KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

Discussion Paper No.855

"Generalized Least Squares Model Averaging"

Qingfeng Liu, Ryo Okui and Arihiro Yoshimura

March 2013

KYOTO UNIVERSITY

KYOTO, JAPAN

Generalized Least Squares Model Averaging*

Qingfeng Liu†, Ryo Okui‡ and Arihiro Yoshimura§

February 6, 2013

Abstract

This paper proposes a method of averaging generalized least squares (GLS) estimators for linear regression models with heteroskedastic errors. We derive two kinds of Mallows' C_p criteria, calculated from the estimates of the mean of the squared errors of the fitted value based on the averaged GLS estimators, for this class of models. The averaging weights are chosen by minimizing Mallows' C_p criterion. We show that this method achieves asymptotic optimality. It is also shown that the asymptotic optimality holds even when the variances of the error terms are estimated and the feasible generalized least squares (FGLS) estimators are averaged. Monte Carlo simulations demonstrate that averaging FGLS estimators yields an estimate that has a remarkably lower level of risk compared with averaging least squares estimators in the presence of heteroskedasticity, and it also works when heteroskedasticity is not present, in finite samples.

JEL classification: C51, C52
Keywords: model averaging, GLS, FGLS, asymptotic optimality, Mallows' C_p

1 Introduction

Model averaging is an important research topic in recent econometrics and statistics literature. It is often difficult to specify a "correct" or "best" statistical model a priori when we conduct data analysis. For example, it is often uncertain which variables should be used as explanatory variables in regression analysis. Numerous model selection techniques, such as the Akaike information criterion by Akaike (1973), Mallows' C_p criterion by Mallows (1973) and the Bayesian information criterion by Schwarz (1978), have been proposed and many researchers have investigated their properties. Relatively recently, model averaging techniques, which combine several candidate models rather than selecting one particular model as done in model selection, have gained the attention of

*Liu and Okui acknowledge financial support from the Joint Usage and Research Center at the Institute of Economic Research, Kyoto University. The authors are solely responsible for any errors.

†Otaru University of Commerce, Email: qliu@res.otaru-u.ac.jp

‡Corresponding author. Institute of Economic Research, Kyoto University, Yoshida-Hommachi, Sakyo, Kyoto, Kyoto, 606-8501, Japan. Tel: +81-75-753-7191. Fax: +81-75-753-7118. Email: okui@kier.kyoto-u.ac.jp

§Kyoto University, Email: yoshimura.a@hy2.ecs.kyoto-u.ac.jp

many researchers. Many works have examined Bayesian model averaging, such as Draper (1995). See, e.g., Hoeting, Madigan, Raftery and Volinsky (1999) for a literature review. Frequentist model averaging has also become important in the recent literature. Hjort and Claeskens (2003) propose model averaging estimators in parametric models and study their asymptotic properties using the local misspecification approach. Liang, Zou, Wan and Zhang (2011) propose a method of selecting model weights based on the exact mean squared error of the estimator of parameter of interest in the context of linear regression. The book by Claeskens and Hjort (2008) provides an excellent overview of the literature on model selection and model averaging. Yuan and Yang (2005) and Zhang, Wan and Zhou (2012) discuss under what circumstances model averaging performs better than model selection.

Hansen (2007) proposes a method of averaging least squares estimators for linear regression models with homoskedastic errors in the presence of many regressors. His method uses Mallows' criterion for choosing the model weights and the objective is to minimize the mean squared error (MSE) from the fit of the averaging estimator. He shows that the proposed method has asymptotic optimality in the sense of Li (1987). Wan, Zhang and Zou (2010) provide an alternative proof of the optimality of Hansen's method in different situations. Hansen and Racine (2010) and Liu and Okui (2012) extend the method of Hansen (2007) to models with possibly heteroskedastic errors.

Heteroskedasticity is a common phenomenon in economic applications and it is important to develop methods that take into account the possibility of the presence of heteroskedasticity. The methods of Hansen and Racine (2010) and Liu and Okui (2012) are robust to heteroskedasticity but still average least squares estimators. However, in the presence of heteroskedasticity, generalized least squares (GLS) estimators should provide better prediction than least squares estimators because GLS estimators have smaller variances. This observation motivates the research reported in this paper.

This paper proposes a method of averaging GLS estimators for linear regression models with heteroskedastic errors. Our aim is to construct an estimate that best predicts the value of the dependent variable in the presence of a large number of regressors. We propose using a weighted average of various GLS estimators, where each estimator uses a different set of regressors. We derive two kinds of Mallows' (1973) C_p-like criteria. The first criterion is an estimate of the MSE of the model fit, and the other is an estimate of the weighted MSE. The averaging weights are selected by minimizing these Mallows criteria. The minimization problems are standard quadratic programming problems and can be easily implemented by many numerical or statistical programming packages. We show that these methods have asymptotic optimality in the sense of Li (1987), such that the weights chosen by the proposed methods can attain the same MSE or weighted MSE as the optimal one asymptotically. This optimality result is one of our main theoretical contributions.

We also consider averaging feasible generalized least squares (FGLS) estimators. As the variances of error terms are unknown in many applications, it is important to consider FGLS estimators. We assume that the variances can be specified by a finite number of parameters that can be estimated at the rate $\sqrt{n}$. We discuss briefly the cases in which the variances are estimated by nonparametric methods. We also assume that all the FGLS estimators to be averaged use the same variance estimator. The weights are chosen by minimizing the Mallows criteria. We prove that this method based on FGLS estimators also achieves asymptotic optimality. The optimality of the model averaging method for the FGLS estimators is new and its proof is nontrivial. It thus constitutes another main theoretical contribution of this paper.

Several model averaging methods have been proposed for linear regression models with heteroskedastic errors. We discuss briefly the relationship between our new method and other alternative methods. As mentioned previously, the methods of Hansen and Racine (2010) and Liu and Okui (2012) are robust to heteroskedasticity, although they average least squares estimators. Our method combines GLS estimators with the aim of reducing the variance of the estimator. Magnus, Wan and Zhang (2011) extend the approach of Magnus, Powell and Prüfer (2010) and propose a method to average GLS estimators by combining the Bayesian and frequentist approaches. Our objective, however, is to obtain an estimator whose fit achieves a small MSE and we choose the weights directly by minimizing an estimate of the MSE. Liu (2011) proposes an alternative model averaging estimator for heteroskedastic models. His method follows the approach of Hjort and Claeskens (2003) and examines the MSE of some finite-dimension parameters. Thus, the objectives of our model averaging differ from those of Liu (2011).

We conduct Monte Carlo simulations to examine the performance of the proposed methods in finite samples. In particular, we compare the proposed methods with other alternative existing methods. The simulation results demonstrate that our methods perform better than the methods of Hansen and Racine (2010) and Liu and Okui (2012), which are robust to heteroskedasticity, but which average least squares estimators and not GLS estimators. Our methods show superior performance compared with the model averaging estimator of Magnus, Wan and Zhang (2011), which also averages FGLS estimators. However, note that our simulations use as a performance measure the MSE from the fitted model, the reduction of which is the aim of our proposed methods, whereas Magnus, Wan and Zhang (2011), in developing their model averaging methods, consider reducing a measure of risk or regret of the *estimator* of the coefficient, not the fit of the model.

For the settings in which the error terms are homoskedastic, our method works as well as the method of Hansen (2007), which is designed for homoskedastic models. Thus, even in the absence of heteroskedasticity, the cost of using our method is not severe. Finally, we examine the performance of the estimator that averages the FGLS estimators

of Robinson (1987), which are based on nonparametric variance estimates. Note that we do not provide theoretical support for this nonparametric method. Nonetheless, we find that averaging the nonparametric version of the FGLS estimators provides good estimates in the simulations. This result indicates that it is worthwhile considering averaging FGLS estimators even when the form of the heteroskedasticity is not known.

The rest of the paper is organized as follows. Section 2 describes our setting and the GLS and FGLS estimators. Section 3 introduces Mallows' C_p criterion and shows its optimality. Section 4 provides the results of Monte Carlo simulations. Section 5 concludes. The proofs of the theorems are given in the Appendix.

2 Model averaging based on GLS estimators

Suppose that we observe the random sample, (y_i, x_i) for $i = 1, \ldots, n$, where y_i is a real-valued scalar random variable and $x_i = (x_{i1}, x_{i2}, \ldots,)$ is a countably infinite real-valued vector.[1] The relationship between y_i and x_i is described by the following linear regression model:

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

where:

$$\mu_i = \sum_{j=1}^{\infty} \theta_j x_{ij},$$

and e_i is an unobserved error term that satisfies:

$$\begin{aligned} E(e_i|x_i) &= 0, \\ E(e_i^2|x_i) &= \sigma_i^2. \end{aligned}$$

The sequence θ_j, $j = 1, \ldots$ is a sequence of unknown parameters, and we assume that $\sum_{j=1}^{J} \theta_j x_{ij}$ converges in mean square to μ_i as $J \to \infty$. Our setup allows heteroskedasticity as in Hansen and Racine (2010) and Liu and Okui (2012). Most of the theoretical results we present are for distributions conditional on X, where $X = (x_1, \ldots, x_n)$. For simplicity, we omit the conditional expressions hereafter.

Our objective is to estimate μ_i. For this purpose, we consider using models that contain finite subsets of the elements of x_i. We consider M candidate models. The mth approximation model has $k_m > 0$ regressors. We assume that $k_1 \leq k_2 \leq \cdots \leq k_M$ and the sequence of the models is nested in the sense that the mth model contains all

[1] We do not need to observe all the elements of x_i, because it is sufficient to observe only those elements of x_i that are used in the estimation. However, it is convenient to describe the procedure and the theory by assuming that every element is observable. An example in which all the elements of x_i are observed is nonparametric sieve estimation where x_i consists of the basis functions of a sieve.

the regressors appearing in the jth model for $j \leq m$, as in Hansen (2007). The mth approximation model of (1) is:

$$y_i = \sum_{j=1}^{k_m} \theta_j x_{ij} + b_{mi} + e_i, \tag{2}$$

for $m = 1, 2, \ldots, M$, where $b_{mi} = \mu_i - \sum_{j=1}^{k_m} \theta_j x_{ij} = \sum_{j=k_m+1}^{\infty} \theta_j x_{ij}$ is the approximation error.

The matrix representation of (1) is $Y = \mu + e$, where $Y = (y_1, \ldots, y_n)'$, $\mu = (\mu_1, \ldots, \mu_n)'$, and $e = (e_1, \ldots, e_n)'$. Let $\Omega = diag(\sigma_1^2, \ldots, \sigma_n^2)$ denote the $n \times n$ positive definite variance–covariance matrix of the error terms. The matrix form of the approximation model (2) is:

$$Y = X_m \Theta_m + b_m + e,$$

where X_m is an $n \times k_m$ matrix of the regressors with ij-th element x_{ij}, $\Theta_m = (\theta_1, \ldots, \theta_{k_m})'$, and $b_m = (b_{m1}, \ldots, b_{mn})'$.

In this paper, we use GLS estimators to construct an estimate of μ. In the case of standard linear regression models, it is known that the GLS procedure provides a more efficient estimator than does the least squares estimator in the presence of heteroskedasticity. We therefore expect the approach based on GLS estimators to yield a better estimate of μ even in the infinite dimensional linear regression models considered in this paper.

