 0, \tag{2.2}$$

$$\mathrm{E}(e_i^2|\mathbf{x}_i, \mathbf{z}_i) = \sigma^2(\mathbf{x}_i, \mathbf{z}_i), \tag{2.3}$$

where y_i is a scalar dependent variable, $\mathbf{x}_i = (x_{1i}, ..., x_{pi})'$ and $\mathbf{z}_i = (z_{1i}, ..., z_{qi})'$ are vectors of regressors, e_i is an unobservable regression error, and $\boldsymbol{\beta}(p \times 1)$ and $\boldsymbol{\gamma}(q \times 1)$ are unknown parameter vectors. The error term is allowed to be heteroskedastic, and there is no further assumption on the distribution of the error term. Here, $\mathbf{x}_i$ are the core regressors, which must be included in the model based on theoretical grounds, while $\mathbf{z}_i$ are the auxiliary regressors, which may or may not be included in the model.[1] Note that $\mathbf{x}_i$ may only include a constant term or even an empty matrix.

Let $\mathbf{y} = (y_1, ..., y_n)'$, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)'$, $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)'$, and $\mathbf{e} = (e_1, ..., e_n)'$. In matrix notation, we write the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e} \tag{2.4}$$

where $\mathbf{H} = (\mathbf{X}, \mathbf{Z})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$.

Suppose that we have a set of M submodels. Let $\boldsymbol{\Pi}_m$ be the $q_m \times q$ selection matrix which selects the included auxiliary regressors. The m'th submodel includes all core regressors $\mathbf{X}$ and a subset of auxiliary regressors $\mathbf{Z}_m$ where $\mathbf{Z}_m = \mathbf{Z}\boldsymbol{\Pi}_m'$. Note that the m'th submodel has $p + q_m$ regressors and q_m is the number of auxiliary regressors $\mathbf{z}_i$ in the submodel m. The set of models could be nested or non-nested.[2] If we consider a sequence of nested models, then $M = q + 1$. If we consider all possible subsets of auxiliary regressors, then $M = 2^q$.

The least squares estimator of $\boldsymbol{\theta}$ for the full model, i.e., all auxiliary regressors are included in the model, is

$$\hat{\boldsymbol{\theta}}_f = \begin{pmatrix} \hat{\boldsymbol{\beta}}_f \\ \hat{\boldsymbol{\gamma}}_f \end{pmatrix} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y}, \tag{2.5}$$

[1] The auxiliary regressors can include any nonlinear transformations of the original variables and the interaction terms between the regressors.

[2] The non-nested models include both the overlapping and the non-overlapping cases. The submodels m and ℓ are called overlapping if $\mathbf{Z}_m \cap \mathbf{Z}_\ell \neq \emptyset$, and non-overlapping otherwise.

and the estimator for the submodel m is

$$\hat{\boldsymbol{\theta}}_m = \begin{pmatrix} \hat{\boldsymbol{\beta}}_m \\ \hat{\boldsymbol{\gamma}}_m \end{pmatrix} = (\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{y}, \tag{2.6}$$

where $\mathbf{H}_m = (\mathbf{X}, \mathbf{Z}_m)$. Let $\mathbf{I}$ denote an identity matrix and $\mathbf{0}$ a zero matrix. If $\boldsymbol{\Pi}_m = \mathbf{I}_q$, then we have $\hat{\boldsymbol{\theta}}_m = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y} = \hat{\boldsymbol{\theta}}_f$, the least squares estimator for the full model. If $\boldsymbol{\Pi}_m = \mathbf{0}$, then we have $\hat{\boldsymbol{\theta}}_m = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the least squares estimator for the narrow model, or the smallest model among all possible submodels.

The parameter of interest is $\mu = \mu(\boldsymbol{\theta}) = \mu(\boldsymbol{\beta}, \boldsymbol{\gamma})$, which is a smooth real-valued function. Let $\hat{\mu}_m = \mu(\hat{\boldsymbol{\theta}}_m) = \mu(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m)$ denote the submodel estimates. Unlike the traditional model selection and model averaging approaches, which assess the global fit of the model, we evaluate the model based on the focus parameter μ. For example, μ may be an individual coefficient or a ratio of two coefficients of regressors.

We now define the averaging estimator of the focus parameter μ. Let $\mathbf{w} = (w_1, ..., w_M)'$ be a weight vector with $w_m \geq 0$ and $\sum_{m=1}^{M} w_m = 1$.[3] That is, the weight vector lies in the unit simplex in $\mathbb{R}^M$:

$$\mathcal{H}_n = \left\{ \mathbf{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}.$$

The sum of the weight vector is required to be one. Otherwise, the averaging estimator is not consistent. The averaging estimator of μ is

$$\bar{\mu}(\mathbf{w}) = \sum_{m=1}^{M} w_m \hat{\mu}_m. \tag{2.7}$$

Note that both Hansen (2007) and Hansen and Racine (2012) consider an infinite-order regression model and make no distinction between core and auxiliary regressors, which is different from our framework. Furthermore, both papers propose an averaging estimator for the conditional mean function instead of the focus parameter μ. The empirical literature tends to focus on one particular parameter instead of assessing the overall properties of the model. In contrast to Hansen (2007) and Hansen and Racine (2012), our method is tailored to the parameter of interest instead of the global fit of the model. We focus attention on a low-dimension function of the model parameters and allow different model weights to be chosen for different parameters of interest.

3 Asymptotic Framework

The least squares estimator for the submodel has omitted variable bias. For nonzero and fixed values of $\boldsymbol{\gamma}$, the asymptotic bias of all models except the full model tends to infinity and hence the

[3] We have fewer restrictions on the weight function than other existing methods. Leung and Barron (2006), Pötscher (2006), Liang, Zou, Wan, and Zhang (2011), and Zhang and Liang (2011) assume the parametric form of the weight function. Hansen (2007) and Hansen and Racine (2012) restrict the weights to be discrete. Contrary to these works, we allow continuous weights without assuming any parametric form, which is more general and applicable than other approaches.

asymptotic approximations break down. We therefore follow Hjort and Claeskens (2003a) and use a local-to-zero asymptotic framework to investigate the asymptotic distribution of the averaging estimator. More precisely, the parameters $\boldsymbol{\gamma}$ are modeled as being a local $n^{-1/2}$ neighborhood of zero.

Assumption 1. $\boldsymbol{\gamma} = \boldsymbol{\gamma}_n = \boldsymbol{\delta}/\sqrt{n}$, where $\boldsymbol{\delta}$ is an unknown constant vector.

Assumption 1 is a technique to ensure that the asymptotic mean squared error of the averaging estimator remains finite.[4] It is a common technique to analyze the asymptotic and finite sample properties of the model selection and averaging estimator, for example, Leeb and Pötscher (2005), Pötscher (2006), Elliott, Gargano, and Timmermann (2013), and Hansen (2013b). This assumption says that the partial correlations between the auxiliary regressors and the dependent variable are weak, which is similar to the definition of the weak instrument, see Staiger and Stock (1997). This assumption implies that as the sample size increases, all of the submodels are close to each other. Under this framework, it is informative to know if we can improve by averaging the candidate models instead of choosing one single model.

The $O(n^{-1/2})$ framework is canonical in the sense that both squared bias and variance have the same order $O(n^{-1})$. Hence, in this context the optimal model is the one that achieves the best trade-off between squared model biases and estimator variances. As shown in the proof of Lemma 1, we can decompose the least squares estimator for the submodel m as

$$\hat{\boldsymbol{\theta}}_m = \boldsymbol{\theta}_m + \left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{H}_m'\mathbf{Z}\left(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}_n + \left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{H}_m'\mathbf{e}$$

where the second term represents the omitted variable bias and $(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m)$ is the selection matrix that chooses the omitted auxiliary regressors. If $\boldsymbol{\gamma}_n$ converges to $\mathbf{0}$ slower than $n^{-1/2}$, the asymptotic bias goes to infinity, which suggests that the full model is the only one we should choose. If $\boldsymbol{\gamma}_n$ converges to $\mathbf{0}$ faster than $n^{-1/2}$, the asymptotic bias goes to zero, which implies that the narrow model is the only one we should consider. In both cases, there is no trade-off between omitted variable bias and estimation variance in the asymptotic theory.[5]

The following assumption is a high-level condition that permits the application of cross-section, panel, and time-series data. Let $\mathbf{h}_i = (\mathbf{x}_i', \mathbf{z}_i')'$ and $\mathbf{Q} = \mathrm{E}\left(\mathbf{h}_i\mathbf{h}_i'\right)$ partitioned so that $\mathrm{E}\left(\mathbf{x}_i\mathbf{x}_i'\right) = \mathbf{Q}_{\mathbf{xx}}$, $\mathrm{E}\left(\mathbf{x}_i\mathbf{z}_i'\right) = \mathbf{Q}_{\mathbf{xz}}$, and $\mathrm{E}\left(\mathbf{z}_i\mathbf{z}_i'\right) = \mathbf{Q}_{\mathbf{zz}}$. Let $\boldsymbol{\Omega} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\big(\mathbf{h}_i\mathbf{h}_j'e_ie_j\big)$ partitioned so that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\big(\mathbf{x}_i\mathbf{x}_j'e_ie_j\big) = \boldsymbol{\Omega}_{\mathbf{xx}}$, $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\big(\mathbf{x}_i\mathbf{z}_j'e_ie_j\big) = \boldsymbol{\Omega}_{\mathbf{xz}}$, and $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\big(\mathbf{z}_i\mathbf{z}_j'e_ie_j\big) = \boldsymbol{\Omega}_{\mathbf{zz}}$. Note that if the error term e_i is serially uncorrelated and identically distributed, $\boldsymbol{\Omega}$ can be simplified as $\boldsymbol{\Omega} = \mathrm{E}\left(\mathbf{h}_i\mathbf{h}_i'e_i^2\right)$, and if the error term is i.i.d. and homoskedastic, then $\boldsymbol{\Omega} = \sigma^2\mathbf{Q}$.

[4] There has been a discussion about the realism of the local asymptotic framework, see Hjort and Claeskens (2003b) and Raftery and Zheng (2003).

[5] The standard asymptotics for nonzero and fixed parameters $\boldsymbol{\gamma}$ correspond to $\boldsymbol{\delta} = \pm\infty$, which is the first case. The zero partial correlations between the auxiliary regressors and the dependent variable correspond to $\boldsymbol{\delta} = \mathbf{0}$, which is the second case.

Assumption 2. As $n \to \infty$, $n^{-1}\mathbf{H}'\mathbf{H} \xrightarrow{p} \mathbf{Q}$ and $n^{-1/2}\mathbf{H}'\mathbf{e} \xrightarrow{d} \mathbf{R} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega})$.

This condition holds under appropriate primitive assumptions. For example, if y_i is a stationary and ergodic martingale difference sequence with finite fourth moments, then the condition follows from the weak law of large numbers and the central limit theorem for martingale difference sequences.

Let

$$\mathbf{S}_0 = \begin{pmatrix} \mathbf{0}_{p\times q} \\ \mathbf{I}_q \end{pmatrix} \text{ and } \mathbf{S}_m = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p\times q_m} \\ \mathbf{0}_{q\times p} & \boldsymbol{\Pi}'_m \end{pmatrix}$$

be selection matrices of dimension $(p+q)\times q$ and $(p+q)\times(p+q_m)$, respectively. Since the extended selection matrix $\mathbf{S}_m$ is non-random with elements either 0 or 1, for the submodel m we have $n^{-1}\mathbf{H}'_m\mathbf{H}_m \xrightarrow{p} \mathbf{Q}_m$ where $\mathbf{Q}_m$ is nonsingular with

$$\mathbf{Q}_m = \mathbf{S}'_m\mathbf{Q}\mathbf{S}_m = \begin{pmatrix} \mathbf{Q}_{\mathbf{xx}} & \mathbf{Q}_{\mathbf{xz}}\boldsymbol{\Pi}'_m \\ \boldsymbol{\Pi}_m\mathbf{Q}_{\mathbf{zx}} & \boldsymbol{\Pi}_m\mathbf{Q}_{\mathbf{zz}}\boldsymbol{\Pi}'_m \end{pmatrix},$$

and $n^{-1/2}\mathbf{H}'_m\mathbf{e} \xrightarrow{d} \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega}_m)$ with

$$\boldsymbol{\Omega}_m = \mathbf{S}'_m\boldsymbol{\Omega}\mathbf{S}_m = \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{xx}} & \boldsymbol{\Omega}_{\mathbf{xz}}\boldsymbol{\Pi}'_m \\ \boldsymbol{\Pi}_m\boldsymbol{\Omega}_{\mathbf{zx}} & \boldsymbol{\Pi}_m\boldsymbol{\Omega}_{\mathbf{zz}}\boldsymbol{\Pi}'_m \end{pmatrix}.$$

The following lemma describes the asymptotic distributions of the least squares estimators. Let $\boldsymbol{\theta}_m = \mathbf{S}'_m\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}'\boldsymbol{\Pi}'_m)' = (\boldsymbol{\beta}', \boldsymbol{\gamma}'_m)'$.

Lemma 1. *Suppose Assumptions 1-2 hold. As $n \to \infty$, we have*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}\right) \xrightarrow{d} \mathbf{Q}^{-1}\mathbf{R} \sim \mathbf{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\right),$$
$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m\right) \xrightarrow{d} \mathbf{A}_m\boldsymbol{\delta} + \mathbf{B}_m\mathbf{R} \sim \mathbf{N}\left(\mathbf{A}_m\boldsymbol{\delta},\ \mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\mathbf{Q}_m^{-1}\right),$$

where $\mathbf{A}_m = \mathbf{Q}_m^{-1}\mathbf{S}'_m\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)$ and $\mathbf{B}_m = \mathbf{Q}_m^{-1}\mathbf{S}'_m$.

Lemma 1 implies that both $\hat{\boldsymbol{\theta}}_f$ and $\hat{\boldsymbol{\theta}}_m$ are consistent. $\mathbf{A}_m\boldsymbol{\delta}$ represents the asymptotic bias of submodel estimators. For the full model, the asymptotic bias is zero since $(\mathbf{I}_q - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m) = \mathbf{0}$. For the submodels, the asymptotic bias is zero if the coefficients of the auxiliary regressors are zero, i.e., $\boldsymbol{\gamma} = 0$, or the auxiliary regressors are uncorrelated, i.e., $\mathbf{Q}$ is a diagonal matrix. The magnitude of the asymptotic bias is determined by two components, the local parameter $\boldsymbol{\delta}$ and the covariance matrix $\mathbf{Q}$, which is illustrated in Figure 1.

Figure 1 shows the asymptotic mean squared error (AMSE) of $\sqrt{n}(\hat{\beta}_2 - \beta_2)$ of the narrow model estimator, the middle model estimator, the full model estimator, and the averaging estimator in a three-nested-model framework. The left panel shows that the best submodel, which has the lowest

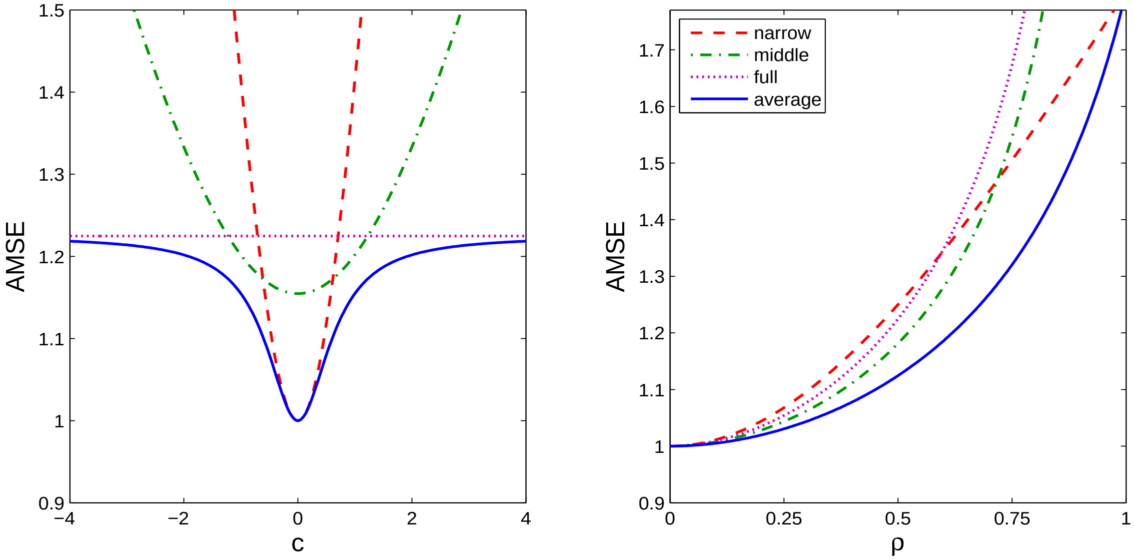

Figure 1: The AMSE of $\sqrt{n}(\hat{\beta}_2 - \beta_2)$ of submodel estimators and the averaging estimator in a three-nested-model framework. The situation is that of $p = 2$, $q = 2$, $M = 3$, $\boldsymbol{\delta} = (c, c)'$, and $\boldsymbol{\Omega} = \sigma^2\mathbf{Q}$. The diagonal elements of $\mathbf{Q}$ are 1, and off-diagonal elements are ρ. The left panel corresponds to $\rho = 0.5$, and the right panel corresponds to $c = 0.75$.

AMSE, varies with $\boldsymbol{\delta}$. When $|\boldsymbol{\delta}|$ is small, the omitted variable bias is relatively small. Therefore, we prefer the narrow model which has an omitted variable bias but a much smaller estimation variance. On the other hand, when $|\boldsymbol{\delta}|$ is large we should prefer the full model. Note that the standard asymptotics for nonzero and fixed parameters $\boldsymbol{\gamma}$ correspond to $\boldsymbol{\delta} = \pm\infty$. The left panel implies that we should always choose the full model if all regression coefficients are modeled as fixed.

The right panel of Figure 1 shows that the best submodel varies with $\boldsymbol{\rho}$, and the full model is not always better in the local asymptotic framework. When the auxiliary regressors are uncorrelated, i.e., $\rho = 0$, all submodel estimators have the same AMSE. For larger ρ, the asymptotic variance increases much faster than asymptotic bias. Therefore, we should consider the smaller model. We also compare the AMSE of the submodel estimators with the AMSE of the averaging estimator with the optimal weight derived in (4.6). The striking feature is that the averaging estimator achieves a much lower AMSE than all submodel estimators in both panels.

4 Focused Information Criterion and Plug-In Averaging Estimator

In this section, we derive a focused information criterion (FIC) model selection for the focus parameter. We also characterize the optimal weights of the averaging estimator and present a plug-in method to estimate the infeasible optimal weights.

4.1 Focused Information Criterion

Let $\mathbf{D}_{\boldsymbol{\theta}_m} = \left(\mathbf{D}_{\boldsymbol{\beta}}', \mathbf{D}_{\boldsymbol{\gamma}_m}'\right)'$, $\mathbf{D}_{\boldsymbol{\beta}} = \partial\mu/\partial\boldsymbol{\beta}$, and $\mathbf{D}_{\boldsymbol{\gamma}_m} = \partial\mu/\partial\boldsymbol{\gamma}_m$, with partial derivatives evaluated at the null points $(\boldsymbol{\beta}', \mathbf{0}')'$. Assume the partial derivatives are continuous in a neighborhood of the null points. Lemma 1 and the delta method imply the following theorem.

Theorem 1. *Suppose Assumptions 1-2 hold. As $n \to \infty$, we have*

$$\sqrt{n}\left(\mu(\hat{\boldsymbol{\theta}}_m) - \mu(\boldsymbol{\theta})\right) \xrightarrow{d} \Lambda_m = \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_m\boldsymbol{\delta} + \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m'\mathbf{R} \sim \mathbf{N}\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_m\boldsymbol{\delta},\ \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_m\mathbf{D}_{\boldsymbol{\theta}}\right)$$

where $\mathbf{C}_m = \left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0$ and $\mathbf{P}_m = \mathbf{S}_m\left(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'$.

Theorem 1 implies joint convergence in distribution of all submodels since all asymptotic distributions can be expressed in terms of the same normal random vector $\mathbf{R}$. A direct calculation yields

$$\mathrm{AMSE}(\hat{\mu}_m) = \mathbf{D}_{\boldsymbol{\theta}}'\left(\mathbf{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{C}_m' + \mathbf{P}_m'\boldsymbol{\Omega}\mathbf{P}_m\right)\mathbf{D}_{\boldsymbol{\theta}}. \tag{4.1}$$

Since $\mathbf{D}_{\boldsymbol{\theta}}$ depends on the focus parameter μ, we can use (4.1) to select a proper submodel depending on the parameter of interest. This is the idea behind FIC proposed by Claeskens and Hjort (2003).

To use (4.1) for model selection, we need to estimate the unknown parameters $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{C}_m$, $\mathbf{P}_m$, $\boldsymbol{\Omega}$, and $\boldsymbol{\delta}$. Define $\hat{\mathbf{D}}_{\boldsymbol{\theta}} = \partial\mu(\hat{\boldsymbol{\theta}}_f)/\partial\boldsymbol{\theta}$ where $\hat{\boldsymbol{\theta}}_f$ is the estimate from the full model. Since $\hat{\boldsymbol{\theta}}_f$ is a consistent estimator of $\boldsymbol{\theta}$, it follows that $\hat{\mathbf{D}}_{\boldsymbol{\theta}}$ is a consistent estimator of $\mathbf{D}_{\boldsymbol{\theta}}$. Note that both $\mathbf{C}_m$ and $\mathbf{P}_m$ are functions of $\mathbf{Q}$ and selection matrices, which can be consistently estimated by the sample analogue.[6] The consistent estimator for $\boldsymbol{\Omega}$ is also available.[7]

We now consider the estimator for the local parameter $\boldsymbol{\delta}$. Unlike $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{C}_m$, $\mathbf{P}_m$, and $\boldsymbol{\Omega}$, there is no consistent estimator for the parameter $\boldsymbol{\delta}$ due to the local asymptotic framework. We can, however, construct an asymptotically unbiased estimator of $\boldsymbol{\delta}$ by using the estimator from the full model. That is, $\hat{\boldsymbol{\delta}} = \sqrt{n}\hat{\boldsymbol{\gamma}}_f$ where $\hat{\boldsymbol{\gamma}}_f$ is the estimate from the full model. From Lemma 1, we know that

$$\hat{\boldsymbol{\delta}} = \sqrt{n}\hat{\boldsymbol{\gamma}}_f \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{R} \sim \mathbf{N}(\boldsymbol{\delta}, \mathbf{S}_0'\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0'). \tag{4.2}$$

As shown above, $\hat{\boldsymbol{\delta}}$ is an asymptotically unbiased estimator for $\boldsymbol{\delta}$ and converges in distribution to a linear function of the normal random vector $\mathbf{R}$. Since the mean of $\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}'$ is $\boldsymbol{\delta}\boldsymbol{\delta}' + \mathbf{S}_0'\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0'$, $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}' - \mathbf{S}_0'\hat{\mathbf{Q}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\mathbf{Q}}^{-1}\mathbf{S}_0$ provides an asymptotically unbiased estimator of $\boldsymbol{\delta}\boldsymbol{\delta}'$.

[6] Let $\hat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i\mathbf{h}_i'$ and then $\hat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$ under Assumption 2.

[7] If the error term is serially uncorrelated and identically distributed, then $\boldsymbol{\Omega}$ can be consistently estimated by the heteroskedasticity-consistent covariance matrix estimator proposed by White (1980). The estimator is $\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i\mathbf{h}_i'\hat{e}_i^2$ where $\hat{e}_i$ is the least squares residual from the full model. If the error term e_i is serially correlated and identically distributed, then $\boldsymbol{\Omega}$ can be estimated consistently by the heteroskedasticity and autocorrelation consistent covariance matrix estimator. The estimator is defined as $\hat{\boldsymbol{\Omega}} = \sum_{j=-n}^{n} k(j/S_n)\hat{\boldsymbol{\Gamma}}(j)$, $\hat{\boldsymbol{\Gamma}}(j) = \frac{1}{n}\sum_{i=1}^{n-j}\mathbf{h}_i\mathbf{h}_{i+j}'\hat{e}_i\hat{e}_{i+j}$ for $j \geq 0$, and $\hat{\boldsymbol{\Gamma}}(j) = \hat{\boldsymbol{\Gamma}}(-j)'$ for $j < 0$, where $k(\cdot)$ is a kernel function and S_n the bandwidth. Under some regularity conditions, it follows that $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$; for serially uncorrelated errors, see White (1980) and White (1984), and for serially correlated errors, see Newey and West (1987) and Andrews (1991b).

Following Claeskens and Hjort (2003), we define the FIC of the m'th submodel as

$$\mathrm{FIC}_m = \hat{\mathbf{D}}_{\boldsymbol{\theta}}' \left(\hat{\mathbf{C}}_m \left(\hat{\boldsymbol{\delta}} \hat{\boldsymbol{\delta}}' - \mathbf{S}_0' \hat{\mathbf{Q}}^{-1} \hat{\boldsymbol{\Omega}} \hat{\mathbf{Q}}^{-1} \mathbf{S}_0 \right) \hat{\mathbf{C}}_m' + \hat{\mathbf{P}}_m \hat{\boldsymbol{\Omega}} \hat{\mathbf{P}}_m \right) \hat{\mathbf{D}}_{\boldsymbol{\theta}}, \tag{4.3}$$

which is an asymptotically unbiased estimator of AMSE($\hat{\mu}_m$). We then select the model with the lowest FIC.

4.2 Plug-In Averaging Estimator

We extend the idea of FIC to the averaging estimator.[8] Instead of comparing the AMSE of each submodel, we derive the AMSE of the averaging estimator with fixed weight in a local asymptotic framework. This result allows us to characterize the optimal weights of the averaging estimator under the quadratic loss function. We then propose a plug-in estimator to estimate the infeasible optimal weights. The following theorem shows the asymptotic normality of the averaging estimator with fixed weights.

Theorem 2. *Suppose Assumptions 1-2 hold. As $n \to \infty$, we have*

$$\sqrt{n} \left(\bar{\mu}(\mathbf{w}) - \mu \right) \xrightarrow{d} \mathbf{N} \left(\mathbf{D}_{\boldsymbol{\theta}}' \mathbf{C}_{\mathbf{w}} \boldsymbol{\delta}, V \right)$$

where $\mathbf{C}_{\mathbf{w}} = \sum_{m=1}^{M} w_m \mathbf{C}_m$ *and* $V = \sum_{m=1}^{M} w_m^2 \mathbf{D}_{\boldsymbol{\theta}}' \mathbf{P}_m \boldsymbol{\Omega} \mathbf{P}_m \mathbf{D}_{\boldsymbol{\theta}} + 2 \sum \sum_{m \neq \ell} w_m w_\ell \mathbf{D}_{\boldsymbol{\theta}}' \mathbf{P}_m \boldsymbol{\Omega} \mathbf{P}_\ell \mathbf{D}_{\boldsymbol{\theta}}$.

The asymptotic bias and variance of the averaging estimator are $\mathbf{D}_{\boldsymbol{\theta}}' \mathbf{C}_{\mathbf{w}} \boldsymbol{\delta}$ and V, respectively. The asymptotic variance has two components. The first component is the weighted average of the variance of each model, and the second component is the weighted average of the covariance between any two models.

Theorem 2 implies that the AMSE of the averaging estimator $\bar{\mu}(\mathbf{w})$ is

$$\mathrm{AMSE}(\bar{\mu}(\mathbf{w})) = \mathbf{w}' \boldsymbol{\Psi} \mathbf{w} \tag{4.4}$$

where $\boldsymbol{\Psi}$ is an $M \times M$ matrix with the (m, ℓ)th element

$$\Psi_{m,\ell} = \mathbf{D}_{\boldsymbol{\theta}}' \left(\mathbf{C}_m \boldsymbol{\delta} \boldsymbol{\delta}' \mathbf{C}_\ell' + \mathbf{P}_m \boldsymbol{\Omega} \mathbf{P}_\ell \right) \mathbf{D}_{\boldsymbol{\theta}}. \tag{4.5}$$

The optimal fixed-weight vector is the value that minimizes AMSE($\bar{\mu}(\mathbf{w})$) over $\mathbf{w} \in \mathcal{H}_n$:

$$\mathbf{w}^o = \underset{\mathbf{w} \in \mathcal{H}_n}{\operatorname{argmin}}\, \mathbf{w}' \boldsymbol{\Psi} \mathbf{w}. \tag{4.6}$$

[8] Hjort and Claeskens (2003a) propose a smoothed FIC averaging estimator, which assigns the weights of each candidate model by using the exponential FIC. The weight function is a parametric form and is defined as $\hat{w} = exp\left(-\alpha \mathrm{FIC}_m / 2\kappa^2\right) / \sum_{\ell=1}^{M} exp\left(-\alpha \mathrm{FIC}_\ell / 2\kappa^2\right)$ where $\kappa^2 = \mathbf{D}_{\boldsymbol{\theta}}' \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathbf{Q}^{-1} \mathbf{D}_{\boldsymbol{\theta}}$. The simulation shows that the performance of the smoothed FIC averaging estimator is sensitive to the choice of the nuisance parameter α and there is no data-driven method available to choose α. They also consider the averaging estimator, which selects weights to minimize the estimated risk in the likelihood framework for a two-model case, the full model and the narrow model.

Since the optimal weights depend on the covariance matrix $\mathbf{\Omega}$, it is quite easy to model the heteroskedasticity. When we have more than two submodels, there is no closed-form solution to (4.6). In this case, the weight vector can be found numerically via quadratic programming for which numerical algorithms are available for most programming languages.

The optimal weights are infeasible because they depend on the unknown parameters $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{C}_m$, $\mathbf{P}_m$, $\mathbf{\Omega}$, and $\boldsymbol{\delta}$. Furthermore, we cannot estimate the optimal weights directly because there is no closed form expression when the number of models is greater than two. A straightforward solution is to estimate the AMSE of the averaging estimator given in (4.4) and (4.5), and to choose the data-dependent weights by minimizing the sample analog of the AMSE.

As mentioned by Hjort and Claeskens (2003a), we can estimate AMSE$(\bar{\mu}(\mathbf{w}))$ by inserting $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$ or using unbiased $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}' - \mathbf{S}_0'\hat{\mathbf{Q}}^{-1}\hat{\mathbf{\Omega}}\hat{\mathbf{Q}}^{-1}\mathbf{S}_0$ for $\boldsymbol{\delta}\boldsymbol{\delta}'$. The plug-in estimator of (4.4) is $\mathbf{w}'\hat{\mathbf{\Psi}}\mathbf{w}$ where $\hat{\mathbf{\Psi}}$ is the sample analog of $\mathbf{\Psi}$ with the (m,ℓ)th element

$$\hat{\Psi}_{m,\ell} = \hat{\mathbf{D}}_{\boldsymbol{\theta}}' \left(\hat{\mathbf{C}}_m \hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}' \hat{\mathbf{C}}_\ell' + \hat{\mathbf{P}}_m \hat{\mathbf{\Omega}} \hat{\mathbf{P}}_\ell \right) \hat{\mathbf{D}}_{\boldsymbol{\theta}}. \tag{4.7}$$

The plug-in averaging estimator is defined as

$$\bar{\mu}(\hat{\mathbf{w}}) = \sum_{m=1}^{M} \hat{w}_m \hat{\mu}_m \quad \text{and} \quad \hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\, \mathbf{w}'\hat{\mathbf{\Psi}}\mathbf{w}. \tag{4.8}$$

The alternative estimator of $\Psi_{m,\ell}$ is

$$\hat{\Psi}_{m,\ell} = \hat{\mathbf{D}}_{\boldsymbol{\theta}}' \left(\hat{\mathbf{C}}_m \left(\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}' - \mathbf{S}_0'\hat{\mathbf{Q}}^{-1}\hat{\mathbf{\Omega}}\hat{\mathbf{Q}}^{-1}\mathbf{S}_0 \right) \hat{\mathbf{C}}_\ell' + \hat{\mathbf{P}}_m \hat{\mathbf{\Omega}} \hat{\mathbf{P}}_\ell \right) \hat{\mathbf{D}}_{\boldsymbol{\theta}}. \tag{4.9}$$

As shown in the next section, the estimator (4.7) has a simpler limiting distribution than the estimator (4.9). Also, the simulation shows that the estimator (4.7) has better finite sample performance than the estimator (4.9).

5 Asymptotic Distributions of Averaging Estimators

In this section, we present the asymptotic distributions of the FIC model selection estimator, the plug-in averaging estimator, the Mallows model averaging (MMA) estimator, and the jackknife model averaging (JMA) estimator.[9] We also propose a valid confidence interval for the model averaging estimator.

5.1 Asymptotic Distributions of FIC and Plug-In Averaging Estimator

The model selection estimator based on information criteria is a special case of the model averaging estimator. The model selection puts the whole weight on the model with the smallest value of the information criterion and gives other models zero weight. The weight function of the model selection estimator can be expressed by the indicator function.

[9] In an earlier version of this paper, we also obtained the distribution results for the AIC model selection estimator and S-AIC model averaging estimator.