The GLS estimator of Θ_m in the mth model is $\hat{\Theta}_m = (X_m' \Omega^{-1} X_m)^{-1} X_m' \Omega^{-1} Y$. The estimator of μ from the mth model is:

$$\hat{\mu}_m = X_m \hat{\Theta}_m = X_m (X_m' \Omega^{-1} X_m)^{-1} X_m' \Omega^{-1} Y.$$

When the variance matrix Ω is unknown, we use the FGLS estimators. Let $\hat{\Omega}$ be an estimate of Ω. The estimation of Ω is discussed in Section 3.2. The FGLS estimator of μ from the mth model is:

$$\hat{\mu}_m^F = X_m \hat{\Theta}_m^F = X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m' \hat{\Omega}^{-1} Y.$$

We propose that averaging the estimators of μ from various models would produce a better estimator than the estimator from each individual model. Let $W = (w_1, \ldots, w_M)'$ be an $M \times 1$ weighting vector in:

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

The generalized least squares model averaging (GLSMA, hereafter) estimator of μ is

defined as:

$$\begin{aligned}\hat{\mu}(W) &= \sum_{m=1}^{M} w_m \hat{\mu}_m \\ &= \sum_{m=1}^{M} w_m X_m (X_m' \Omega^{-1} X_m)^{-1} X_m' \Omega^{-1} Y \\ &= G(W) Y,\end{aligned}$$

for some $W \in \mathcal{H}_M$, where the matrix $G(W) = \sum_{m=1}^{M} w_m X_m (X_m' \Omega^{-1} X_m)^{-1} X_m' \Omega^{-1}$ is introduced for later use. Similarly, the feasible generalized least squares model averaging (FGLSMA) estimator is:

$$\begin{aligned}\hat{\mu}(W) &= \sum_{m=1}^{M} w_m \hat{\mu}_m^F \\ &= \sum_{m=1}^{M} w_m X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m' \hat{\Omega}^{-1} Y \\ &= G^F(W) Y,\end{aligned}$$

where $G^F(W)$ is defined similarly to $G(W)$. Note that we assume that all FGLS estimators to be averaged use the same variance estimator, $\hat{\Omega}$.

3 Choice of the weighting vector

This section discusses the choice of the weighting vector and the properties of the proposed procedure. We propose choosing the weighting vector by minimizing an estimate of the MSE of the fit from the model averaging estimator. This approach yields criteria similar to Mallows' (1978) criterion, and this procedure is easy computationally. Section 3.1 discusses the estimation method of the GLSMA estimator. The method of the FGLSMA estimator is discussed in Section 3.2.

3.1 Mallows' criterion for GLSMA

We would like to use a weight vector that achieves a small MSE for the fitted model. The relevant loss function is therefore:

$$\begin{aligned}L_H(W) &= \|\mu - \hat{\mu}(W)\|_H^2 \\ &= (\hat{\mu}(W) - \mu)' H^{-1} (\hat{\mu}(W) - \mu)\end{aligned}$$

for some $n \times n$ positive definite matrix H. The risk function is:

$$R_H(W) = E(L_H(W)).$$

In this paper, we consider two choices of the weighting matrix H: $H = I_n$ and $H = \Omega$. Although we could consider alternative weighting matrices, we believe that these two are the most natural choices.

We estimate the loss function and use the estimated loss function as the criterion for choosing the weighting vector. This results in a criterion in the spirit of Mallows (1973). Mallows' criterion C_{I_n} for $L_{I_n}(W)$ is defined as:

$$\begin{aligned} C_{I_n}(W) &= \|Y - \hat{\mu}(W)\|_{I_n}^2 + 2tr\left(\sum_{m=1}^{M} w_m X_m (X_m'\Omega^{-1}X_m)^{-1}X_m'\right) \\ &= \|Y - \hat{\mu}(W)\|_{I_n}^2 + 2tr\left(G(W)\Omega\right), \end{aligned}$$

and C_Ω for $L_\Omega(W)$ is defined as:

$$\begin{aligned} C_\Omega(W) &= \|Y - \hat{\mu}(W)\|_\Omega^2 + 2tr(\sum_{m=1}^{M} w_m \Omega^{-1/2} X_m (X_m'\Omega^{-1}X_m)^{-1}X_m'\Omega^{-1/2\prime}) \\ &= \left\|\Omega^{-1/2}Y - \Omega^{-1/2}\hat{\mu}(W)\right\|_{I_n}^2 + 2\sum_{m=1}^{M} w_m k_m. \end{aligned}$$

Note that the C_Ω criterion is equivalent to the Mallows criterion derived by Hansen (2007) applied to the data $(\Omega^{-1/2}Y, \Omega^{-1/2}X)$. Another important observation is that $E(C_{I_n}(W)) = E(L_{I_n}(W)) + tr(\Omega)$ and $E(C_\Omega(W)) = E(L_\Omega(W)) + n$. These criteria are unbiased estimators of the corresponding loss functions plus some constants that do not depend on the weights.

We choose the weighting vector by minimizing Mallows' criterion. Let

$$\hat{W}_{I_n} = \arg\min_{W \in \mathcal{H}_M} C_{I_n}(W).$$

and

$$\hat{W}_\Omega = \arg\min_{W \in \mathcal{H}_M} C_\Omega(W).$$

We then use $\hat{\mu}(\hat{W}_{I_n})$ or $\hat{\mu}(\hat{W}_\Omega)$ as our estimate of μ. These minimization problems are standard quadratic programming problems and can be implemented easily by many statistical or numerical packages.

We note that, although this paper focuses on model averaging, these criteria may also be used for model selection.

Next, we establish an optimal property of the proposed procedure for weight choice. The optimality considered in this paper is similar to that of Li (1987). We examine whether the loss evaluated at the chosen weights asymptotically achieves the minimum loss. To demonstrate optimality, we follow the proof strategy of Hansen (2007) and we restrict the weighting vector to be discrete, i.e., for some integer $N < \infty$, the elements of the weighting vector are restricted to be one of the set $\{0, 1/N, 2/N, \ldots, 1\}$ and let

$\mathcal{H}_M(N)$ be the subset of $\mathcal{H}_M$ restricted to this set of weighting vectors. We note that in practice this restriction may be ignored because the weights chosen from $\mathcal{H}_M$ can be arbitrarily close to that from $\mathcal{H}_M(N)$ by making N sufficiently large, although a large N imposes a heavy restriction on the moments of e.

We first consider $C_{I_n}(W)$. We use the following assumptions.

Assumption 1. $E(|e_i|^{4(N+1)}) \leq \kappa < \infty$ *for some* κ.

Assumption 2. $\xi_n \equiv \inf_{W \in \mathcal{H}_n(N)} R_{I_n}(W) \to \infty$ *as* $n \to \infty$.

Assumption 3. $\lim_{n\to\infty} \sup_{W \in \mathcal{H}_n(N)} \lambda(G(W)G(W)') < \infty$ *where* $\lambda(A)$ *denotes the maximum eigenvalue of the matrix* A.

Assumption 4. $0 < \inf_i \sigma_i^2 \leq \sup_i \sigma_i^2 < \infty$.

Assumption 1 imposes a bound on the moments of e. It may look strong but such an assumption is made in most of the model averaging literature (see, e.g., Hansen (2007)). Assumption 2 requires that any finite approximation of the true model is misspecified, i.e., modeling biases are not zero. This assumption is standard in the nonparametric regression literature. Assumption 3 is quite natural (see, for example, Li (1987)). Assumption 4 excludes degeneracy and divergence of the variances.

We now have the following theorem about the optimality of the weighting vector chosen by minimizing $C_{I_n}(W)$.

Theorem 1. *We assume 1, 2, 3 and 4. Then, as* $n \to \infty$,

$$\frac{L_{I_n}(\tilde{W}_{I_n})}{\inf_{W \in \mathcal{H}_n(N)} L_{I_n}(W)} \to_p 1,$$

where $\tilde{W}_{I_n} = \arg\min_{W \in \mathcal{H}_M(N)} C_{I_n}(W)$.

For $C_\Omega(W)$, we use the following assumption instead of Assumption 2.

Assumption 5. $\xi_n^\Omega \equiv \inf_{W \in \mathcal{H}_M(N)} R_\Omega(W) \to \infty$ *as* $n \to \infty$.

We then obtain the following theorem for the optimality of $C_\Omega(W)$.

Theorem 2. *We now assume 1, 4 and 5. Then, as* $n \to \infty$,

$$\frac{L_\Omega(\tilde{W}_\Omega)}{\inf_{W \in \mathcal{H}_n(N)} L_\Omega(W)} \to_p 1,$$

where $\tilde{W}_\Omega = \arg\min_{W \in \mathcal{H}_M(N)} C_\Omega(W)$.

These theorems show that the squared error evaluated at the weighting vector chosen by using the GLS-based Mallows criteria is asymptotically equivalent to that evaluated at the infeasible optimal weighting vector.

The proof of Theorem 1 is in the Appendix. It is an application of Theorem 2.1 of Li (1987) and follows the steps taken by the proof of Theorem 1 of Hansen (2007). The proof of Theorem 2 is omitted because $C_\Omega(W)$ is equivalent to the criterion considered by Hansen (2007) applied to the data $(\Omega^{-1/2}Y, \Omega^{-1/2}X)$. These transformed data are homoskedastic so this theorem follows immediately from Theorem 1 of Hansen (2007).

Although considering the restricted set $\mathcal{H}_M(N)$ may not cause a serious problem in applications and is the approach chosen by Hansen (2007) and Hansen and Racine (2010), this restriction may still appear to be somewhat unsatisfactory. We can relax this restriction and consider the unrestricted set $\mathcal{H}_M$ by imposing different sets of assumptions. This approach is selected by Wan, Zhang and Zou (2010) and Liu and Okui (2012). Kuersteiner and Okui (2010) also provide a different proof strategy.

3.2 Mallows' criterion for FGLSMA

In this subsection, we extend the Mallows criteria described in the previous subsection to the FGLSMA estimator. The loss function is defined to be:

$$\begin{aligned} L_H^F(W) &= \left\|\mu - \hat{\mu}^F(W)\right\|_H^2 \\ &= (\hat{\mu}^F(W) - \mu)' H^{-1} (\hat{\mu}^F(W) - \mu). \end{aligned}$$

Analogous to the case for GLSMA, we consider the Mallows criteria for $H = I_n$ and $H = \Omega$, which are:

$$C_{I_n}^F(W) = \left\|Y - \hat{\mu}^F(W)\right\|_{I_n}^2 + 2tr\left(\sum_{m=1}^M w_m X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m'\right)$$

and

$$\begin{aligned} C_\Omega^F(W) &= \|Y - \hat{\mu}_F(W)\|_{\hat{\Omega}}^2 + 2\sum_{m=1}^M w k_m \\ &= \left\|\hat{\Omega}^{-1/2}Y - \hat{\Omega}^{-1/2}\hat{\mu}_F(W)\right\|_{I_n}^2 + 2\sum_{m=1}^M w k_m. \end{aligned}$$

We choose the weights for the FGLSMA estimator by minimizing these criteria.

The main purpose of this subsection is to show the asymptotic optimality of the criteria $C_{I_n}^F(W)$ and $C_\Omega^F(W)$. We use the following additional assumptions.

Assumption 6. $\lim_{n\to\infty} \sum_{i=1}^n x_{m,j,i}^2/n$ *is bounded uniformly in* m *and* j.

Assumption 7. *The maximum eigenvalue of* $\sum_{i=1}^n x_{mi}x_{mi}'/n$ *is bounded uniformly in* n *and* m. *The minimum eigenvalue of* $\sum_{i=1}^n x_{mi}x_{mi}'/n$ *is away from zero uniformly in* n *and* m.

Assumption 8. *There exists* $C < \infty$ *such that* $lim_{n\to\infty} \sum_{i=1}^n \mu_i^2/n < C$.

Assumption 9. $\sup_i |\hat{\sigma}_i^{-2} - \sigma_i^{-2}| = O_p(n^{-1/2})$, *where* $\hat{\sigma}_i^2$ *is the ith diagonal element of* $\hat{\Omega}$.