The weight function of the FIC estimator is thus

$$\hat{w}_m = \mathbf{1}\left\{ \mathrm{FIC}_m = \min(\mathrm{FIC}_1, \mathrm{FIC}_2, ..., \mathrm{FIC}_M) \right\},$$

where $\mathbf{1}\{\cdot\}$ is an indicator function that takes value 1 if $\mathrm{FIC}_m = \min(\mathrm{FIC}_1, \mathrm{FIC}_2, ..., \mathrm{FIC}_M)$ and 0 otherwise.

Note that $\hat{\mathbf{D}}_{\boldsymbol{\theta}}$, $\hat{\mathbf{C}}_m$, $\hat{\mathbf{P}}_m$, and $\hat{\mathbf{\Omega}}$ are consistent estimators. Since $\hat{\boldsymbol{\delta}} = \sqrt{n} \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{R}$, we can show that

$$\mathrm{FIC}_m \xrightarrow{d} \mathbf{D}_{\boldsymbol{\theta}}' \left(\mathbf{C}_m \left(\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}' - \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0 \right) \mathbf{C}_m' + \mathbf{P}_m\mathbf{\Omega}\mathbf{P}_m \right) \mathbf{D}_{\boldsymbol{\theta}}.$$

This result implies that the FIC estimator has a nonstandard limiting distribution. The following theorem presents the asymptotic distribution of the plug-in averaging estimator defined in (4.7) and (4.8).[10]

Theorem 3. *Let $\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{H}_n} \mathbf{w}'\hat{\mathbf{\Psi}}\mathbf{w}$ be the plug-in weights. Assume $\hat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$. Suppose Assumptions 1-2 hold. As $n \to \infty$, we have*

$$\mathbf{w}'\hat{\mathbf{\Psi}}\mathbf{w} \xrightarrow{d} \mathbf{w}'\mathbf{\Psi}^*\mathbf{w} \tag{5.1}$$

where $\mathbf{\Psi}^*$ is an $M \times M$ matrix with the (m, ℓ)th element

$$\Psi_{m,\ell}^* = \mathbf{D}_{\boldsymbol{\theta}}' \left(\mathbf{C}_m\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}'\mathbf{C}_\ell' + \mathbf{P}_m\mathbf{\Omega}\mathbf{P}_\ell \right) \mathbf{D}_{\boldsymbol{\theta}}. \tag{5.2}$$

Also, we have

$$\hat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{H}_n} \mathbf{w}'\mathbf{\Psi}^*\mathbf{w}, \tag{5.3}$$

and

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}) - \mu\big) \xrightarrow{d} \sum_{m=1}^{M} w_m^*\Lambda_m \tag{5.4}$$

where Λ_m is defined in Theorem 1.

Rather than impose regularity conditions, we assume there exists a consistent estimator for $\mathbf{\Omega}$. The sufficient condition for the consistency is that e_i is i.i.d. or a martingale difference sequence with finite fourth moment. For serial correlation, data is a mean zero α-mixing or φ-mixing sequence. Theorem 3 shows that the estimated weights are asymptotically random under the local asymptotic assumption. This is because the local parameter $\boldsymbol{\delta}$ cannot be consistently estimated and thus the estimate $\hat{\boldsymbol{\delta}}$ is random in the limit.

In order to derive the asymptotic distribution of the plug-in averaging estimator, we show that there is joint convergence in distribution of all submodel estimators $\hat{\mu}_m$ and estimated weights $\hat{\mathbf{w}}$.

[10] For the plug-in averaging estimator defined in (4.9), the limiting distribution is the same except (5.2) is replaced by $\Psi_{m,\ell}^* = \mathbf{D}_{\boldsymbol{\theta}}' \left(\mathbf{C}_m \left(\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}' - \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0 \right) \mathbf{C}_\ell' + \mathbf{P}_m\mathbf{\Omega}\mathbf{P}_\ell \right) \mathbf{D}_{\boldsymbol{\theta}}$.

The joint convergence in distribution comes from the fact that both Λ_m and w_m^* can be expressed in terms of the normal random vector $\mathbf{R}$. It turns out the limiting distribution of the plug-in averaging estimator is not normally distributed. Instead, it is a nonlinear function of the normal random vector $\mathbf{R}$. The non-normal nature of the limiting distribution of the averaging estimator with data-dependent weights is also pointed out by Hjort and Claeskens (2003a) and Claeskens and Hjort (2008).

5.2 Mallows Model Averaging Estimator

Hansen (2007) proposes the Mallows model averaging estimator for the homoskedastic linear regression model. He extends the asymptotic optimality from model selection in Li (1987) to model averaging. He shows that the average squared error of the MMA estimator is asymptotic equivalent to the lowest expected squared error. The MMA estimator, however, is not asymptotically optimal in our framework. This is because the condition (15) of Hansen (2007) does not hold in the local asymptotic framework. The condition requires that there is no submodel m for which the bias is zero, which does not hold in our framework since the full model has no bias.

Let $\hat{\mathbf{e}}(\mathbf{w}) = \mathbf{y} - \mathbf{H}\bar{\boldsymbol{\theta}}(\mathbf{w})$ be the averaging residual vector and $\bar{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{S}_m \hat{\boldsymbol{\theta}}_m$ the averaging estimator of $\boldsymbol{\theta}$. Hansen (2007) suggests selecting the model weights by minimizing the Mallow's criterion:

$$C_n(\mathbf{w}) = \hat{\mathbf{e}}(\mathbf{w})'\hat{\mathbf{e}}(\mathbf{w}) + 2\sigma^2 \mathbf{k}'\mathbf{w}, \tag{5.5}$$

where $\sigma^2 = \mathrm{E}(e_i^2)$, $\mathbf{k} = (k_1, ..., k_M)'$, and $k_m = p + q_m$.

Let $\hat{\mathbf{e}}_f = \mathbf{y} - \mathbf{H}\hat{\boldsymbol{\theta}}_f$ and $\hat{\mathbf{e}}_m = \mathbf{y} - \mathbf{H}_m\hat{\boldsymbol{\theta}}_m$ be the residual vectors from the full model and the submodel m, respectively. To derive the asymptotic distribution of the MMA estimator, we add and subtract the sum of squared residuals of the full model and rewrite the Mallow's criterion (5.5) as

$$C_n(\mathbf{w}) = \mathbf{w}'\boldsymbol{\zeta}_n\mathbf{w} + 2\sigma^2\mathbf{k}'\mathbf{w} + \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f, \tag{5.6}$$

where $\boldsymbol{\zeta}_n$ is an $M \times M$ matrix with the (m, ℓ)th element $\zeta_{m,\ell} = \hat{\mathbf{e}}_m'\hat{\mathbf{e}}_\ell - \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f$. Note that $\hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f$ is not related to the weight vector $\mathbf{w}$. Therefore, minimizing (5.6) over $\mathbf{w} = (w_1, ..., w_M)$ is equivalent to minimizing

$$\widetilde{C}_n(\mathbf{w}) = \mathbf{w}'\boldsymbol{\zeta}_n\mathbf{w} + 2\sigma^2\mathbf{k}'\mathbf{w}. \tag{5.7}$$

Since the criterion function $\widetilde{C}_n(\mathbf{w})$ is a quadratic function of the weight vector, the MMA weights can be found by quadratic programming as the optimal fixed-weight vector and the plug-in weight vector. However, unlike the plug-in averaging estimator where the weights are tailored to the parameter of interest, the MMA estimator selects the weights based on the conditional mean function. In practice, we use $s^2 = \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f/(n - p - q)$ to estimate σ^2. Under some regularity conditions, it follows that s^2 is consistent for σ^2. The following theorem shows the limiting distribution of the MMA estimator.[11]

Theorem 4. *Let* $\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}} \widetilde{C}_n(\mathbf{w})$ *be the MMA weights. Suppose Assumptions 1-2 hold. As* $n \to \infty$*, we have*

$$\widetilde{C}_n(\mathbf{w}) = \mathbf{w}'\boldsymbol{\zeta}_n\mathbf{w} + 2\sigma^2\mathbf{k}'\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w} + 2\sigma^2\mathbf{k}'\mathbf{w} \tag{5.8}$$

where $\boldsymbol{\zeta}^*$ *is an* $M \times M$ *matrix with the* (m,ℓ)*th element*

$$\zeta^*_{m,\ell} = \mathbf{R}'_m\mathbf{Q}\mathbf{R}_\ell \quad \text{and} \quad \mathbf{R}_m = \mathbf{C}_m\boldsymbol{\delta} + \left(\mathbf{P}_m - \mathbf{Q}^{-1}\right)\mathbf{R}. \tag{5.9}$$

Also, we have

$$\hat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^* = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}} \left(\mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w} + 2\sigma^2\mathbf{k}'\mathbf{w}\right) \tag{5.10}$$

and

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}) - \mu\big) \xrightarrow{d} \sum_{m=1}^{M} w^*_m\Lambda_m \tag{5.11}$$

where Λ_m *is defined in Theorem 1.*

The main difference between Theorem 3 and 4 is the limiting behavior of the weight vector. Since the plugin averaging estimator chooses the weight based on the focus parameter, the asymptotic distribution of the selected weight involves the partial derivatives $\mathbf{D}_{\boldsymbol{\theta}}$. Therefore, for a different parameter of interests, we have different asymptotic distributions. Unlike the plug-in averaging estimator, the MMA estimator selects the weights based on the conditional mean function. As a result, the limiting distribution of the weight function does not depend on the parameter of interest.

5.3 Jackknife Model Averaging Estimator

Hansen and Racine (2012) propose the jackknife model averaging estimator for the linear regression model and demonstrate the asymptotic optimality of the JMA estimator in the presence of heteroskedasticity. They extend the asymptotic optimality from model selection for heteroskedastic regressions in Andrews (1991a) to model averaging. Similar to the MMA estimator, the JMA estimator is not asymptotically optimal in the linear regression model with a finite number of regressors.

Hansen and Racine (2012) suggest selecting the weights by minimizing a leave-one-out cross-validation criterion:

$$CV_n(\mathbf{w}) = \frac{1}{n}\mathbf{w}'\tilde{\mathbf{e}}'\tilde{\mathbf{e}}\mathbf{w} \tag{5.12}$$

[11] Hansen (2013b) also derives the asymptotic distribution of the MMA estimator. He derives the asymptotic distribution of the MMA estimator in a nested model framework where the regressors can be partitioned into groups, while our results can apply to both nested or non-nested models.

where $\tilde{\mathbf{e}} = (\tilde{\mathbf{e}}_1, ..., \tilde{\mathbf{e}}_M)$ is a $n \times M$ matrix of leave-one-out least squares residuals and $\tilde{\mathbf{e}}_m$ are the residuals of submodel m obtained by least squares estimation without the $i'th$ observation.

To derive the asymptotic distribution of the JMA estimator, we adopt the same strategy and rewrite (5.12) as

$$CV_n(\mathbf{w}) = \frac{1}{n}\mathbf{w}'\boldsymbol{\xi}_n\mathbf{w} + \frac{1}{n}\hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f \tag{5.13}$$

where $\boldsymbol{\xi}_n$ is an $M \times M$ matrix with the (m, ℓ)th element $\xi_{m,\ell} = \tilde{\mathbf{e}}_m'\tilde{\mathbf{e}}_\ell - \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f$. Note that minimizing $CV_n(\mathbf{w})$ over $\mathbf{w} = (w_1, ..., w_M)$ is equivalent to minimizing

$$\widetilde{CV_n}(\mathbf{w}) = \mathbf{w}'\boldsymbol{\xi}_n\mathbf{w}. \tag{5.14}$$

Like the MMA estimator, the JMA estimator chooses the weights based on the conditional mean function instead of the focus parameter. Similar to the plug-in averaging estimator and the MMA estimator, the weight vector of the JMA estimator can be found by quadratic programming.[12] The following assumption is imposed on the data generating process.

Assumption 3. (a) $\{(y_i, \mathbf{x}_i, \mathbf{z}_i) : i = 1, ..., n\}$ are i.i.d. (b) $\mathrm{E}(e_i^4) < \infty$, $\mathrm{E}(x_{ji}^4) < \infty$ for $j = 1, ..., p$, and $\mathrm{E}(z_{ji}^4) < \infty$ for $j = 1, ..., q$.

Condition (a) in Assumption 3 is the i.i.d. assumption, which is also made in Hansen and Racine (2012). The result in Theorem 5 can be extended to the stationary case. Condition (b) is the standard assumption for the linear regression model. Note that Assumption 3 implies Assumption 2. Therefore, the results in Lemma 1, Theorem 1, and Theorem 2 hold under Assumptions 1 and 3.

Theorem 5. *Let* $\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\,\widetilde{CV_n}(\mathbf{w})$ *be the JMA weights. Suppose Assumptions 1 and 3 hold. As* $n \to \infty$, *we have*

$$\widetilde{CV_n}(\mathbf{w}) = \mathbf{w}'\boldsymbol{\xi}_n\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\xi}^*\mathbf{w} \tag{5.15}$$

where $\boldsymbol{\xi}^*$ *is an* $M \times M$ *matrix with the* (m, ℓ)*th element*

$$\xi^*_{m,\ell} = \mathbf{R}_m'\mathbf{Q}\mathbf{R}_\ell + tr\left(\mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\right) + tr\left(\mathbf{Q}_\ell^{-1}\boldsymbol{\Omega}_\ell\right), \tag{5.16}$$

where $\mathbf{R}_m$ *is defined in Theorem 4. Also, we have*

$$\hat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^* = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\,\mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}, \tag{5.17}$$

and

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}) - \mu\big) \xrightarrow{d} \sum_{m=1}^{M} w_m^*\Lambda_m \tag{5.18}$$

where Λ_m *is defined in Theorem 1.*

[12] However, the computational burden of the JMA estimator is heavier than the plug-in averaging estimator and MMA estimator when both the sample size and the number of regressors are large.

Note that the first term of $\xi^*_{m,\ell}$ in (5.16) is the same as $\zeta^*_{m,\ell}$ in (5.9). This is because both the JMA and MMA estimators select weights based on the conditional mean function. Under conditional homoskedasicity $\mathrm{E}(e_i^2|\mathbf{x}_i,\mathbf{z}_i)=\sigma^2$, we have $\mathbf{\Omega}=\sigma^2\mathbf{Q}$. Thus, in this case, the second and third terms in (5.16) are simplified as $\sigma^2 k_m$ and $\sigma^2 k_\ell$.

5.4 Valid Confidence Interval

We now discuss how to make inference based on the distribution results derived from previous sections. Let $w(m|\hat{\boldsymbol{\delta}})$ denote a data-dependent weight function for the m'th model. Consider an averaging estimator of the focus parameter μ as

$$\bar{\mu}=\sum_{m=1}^{M} w(m|\hat{\boldsymbol{\delta}})\hat{\mu}_m \tag{5.19}$$

where the weights $w(m|\hat{\boldsymbol{\delta}})$ take the values in the interval $[0,1]$ and the sum of the weights is required to sum to 1. Following Theorem 2, we define the standard error of $\bar{\mu}$ as $s(\bar{\mu})=n^{-1/2}\sqrt{\hat{V}}$ where

$$\hat{V}=\sum_{m=1}^{M} w(m|\hat{\boldsymbol{\delta}})^2\hat{\mathbf{D}}_{\boldsymbol{\theta}}'\hat{\mathbf{P}}_m\hat{\mathbf{\Omega}}\hat{\mathbf{P}}_m\hat{\mathbf{D}}_{\boldsymbol{\theta}}+2\sum\sum_{m\neq\ell} w(m|\hat{\boldsymbol{\delta}})w(\ell|\hat{\boldsymbol{\delta}})\hat{\mathbf{D}}_{\boldsymbol{\theta}}'\hat{\mathbf{P}}_m\hat{\mathbf{\Omega}}\hat{\mathbf{P}}_\ell\hat{\mathbf{D}}_{\boldsymbol{\theta}}. \tag{5.20}$$

Since μ is a scalar, we can construct the confidence interval by using the t-statistic. Consider the t-statistic of the averaging estimator of μ

$$t_n(\mu)=\frac{\bar{\mu}-\mu}{s(\bar{\mu})}. \tag{5.21}$$

Unfortunately, the asymptotic distribution of the t-statistic $t_n(\mu)$ is nonstandard. Furthermore, $t_n(\mu)$ is not asymptotically pivotal. Suppose $w(m|\hat{\boldsymbol{\delta}})\xrightarrow{d} w(m|\mathbf{R}_{\boldsymbol{\delta}})$ where $\mathbf{R}_{\boldsymbol{\delta}}=\boldsymbol{\delta}+\mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{R}$.[13] Then we can show that

$$t_n(\mu)\xrightarrow{d}(V(\mathbf{R}_{\boldsymbol{\delta}}))^{-1/2}\sum_{m=1}^{M} w(m|\mathbf{R}_{\boldsymbol{\delta}})\Lambda_m \tag{5.22}$$

where Λ_m is defined in Theorem 1 and

$$V(\mathbf{R}_{\boldsymbol{\delta}})=\sum_{m=1}^{M} w(m|\mathbf{R}_{\boldsymbol{\delta}})^2\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathbf{\Omega}\mathbf{P}_m\mathbf{D}_{\boldsymbol{\theta}}+2\sum\sum_{m\neq\ell} w(m|\mathbf{R}_{\boldsymbol{\delta}})w(\ell|\mathbf{R}_{\boldsymbol{\delta}})\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathbf{\Omega}\mathbf{P}_\ell\mathbf{D}_{\boldsymbol{\theta}}.$$

Equation (5.22) shows that the limiting distribution of the t-statistic $t_n(\mu)$ is a nonlinear function of the normal random vector $\mathbf{R}$ and the local parameter $\boldsymbol{\delta}$. In Figure 2, we simulate the asymptotic distribution of the model averaging t-statistic in a three-nested-model framework for three different ρ. The density functions are computed by kernel estimation using 5000 random samples. The figure shows that the asymptotic distributions of $t_n(\mu)$ for large ρ are quite different from the standard normal probability density function. As a result, the traditional confidence intervals based on normal approximations lead to distorted inference.