Assumption 10. $k_M^2/n \to 0$ *and* $k_M/\xi_n \to 0$, *as* $n \to \infty$.

Assumptions 6 and 7 state that the regressors are uniformly bounded. Assumption 8 is a standard moment condition.

Assumption 9 require that the variances can be estimated at a $\sqrt{N}$ parametric rate. These assumptions do not allow us to use the FGLS estimator based on nonparametric variance estimates as considered by Robinson (1987). An extension to such an estimator seems to be a promising future research agenda.

Assumption 10 may look strong as a condition that restricts the size of the largest model M. For example, if the coefficients $\theta_j = j^{-\alpha}$ for some α and the errors are homoskedastic, we have $\xi_n = O(n^{1/(1+2\alpha)})$ (see, e.g., Hansen (2007)). In this case, Assumption 10 requires $k_M = o(n^{1/(1+2\alpha)})$. This restriction is the cost of using estimated variances in the GLS estimation. However, as long as α is small (i.e., as long as the coefficients do not decline drastically as the model gets large), we can consider a large number of regressors.

The following theorems show that the optimality holds even when the variances are unknown and must be estimated.

Theorem 3. *Suppose that Assumptions 1, 2, 3 4, 6, 7, 8, 9, and 10 hold. Then, as* $n \to \infty$*:*

$$\frac{L_{I_n}^F(\tilde{W}_{I_n})}{\inf_{W \in \mathcal{H}_n(N)} L_{I_n}(W)} \to_p 1,$$

where $\tilde{W}_{I_n} = \arg\min_{W \in \mathcal{H}_M(N)} C_{I_n}^F(W)$.

Theorem 4. *Suppose that Assumptions 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 hold. Then, as* $n \to \infty$*:*

$$\frac{L_{\Omega}^F(\tilde{W}_{\Omega})}{\inf_{W \in \mathcal{H}_n(N)} L_{\Omega}(W)} \to_p 1,$$

where $\tilde{W}_{\Omega} = \arg\min_{W \in \mathcal{H}_M(N)} C_{\Omega}^F(W)$.

The proofs of these theorems are in the Appendix. The main part of the proof is to show that the estimation errors in the variances do not affect the squared errors asymptotically. The additional assumptions are used for this purpose. The proofs are new and they are the main technical contribution of this paper. Note that Theorem 4 uses Assumptions 2 and 3, which are not needed for Theorem 2. This is because the proof of Theorem 4 uses some results from the proofs of Theorems 1 and 3. These assumptions can be relaxed at the cost of lengthening the proofs greatly.

These theorems are remarkable in the sense that the squared error attained by Mallows' criteria with estimated variances is equivalent asymptotically to the minimum of the loss function with true variance. In other words, even if we need to estimate the variance, we can achieve the minimum loss that is computed by assuming that the variances are known. This result provides strong support for the use of the FGLSMA estimator. The cost is that we need to impose additional assumptions whose main practical implication is that the number of regressors should be small. We investigate to what extent this cost is important in the simulations.

4 Monte Carlo studies

We conduct Monte Carlo simulations to investigate the finite sample performance of our method and compare it with those of existing methods. In particular, we examine the performance of the FGLSMA estimator.

4.1 Design

The following simulation design is almost the same as that of Liu and Okui (2012). We draw a random sample of $\{x_i, e_i\}$ for each replication such that $x_i = (x_{i1}, \ldots, x_{i10000})$, $x_{i1} = 1$ and other x_{ij}s are independent over j with $x_{ij} \sim N(0,1)$. The error term e_i is independent of x_{ij} for any j and $e_i \sim N(0, \sigma_i^2)$, where $\sigma_i^2 = x_{2i}^4 + 0.01$. This specification of variances is similar to that of Hansen and Racine (2010), but we add 0.01 in order to guarantee that the variances are strictly positive. The dependent variable y_i is generated by:

$$y_i = \sum_{j=1}^{10000} \theta_j x_{ij} + e_i,$$

where the parameters are specified as $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$. The parameter α, which determines how quickly the magnitude of θ_j decays as j increases, is set at $\alpha = 0.5$ and we vary the values of c so that the population R^2 increases with c from 0.1 to 0.9. We note that the population R^2 is defined as $R^2 = (var(y_i) - var(e_i))/var(y_i)$ and it is $(\sum_{j=1}^{\infty} \theta_j^2)/(\sum_{j=1}^{\infty} \theta_j^2 + var(e_i))$, where $var(e_i)$ is the unconditional variance and $var(e_i) = E(\sigma_i^2) = 3.01$ in our case. The sample size is $n = 150$. The number of observable regressors is $K = 10, 20, 30$ and 40. We consider K different models so that $M = K$ and $k_m = m$ for any m. The kth model includes the first k regressors and the $(k+1)$th model nests the kth model. The number of replications is 1000.

To compute the FGLSMA estimator, we need to estimate the variance matrix Ω. We consider the following model:

$$\sigma_i^2 = \beta_0 + \beta_1 x_{i2}^2 + \beta_2 x_{i2}^4.$$

The parameters $(\beta_0, \beta_1\beta_2)$ are estimated by maximum likelihood using the model whose mean function includes all the available regressors. We then obtain the variance estimate for i by $\hat{\sigma}_i^2 = (\hat{\beta}_0 + \hat{\beta}_1 x_{i2}^2 + \hat{\beta}_2 x_{i2}^4)(n/(n-K))$, where $n/(n-K)$ is the degrees of freedom correction because we essentially estimate the variance using the residuals from the regression with K regressors. We note that this specification is correct in the sense that the model includes the data-generating process by setting $(\beta_0, \beta_1, \beta_2) = (0.01, 0, 1)$. The FGLSMA estimator is constructed by computing K different FGLS estimators and averaging them using the weights chosen by minimizing $C_{I_n}^F(W)$ or $C_{\Omega}^F(W)$. We use the same variance estimates (which are estimated from the largest model) for all the FGLS estimators to be averaged and to compute the criteria.

We also compute the following six alternative model averaging estimators. The generalized C_p (GC) method by Liu and Okui (2012), the weighted average least squares (WALS) estimator by Magnus, Wan and Zhang (2011), the jackknife model averaging (JMA) method by Hansen and Racine (2010), Mallows' model averaging (MMA) procedure by Hansen (2007) and the estimator based on the smoothed Akaike information criterion (SAIC) and the smoothed Bayesian information criterion (SBIC) by Buckland, Burnham and Augustin (1997).[2] We note that MMA, SAIC and SBIC are not designed for heteroskedastic data. GC and JMA are robust to heteroskedasticity, but they average least squares, not GLS, estimators. The WALS estimator is based on the FGLS estimator, but its averaging procedure and averaging objective are different from ours. For the WALS estimator, we use the same model for the variance as that for the FGLSMA estimator.

The measure of performance of each estimator is the MSE. Let μ^r be the vector of the true value of μ in the rth replication and $\hat{\mu}^{(r)}$ be the vector of the estimator of $\mu^{(r)}$. The MSE is computed as $\sum_{r=1}^{1000}(\hat{\mu}^{(r)} - \mu^{(r)})'(\hat{\mu}^{(r)} - \mu^{(r)})/1000 = \sum_{r=1}^{1000}\sum_{i=1}^{n}(\hat{\mu}_i^{(r)} - \mu_i^{(r)})^2/1000$. The sample weighted mean squared error (WMSE) defined as $WMSE = \sum_{r=1}^{1000}\sum_{i=1}^{n}((\hat{\mu}_i^{(r)} - \mu_i^{(r)})^2/\sigma_i^2)/1000$ is used for the comparison with the FGLSMA estimator based on $C_{\Omega}^F(W)$. We report the ratios of the MSEs and WMSEs that are calculated using the MSE and WMSE of the FGLSMA estimators based on $C_{I_n}^F(W)$ and $C_{\Omega}^F(W)$ as the denominator, respectively.

4.2 Results

The simulation results are summarized in Figures 1 and 2. Figure 1 plots the sample MSE ratios against the population R^2 for the comparison with $C_{I_n}^F(W)$, and Figure 2 plots the sample WMSE ratios against the population R^2 for the comparison with $C_{\Omega}^F(W)$.

The FGLSMA estimator performs remarkably well, particularly when the cardinality of the set of models is large and/or the population R^2 is small. This result illustrates that

[2]SAIC was originally suggested by Akaike (1979) and extended by Buckland, Burnham and Augustin (1997).

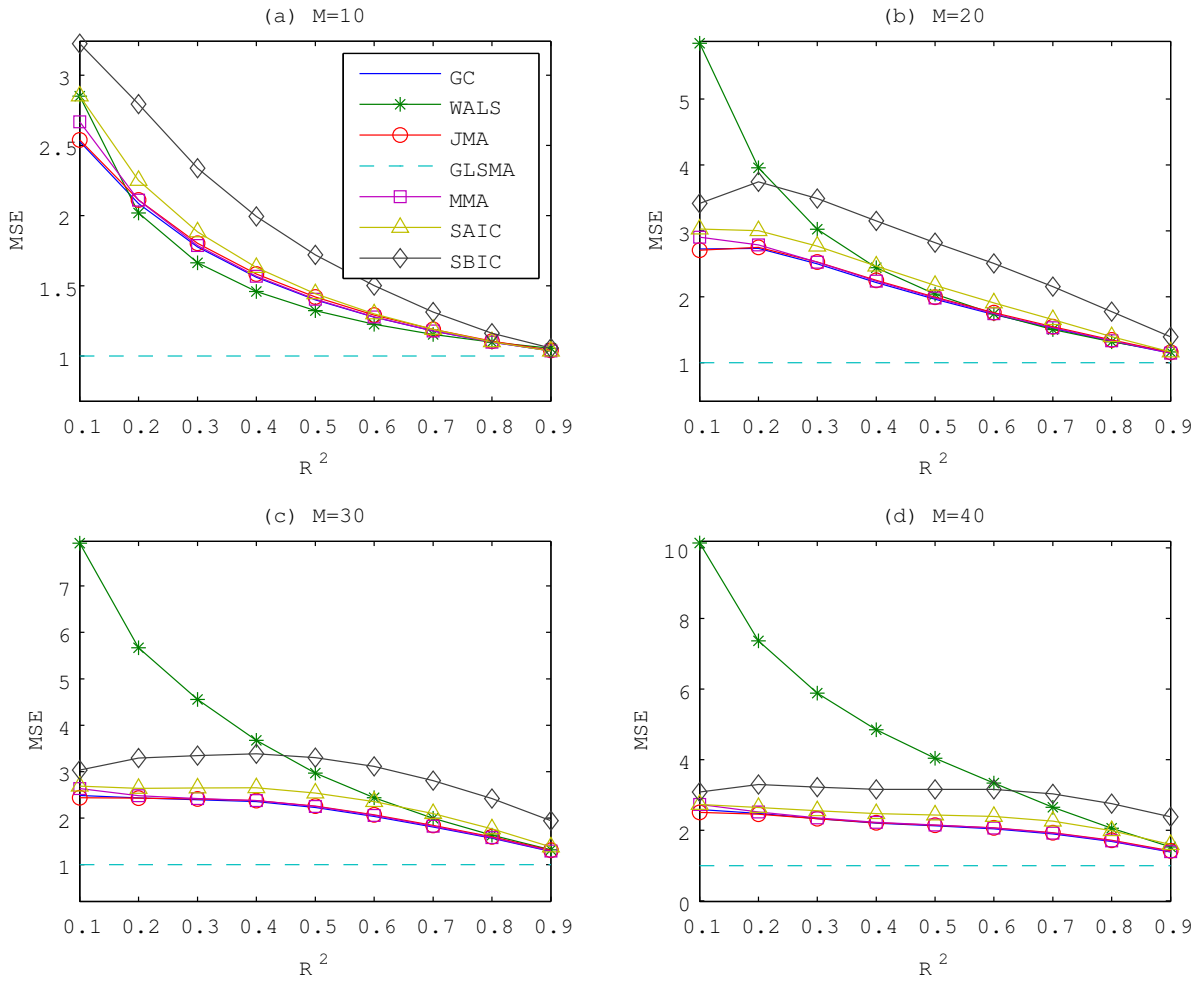

Figure 1: Performance of the FGLSMA estimator based on $C_{I_n}^F(W)$ and alternative estimators in the heteroskedastic design with $n = 150$.

the GLS procedure effectively reduces the variability of the estimator and provides a much more precise estimate. Moreover, the result indicates that the FGLSMA estimator works well even when the variances are estimated and the number of regressors is not small, while the asymptotic theory that supports the use of the FGLSMA estimator imposes a rather stringent condition on the number of regressors. Among the alternative procedures, the performance of SAIC and SBIC is not encouraging. This result is not surprising because these procedures are made for homoskedastic data but here we use heteroskedastic data. It is somewhat surprising that the performance of MMA is comparable to that of GC or JMA. WALS does not perform well particularly when R^2 is small and M is large. Note that the theory of WALS assumes that M is fixed so it is designed for situations with a small M. Our procedure outperforms WALS even when $M = 10$. Note that the WALS estimator is developed with a different objective (not minimizing the MSE of the fit) and it may be natural that its performance is inferior to our procedure, which minimizes the estimate of the MSE, when the measure of performance is the MSE.

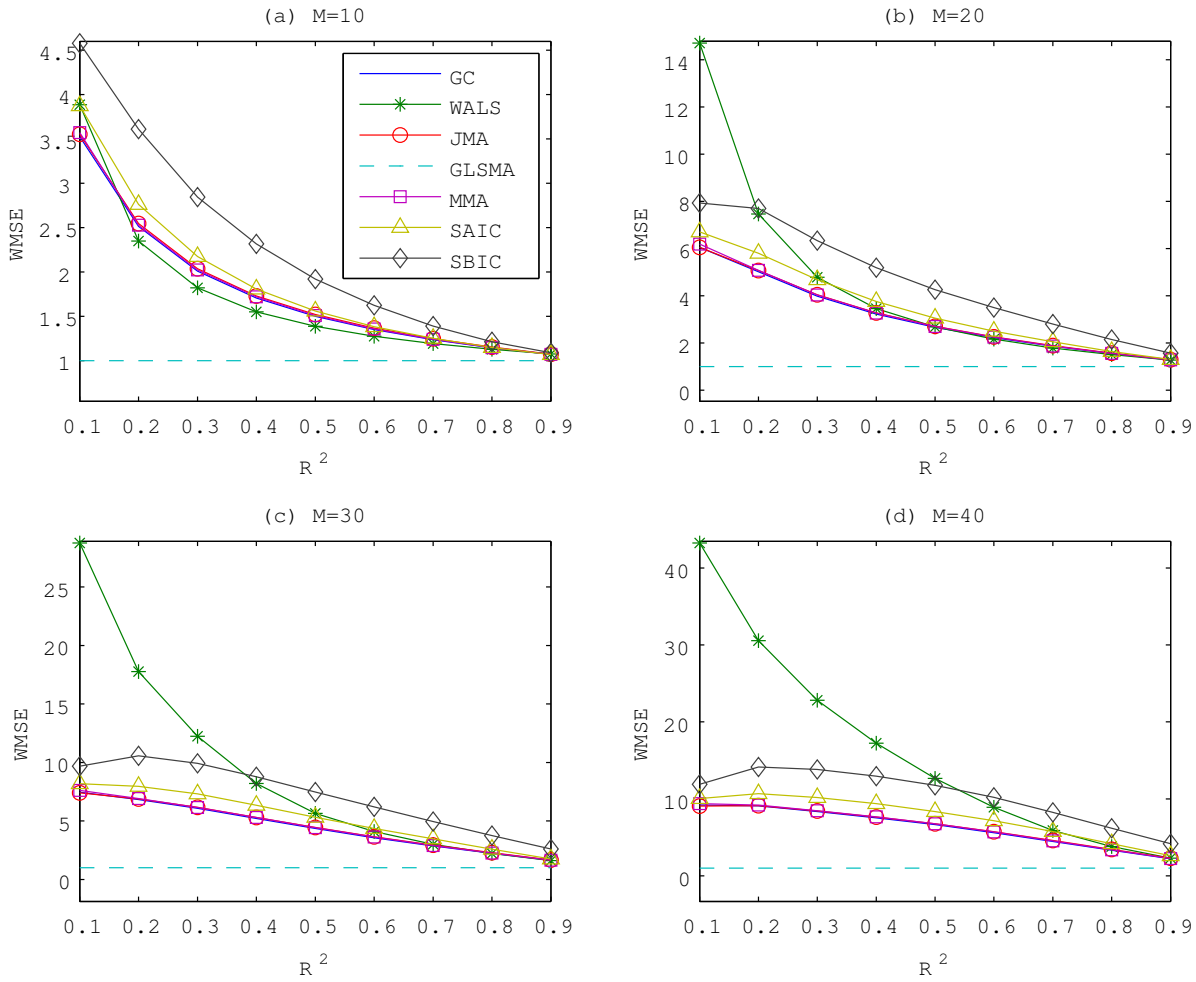

Figure 2: Performance of the FGLSMA estimator based on $C_{\Omega}^{F}(W)$ and alternative estimators in the heteroskedastic design with $n = 150$.

4.3 Homoskedastic cases

In this section, we present the results of additional simulations in which the error terms are homoskedastic. The purpose of this simulation exercise is to evaluate a possible loss in terms of the MSE associated with using our FGLSMA estimator when there is no need to use the FGLS estimators. The design of the experiments is the same as the one presented in Section 4.1 except that we now set $\sigma_i^2 = 1$ for any i. We compare the FGLSMA estimator with MMA, SAIC, SBIC and WALS.

The results are summarized in Figures 3 and 4. It is remarkable that the performance of the FGLSMA estimator is similar to that of MMA in most of the cases. When M is large and R^2 is small, MMA performs noticeably better than FGLSMA based on $C_{\Omega}(W)$, but they are almost indistinguishable when $C_{I_n}(W)$ is used. This result indicates that the estimation of the variance does not worsen performance significantly, even when the estimation is not necessary. Other procedures do not perform well compared with MMA and FGLSMA in terms of the MSE in many cases.

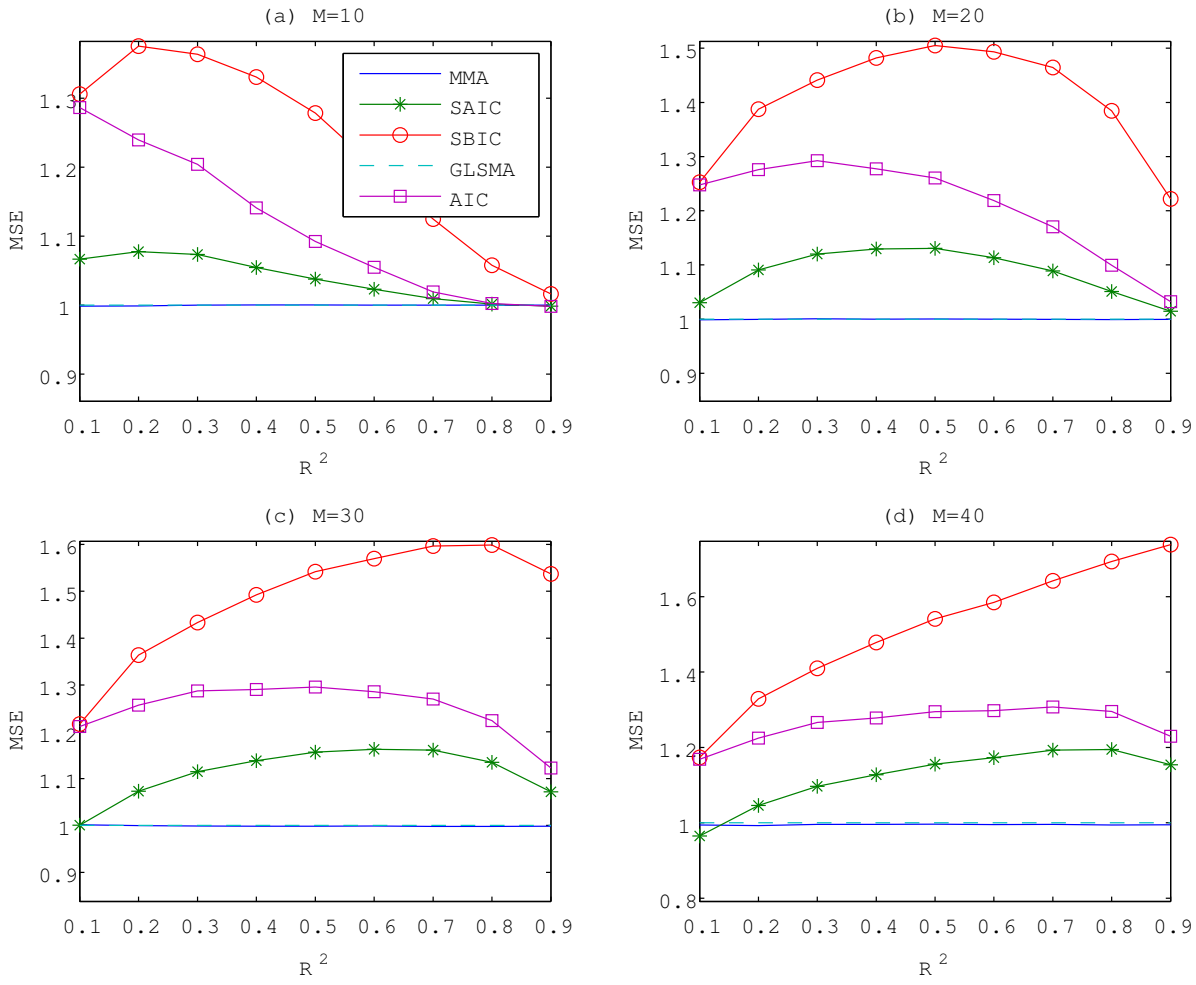

Figure 3: Performance of the FGLSMA estimator based on $C^F_{I_n}(W)$ and alternative estimators in the homoskedastic design with $n = 150$.

In summary, we observe that using the FGLSMA estimator may not cause a serious loss even when homoskedasticity is plausible as long as M is not very large or R^2 is not very small.

4.4 Nonparametric variance estimation

Finally, we examine the performance of the FGLSMA estimator based on nonparametric variance estimates. We compute the semiparametric GLS estimators proposed by Robinson (1987) and average them using Mallows' criteria. This procedure is not supported by the theory presented in this paper because the variance estimators are not $\sqrt{n}$ consistent. Nonetheless, it is sometimes difficult to specify the model correctly for the variances and it would be worthwhile to examine the performance of the procedure that does not require correct modeling of the variance function. We use the same data-generating process as that in Section 4.1.