[13] For example, if $\mathbf{w}(\hat{\boldsymbol{\delta}})=(w(1|\hat{\boldsymbol{\delta}}),...,w(M|\hat{\boldsymbol{\delta}}))$ are the plug-in weights, then $\mathbf{w}(\hat{\boldsymbol{\delta}})\xrightarrow{d}\mathbf{w}(\mathbf{R}_{\boldsymbol{\delta}})=\underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\,\mathbf{w}'\mathbf{\Psi}^*\mathbf{w}$ as shown in Theorem 3.

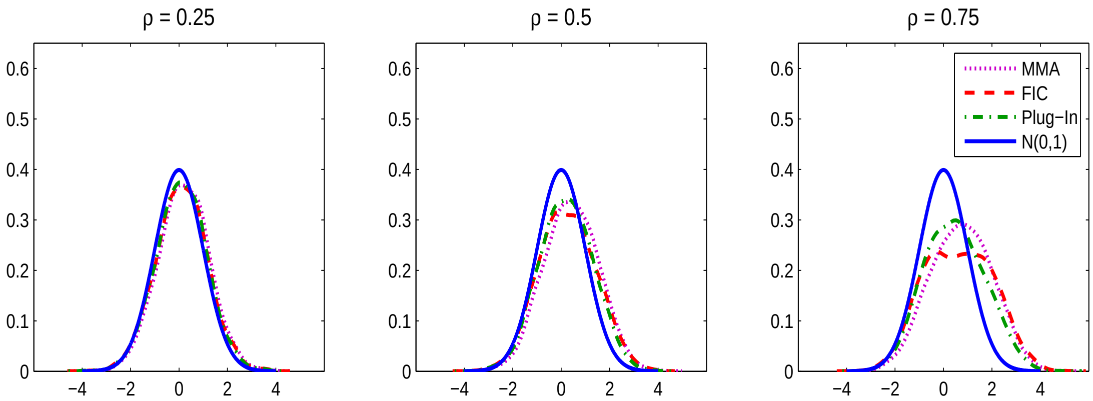

Figure 2: Density functions of the model averaging t-statistic in a three-nested-model framework. The situation is that of $p = 2$, $q = 2$, $M = 3$, $\boldsymbol{\delta} = (1, 1)'$, and $\mathbf{\Omega} = \sigma^2 \mathbf{Q}$. The diagonal elements of $\mathbf{Q}$ are 1 and off-diagonal elements are ρ. The three situations correspond to $\rho = 0.25$, $\rho = 0.50$, and $\rho = 0.75$.

As shown above, the asymptotic distribution of the t-statistic of the averaging estimator depends on unknown parameters, and thus cannot directly be used for inference. Furthermore, we cannot simulate the asymptotic distribution of $t_n(\mu)$ since the local parameters are unknown and cannot be estimated consistently. To address this issue, we propose a simple procedure for constructing valid confidence intervals. The following theorem presents a general distribution theorem for the averaging estimator with data-dependent weights.

Theorem 6. *Assume* $w(m|\hat{\boldsymbol{\delta}}) \xrightarrow{d} w(m|\mathbf{R}_{\boldsymbol{\delta}})$. *Suppose Assumptions 1-2 hold. As* $n \to \infty$, *we have*

$$\sqrt{n}\,(\bar{\mu} - \mu) \xrightarrow{d} \mathbf{D}_{\boldsymbol{\theta}}' \mathbf{Q}^{-1} \mathbf{R} + \mathbf{D}_{\boldsymbol{\theta}}' \left(\sum_{m=1}^{M} w(m|\mathbf{R}_{\boldsymbol{\delta}}) \mathbf{C}_m \right) \mathbf{R}_{\boldsymbol{\delta}}$$

where $\mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{S}_0' \mathbf{Q}^{-1} \mathbf{R}$.

Theorem 6 shows that the limiting distribution of the averaging estimator with data-dependent weights is nonstandard in general since the estimated weights are asymptotically random. As discussed above, a direct construction of a confidence interval based on the t-statistic is not valid since the limiting distribution of $\sqrt{n}\,(\bar{\mu} - \mu)$ is a nonlinear function of the normal random vector $\mathbf{R}$ and the local parameters $\boldsymbol{\delta}$.

We follow Hjort and Claeskens (2003a), Claeskens and Carroll (2007), and Zhang and Liang (2011) to construct a valid confidence interval as follows. Let $\hat{\kappa}^2$ be a consistent estimator of $\mathbf{D}_{\boldsymbol{\theta}}' \mathbf{Q} \mathbf{\Omega} \mathbf{Q} \mathbf{D}_{\boldsymbol{\theta}}$. Since there is a simultaneous convergence in distribution, it follows that

$$\left[\sqrt{n}\,(\bar{\mu} - \mu) - \hat{\mathbf{D}}_{\boldsymbol{\theta}}' \left(\sum_{m=1}^{M} w(m|\hat{\boldsymbol{\delta}}) \hat{\mathbf{C}}_m \right) \hat{\boldsymbol{\delta}} \right] \Big/ \hat{\kappa} \xrightarrow{d} \mathbf{N}\,(0, 1)\,.$$

Let $b(\hat{\boldsymbol{\delta}}) = \hat{\mathbf{D}}'_{\boldsymbol{\theta}} \left(\sum_{m=1}^{M} w(m|\hat{\boldsymbol{\delta}})\hat{\mathbf{C}}_m\right) \hat{\boldsymbol{\gamma}}_f$. Then, we define the confidence interval for μ as

$$\mathrm{CI}_n = \left[\bar{\mu} - b(\hat{\boldsymbol{\delta}}) - z_{1-\alpha/2}\frac{\hat{\kappa}}{\sqrt{n}},\ \bar{\mu} - b(\hat{\boldsymbol{\delta}}) + z_{1-\alpha/2}\frac{\hat{\kappa}}{\sqrt{n}}\right] \tag{5.23}$$

where $z_{1-\alpha/2}$ is $1-\alpha/2$ quantile of the standard normal distribution. Thus, we have $Pr(\mu \in \mathrm{CI}_n) \to 2\Phi(z_{1-\alpha/2}) - 1$ where $\Phi(\cdot)$ is a standard normal distribution function, which means the proposed confidence interval (5.23) has asymptotically the correct coverage probability.

6 Simulation Study

In this section, we investigate the finite sample mean square error of the averaging estimators via Monte Carlo experiments. We also provide the comparison of the coverage probability between the proposed confidence intervals and traditional confidence intervals.

6.1 Simulation Setup

We consider a linear regression model with a finite number of regressors

$$y_i = \sum_{j=1}^{k} \theta_j x_{ji} + e_i, \quad i = 1, ..., n, \tag{6.1}$$

where $x_{1i} = 1$ and $(x_{2i}, ..., x_{ki})' \sim N(0, \mathbf{Q})$. The diagonal elements of $\mathbf{Q}$ are 1, and off-diagonal elements are ρ. The error term is generated from a normal distribution $N(0, \sigma_i^2)$ where $\sigma_i = 1$ for the homoskedastic simulation and $\sigma_i = (1 + 6x_{2i}^2)/11$ for the heteroskedastic simulation. We let x_{1i}, x_{2i}, and x_{3i} be the core regressors and consider all other regressors auxiliary. The regression coefficients are determined by the rule

$$\boldsymbol{\theta} = c\left(\frac{1}{a}, \frac{1}{a}, \frac{1}{a}, \frac{1}{\sqrt{n}}\left(1, \frac{q-1}{q}, ..., \frac{1}{q}\right)\right) \tag{6.2}$$

where q is the number of the auxiliary regressors. The parameter c is selected to control the population $R^2 = \tilde{\boldsymbol{\theta}}'\mathbf{Q}\tilde{\boldsymbol{\theta}}/(1 + \tilde{\boldsymbol{\theta}}'\mathbf{Q}\tilde{\boldsymbol{\theta}})$ where $\tilde{\boldsymbol{\theta}} = (\theta_2, ..., \theta_k)'$ and R^2 varies on a grid between 0.1 and 0.9. The local parameters are determined by $\delta_j = \sqrt{n}\theta_j = c(k - j + 1)/q$ for $j \geq 4$. We consider all possible submodels, that is, the number of models is $M = 2^{k-3}$.

We consider five estimators: (1) optimal frequentist model averaging estimator (labeled OFMA), (2) Mallows model averaging estimator (labeled MMA), (3) jackknife model averaging estimator (labeled JMA), (4) focused information criterion model selection (labeled FIC), and (5) plug-in averaging estimator (labeled Plug-In).[14] The optimal frequentist model averaging estimator is

[14] We only report the results of the plug-in averaging estimator defined in (4.7) since the estimator (4.7) outperforms the estimator (4.9) in most simulations.

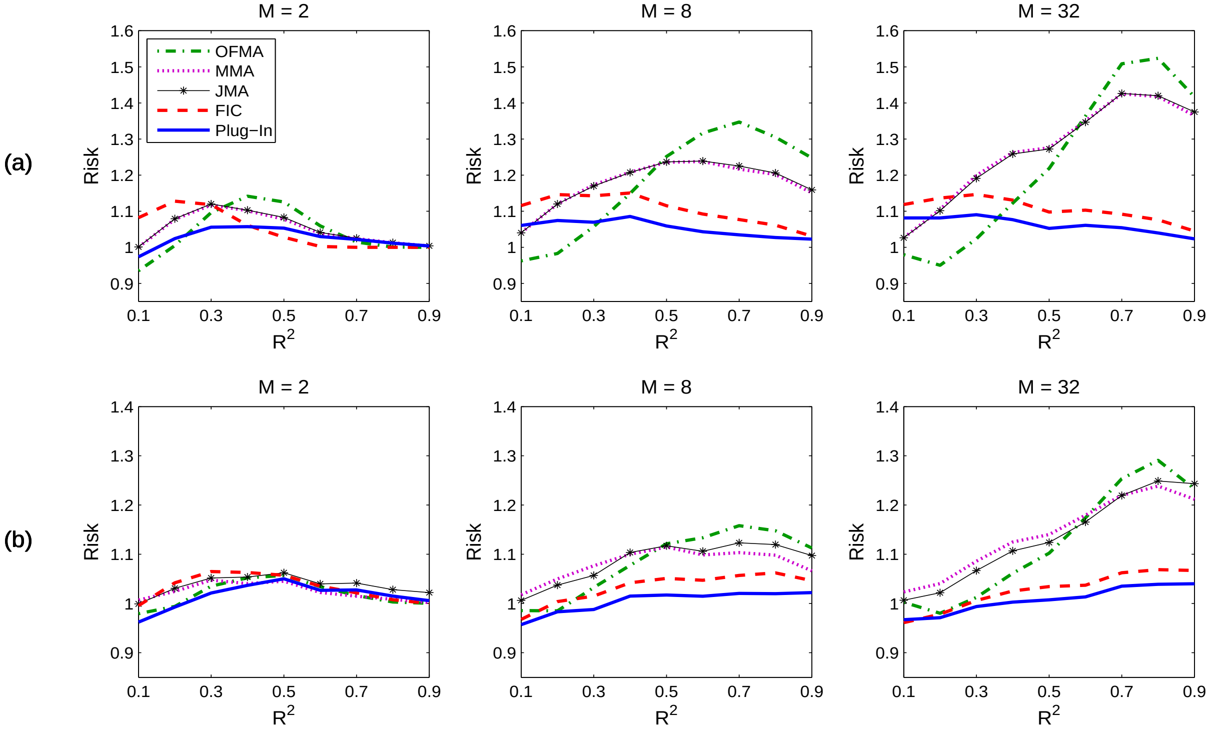

Figure 3: Normalized risk functions for averaging estimators under homoskedastic errors in row (a) and under heteroskedastic errors in row (b). The situation corresponds to $a = 12$, $\rho = 0.5$, and $n = 100$.

proposed by Liang, Zou, Wan, and Zhang (2011), and suggests selecting the weights by minimizing the trace of an unbiased estimator of the mean squared error of the averaging estimator.[15]

Our parameter of interest is $\mu = \theta_1 + \theta_2 + \theta_3$, the sum of the coefficients of the core regressors. To evaluate the finite behavior of the averaging estimators, we compute the risk based on the quadratic loss function. The risk (expected squared error) is calculated by averaging across 5000 random samples. We follow Hansen (2007) and normalize the risk by dividing by the risk of the infeasible optimal least squares estimator, i.e., the risk of the best-fitting submodel m.

[15] Liang, Zou, Wan, and Zhang (2011) consider a parametric form of the weight function. The weight function is defined as $w_m = \left(a^{k_m}(n-k_m)^b(\hat{\sigma}_m^2)\right) / \left(\sum_{\ell=1}^{M} a^{k_\ell}(n-k_\ell)^b(\hat{\sigma}_\ell^2)\right)$ where $k_m = p + q_m$ and parameters (a, b, c) are chosen by minimizing the criterion function $C_n(a,b,c) = \hat{\sigma}^2 tr(\mathbf{X}'\mathbf{X})^{-1} - \hat{\sigma}^2 tr(\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}') + \mathbf{w}'(a,b,c)\mathbf{C}_1\mathbf{w}(a,b,c) - \frac{4}{n}c\hat{\sigma}^2\mathbf{w}'(a,b,c)\mathbf{C}_2\mathbf{w}(a,b,c) + 2\hat{\sigma}^2\mathbf{w}'(a,b,c)\boldsymbol{\phi} + \frac{4}{n}c\hat{\sigma}^2\mathbf{w}'(a,b,c)diag(\mathbf{C}_2)$ where $\mathbf{w}(a,b,c) = (w_1, ..., w_M)'$, $\tilde{\mathbf{Q}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{M}_x\mathbf{Z})^{-1/2}$, $\mathbf{M}_x = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$, $\mathbf{C}_1$ is an $M \times M$ matrix with (m,ℓ) element $C_{m,\ell}^1 = \tilde{\boldsymbol{\theta}}'(\mathbf{I}_q - \tilde{\mathbf{W}}_m)\tilde{\mathbf{Q}}'\tilde{\mathbf{Q}}(\mathbf{I}_q - \tilde{\mathbf{W}}_\ell)\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}} = (\mathbf{Z}'\mathbf{M}_x\mathbf{Z})^{1/2}\hat{\boldsymbol{\gamma}}_f$, $\tilde{\mathbf{W}}_m = \mathbf{I}_q - \tilde{\mathbf{P}}_m$, $\tilde{\mathbf{P}}_m = (\mathbf{Z}'\mathbf{M}_x\mathbf{Z})^{-1/2}\boldsymbol{\Pi}_m'(\boldsymbol{\Pi}_m(\mathbf{Z}'\mathbf{M}_x\mathbf{Z})^{-1}\boldsymbol{\Pi}_m')^{-1}\boldsymbol{\Pi}_m(\mathbf{Z}'\mathbf{M}_x\mathbf{Z})^{-1/2}$, $\mathbf{C}_2$ is an $M \times M$ matrix with (m,ℓ) element $C_{m,\ell}^2 = (\hat{\sigma}_m^2)^{-1}\tilde{\boldsymbol{\theta}}'\tilde{\mathbf{W}}_m'\tilde{\mathbf{Q}}'\tilde{\mathbf{Q}}(\mathbf{I}_q - \tilde{\mathbf{W}}_\ell)\tilde{\boldsymbol{\theta}}$, $\boldsymbol{\phi} = (\phi_1, ..., \phi_M)$ with $\phi_m = tr(\tilde{\mathbf{Q}}\tilde{\mathbf{W}}_m\tilde{\mathbf{Q}}')$, and $diag(\mathbf{C}_2)$ is the diagonal of $\mathbf{C}_2$.