The variances are estimated by the kth nearest neighbor (k-NN) estimator as sug-

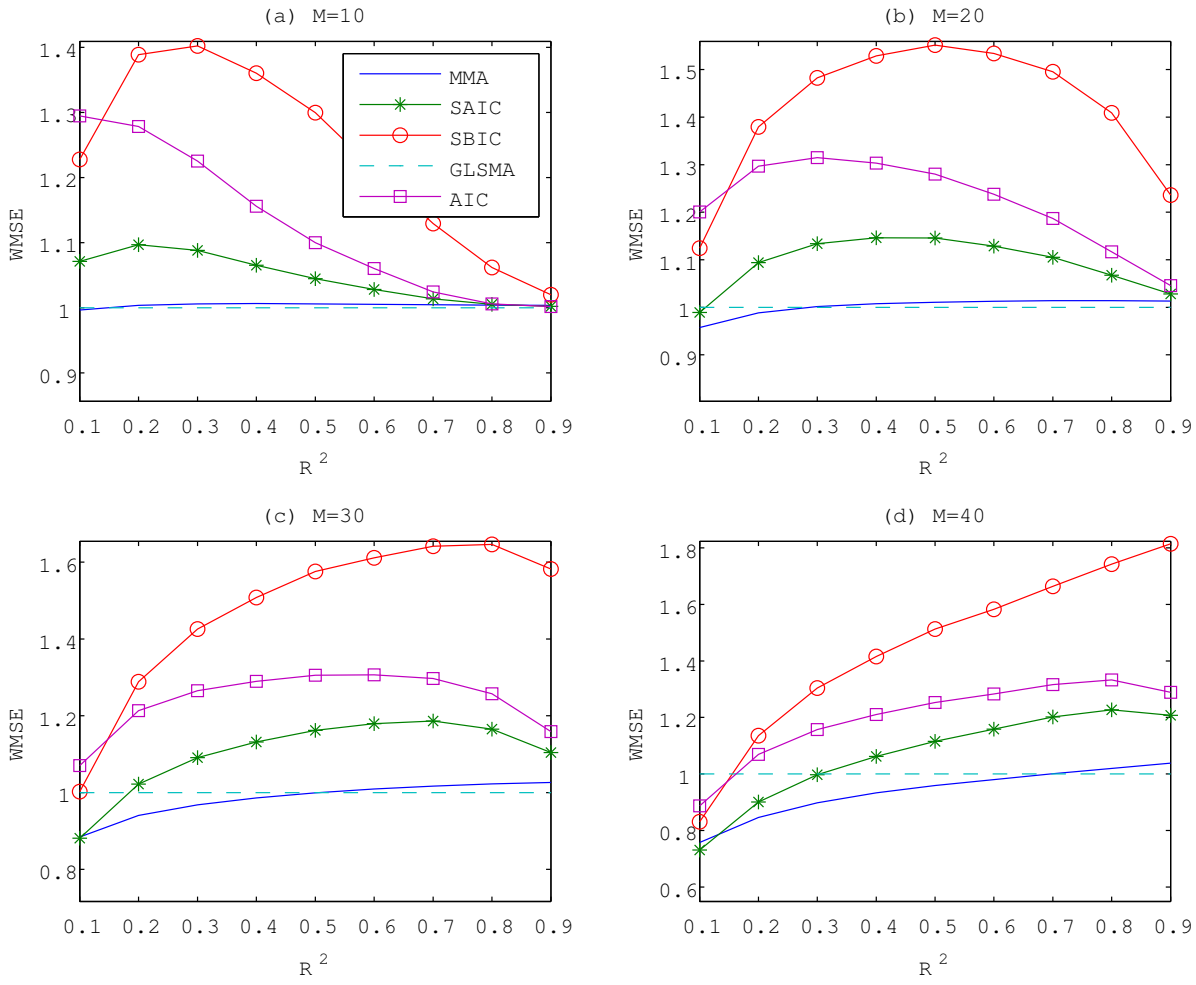

Figure 4: Performance of the FGLSMA estimator based on $C_{\Omega}^{F}(W)$ and alternative estimators in the homoskedastic design with $n = 150$.

gested by Robinson (1987). We estimate the largest model by least squares and use the squared residuals from this regression as the dependent variable when we estimate the variance function nonparametrically. We use all available regressors as regressors in the nonparametric estimation. For k-NN, we need to determine the tuning parameter, called the k nearest neighbors and denoted as k_{NN}, used in the estimation. We use the leave-one-out method to select k_{NN}. As a value of k_{NN} that is too large will not give us a good estimate, we restrict $1 \leq k_{NN} \leq 11$. We employ the triangular k-NN weights defined in Robinson (1987) to perform k-NN estimation.[3]

The summarized results are presented as Figures 5 and 6. Note that the WALS estimator is not based on the nonparametric variance estimates and is the same as that examined above. The results are encouraging. The FGLSMA estimators still perform well

[3]To simplify the derivation of the properties of the estimator, Robinson (1987) does not use the ith observation for the estimation of the ith diagonal element of the variance matrix. We find that this type of data split technique causes poor performance of the estimator. Hence we include the ith observation in the estimation.

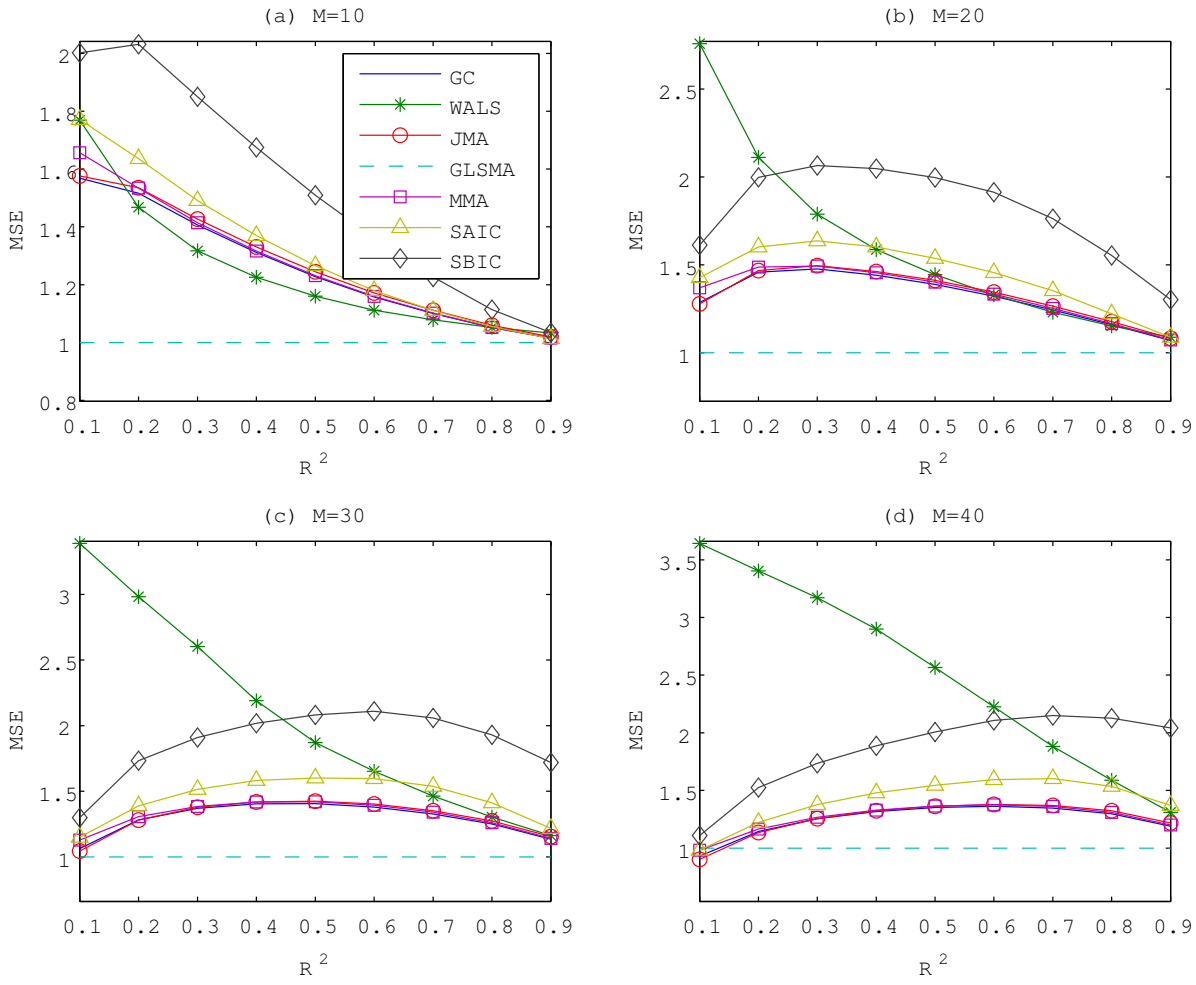

Figure 5: Performance of the FGLSMA estimator with nonparametric variance estimates based on $C_{I_n}^F(W)$ and alternative estimators in the heteroskedastic design with $n = 150$.

even when the variances are estimated nonparametrically. The performance is inferior to that where the variances are estimated at the parametric rate. (Note that the scales of the horizontal axes are different from those in Figures 1 and 2.) Nonetheless, it outperforms other procedures.

We also examine the performance of the FGLSMA estimator with nonparametric variance estimates in homoskedastic cases using the design in Section 4.3. Figures 7 and 8 summarize the results. When the errors are homoskedastic, we see that unnecessarily estimating the variances can deteriorate the performance of the estimators. In particular, the performance of the FGLSMA estimator is substantially worse than those of other estimators when $M = 30$ and $M = 40$ and R^2 is small. Nonetheless, when $M = 10$, the performance can be comparable to MMA and can be better than other estimators. Therefore, the FGLSMA estimator based on nonparametrically estimated variances may be considered when M is small and its theoretical justification would be a promising area of future research.