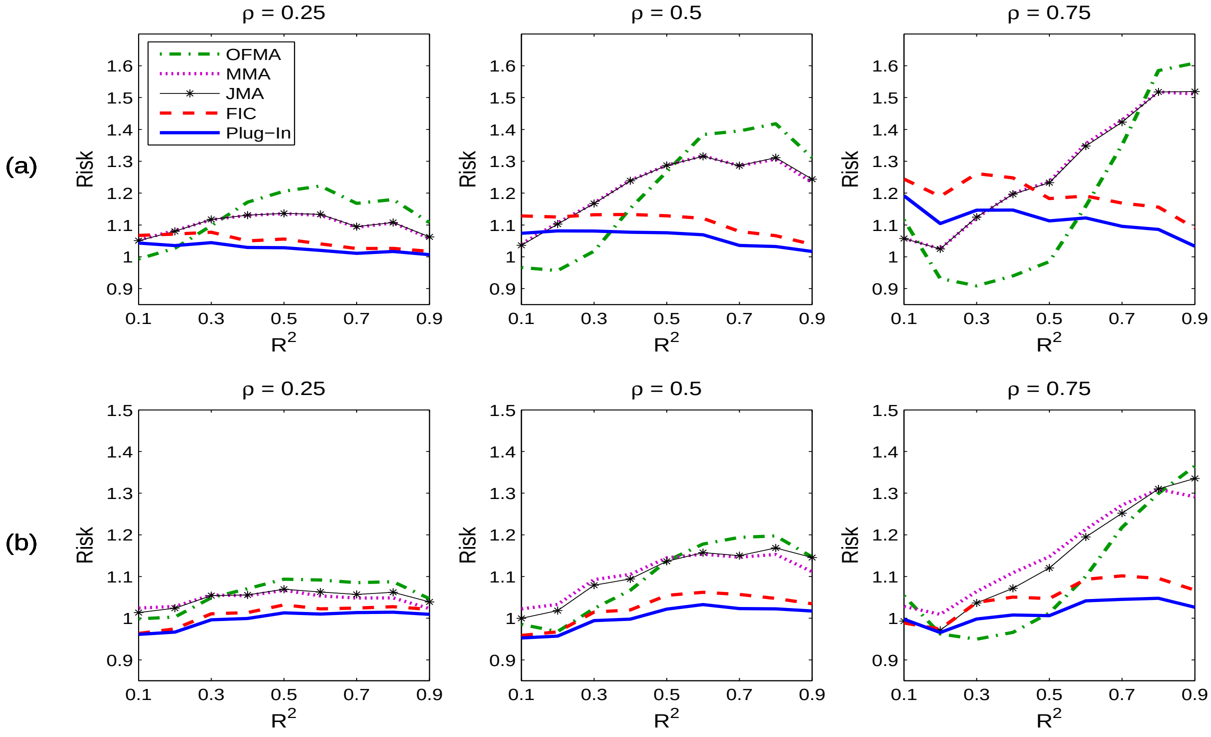

Figure 4: Normalized risk functions for averaging estimators under homoskedastic errors in row (a) and under heteroskedastic errors in row (b). The situation corresponds to $a = 12$, $M = 16$, and $n = 100$.

6.2 Simulation Results

The normalized risk functions are displayed in Figures 3-6. In each figure, the homoskedastic and heteroskedastic simulations are displayed in row (a) and (b), respectively. The main observations from the simulations are (i) MMA and JMA have similar normalizes risk in both homoskedastic and heteroskedastic setups; (ii) Plug-In achieves lower normalized risk than FIC, and both FIC and Plug-In have much lower normalized risk than MMA and JMA in most cases; (iii) OFMA performs noticeably better than other estimators when R^2 is small but performs worse than other estimators when R^2 is large under homoskedastic errors.

Figure 3 shows the effect of the number of models on the normalized risk. When we only consider two models, the restricted and nonrestricted models, all estimators have similar normalized risk in both homoskedastic and heteroskedastic simulations. The normalized risk of most estimators increases as the number of models increases, while the risk of Plug-In is close to that of the infeasible optimal least squares estimator in most ranges of the parameter space. Figure 4 shows the effect of the correlation between regressors on the normalized risk. All estimators have larger risk when ρ and R^2 are larger. JMA has lower normalized risk than MMA for larger ρ under heteroskedastic errors.

Figure 5 shows the effect of the sample size on the normalized risk. As the sample size increases,

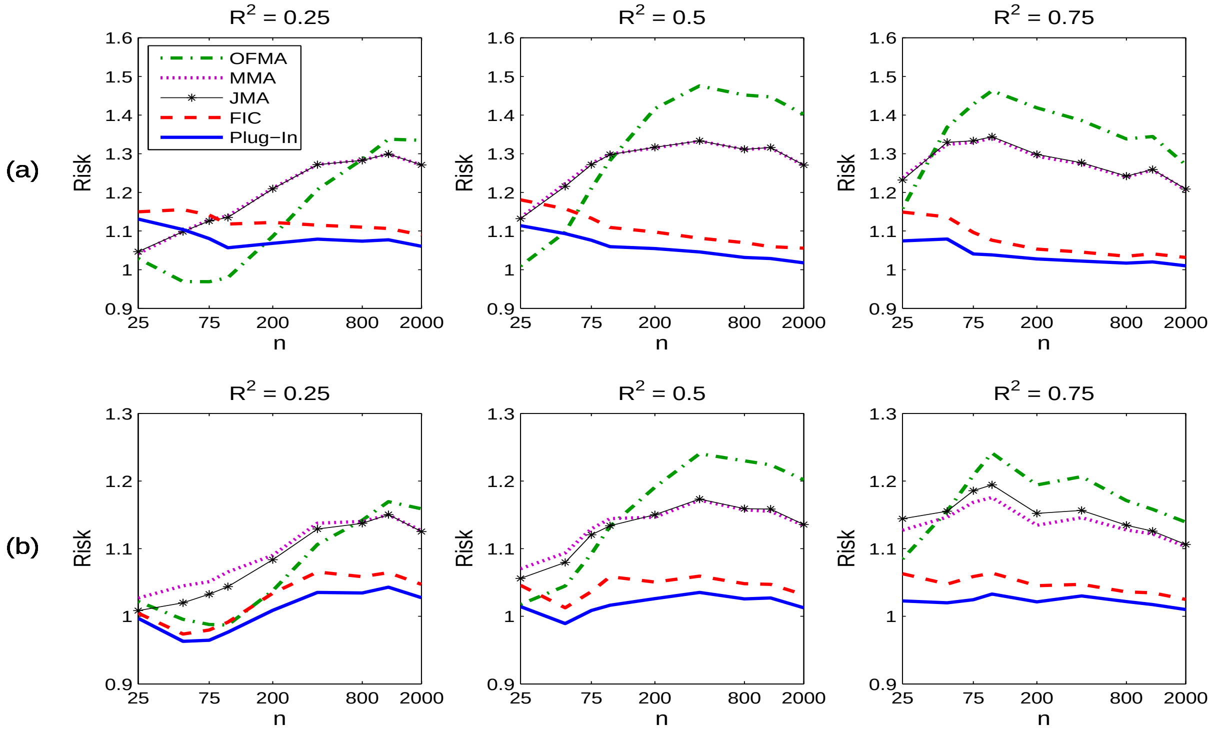

Figure 5: Normalized risk functions for averaging estimators under homoskedastic errors in row (a) and under heteroskedastic errors in row (b). The situation corresponds to $a = 12$, $M = 16$, and $\rho = 0.5$.

the normalized risk of both MMA and JMA increase. Therefore, it shows that both estimators are not asymptotically optimal in a linear regression model with a finite number of regressors. Unlike, MMA, JMA, and OFMA, the normalized risk of FIC and Plug-In are getting close to one as n increases. Figure 6 shows the effect of the importance of the auxiliary regressors on the normalized risk. Note that the parameter a measures the importance of the auxiliary regressors relative to the core regressors. The larger a implies that the auxiliary regressors have a greater influence on the model. The result shows that FIC and Plug-In are relatively unaffected by the value of a and R^2, while OFMA, MMA, and JMA have larger normalized risk when a and R^2 are larger.

6.3 Coverage Probabilities

We now examine the finite sample performance of proposed and traditional confidence intervals. The traditional confidence intervals of OFMA, MMA, JMA, FIC, and Plug-In estimators are constructed by inverting the model averaging t-statistic defined in (5.21), that is,

$$\mathrm{CI}_n = \left[\bar{\mu} - z_{1-\alpha/2} s(\bar{\mu}),\ \bar{\mu} + z_{1-\alpha/2} s(\bar{\mu})\right]$$

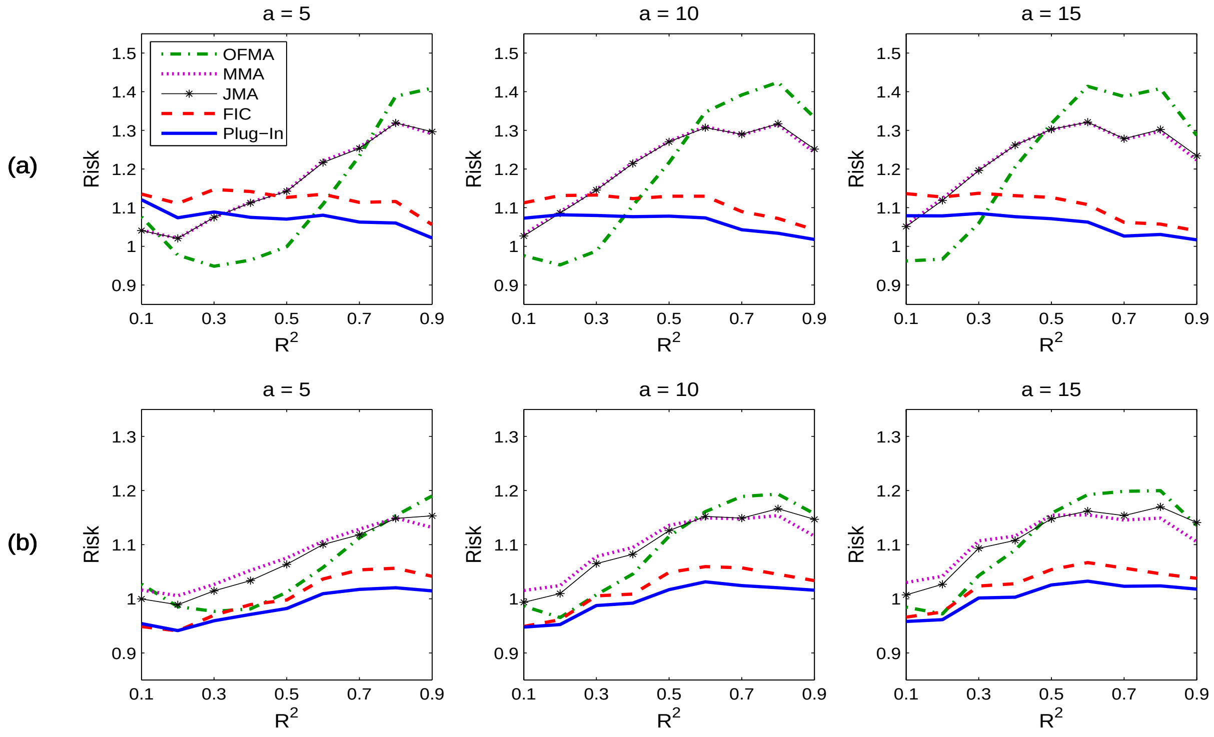

Figure 6: Normalized risk functions for averaging estimators under homoskedastic errors in row (a) and under heteroskedastic errors in row (b). The situation corresponds to $M = 16$, $\rho = 0.5$, and $n = 100$.

while the proposed valid confidence intervals (labeled Valid) are computed based on (5.23).[16] The data generating process is based on (6.1) and (6.2). The number of simulations is 5000.

The finite-sample coverage probabilities of the 90% confidence intervals for homoskedastic errors and heteroskedastic errors are reported in Tables 1 and 2, respectively. Overall, the coverage probabilities of the valid confidence intervals are generally close to the nominal values, while the traditional confidence intervals are much lower than the level 90%. When ρ gets bigger, the coverage probabilities of the traditional confidence intervals are substantially smaller than the nominal values. Among these averaging estimators, the coverage probabilities of Plug-In are closer to the nominal level than other estimators. It is also worth mentioning that the coverage probabilities of OFMA are close to the level 90% when R^2 is small but are lower than other estimators when both R^2 and ρ are large.

[16] Since the coverage probabilities of the valid confidence intervals of OFMA, MMA, JMA, FIC, and Plug-In are quite similar, we only report the results of the valid confidence intervals of the plug-in averaging estimator for space considerations.

Table 1: Coverage Probabilities of 90% Confidence Intervals under homoskedastic errors

n	R^2	ρ	OFMA	MMA	JMA	FIC	Plug-In	Valid
100	0.25	0.00	0.867	0.866	0.868	0.861	0.863	0.874
		0.25	0.852	0.842	0.842	0.853	0.853	0.875
		0.50	0.861	0.793	0.795	0.816	0.826	0.888
		0.75	0.883	0.723	0.724	0.702	0.730	0.876
	0.50	0.00	0.863	0.868	0.867	0.864	0.863	0.869
		0.25	0.824	0.838	0.840	0.856	0.857	0.877
		0.50	0.774	0.773	0.773	0.818	0.826	0.863
		0.75	0.807	0.698	0.699	0.777	0.777	0.877
	0.75	0.00	0.865	0.871	0.868	0.867	0.866	0.873
		0.25	0.836	0.848	0.848	0.863	0.867	0.877
		0.50	0.761	0.787	0.781	0.849	0.853	0.877
		0.75	0.707	0.719	0.715	0.820	0.825	0.875
500	0.25	0.00	0.899	0.898	0.899	0.900	0.900	0.901
		0.25	0.836	0.853	0.851	0.876	0.879	0.892
		0.50	0.804	0.793	0.793	0.844	0.848	0.887
		0.75	0.869	0.743	0.743	0.801	0.793	0.895
	0.50	0.00	0.901	0.901	0.901	0.901	0.900	0.901
		0.25	0.854	0.872	0.870	0.892	0.895	0.903
		0.50	0.788	0.814	0.814	0.873	0.876	0.892
		0.75	0.736	0.731	0.731	0.844	0.849	0.902
	0.75	0.00	0.896	0.897	0.897	0.894	0.896	0.898
		0.25	0.872	0.879	0.879	0.892	0.892	0.895
		0.50	0.815	0.835	0.835	0.884	0.884	0.897
		0.75	0.731	0.750	0.749	0.865	0.868	0.894

7 An Empirical Example

In this section, we apply the model averaging methods to cross-country growth regressions. The challenge of empirical research on economic growth is that one does not know exactly what explanatory variables should be included in the true model. Many studies attempt to identify the variables explaining the differences in growth rates across countries by regressing the average growth rate of GDP per capita on a large set of potentially relevant variables, see Durlauf, Johnson, and Temple (2005) for a literature review. Due to limited number of the observations and a large amount of the candidate variables, the empirical growth literature has been heavily criticized for its kitchen-sink approach.

In order to take into account the model uncertainty, Bayesian model averaging techniques have been applied to empirical growth, including Fernandez, Ley, and Steel (2001), Sala-i Martin, Doppelhofer, and Miller (2004), Durlauf, Kourtellos, and Tan (2008), and Magnus, Powell, and Prufer (2010). We apply frequentist model averaging approaches as an alternative to Bayesian model averaging techniques to economic growth. We estimate the following cross-country growth

Table 2: Coverage Probabilities of 90% Confidence Intervals under heteroskedastic errors

n	R^2	ρ	OFMA	MMA	JMA	FIC	Plug-In	Valid
100	0.25	0.00	0.845	0.844	0.845	0.836	0.838	0.847
		0.25	0.822	0.824	0.824	0.822	0.825	0.852
		0.50	0.832	0.801	0.804	0.805	0.807	0.861
		0.75	0.863	0.761	0.769	0.756	0.757	0.866
	0.50	0.00	0.846	0.846	0.847	0.839	0.838	0.847
		0.25	0.825	0.828	0.829	0.828	0.830	0.857
		0.50	0.784	0.780	0.782	0.795	0.798	0.848
		0.75	0.800	0.732	0.747	0.764	0.769	0.860
	0.75	0.00	0.846	0.846	0.844	0.837	0.837	0.847
		0.25	0.822	0.828	0.827	0.832	0.832	0.850
		0.50	0.785	0.799	0.796	0.816	0.825	0.853
		0.75	0.728	0.729	0.725	0.784	0.786	0.857
500	0.25	0.00	0.895	0.895	0.895	0.896	0.894	0.894
		0.25	0.860	0.871	0.871	0.869	0.869	0.885
		0.50	0.843	0.842	0.841	0.852	0.854	0.883
		0.75	0.867	0.795	0.797	0.815	0.820	0.892
	0.50	0.00	0.895	0.895	0.895	0.892	0.894	0.896
		0.25	0.870	0.881	0.881	0.881	0.883	0.895
		0.50	0.819	0.837	0.836	0.855	0.859	0.883
		0.75	0.789	0.782	0.782	0.846	0.850	0.897
	0.75	0.00	0.890	0.890	0.890	0.888	0.888	0.890
		0.25	0.875	0.874	0.874	0.882	0.881	0.888
		0.50	0.840	0.853	0.853	0.867	0.871	0.881
		0.75	0.779	0.794	0.791	0.859	0.863	0.894

regression

$$g_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + e_i \tag{7.1}$$

where g_i is average growth rate of GDP per capita between 1960 and 1996, $\mathbf{x}_i$ are the Solow variables from the neoclassical growth theory, and $\mathbf{z}_i$ are fundamental growth determinants such as geography, institutions, religion, and ethnic fractionalization from the new fundamental growth theory. Here, $\mathbf{x}_i$ are core regressors, which appear in every submodel, while $\mathbf{z}_i$ are the auxiliary regressors, which serve as controls of the neoclassical growth theory and may or may not be included in the submodels.

We follow Magnus, Powell, and Prufer (2010) and consider two model specifications to compare the neoclassical growth theory with the fundamental new growth theory. Model Setup A includes six core regressors and four auxiliary regressors. The six core regressors are the constant term (CONSTANT), the log of GDP per capita in 1960 (GDP60), the 1960-1985 equipment investment share of GDP (EQUIPINV), the primary school enrollment rate in 1960 (SCHOOL60), the life expectancy at age zero in 1960 (LIFE60), and the population growth rate between 1960 and 1990 (DPOP). The four auxiliary regressors are a rule of law index (LAW), a country's fraction of tropical area (TROPICS), an average index of ethnolinguistic fragmentation in a country (AVELF), and the fraction of Confucian population (CONFUC), see Magnus, Powell, and Prufer (2010) for a detailed description of the data. Model Setup B contains two core regressors, the constant term and GDP60, and all other variables in Model Setup A are auxiliary regressors.[17] The parameter of interest is the convergence term of the Solow growth model, that is, the coefficient of the log GDP per capita in 1960. The total number of observations is 74. We consider all possible submodels, that is, we have 16 submodels in Model Setup A and 128 submodels in Model Setup B.

We consider seven estimators: (1) the least squares estimator for the full model (labeled Full), (2) the averaging estimator with equal weights (labeled Equal), (3) optimal frequentist model averaging estimator (labeled OFMA), (4) Mallows model averaging estimator (labeled MMA), (5) jackknife model averaging estimator (labeled JMA), (6) focused information criterion model selection (labeled FIC), and (7) plug-in averaging estimator (labeled Plug-In). The standard errors of data-dependent model averaging estimators are calculated by the equation (5.20).

The estimation results for Model Setup A and B are given in Tables 3 and 4, respectively. We also report the estimation results for the weighted-average least squares (WALS) estimator proposed by Magnus, Powell, and Prufer (2010) for comparison. The WALS estimator is a Bayesian model averaging technique that uses a Laplace distribution instead of the normal prior as the parameter prior. The results in Tables 3 and 4 show that all coefficients have the same signs across different estimation methods except the estimated coefficient of DPOP by FIC in Model Setup A.

In Model Setup A, the coefficient estimate and standard error of GDP60 are similar across different estimators while OFMA has a relative lower coefficient estimate of GDP60. In Model Setup B, the plug-in averaging estimate of GDP60 is quite close to the least squares estimate from the full model and is higher in absolute value than other estimates. As we expected, the 90% confidence interval of the plug-in averaging estimate for GDP60 calculated by the proposed method $(-0.0213, -0.0097)$ is wider than the traditional confidence interval $(-0.0193, -0.0115)$. The important finding from our results is that the plug-in averaging estimator has the smaller standard error of GDP60 as compared to other estimators.

It is also instructive to contrast the results of Plug-In and WALS estimators. In Model Setup A, the estimation results are similar between Plug-In and WALS. In Model Setup B, the estimated coefficient of GDP60 is slightly higher in absolute value for Plug-In than for WALS, while the estimated standard error of GDP60 is smaller for Plug-In than for WALS. Therefore, the convergence speed of the growth model implied by our result is higher than that found by Magnus, Powell, and Prufer (2010). Comparing the results between Model Setup A and Model Setup B, we find that the plug-in averaging estimator chooses different fundamental growth determinants in different model specifications. Therefore, our results support the findings of Durlauf, Kourtellos, and Tan (2008) and Magnus, Powell, and Prufer (2010) that the fundamental variables are not robustly correlated with growth.

[17] Model Setup B is slightly different than that in Magnus, Powell, and Prufer (2010). They treat the constant term as the only core regressor. Since GDP60 is the parameter of interest, as suggested by one referee, we also include GDP60 as the core regressor in Model Setup B.

Table 3: Coefficient estimates and standard errors, Model Setup A

	Full	Equal	OFMA	MMA	JMA	FIC	Plug-In	WALS
CONSTANT	0.0609	0.0603	0.0489	0.0558	0.0559	0.0587	0.0641	0.0594
	(0.0193)	(0.0192)	(0.0203)	(0.0199)	(0.0201)	(0.0202)	(0.0182)	(0.0221)
GDP60	-0.0155	-0.0157	-0.0138	-0.0150	-0.0156	-0.0160	-0.0156	-0.0156
	(0.0030)	(0.0028)	(0.0030)	(0.0029)	(0.0029)	(0.0028)	(0.0027)	(0.0033)
EQUIPINV	0.1366	0.1835	0.1623	0.1526	0.1511	0.2405	0.2263	0.1555
	(0.0400)	(0.0361)	(0.0369)	(0.0382)	(0.0390)	(0.0353)	(0.0349)	(0.0551)
SCHOOL60	0.0170	0.0173	0.0161	0.0173	0.0181	0.0184	0.0137	0.0175
	(0.0085)	(0.0081)	(0.0081)	(0.0081)	(0.0081)	(0.0079)	(0.0085)	(0.0097)
LIFE60	0.0008	0.0009	0.0008	0.0008	0.0009	0.0010	0.0010	0.0009
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)
DPOP	0.3466	0.1736	0.1707	0.2596	0.2465	-0.0341	0.0055	0.2651
	(0.1911)	(0.1706)	(0.1722)	(0.1788)	(0.1760)	(0.1635)	(0.1718)	(0.2487)
LAW	0.0174	0.0094	0.0113	0.0144	0.0166			0.0147
	(0.0058)	(0.0028)	(0.0039)	(0.0047)	(0.0052)			(0.0065)
TROPICS	-0.0075	-0.0040	-0.0036	-0.0057	-0.0043			-0.0055
	(0.0036)	(0.0018)	(0.0016)	(0.0025)	(0.0018)			(0.0037)
AVELF	-0.0077	-0.0048	-0.0019	-0.0039	-0.0026		-0.0104	-0.0053
	(0.0066)	(0.0033)	(0.0015)	(0.0025)	(0.0016)		(0.0065)	(0.0048)
CONFUC	0.0562	0.0317	0.0622	0.0521	0.0430		0.0251	0.0443
	(0.0129)	(0.0062)	(0.0124)	(0.0108)	(0.0088)		(0.0045)	(0.0163)

Note: Standard errors are reported in parentheses. The column labeled WALS displays the weighted-average least squares estimates of Magnus, Powell, and Prufer (2010, Table 2).

Table 4: Coefficient estimates and standard errors, Model Setup B

	Full	Equal	OFMA	MMA	JMA	FIC	Plug-In	WALS
CONSTANT	0.0609	0.0575	0.0606	0.0554	0.0533	0.0856	0.0801	0.0691
	(0.0193)	(0.0154)	(0.0177)	(0.0149)	(0.0149)	(0.0139)	(0.0133)	(0.0212)
GDP60	-0.0155	-0.0120	-0.0149	-0.0134	-0.0139	-0.0150	-0.0154	-0.0148
	(0.0030)	(0.0023)	(0.0029)	(0.0025)	(0.0025)	(0.0022)	(0.0020)	(0.0031)
EQUIPINV	0.1366	0.1080	0.1415	0.1271	0.1315		0.1389	0.1246
	(0.0400)	(0.0171)	(0.0375)	(0.0190)	(0.0212)		(0.0144)	(0.0470)
SCHOOL60	0.0170	0.0131	0.0153	0.0155	0.0144	0.0406		0.0153
	(0.0085)	(0.0035)	(0.0067)	(0.0034)	(0.0027)	(0.0069)		(0.0082)
LIFE60	0.0008	0.0006	0.0008	0.0007	0.0008		0.0008	0.0007
	(0.0003)	(0.0001)	(0.0002)	(0.0001)	(0.0001)		(0.0001)	(0.0003)
DPOP	0.3466	0.0094	0.2046	0.1486	0.1764			0.1038
	(0.1911)	(0.0788)	(0.1207)	(0.0463)	(0.0692)			(0.2171)
LAW	0.0174	0.0112	0.0155	0.0131	0.0152	0.0348	0.0165	0.0149
	(0.0058)	(0.0024)	(0.0052)	(0.0026)	(0.0033)	(0.0039)	(0.0031)	(0.0058)
TROPICS	-0.0075	-0.0042	-0.0058	-0.0053	-0.0041		-0.0026	-0.0065
	(0.0036)	(0.0017)	(0.0029)	(0.0020)	(0.0016)		(0.0020)	(0.0035)
AVELF	-0.0077	-0.0056	-0.0057	-0.0045	-0.0033	-0.0137	-0.0152	-0.0071
	(0.0066)	(0.0031)	(0.0046)	(0.0023)	(0.0017)	(0.0063)	(0.0061)	(0.0052)
CONFUC	0.0562	0.0374	0.0594	0.0524	0.0443			0.0471
	(0.0129)	(0.0060)	(0.0126)	(0.0092)	(0.0081)			(0.0140)

Note: Standard errors are reported in parentheses.

Table 5: Weights placed on each submodel, Model Setup A

Model	MMA	JMA	FIC	Plug-In
1	0.000	0.000	1.000	0.000
4	0.000	0.070	0.000	0.000
5	0.000	0.000	0.000	0.624
6	0.069	0.000	0.000	0.000
8	0.076	0.243	0.000	0.000
9	0.000	0.071	0.000	0.000
10	0.000	0.424	0.000	0.000
11	0.173	0.000	0.000	0.000
12	0.450	0.192	0.000	0.000
13	0.000	0.000	0.000	0.376
14	0.232	0.000	0.000	0.000

Table 6: Weights placed on each submodel, Model Setup B

Model	MMA	JMA	FIC	Plug-In
36	0.000	0.088	0.000	0.000
66	0.000	0.000	0.000	0.309
82	0.000	0.000	0.000	0.122
83	0.000	0.000	1.000	0.000
84	0.000	0.262	0.000	0.000
117	0.000	0.000	0.000	0.570
125	0.241	0.000	0.000	0.000
134	0.116	0.210	0.000	0.000
148	0.149	0.054	0.000	0.000
164	0.316	0.000	0.000	0.000
179	0.032	0.000	0.000	0.000
189	0.017	0.386	0.000	0.000
213	0.128	0.000	0.000	0.000

Table 7: Regressor set of the submodel, Model Setup A

Model	Regressor Set
1	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP
4	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, LAW, TROPICS
5	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, AVELF
6	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, LAW, AVELF
8	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, LAW, TROPICS, AVELF
9	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, CONFUC
10	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, LAW, CONFUC
11	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, TROPICS, CONFUC
12	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, LAW, TROPICS, CONFUC
13	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, AVELF, CONFUC
14	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LIFE60, DPOP, LAW, AVELF, CONFUC

Table 8: Regressor set of the submodel, Model Setup B

Model	Regressor Set
36	CONSTANT, GDP60, EQUIPINV, SCHOOL60, TROPICS
66	CONSTANT, GDP60, EQUIPINV, AVELF
82	CONSTANT, GDP60, EQUIPINV, LAW, AVELF
83	CONSTANT, GDP60, SCHOOL60, LAW, AVELF
84	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LAW, AVELF
117	CONSTANT, GDP60, LIFE60, LAW, TROPICS, AVELF
125	CONSTANT, GDP60, LIFE60, DPOP, LAW, TROPICS, AVELF
134	CONSTANT, GDP60, EQUIPINV, LIFE60, CONFUC
148	CONSTANT, GDP60, EQUIPINV, SCHOOL60, LAW, CONFUC
164	CONSTANT, GDP60, EQUIPINV, SCHOOL60, TROPICS, CONFUC
179	CONSTANT, GDP60, SCHOOL60, LAW, TROPICS, CONFUC
189	CONSTANT, GDP60, LIFE60, DPOP, LAW, TROPICS, CONFUC
213	CONSTANT, GDP60, LIFE60, LAW, AVELF, CONFUC

Tables 5 and 6 report the weights placed on each submodel, and Tables 7 and 8 report the regressor sets for each submodel. We only report the results of MMA, JMA, FIC, and Plug-In estimators, since OFMA weights are spread out across all submodels. One interesting observation is that the submodels chosen by Plug-In are completely different from those chosen by MMA and JMA in both Model Setup A and B. The submodels chosen by MMA and JMA cover the entire regressor set, while Plug-In excludes the regressors LAW and TROPICS in Model Setup A and the regressors SCHOOL60, DPOP, and CONFUC in Model Setup B.

8 Conclusion

In this paper we study the limiting distributions of least squares averaging estimators for heteroskedastic regressions. We show that the asymptotic distributions of averaging estimators with data-dependent weights are nonstandard in the local asymptotic framework. To address the inference after model selection and averaging, we provide a formula to calculate the standard error and a simple procedure to construct valid confidence intervals. Simulation results show that the coverage probability of proposed confidence intervals achieves the nominal level while the coverage probability of traditional confidence intervals is generally too low.

While this paper has focused on the least squares estimator, the proposed averaging method can be easily extended to the generalized least squares procedure.[18] It would be greatly desirable to extend the methodology to average across different candidate models and different procedures. Yang (2000), Yang (2001), and Yuan and Yang (2005) propose an adaptive regression to combine multiple regression models or procedures under the normality assumption. However, it is still unclear how to extend the analysis to the general setup. Another possible extension would be to investigate the asymptotic risk of least squares averaging estimators and to study the minimax efficient bound. Recently, Hansen (2013b) applies Stein's Lemma to examine the asymptotic risk of averaging estimators in a nested model framework. It would be an important research topic to extend the analysis to a more general model setting.

[18] Let $\mathbf{V} = diag(\sigma_1^2, ..., \sigma_n^2)$ denote the $n \times n$ positive definite variance-covariance matrix of the error terms. Then, the generalized least squares (GLS) estimator for the submodel m is $\hat{\boldsymbol{\theta}}_m = (\mathbf{H}_m' \mathbf{V}^{-1} \mathbf{H}_m)^{-1} \mathbf{H}_m' \mathbf{V}^{-1} \mathbf{y}$, and the asymptotic distribution of the GLS estimator is $\sqrt{n}\big(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m\big) \xrightarrow{d} \mathbf{A}_m \boldsymbol{\delta} + \mathbf{B}_m \mathbf{R} \sim \mathbf{N}\big(\mathbf{A}_m \boldsymbol{\delta},\ (\mathbf{S}_m' \boldsymbol{\Omega} \mathbf{S}_m)^{-1}\big)$, where $\boldsymbol{\Omega} = \mathrm{E}\left(\sigma_i^{-2} \mathbf{h}_i \mathbf{h}_i'\right)$, $\mathbf{A}_m = (\mathbf{S}_m' \boldsymbol{\Omega} \mathbf{S}_m)^{-1} \mathbf{S}_m' \boldsymbol{\Omega} \mathbf{S}_0 (\mathbf{I}_q - \boldsymbol{\Pi}_m' \boldsymbol{\Pi}_m)$, and $\mathbf{B}_m = (\mathbf{S}_m' \boldsymbol{\Omega} \mathbf{S}_m)^{-1} \mathbf{S}_m'$. Similarly, the results in Theorems 1-3 still hold except the definitions of $\mathbf{C}_m$ and $\mathbf{P}_m$ are replaced by $\mathbf{C}_m = (\mathbf{P}_m \boldsymbol{\Omega} - \mathbf{I}_{p+q}) \mathbf{S}_0$ and $\mathbf{P}_m = \mathbf{S}_m (\mathbf{S}_m' \boldsymbol{\Omega} \mathbf{S}_m)^{-1} \mathbf{S}_m'$. Thus, we can construct the plug-in averaging estimator in the same way as (4.8).