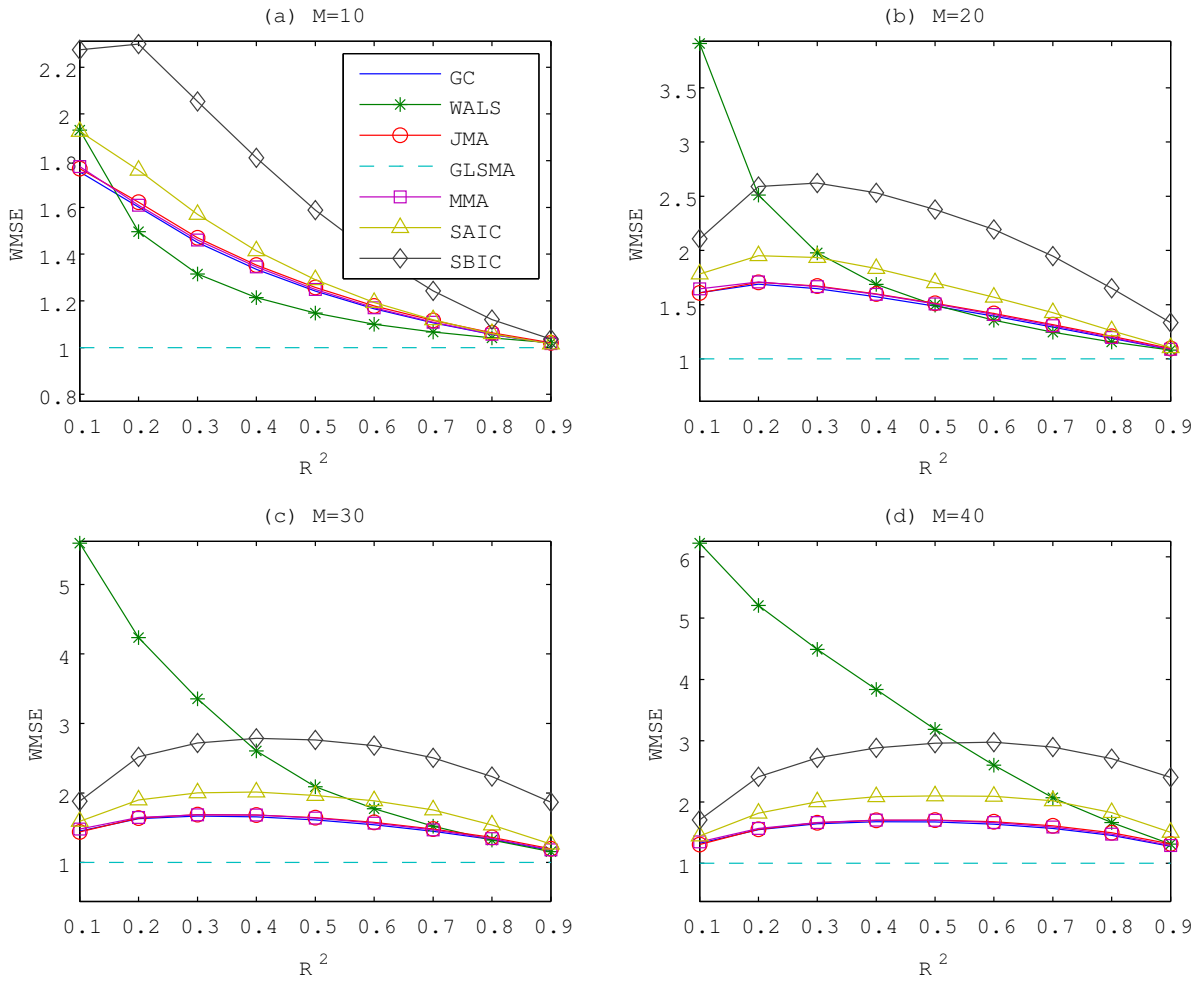

Figure 6: Performance of the FGLSMA estimator with nonparametric variance estimates based on $C_{\Omega}^{F}(W)$ and alternative estimators in the heteroskedastic design with $n = 150$.

5 Conclusion

This paper proposes a model averaging estimator for linear regression models with possibly heteroskedastic errors. The estimator is a weighted average of the GLS estimators. The weighting vector is chosen by minimizing Mallows' criteria, calculated from the estimates of the MSE of the fit of the estimator. The criteria are shown to achieve asymptotic optimality in the sense that the squared error evaluated at the chosen weight is asymptotically indistinguishable from the minimum of the squared error. We also consider a weighted average of the FGLS estimators, derive Mallows' criteria for it and prove its asymptotic optimality. The Monte Carlo simulations show that our methods work well compared with existing procedures.

There are several promising areas for future research. The Monte Carlo simulations indicate that the model averaging estimator based on the FGLS estimators with nonparametrically estimated variances can perform well. As such an estimator is not supported by the theoretical argument provided in this paper, some theoretical work on it would be

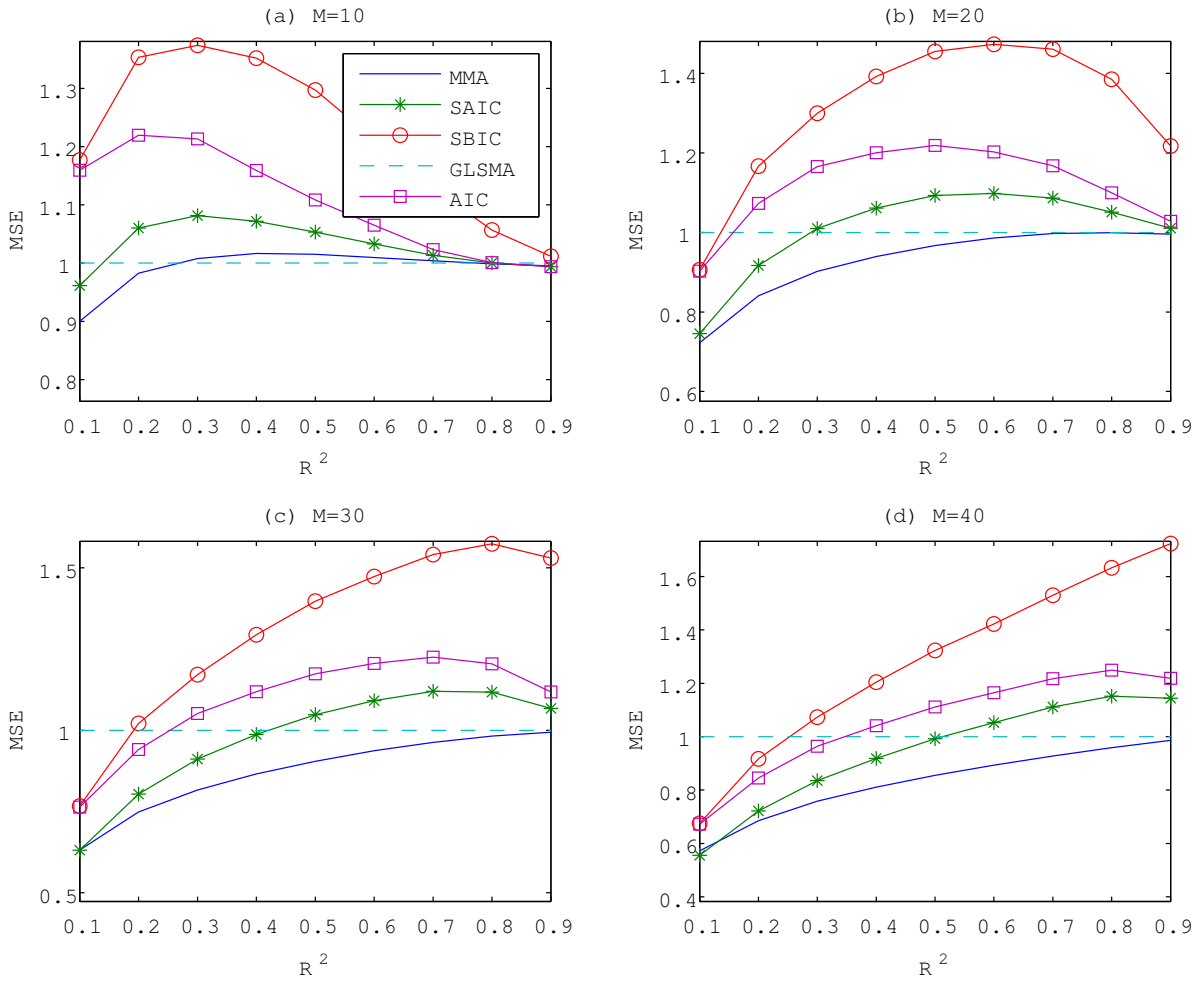

Figure 7: Performance of the FGLSMA estimator with nonparametric variance estimates based on $C^F_{I_n}(W)$ and alternative estimators in the homoskedastic design with $n = 150$.

an interesting research topic. Other future research topics include an extension to time series data analysis, an extension to nonlinear models such as probit models or hazard models and an extension to models defined by moment conditions.

A Mathematical appendix

This appendix presents the proofs of the theorems in the paper. Let $||\cdot||$ denote the Euclidean norm so that $||\cdot||_{I_n} = ||\cdot||$. Let $\sup_W$ be an abbreviation of $\sup_{W \in \mathcal{H}_M(N)}$.

A.1 Proof of Theorem 1

Proof. The proof consists of two steps: first, we propose the sufficient conditions for the results of Theorem 2.1 of Li (1987) to be applied to the current case, and second, we show that these sufficient conditions are met in our setup following the same steps as those for the proof of Theorem 1 of Hansen (2007).

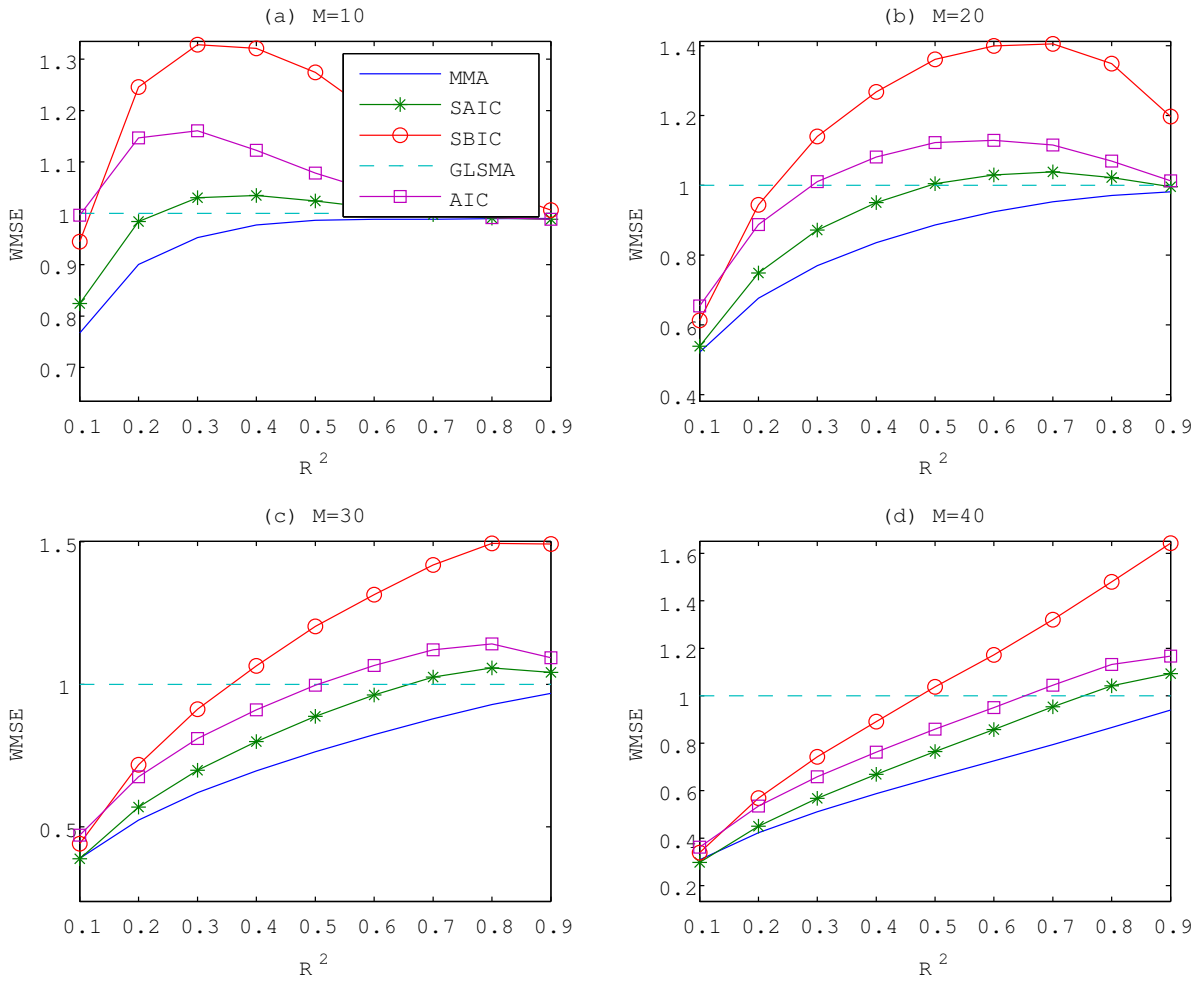

Figure 8: Performance of the FGLSMA estimator with nonparametric variance estimates based on $C_{\Omega}^{F}(W)$ and alternative estimators in the homoskedastic design with $n = 150$.