Appendix

A Proofs

Proof of Lemma 1: We first show the asymptotic distribution of the least squares estimator for the full model. By Assumption 2 and the application of the continuous mapping theorem, it follows that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_f-\boldsymbol{\theta}\right)=\left(\frac{1}{n}\mathbf{H}'\mathbf{H}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{H}'\mathbf{e}\right)\xrightarrow{d}\mathbf{Q}^{-1}\mathbf{R}\sim\mathbf{N}(\mathbf{0},\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}).$$

We next show the asymptotic distribution of the least squares estimator for each submodel. Note that $\mathbf{H}_m=(\mathbf{X},\mathbf{Z}\boldsymbol{\Pi}_m')=\mathbf{H}\mathbf{S}_m$ and $\mathbf{Z}=\mathbf{H}\mathbf{S}_0$. By some algebra, it follows that

$$\begin{aligned}\hat{\boldsymbol{\theta}}_m&=(\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{y}\\&=\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\left(\mathbf{H}_m'\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\boldsymbol{\gamma}+\mathbf{Z}(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m)\boldsymbol{\gamma}+\mathbf{e}\right)\right)\\&=\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{H}_m'\mathbf{H}_m\boldsymbol{\theta}_m+\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{H}_m'\mathbf{Z}\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}+\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{H}_m'\mathbf{e}\\&=\boldsymbol{\theta}_m+\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{S}_m'\mathbf{H}'\mathbf{H}\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}+\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{S}_m'\mathbf{H}'\mathbf{e}.\end{aligned}$$

Therefore, by Assumptions 1-2 and the application of the continuous mapping theorem, we have

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m\right)&=\left(\frac{1}{n}\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\left(\frac{1}{n}\mathbf{S}_m'\mathbf{H}'\mathbf{H}\mathbf{S}_0\right)\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\sqrt{n}\boldsymbol{\gamma}+\left(\frac{1}{n}\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{S}_m'\left(\frac{1}{\sqrt{n}}\mathbf{H}'\mathbf{e}\right)\\&\xrightarrow{d}\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta}+\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{R}\\&=\mathbf{A}_m\boldsymbol{\delta}+\mathbf{B}_m\mathbf{R}\sim\mathbf{N}\left(\mathbf{A}_m\boldsymbol{\delta},\ \mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\mathbf{Q}_m^{-1}\right)\end{aligned}$$

where $\mathbf{A}_m=\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)$ and $\mathbf{B}_m=\mathbf{Q}_m^{-1}\mathbf{S}_m'$. This completes the proof. ■

Proof of Theorem 1: Define $\boldsymbol{\gamma}_{m^c}=\{\boldsymbol{\gamma}:\gamma_j\notin\boldsymbol{\gamma}_m,\ for\ j=1,...,q\}$. That is, $\boldsymbol{\gamma}_{m^c}$ is the set of parameters γ_j which are not included in submodel m. Hence, we can write $\mu(\boldsymbol{\theta})$ as $\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\boldsymbol{\gamma}_{m^c})$. Also, $\mu(\boldsymbol{\theta}_m)=\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\mathbf{0})$.

Note that $\boldsymbol{\gamma}=O(n^{-1/2})$ by Assumption 1. Then by a standard Taylor series expansion of $\mu(\boldsymbol{\theta})$ about $\boldsymbol{\gamma}_{m^c}=\mathbf{0}$, it follows that

$$\begin{aligned}\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\boldsymbol{\gamma}_{m^c})&=\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\mathbf{0})+\mathbf{D}_{\boldsymbol{\gamma}_{m^c}}'\boldsymbol{\gamma}_{m^c}+O(n^{-1})\\&=\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\mathbf{0})+\mathbf{D}_{\boldsymbol{\gamma}}'\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}+O(n^{-1}).\end{aligned}$$

That is, $\mu(\boldsymbol{\theta})-\mu(\boldsymbol{\theta}_m)=\mathbf{D}_{\boldsymbol{\gamma}}'\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}+O(n^{-1})$.

Let $\mathbf{P}_m=\mathbf{S}_m\left(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'$. By Assumptions 1-2 and the application of the delta method, we have

$$\begin{aligned}\sqrt{n}\left(\mu(\hat{\boldsymbol{\theta}}_m)-\mu(\boldsymbol{\theta})\right) &= \sqrt{n}\left(\mu(\hat{\boldsymbol{\theta}}_m)-\mu(\boldsymbol{\theta}_m)\right)-\sqrt{n}\Big(\mu(\boldsymbol{\theta})-\mu(\boldsymbol{\theta}_m)\Big)\\ &\xrightarrow{d} \mathbf{D}_{\boldsymbol{\theta}_m}'\left(\mathbf{A}_m\boldsymbol{\delta}+\mathbf{B}_m\mathbf{R}\right)-\mathbf{D}_{\gamma}'\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta}\\ &= \mathbf{D}_{\boldsymbol{\theta}_m}'\mathbf{A}_m\boldsymbol{\delta}-\mathbf{D}_{\gamma}'\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}_m}'\mathbf{B}_m\mathbf{R}\\ &= \left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{S}_m\left(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0-\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{S}_0\right)\left(\mathbf{I}_q-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{R}\\ &= \left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{S}_m\left(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0-\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{S}_0\right)\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{S}_m\left(\mathbf{S}_m'\mathbf{P}_m\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{R}\\ &= \mathbf{D}_{\boldsymbol{\theta}}'\left(\mathbf{P}_m\mathbf{Q}-\mathbf{I}_{p+q}\right)\mathbf{S}_0\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathbf{R}\\ &\equiv \Lambda_m\sim\mathbf{N}\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_m\boldsymbol{\delta},\ \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_m\mathbf{D}_{\boldsymbol{\theta}}\right),\end{aligned}$$

where the fifth equality holds by the fact that $\mathbf{S}_0\boldsymbol{\Pi}_m'=\mathbf{S}_m\left(\mathbf{0}_{p\times q_m}',\mathbf{I}_{q_m}\right)'$.
This completes the proof. ■

Proof of Theorem 2: From Theorem 1, there is joint convergence in distribution of all $\sqrt{n}\left(\mu(\hat{\boldsymbol{\theta}}_m)-\mu(\boldsymbol{\theta})\right)$ to Λ_m since all of Λ_m can be expressed in terms of $\mathbf{R}$. Since the weights are non-random, it follows that

$$\sqrt{n}\left(\bar{\mu}(\mathbf{w})-\mu\right)=\sum_{m=1}^{M}w_m\sqrt{n}\left(\hat{\mu}_m-\mu\right)\xrightarrow{d}\sum_{m=1}^{M}w_m\Lambda_m\equiv\Lambda.$$

Therefore, the asymptotic distribution of the averaging estimator is a weighted average of the normal distributions, which is also a normal distribution.

By Theorem 1 and standard algebra, we can show the mean of Λ as

$$\mathrm{E}\left(\sum_{m=1}^{M}w_m\Lambda_m\right)=\sum_{m=1}^{M}w_m\mathrm{E}\left(\Lambda_m\right)=\sum_{m=1}^{M}w_m\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_m\boldsymbol{\delta}=\mathbf{D}_{\boldsymbol{\theta}}'\sum_{m=1}^{M}w_m\mathbf{C}_m\boldsymbol{\delta}=\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_{\mathbf{w}}\boldsymbol{\delta}$$

where $\mathbf{C}_{\mathbf{w}}=\sum_{m=1}^{M}w_m\mathbf{C}_m$.

Next we show the variance of Λ. For any two submodels, we have

$$\begin{aligned}Cov(\Lambda_m,\Lambda_\ell) &= \mathrm{E}\left[\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_m\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathbf{R}-\mathrm{E}\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_m\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathbf{R}\right)\right)\right.\\ &\qquad\left.\times\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_\ell\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_\ell\mathbf{R}-\mathrm{E}\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_\ell\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_\ell\mathbf{R}\right)\right)\right]\\ &= \mathrm{E}\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathbf{R}\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_\ell\mathbf{R}\right)=\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathrm{E}\left(\mathbf{R}\mathbf{R}'\right)\mathbf{P}_\ell'\mathbf{D}_{\boldsymbol{\theta}}=\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell'\mathbf{D}_{\boldsymbol{\theta}}\end{aligned}$$

where the second equality holds by the fact that $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{C}_m$, $\mathbf{P}_m$, and $\boldsymbol{\delta}$ are constant vectors and $\mathbf{R}\sim\mathbf{N}(\mathbf{0},\boldsymbol{\Omega})$. Therefore, variance of Λ is

$$\begin{aligned}var\left(\sum_{m=1}^{M}w_m\Lambda_m\right) &= \sum_{m=1}^{M}w_m^2Var(\Lambda_m)+2\sum\sum_{m\neq\ell}w_mw_\ell Cov(\Lambda_m,\Lambda_p)\\ &= \sum_{m=1}^{M}w_m^2\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_m'\mathbf{D}_{\boldsymbol{\theta}}+2\sum\sum_{m\neq\ell}w_mw_\ell\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell'\mathbf{D}_{\boldsymbol{\theta}}\equiv V.\end{aligned}$$

This completes the proof. ■

Proof of Theorem 3: We first show the limiting distribution of $\hat{\Psi}_{m,\ell}$. By Lemma 1, we have $\hat{\boldsymbol{\theta}}_f \xrightarrow{p} \boldsymbol{\theta}$, which implies that $\hat{\mathbf{D}}_{\boldsymbol{\theta}} \xrightarrow{p} \mathbf{D}_{\boldsymbol{\theta}}$. Since $\hat{\mathbf{D}}_{\boldsymbol{\theta}}$, $\hat{\mathbf{Q}}$, and $\hat{\boldsymbol{\Omega}}$ are consistent estimators for $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{Q}$, and $\boldsymbol{\Omega}$, we have $\hat{\mathbf{D}}_{\boldsymbol{\theta}}'\hat{\mathbf{P}}_m\hat{\boldsymbol{\Omega}}\hat{\mathbf{P}}_\ell\hat{\mathbf{D}}_{\boldsymbol{\theta}} \xrightarrow{p} \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell\mathbf{D}_{\boldsymbol{\theta}}$ by the continuous mapping theorem. Recall that $\hat{\boldsymbol{\delta}} \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{R}$. Then by the application of Slutsky's theorem, we have

$$\hat{\Psi}_{m,\ell} = \hat{\mathbf{D}}_{\boldsymbol{\theta}}'\left(\hat{\mathbf{C}}_m\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}'\hat{\mathbf{C}}_\ell' + \hat{\mathbf{P}}_m\hat{\boldsymbol{\Omega}}\hat{\mathbf{P}}_\ell\right)\hat{\mathbf{D}}_{\boldsymbol{\theta}} \xrightarrow{d} \mathbf{D}_{\boldsymbol{\theta}}'\left(\mathbf{C}_m\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}'\mathbf{C}_\ell' + \mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell\right)\mathbf{D}_{\boldsymbol{\theta}} = \Psi_{m,\ell}^*.$$

Since all of $\Psi_{m,\ell}^*$ can be expressed in terms of the normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\hat{\Psi}_{m,\ell}$ to $\Psi_{m,\ell}^*$. Hence, it follows that $\mathbf{w}'\hat{\boldsymbol{\Psi}}\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\Psi}^*\mathbf{w}$.

We next show the limiting distribution of $\hat{\mathbf{w}}$. Note that $\mathbf{w}'\boldsymbol{\Psi}^*\mathbf{w}$ is a convex minimization problem since $\mathbf{w}'\boldsymbol{\Psi}^*\mathbf{w}$ is quadratic and $\boldsymbol{\Psi}^*$ is positive definite. Hence, the limiting process $\mathbf{w}'\boldsymbol{\Psi}^*\mathbf{w}$ is continuous in $\mathbf{w}$ and has a unique minimum. Also note that $\hat{\mathbf{w}} = O_p(1)$ by the fact that $\mathcal{H}_n$ is convex. Therefore, by Theorem 3.2.2 of Van der Vaart and Wellner (1996) or Theorem 2.7 of Kim and Pollard (1990), the minimizer $\hat{\mathbf{w}}$ converges in distribution to the minimizer of $\mathbf{w}'\boldsymbol{\Psi}^*\mathbf{w}$, which is $\mathbf{w}^*$.

Finally, we show the asymptotic distribution of the plug-in averaging estimator. Since both Λ_m and w_m^* can be expressed in terms of the same normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\hat{\mu}_m$ and $\hat{w}_m$. By Theorem 1, (4.8), and (5.3), it follows that

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}) - \mu\big) = \sum_{m=1}^{M}\hat{w}_m\sqrt{n}\left(\hat{\mu}_m - \mu\right) \xrightarrow{d} \sum_{m=1}^{M} w_m^*\Lambda_m.$$

This completes the proof. ■

Proof of Theorem 4: We first show the limiting distribution of $\zeta_{m,\ell}$. Since $\hat{\mathbf{e}}_m'\hat{\mathbf{e}}_f = \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f$ and $\hat{\mathbf{e}}_m - \hat{\mathbf{e}}_f = -\mathbf{H}(\mathbf{S}_m\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_f)$, we have

$$\zeta_{m,\ell} = \hat{\mathbf{e}}_m'\hat{\mathbf{e}}_\ell - \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f = \left(\hat{\mathbf{e}}_m - \hat{\mathbf{e}}_f\right)'\left(\hat{\mathbf{e}}_\ell - \hat{\mathbf{e}}_f\right) = \sqrt{n}(\mathbf{S}_m\hat{\boldsymbol{\theta}}_{\boldsymbol{m}} - \hat{\boldsymbol{\theta}}_f)'\left(\frac{1}{n}\mathbf{H}'\mathbf{H}\right)\sqrt{n}(\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell - \hat{\boldsymbol{\theta}}_f).$$

From Lemma 1, it follows that

$$\begin{aligned}
\sqrt{n}(\mathbf{S}_m\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_f) &= \mathbf{S}_m\sqrt{n}(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m) + \sqrt{n}(\mathbf{S}_m\boldsymbol{\theta}_m - \boldsymbol{\theta}) - \sqrt{n}(\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}) \\
&\xrightarrow{d} \left(\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0 - \mathbf{S}_0\right)\left(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta} + \left(\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m' - \mathbf{Q}^{-1}\right)\mathbf{R} \\
&= \left(\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0 - \mathbf{S}_0\right)\boldsymbol{\delta} + \left(\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m' - \mathbf{Q}^{-1}\right)\mathbf{R} \\
&= \mathbf{C}_m\boldsymbol{\delta} + \left(\mathbf{P}_m - \mathbf{Q}^{-1}\right)\mathbf{R} = \mathbf{R}_m
\end{aligned}$$

where the third equality holds by the fact that $\mathbf{S}_0\boldsymbol{\Pi}_m' = \mathbf{S}_m\left(\mathbf{0}_{p\times q_m}', \mathbf{I}_{q_m}\right)'$. Then, by the application of Slutskys theorem, we have $\zeta_{m,\ell} \xrightarrow{d} \mathbf{R}_m'\mathbf{Q}\mathbf{R}_\ell = \zeta_{m,\ell}^*$. Since all of $\zeta_{m,\ell}^*$ can be expressed in terms of the normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\zeta_{m,\ell}$ to $\zeta_{m,\ell}^*$. This implies (5.8). Following a similar argument to the proof of Theorem 3, we can show (5.10) and (5.11). This completes the proof. ■

Proof of Theorem 5: We first show the limiting distribution of $\xi_{m,\ell}$. Define $h_i = \mathbf{h}_i'(\mathbf{H}'\mathbf{H})^{-1}\mathbf{h}_i$. Note that $h_i = o_p(1)$, see Theorem 6.20.1 of Hansen (2013a). Then it follows that $\tilde{e} = \hat{e}_i(1-h_i)^{-1} \approx \hat{e}_i(1+h_i)$ where $\hat{e}_i$ is the least squares residual and $\tilde{e}$ is the leave-one-out least squares residual from the full model. For the submodel m, we have $\mathbf{h}_{mi} = \mathbf{S}_m'\mathbf{h}_i$, $h_{mi} = \mathbf{h}_i'\mathbf{S}_m(\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{S}_m'\mathbf{h}_i$, and $\tilde{e}_{mi} \approx \hat{e}_{mi}(1+h_{mi})$. Then it follows that

$$\begin{aligned}
\sum_{i=1}^n \tilde{e}_{mi}\tilde{e}_{\ell i} &\approx \sum_{i=1}^n \hat{e}_{mi}\hat{e}_{\ell i} + \sum_{i=1}^n \hat{e}_{mi}\hat{e}_{\ell i}(h_{mi}+h_{\ell i}) + \sum_{i=1}^n \hat{e}_{mi}\hat{e}_{\ell i}h_{mi}h_{\ell i} \\
&= \sum_{i=1}^n \hat{e}_{mi}\hat{e}_{\ell i} + \sum_{i=1}^n \hat{e}_{mi}\hat{e}_{\ell i}\left(\mathbf{h}_i'\mathbf{S}_m(\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{S}_m'\mathbf{h}_i + \mathbf{h}_i'\mathbf{S}_\ell(\mathbf{H}_\ell'\mathbf{H}_\ell)^{-1}\mathbf{S}_\ell'\mathbf{h}_i\right) + o_p(1) \\
&= \sum_{i=1}^n \hat{e}_{mi}\hat{e}_{\ell i} + tr\left(\left(\mathbf{S}_m\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{S}_m' + \mathbf{S}_\ell\left(\mathbf{H}_\ell'\mathbf{H}_\ell\right)^{-1}\mathbf{S}_\ell'\right)\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i'\hat{e}_{mi}\hat{e}_{\ell i}\right) + o_p(1) \\
&= \sum_{i=1}^n \hat{e}_{mi}\hat{e}_{\ell i} + tr\left(\mathbf{S}_m\hat{\mathbf{Q}}_m^{-1}\mathbf{S}_m'\tilde{\boldsymbol{\Omega}}\right) + tr\left(\mathbf{S}_\ell\hat{\mathbf{Q}}_\ell^{-1}\mathbf{S}_\ell'\tilde{\boldsymbol{\Omega}}\right) + o_p(1),
\end{aligned}$$

where $\hat{\mathbf{Q}}_m = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_{mi}\mathbf{h}_{mi}'$, $\hat{\mathbf{Q}}_\ell = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_{\ell i}\mathbf{h}_{\ell i}'$, and $\tilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i'\hat{e}_{mi}\hat{e}_{\ell i}$. In Lemma 2, we show that $\tilde{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$. By Assumption 3 and the application of the continuous mapping theorem, it follows that $tr\left(\mathbf{S}_m\hat{\mathbf{Q}}_m^{-1}\mathbf{S}_m'\tilde{\boldsymbol{\Omega}}\right) \xrightarrow{p} tr\left(\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\boldsymbol{\Omega}\right) = tr\left(\mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\right)$. Similarly, we have $tr\left(\mathbf{S}_\ell\hat{\mathbf{Q}}_\ell^{-1}\mathbf{S}_\ell'\tilde{\boldsymbol{\Omega}}\right) \xrightarrow{p} tr\left(\mathbf{Q}_\ell^{-1}\boldsymbol{\Omega}_\ell\right)$. As shown in Theorem 4, we have $\hat{\mathbf{e}}_m'\hat{\mathbf{e}}_\ell - \hat{\mathbf{e}}'\hat{\mathbf{e}} \xrightarrow{d} \mathbf{R}_m'\mathbf{Q}\mathbf{R}_p$. Therefore, it follows that

$$\begin{aligned}
\xi_{m,\ell} &= \tilde{\mathbf{e}}_{mi}'\tilde{\mathbf{e}}_{\ell i} - \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f \\
&= \left(\hat{\mathbf{e}}_m'\hat{\mathbf{e}}_\ell - \hat{\mathbf{e}}_f'\hat{\mathbf{e}}_f\right) + tr\left(\mathbf{S}_m\hat{\mathbf{Q}}_m^{-1}\mathbf{S}_m'\tilde{\boldsymbol{\Omega}}\right) + tr\left(\mathbf{S}_\ell\hat{\mathbf{Q}}_\ell^{-1}\mathbf{S}_\ell'\tilde{\boldsymbol{\Omega}}\right) + o_p(1) \\
&\xrightarrow{d} \mathbf{R}_m'\mathbf{Q}\mathbf{R}_\ell + tr\left(\mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\right) + tr\left(\mathbf{Q}_\ell^{-1}\boldsymbol{\Omega}_\ell\right) = \xi_{m,\ell}^*
\end{aligned}$$

Since all of $\xi_{m,\ell}^*$ can be expressed in terms of the normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\xi_{m,\ell}$ to $\xi_{m,\ell}^*$. Hence, it follows that $\mathbf{w}'\boldsymbol{\xi}_n\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}$. Following a similar argument to the proof of Theorem 3, we can show (5.17) and (5.18). This completes the proof. ■

Lemma 2. *For $m, \ell = 1, ..., M$, let $\tilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i'\hat{e}_{mi}\hat{e}_{\ell i}$ where $\hat{e}_{mi}$ and $\hat{e}_{\ell i}$ are the least squares residuals from the submodel m and ℓ. Suppose Assumptions 1 and 3 hold. As $n \to \infty$, we have $\tilde{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega} = E(\mathbf{h}_i\mathbf{h}_i'e_i^2)$.*

Proof of Lemma 2: The proof is similar to that of Theorem 6.7.1 of Hansen (2013a). Let $\|\cdot\|$ be the Euclidean norm. That is, for a $k \times 1$ vector $\mathbf{x}_i$, $\|\mathbf{x}_i\| = (\sum_{j=1}^k x_{ij}^2)^{1/2}$. Observe that

$$\tilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i'\hat{e}_{mi}\hat{e}_{\ell i} = \frac{1}{n}\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i'\hat{e}_i^2 + \frac{1}{n}\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i'\left(\hat{e}_{mi}\hat{e}_{\ell i} - \hat{e}_i^2\right).$$

By Assumption 3 and the weak law of large number, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i\mathbf{h}_i'\hat{e}_i^2 \xrightarrow{p} \mathrm{E}(\mathbf{h}_i\mathbf{h}_i'e_i^2)=\mathbf{\Omega}.$$

We next show the second term converges in probability to zero. By the Triangle Inequality,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i\mathbf{h}_i'\left(\hat{e}_{mi}\hat{e}_{\ell i}-\hat{e}_i^2\right)\right\| \leq \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{h}_i\mathbf{h}_i'\left(\hat{e}_{mi}\hat{e}_{\ell i}-\hat{e}_i^2\right)\right\| = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_i\|^2\left|\hat{e}_{mi}\hat{e}_{\ell i}-\hat{e}_i^2\right|.$$

Note that $\tilde{e}_{mi}=y_i-\mathbf{h}_{mi}'\hat{\boldsymbol{\theta}}_m=e_i-\mathbf{h}_i(\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta})$. Similarly, we have $\tilde{e}_{\ell i}=e_i-\mathbf{h}_i(\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta})$. Thus,

$$\hat{e}_{mi}\hat{e}_{\ell i}=e_i^2-e_i\mathbf{h}_i'\left(\left(\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right)+\left(\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right)\right)+\left(\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right)'\mathbf{h}_i\mathbf{h}_i'\left(\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right).$$

Therefore, by the Triangle Inequality and the Schwarz Inequality, it follows that

$$\begin{aligned}|\hat{e}_{mi}\hat{e}_{\ell i}-\hat{e}_i^2| &\leq \left|e_i\mathbf{h}_i'\left(\left(\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right)+\left(\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right)\right)\right|+\left|\left(\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right)'\mathbf{h}_i\mathbf{h}_i'\left(\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right)\right| \\ &\leq |e_i|\,\|\mathbf{h}_i\|\left(\left\|\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right\|+\left\|\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right\|\right)+\|\mathbf{h}_i\|^2\left\|\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right\|\left\|\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right\|.\end{aligned}$$

Thus, we have

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i\mathbf{h}_i'\left(\hat{e}_{mi}\hat{e}_{\ell i}-\hat{e}_i^2\right)\right\| \leq& \left(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_i\|^3|e_i|\right)\left(\left\|\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right\|+\left\|\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right\|\right) \\ &+\left(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_i\|^4\right)\left\|\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right\|\left\|\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right\| \qquad \text{(A.1)}\end{aligned}$$

By Assumption 1, Lemma 1, the Triangle Inequality, and the Schwarz Inequality,

$$\begin{aligned}\left\|\mathbf{S}_m\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\right\| &\leq \left\|\mathbf{S}_m\left(\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m\right)\right\|+\|\mathbf{S}_m\boldsymbol{\theta}_m-\boldsymbol{\theta}\| \\ &\leq \|\mathbf{S}_m\|\left\|\left(\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m\right)\right\|+\left\|\mathbf{S}_0\left(\mathbf{I}_q-\mathbf{\Pi}_m'\mathbf{\Pi}_m\right)\right\|\|\boldsymbol{\gamma}_n\|=o_p(1) \qquad \text{(A.2)}\end{aligned}$$

Similarly, we have $\left\|\mathbf{S}_\ell\hat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}\right\|=o_p(1)$. Then, by Assumption 3, the weak law of large number, and Hoölder's Inequality, we have $\left(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_i\|^4\right)\xrightarrow{p}\mathrm{E}\,\|\mathbf{h}_i\|^4<\infty$ and

$$\left(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_i\|^3|e_i|\right)\xrightarrow{p}\mathrm{E}\left(\|\mathbf{h}_i\|^3|e_i|\right)\leq\left(\mathrm{E}\,\|\mathbf{h}_i\|^4\right)^{3/4}\left(\mathrm{E}|e_i|^4\right)^{1/4}<\infty. \qquad \text{(A.3)}$$

Combining (A.1), (A.2), and (A.3), we have $\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i\mathbf{h}_i'\left(\hat{e}_{mi}\hat{e}_{\ell i}-\hat{e}_i^2\right)\right\|=o_p(1)$. This completes the proof. ■

Proof of Theorem 6: From Theorem 1, there is joint convergence in distribution of all $\sqrt{n}\left(\mu(\hat{\boldsymbol{\theta}}_m)-\mu(\boldsymbol{\theta})\right)$ to Λ_m since all of Λ_m can be expressed in terms of $\mathbf{R}$. Also, $w(m|\hat{\boldsymbol{\delta}})\xrightarrow{d}$

$w(m|\mathbf{R}_{\boldsymbol{\delta}})$ where $w(m|\mathbf{R}_{\boldsymbol{\delta}})$ is a function of the random vector $\mathbf{R}$. Therefore,

$$\begin{aligned}
\sqrt{n}\left(\bar{\mu}-\mu\right) &= \sum_{m=1}^{M} w(m|\hat{\boldsymbol{\delta}})\sqrt{n}(\hat{\mu}_m-\mu) \\
&\xrightarrow{d} \sum_{m=1}^{M} w(m|\mathbf{R}_{\boldsymbol{\delta}})\left(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{C}_m\boldsymbol{\delta}+\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{P}_m\mathbf{R}\right) \\
&= \mathbf{D}_{\boldsymbol{\theta}}'\sum_{m=1}^{M} w(m|\mathbf{R}_{\boldsymbol{\delta}})\left(\mathbf{P}_m\mathbf{Q}-\mathbf{C}_m\mathbf{S}_0'\right)\mathbf{Q}^{-1}\mathbf{R}+\mathbf{D}_{\boldsymbol{\theta}}'\sum_{m=1}^{M} w(m|\mathbf{R}_{\boldsymbol{\delta}})\mathbf{C}_m\mathbf{R}_{\boldsymbol{\delta}} \\
&= \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{Q}^{-1}\mathbf{R}+\mathbf{D}_{\boldsymbol{\theta}}'\left(\sum_{m=1}^{M} w(m|\mathbf{R}_{\boldsymbol{\delta}})\mathbf{C}_m\right)\mathbf{R}_{\boldsymbol{\delta}}
\end{aligned}$$

where the last equality holds by the fact that

$$\begin{aligned}
\mathbf{P}_m\mathbf{Q}-\mathbf{C}_m\mathbf{S}_0' &= \mathbf{P}_m\mathbf{Q}-\left(\mathbf{P}_m\mathbf{Q}\begin{bmatrix}\mathbf{0}_{p\times p} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{I}_q\end{bmatrix}-\begin{bmatrix}\mathbf{0}_{p\times p} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{I}_q\end{bmatrix}\right) \\
&= \mathbf{P}_m\mathbf{Q}\begin{bmatrix}\mathbf{I}_p & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q}\end{bmatrix}+\begin{bmatrix}\mathbf{0}_{p\times p} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{I}_q\end{bmatrix} \\
&= \mathbf{S}_m\left(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\begin{bmatrix}\mathbf{I}_p & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q_m\times p} & \mathbf{0}_{q_m\times q}\end{bmatrix}+\begin{bmatrix}\mathbf{0}_{p\times p} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{I}_q\end{bmatrix}=\mathbf{I}_{p+q}.
\end{aligned}$$

This completes the proof. ■

References

Andrews, D. W. K. (1991a): "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

——— (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

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

Claeskens, G. and R. J. Carroll (2007): "An Asymptotic Theory for Model Selection Inference in General Semiparametric Problems," *Biometrika*, 94, 249–265.

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

——— (2008): *Model Selection and Model Averaging*, Cambridge University Press.

DiTraglia, F. (2013): "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," Working Paper, University of Pennsylvania.

Durlauf, S., A. Kourtellos, and C. Tan (2008): “Are Any Growth Theories Robust?” *The Economic Journal*, 118, 329–346.

Durlauf, S. N., P. A. Johnson, and J. R. Temple (2005): “Growth Econometrics,” in *Handbook of Economic Growth*, ed. by P. Aghion and S. Durlauf, Elsevier, vol. 1, 555–677.

Elliott, G., A. Gargano, and A. Timmermann (2013): “Complete Subset Regressions,” *Journal of Econometrics*, 177, 357–373.

Fernandez, C., E. Ley, and M. Steel (2001): “Model Uncertainty in Cross-Country Growth Regressions,” *Journal of Applied Econometrics*, 16, 563–576.

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

——— (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.

——— (2013a): “Econometrics,” Unpublished Manuscript, University of Wisconsin.

——— (2013b): “Model Averaging, Asymptotic Risk, and Regressor Groups,” Forthcoming. *Quantitative Economics.*

Hansen, B. E. and J. Racine (2012): “Jackknife Model Averaging,” *Journal of Econometrics*, 167, 38–46.

Hansen, P., A. Lunde, and J. Nason (2011): “The Model Confidence Set,” *Econometrica*, 79, 453–497.

Hausman, J. (1978): “Specification Tests in Econometrics,” *Econometrica*, 46, 1251–1271.

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

——— (2003b): “Rejoinder to “The Focused Information Criterion” and “Frequentist Model Average Estimators”,” *Journal of the American Statistical Association*, 98, 938–945.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999): “Bayesian Model Averaging: A Tutorial,” *Statistical Science*, 14, 382–401.

Kabaila, P. (1995): “The Effect of Model Selection on Confidence Regions and Prediction Regions,” *Econometric Theory*, 11, 537–537.

——— (1998): “Valid Confidence Intervals in Regression after Variable Selection,” *Econometric Theory*, 14, 463–482.

KIM, J. AND D. POLLARD (1990): "Cube Root Asymptotics," *The Annals of Statistics*, 18, 191–219.

LEEB, H. AND B. PÖTSCHER (2003): "The Finite-Sample Distribution of Post-Model-Selection Estimators and Uniform versus Non-Uniform Approximations," *Econometric Theory*, 19, 100–142.

——— (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

——— (2006): "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554–2591.

——— (2008): "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338–376.

——— (2012): "Testing in the Presence of Nuisance Parameters: Some Comments on Tests Post-Model-Selection and Random Critical Values," Working Paper, University of Vienna.

LEUNG, G. AND A. BARRON (2006): "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52, 3396–3410.

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, 958–975.

LIANG, H., G. ZOU, A. WAN, AND X. ZHANG (2011): "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 106, 1053–1066.

MAGNUS, J., O. POWELL, AND P. PRUFER (2010): "A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics," *Journal of Econometrics*, 154, 139–153.

MORAL-BENITO, E. (2013): "Model Averaging in Economics: An Overview," forthcoming *Journal of Economic Surveys*.

NEWEY, W. AND K. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

PÖTSCHER, B. (1991): "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163–185.

——— (2006): "The Distribution of Model Averaging Estimators and an Impossibility Result Regarding its Estimation," *Lecture Notes-Monograph Series*, 52, 113–129.

RAFTERY, A. E. AND Y. ZHENG (2003): "Discussion: Performance of Bayesian Model Averaging," *Journal of the American Statistical Association*, 98, 931–938.

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.

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

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer Verlag.

Wan, A., X. Zhang, and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

——— (1984): *Asymptotic Theory for Econometricians*, Academic Press.

White, H. and X. Lu (2014): "Robustness Checks and Robustness Tests in Applied Economics," *Journal of Econometrics*, 178, Part 1, 194 – 206.

Yang, Y. (2000): "Combining Different Procedures for Adaptive Regression," *Journal of Multivariate Analysis*, 74, 135–161.

——— (2001): "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574–588.

Yuan, Z. and Y. Yang (2005): "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214.

Zhang, X. and H. Liang (2011): "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200.

Zou, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.