Step 1 : First, we show that the following three conditions are sufficient to prove the desired result:

$$\lim_{n\to\infty} \sup_{W} \lambda(G(W)'G(W)) < \infty, \tag{3}$$

$$E(|e_i|^{4(N+1)}) \leq \kappa < \infty, \tag{4}$$

$$\sum_{W\in\mathcal{H}_M(N)} R_{In}(W)^{-(N+1)} \to 0 \quad \text{as} \quad n \to \infty. \tag{5}$$

We observe:

$$C_{I_n}(W) = \|e\|^2 + L_{I_n}(W) + 2\langle e, (I - G(W))\mu\rangle + 2\left(tr(G(W)\Omega) - \langle e, G(W)e\rangle\right)$$

and the first term does not depend on a weight vector W. Lemma A.9 of Donald and

Newey (2001) tells us that it is sufficient to show that:

$$\sup_W \frac{\langle e, (I - G(W))\mu\rangle}{R_{I_n}(W)} \to_p 0, \tag{6}$$

$$\sup_W \frac{|tr(G(W)\Omega) - \langle e, G(W)e\rangle|}{R_{I_n}(W)} \to_p 0, \tag{7}$$

$$\sup_W \left| \frac{L_{I_n}(W)}{R_{I_n}(W)} - 1 \right| \to_p 0. \tag{8}$$

For (6), using the Bonferroni inequality and the Chebyshev inequality, we have, for any $\delta > 0$:

$$\begin{aligned} & P\left(\sup_W |\langle e, (I - G(W))\mu\rangle|/R_{I_n}(W) > \delta\right) \\ \leq & \sum_{W\in\mathcal{H}_M(N)} P\left(|\langle e, (I - G(W))\mu\rangle|/R_{I_n}(W) > \delta\right) \quad \text{(Bonferroni)} \\ \leq & \sum_{W\in\mathcal{H}_M(N)} \frac{E(\langle e, (I - G(W)\mu\rangle^{2(N+1)})}{R_{I_n}(W)^{2(N+1)}\delta^{2(N+1)}}. \quad \text{(Chebyshev)} \end{aligned}$$

Theorem 2 of Whittle (1960) implies that $E(\langle e, (I-G(W)\mu\rangle^{2(N+1)}) \leq C\|(I-G(W))\mu\|^{2(N+1)}$. As we can write $R_{I_n}(W) = \|\mu - G(W)\mu\|^2 + tr(G(W)'G(W)\Omega)$, we have $\|(I-G)(W)\mu\|^2 \leq R_{I_n}$. Thus, it holds that:

$$\sum_{W\in\mathcal{H}_M(N)} \frac{E(\langle e, (I - G(W)\mu\rangle^{2(N+1)})}{R_{I_n}(W)^{2(N+1)}\delta^{2(N+1)}} \leq C\delta^{-2(N+1)} \sum_{W\in\mathcal{H}_M(N)} R_{I_n}(W)^{-(N+1)} \to 0.$$

Similarly, for (7), the Bonferroni inequality, the Chebyshev inequality and Theorem 2 of Whittle (1960)) imply:

$$\begin{aligned} & P\left(\sup_W \frac{|tr(G(W)\Omega) - \langle e, G(W)e\rangle|}{R_{I_n}(W)} > \delta\right) \\ \leq & \frac{C}{\delta^{2(N+1)}} \sum_{W\in\mathcal{H}_M(N)} \frac{E\left(tr(G(W)\Omega) - \langle e, G(W)e\rangle\right)^{2(N+1)}}{R_{I_n}(W)^{2(N+1)}} \\ \leq & \frac{C}{\delta^{2(N+1)}} \sum_{W\in\mathcal{H}_M(N)} \frac{tr(G(W)'G(W))^{(N+1)}}{R_{I_n}(W)^{2(N+1)}}. \end{aligned}$$

Let γ_i be the ith diagonal element of $G(W)'G(W)$. Noting that $\gamma_i \geq 0$, it follows that:

$$tr(G(W)'G(W)) = \sum_{i=1}^n \gamma_i \leq \sum_{i=1}^n \gamma_i \frac{\sigma_i^2}{\inf_j \sigma_j^2} = (\inf_j \sigma_j^2)^{-1} tr(G(W)'G(W)\Omega).$$

Moreover, the definition of $R_{I_n}(W)$ gives $tr(G(W)'G(W)\Omega) \leq R_{I_n}(W)$. Thus, we have:

$$\begin{aligned}
& \frac{C}{\delta^{2(N+1)}} \sum_{W \in \mathcal{H}_M(N)} \frac{tr(G(W)'G(W))^{(N+1)}}{R_{I_n}(W)^{2(N+1)}} \\
\leq & \frac{C}{\delta^{2(N+1)}} \sum_{W \in \mathcal{H}_M(N)} \left(\inf_i \sigma_i^2\right)^{-(N+1)} \frac{tr(G(W)'G(W)\Omega)^{(N+1)}}{R_{I_n}(W)^{2(N+1)}} \\
\leq & \frac{C}{\delta^{2(N+1)}} \left(\inf_i \sigma_i^2\right)^{-(N+1)} \sum_{W \in \mathcal{H}_M(N)} R_{I_n}(W)^{-(N+1)} \to 0.
\end{aligned}$$

To prove (8), we observe:

$$\begin{aligned}
L_{I_n}(W) &= \|\mu - \hat{\mu}(W)\|^2 \\
&= \|(I - G(W))\mu\|^2 + tr(G(W)'G(W)\Omega) + \|G(W)e\|^2 \\
&\quad -tr(G(W)'G(W)\Omega) - 2\langle (I - G(W))\mu, G(W)e\rangle \\
&= R_{I_n}(W) + \|G(W)e\|^2 - tr(G(W)'G(W)\Omega) - 2\langle (I - G(W))\mu, G(W)e\rangle.
\end{aligned}$$

Thus, if we show that:

$$\sup_W \frac{|\|G(W)e\|^2 - tr(G(W)'G(W)\Omega)|}{R_{I_n}(W)} \to_p 0, \tag{9}$$

$$\sup_W \frac{|\langle (I - G(W))\mu, G(W)e\rangle|}{R_{I_n}(W)} \to_p 0, \tag{10}$$

we have (8). Using the same argument as for proving (6) and (7), Chebyshev's inequality, Theorem 2 of Whittle (1960), $\|G(W)e\|^2 = \langle e, G(W)'G(W)e\rangle$ and

$$\begin{aligned}
tr(G(W)'G(W)G(W)'G(W)) &\leq \lambda(G(W)'G(W))tr(G(W)'G(W)) \\
&\leq CR_{I_n}(W)
\end{aligned}$$

implies (9). Similarly to (9), because $\langle (I - G(W))\mu, G(W)e\rangle = \langle G(W)'(I - G(W))\mu, e\rangle$ and

$$\begin{aligned}
\|G(W)'(I - G(W))\mu\|^2 &\leq \lambda(G(W)'G(W))\|(I - G(W))\mu\|^2 \\
&\leq CR_{I_n}(W),
\end{aligned}$$

(10) holds. This completes Step 1.

Step 2 : We show that three conditions, (3), (4) and (5), are satisfied in our setup. As (3) and (4) hold by the assumptions of the theorem, we only need to prove (5). Let $W_{j_1,\ldots,j_N} \in \mathcal{H}_M(N)$ denote the discrete weight set for some fixed integer $N < \infty$ as in Hansen (2007). More precisely, $(j_i, \ldots, j_N)$ is the set of integers satisfying $1 \leq j_1 \leq j_2 \leq$

$\cdots \leq j_N$, and the j_i element of $W_{j_1,\ldots,j_N} = 1/N$ and the other elements are zero. We then observe that:

$$\begin{aligned} R_{I_n}(W_{j_1,\ldots,j_N}) &\geq tr(G(W_{j_1,\ldots,j_N})\Omega G(W_{j_1,\ldots,j_N})') \\ &= \sum_{m=1}^{N}\sum_{l=1}^{N} N^{-2} tr\left(X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}X_{j_m}'\Omega^{-1}X_{j_l}(X_{j_l}'\Omega^{-1}X_{j_l})^{-1}X_{j_l}'\right). \end{aligned}$$

If $j_m \leq j_l$, we can write $X_{j_m} = X_{j_l}F$ for some matrix F and it holds that:

$$\begin{aligned} & X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}X_{j_m}'\Omega^{-1}X_{j_l}(X_{j_l}'\Omega^{-1}X_{j_l})^{-1}X_{j_l}' \\ = \; & X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}F'X_{j_l}'\Omega^{-1}X_{j_l}(X_{j_l}'\Omega^{-1}X_{j_l})^{-1}X_{j_l}' \\ = \; & X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}F'X_{j_l}' \\ = \; & X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}X_{j_m}'. \end{aligned}$$

Let $\tilde{\gamma}_i$ be the ith diagonal element of $X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}X_{j_m}'$. Note that $\tilde{\gamma}_i \geq 0$. Thus, when $j_m \leq j_l$ we have:

$$\begin{aligned} tr(X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}X_{j_m}') &= \sum_{i=1}^{n}\tilde{\gamma}_i \\ &\geq \sum_{i=1}^{n}\tilde{\gamma}_i\frac{\sigma_i^{-2}}{\sup_i \sigma_i^{-2}} \\ &= \left(\inf_i \sigma^2\right) tr(X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}X_{j_m}'\Omega^{-1}) \\ &= \left(\inf_i \sigma^2\right) k_{j_m}. \end{aligned}$$

Thus, we observe that:

$$\begin{aligned} & \sum_{m=1}^{N}\sum_{l=1}^{N} N^{-2} tr\left(X_{j_m}(X_{j_m}'\Omega^{-1}X_{j_m})^{-1}X_{j_m}'\Omega^{-1}X_{j_l}(X_{j_l}'\Omega^{-1}X_{j_l})^{-1}X_{j_l}'\right) \\ \geq \; & (\inf_i \sigma_i^2)\sum_{m=1}^{N}\sum_{l=1}^{N} N^{-2}\min(k_{j_m}, k_{j_l}) \\ \geq \; & (\inf_i \sigma_i^2)N^{-2}k_{j_N} \\ \geq \; & (\inf_i \sigma_i^2)N^{-2}j_N. \end{aligned}$$

If we replace the bound $R_n(W_{j_1,\ldots,j_N}) \geq \sigma^2 j_N/N^2$ in the proof of Theorem 1 of Hansen (2007) with the bound above, the proof of (5) is the same as that of Theorem 1 of Hansen (2007). This completes the proof. □

A.2 Lemmas for the proofs of Theorems 3 and 4

Let $\|\cdot\|$ denote the Euclidean norm ($\|A\| = \sqrt{tr(A'A)}$) and $\|\cdot\|_1$ denote the Banach norm ($\|A\|_1 = \sup_{x\neq 0}(\|Ax\|/\|x\|)$) of the matrices. See Wiener and Masani (1958) and Lewis and Reinsel (1985) for properties of these norms. In particular, we use the following result: $\|AB\| \leq \|A\|_1\|B\|$.

Lemma 1. *Suppose that Assumptions 6 and 9 hold. Then, as $n \to \infty$:*

$$\sup_m \left\| \frac{1}{n} X_m' \hat{\Omega}^{-1} X_m - \frac{1}{n} X_m' \Omega^{-1} X_m \right\| = O_p\left(\frac{k_M}{\sqrt{n}} \right).$$

Proof. We observe that:

$$\begin{aligned} \left\| \frac{1}{n} X_m' \hat{\Omega}^{-1} X_m - \frac{1}{n} X_m' \Omega^{-1} X_m \right\| &= \left\| \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_i^2} \right) x_{mi} x_{mi}' \right\| \\ &\leq \sup_i \left| \frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_i^2} \right| \frac{1}{n} \sum_{i=1}^n \left\| x_{mi} x_{mi}' \right\|. \end{aligned}$$

As Assumption 9 gives:

$$\sup_i \left| \frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_i^2} \right| = O_p\left(\frac{1}{\sqrt{n}} \right)$$

and Assumption 6 implies:

$$\sup_m \frac{1}{n} \sum_{i=1}^n \left\| x_{mi} x_{mi}' \right\| = \sup_m \frac{1}{n} \sum_{i=1}^n \sqrt{tr(x_{mi} x_{mi}' x_{mi} x_{mi}')} \leq \sup_m \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k_m} x_{m,j,1}^2 = O(k_M),$$

we have the desired result. □

Lemma 2. *Suppose that Assumptions 4, 6, 7 and 9 hold. Then, if $n \to \infty$ and $k_M^2/n \to 0$, then:*

$$\sup_m \left\| \left(\frac{1}{n} X_m' \hat{\Omega}^{-1} X_m \right)^{-1} - \left(\frac{1}{n} X_m' \Omega^{-1} X_m \right)^{-1} \right\|_1 = O_p\left(\frac{k_M}{\sqrt{n}} \right) \tag{11}$$

and

$$\sup_m \left\| \left(\frac{1}{n} X_m' \hat{\Omega}^{-1} X_m \right)^{-1} \right\|_1 = O_p(1). \tag{12}$$

Proof. Let

$$\hat{A}_m = \frac{1}{n} X_m' \hat{\Omega}^{-1} X_m, \text{ and } A_m = \frac{1}{n} X_m' \Omega^{-1} X_m.$$

As

$$A_m = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_i^2} x_{mi} x_{mi}',$$

Assumptions 4 and 7 imply that $\sup_m ||A_m^{-1}||_1 \leq F < \infty$ for some constant F. Similarly to Theorem 1 of Lewis and Reinsel (1985) or Lemma 3 of Berk (1974), we can write:

$$\hat{A}_m^{-1} - A_m^{-1} = \hat{A}_m^{-1}(\hat{A}_m - A_m)A_m^{-1} = \left[A_m^{-1} + \left(\hat{A}_m^{-1} - A_m^{-1} \right) \right] (\hat{A}_m - A_m)A_m^{-1},$$

and thus:

$$\begin{aligned} ||\hat{A}_m^{-1} - A_m^{-1}||_1 &\leq \left(||A_m^{-1}||_1 + ||\hat{A}_m^{-1} - A_m^{-1}||_1\right) ||\hat{A}_m - A_m||_1 ||A_m^{-1}||_1 \\ &\leq \left(F + ||\hat{A}_m^{-1} - A_m^{-1}||_1\right) ||\hat{A}_m - A_m||_1 F. \end{aligned}$$

We note that $\sup_m ||\hat{A}_m - A_m||_1 \leq \sup_m ||\hat{A}_m - A_m|| \to_p 0$ by Lemma 1 and the conditions of this lemma. Thus, with probability approaching one, we have $F||\hat{A}_m - A_m||_1 < \gamma < 1$ for any m and for some γ:

$$\sup_m ||\hat{A_m}^{-1} - A_m^{-1}||_1 \leq \sup_m \frac{F^2||\hat{A}_m - A_m||_1}{1 - F||\hat{A}_m - A_m||_1}.$$

The above inequality implies (11).

(12) follows because:

$$\sup_m ||(\hat{A}_m)^{-1}||_1 \leq \sup_m ||A_m^{-1}||_1 + \sup_m ||(\hat{A}_m)^{-1} - A_m^{-1}||_1 \leq F + o_p(1) = O_p(1),$$

where the first inequality is the triangular inequality and the second inequality follows from the assumption of the lemma. □

Lemma 3. *Suppose that Assumptions 4, 6, 7 and 9 hold. As $n \to \infty$ and $k_M^2/n \to 0$, it holds that*

$$\begin{aligned} &\sup_W \left| tr\left(\sum_{m=1}^M w_m X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m'\right) - tr\left(\sum_{m=1}^M w_m X_m (X_m' \Omega^{-1} X_m)^{-1} X_m'\right) \right| \\ = \; & O_p\left(\frac{k_M^2}{\sqrt{n}}\right). \end{aligned}$$

Proof. We observe that:

$$\begin{aligned} &\sup_m \left| tr\left(X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m'\right) - tr\left(X_m (X_m' \Omega^{-1} X_m)^{-1} X_m'\right) \right| \\ = \; &\sup_m \left| \frac{1}{n} \sum_{i=1}^n x_{mi}' \left(\left(\frac{1}{n} X_m' \hat{\Omega}^{-1} X_m\right)^{-1} - \left(\frac{1}{n} X_m' \Omega^{-1} X_m\right)^{-1} \right) x_{mi} \right| \\ \leq \; &\sup_m \left\| \left(\frac{1}{n} X_m' \hat{\Omega}^{-1} X_m\right)^{-1} - \left(\frac{1}{n} X_m' \Omega^{-1} X_m\right)^{-1} \right\|_1 \sup_m \frac{1}{n} \sum_{i=1}^n x_{mi}' x_{mi} \\ = \; &O_p\left(\frac{k_M^2}{\sqrt{n}}\right), \end{aligned}$$

where the inequality stems from the definition of the Banach norm and the last equality follows from Lemma 2. It therefore holds that:

$$\begin{aligned} &\sup_W tr \left| \left(\sum_{m=1}^M w_m X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m'\right) - tr\left(\sum_{m=1}^M w_m X_m (X_m' \Omega^{-1} X_m)^{-1} X_m'\right) \right| \\ = \; &\sup_W \sum_{m=1}^M w_m \left(tr\left(X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m'\right) - tr\left(X_m (X_m' \Omega^{-1} X_m)^{-1} X_m'\right) \right) \\ = \; &O_p\left(\frac{k_M^2}{\sqrt{n}}\right). \end{aligned}$$

□

Lemma 4. *Suppose that Assumptions 4, 6, 7, 8 and 9 hold. As $n \to \infty$ and $k_M^2/n \to 0$, it holds that:*

$$\sup_m \left\| \hat{\Theta}_m - \hat{\Theta}_m^F \right\| = O_p\left(\sqrt{\frac{k_M}{n}} \right).$$

Proof. The difference between $\hat{\Theta}_m$ and $\hat{\Theta}_m^F$ can be written as follows. Let $\hat{e}(m) = Y - X_m' \hat{\Theta}_m$. Then, we have:

$$\hat{\Theta}_m^F = (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m' \hat{\Omega}^{-1} Y = \hat{\Theta}_m + (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m' \hat{\Omega}^{-1} \hat{e}(m).$$

Therefore, it follows that

$$\hat{\Theta}_m - \hat{\Theta}_m^F = -(X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m' \hat{\Omega}^{-1} \hat{e}(m).$$

We now examine the order of the term $X_m' \hat{\Omega}^{-1} \hat{e}(m)$. Since $X_m' \Omega^{-1} \hat{e}(m) = 0$ by the definition of $\hat{e}(m)$, we have:

$$\begin{aligned} \frac{1}{n} X_m' \hat{\Omega}^{-1} \hat{e}(m) &= \frac{1}{n} X_m' \hat{\Omega}^{-1} \hat{e}(m) - \frac{1}{n} X_m' \Omega^{-1} \hat{e}(m) \\ &= \frac{1}{n} X_m' \left(\hat{\Omega}^{-1} - \Omega^{-1} \right) \hat{e}(m) = \frac{1}{n} \sum_{i=1}^n x_{mi} \left(\frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_i^2} \right) \hat{e}_i(m), \end{aligned}$$

where $\hat{e}_i(m)$ is the ith element of $\hat{e}(m)$. By the Cauchy-Schwarz inequality, we have:

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n x_{mi} \left(\frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_i^2} \right) \hat{e}_i(m) \right\| \\ \leq & \sup_i \left| \frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_i^2} \right| \cdot \left(\frac{1}{n} \sum_{i=1}^n \| x_{mi} \|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n e_i(m)^2 \right)^{1/2}. \end{aligned}$$

Assumption 6 implies that $\sup_m \sum_{i=1}^n \| x_{mi} x_{mi}' \|^2 / n = O(k_M)$. Assumptions 1 and 8 imply that $\sup_m \sum_{i=1}^n e_i(m)^2 = O_p(1)$. Therefore, by Assumption 9, it holds that:

$$\sup_m \left\| \frac{1}{n} X_m' \hat{\Omega}^{-1} \hat{e}(m) \right\| = O_p\left(\sqrt{\frac{k_M}{n}} \right).$$

Therefore, by Lemma 2, it follows that:

$$\begin{aligned} \sup_m \left\| \hat{\Theta}_m - \hat{\Theta}_m^F \right\| &= \sup_m \left\| (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m' \hat{\Omega}^{-1} \hat{e}(m) \right\| \\ &\leq \sup_m \left\| \left(\frac{1}{n} X_m' \hat{\Omega}^{-1} X_m \right)^{-1} \right\|_1 \cdot \sup_m \left\| \frac{1}{n} X_m' \hat{\Omega}^{-1} \hat{e}(m) \right\| \\ &= O_p\left(\sqrt{\frac{k_M}{n}} \right). \end{aligned}$$

□

A.3 Proof of Theorem 3

Proof. We show that:

$$\frac{L_{I_n}^F(\bar{W})}{\inf_{W\in\mathcal{H}_M(N)} L_{I_n}^F(W)} \to_p 1,$$

where $\bar{W} = \arg\min_{W\in\mathcal{H}_M(N)} C_{I_n}^F(W)$.

Lemma A.9 of Donald and Newey (2001) demonstrates that it is sufficient to show that:

$$\sup_W \left(\frac{C_{I_n}^F(W) - L_{I_n}^F(W)}{L_{I_n}^F(W)} \right) \to 0.$$

As we have already established that:

$$\sup_W \left| \frac{L_{I_n}(W)}{R_{I_n}(W)} - 1 \right| \to_p 0$$

and

$$\sup_W \left| \frac{C_{I_n}(W) - L_{I_n}(W)}{L_{I_n}(W)} \right| \to_p 0,$$

we need to show that:

$$\sup_W \left| \frac{C_{I_n}^F(W) - C_{I_n}(W)}{R_{I_n}(W)} \right| \to_p 0,$$
$$\sup_W \left| \frac{L_{I_n}^F(W) - L_{I_n}(W)}{R_{I_n}(W)} \right| \to_p 0.$$

To do so, we observe that:

$$\begin{aligned} & C_{I_n}^F(W) - C_{I_n}(W) \\ = \; & 2(Y - \hat{\mu}(W))'(\hat{\mu}_F(W) - \hat{\mu}(W)) + \|\hat{\mu}_F(W) - \hat{\mu}(W)\|^2 \\ & +2\left(tr\left(\sum_{m=1}^M w_m X_m (X_m'\hat{\Omega}^{-1}X_m)^{-1}X_m' \right) - tr\left(\sum_{m=1}^M w_m X_m (X_m'\Omega^{-1}X_m)^{-1}X_m' \right) \right), \end{aligned}$$

and

$$L_{I_n}^F(W) - L_{I_n}(W) \;=\; \left\|\hat{\mu}(W) - \hat{\mu}^F(W)\right\|^2 + (\mu - \hat{\mu}(W))'(\hat{\mu}(W) - \hat{\mu}^F(W)).$$

Thus, it is sufficient to show the following three conditions:

$$\sup_W \left| \frac{(Y - \hat{\mu}(W))'(\hat{\mu}^F(W) - \hat{\mu}(W))}{R_{I_n}(W)} \right| \to_p 0, \tag{13}$$

$$\sup_W \left| \frac{\|\hat{\mu}^F(W) - \hat{\mu}(W)\|^2}{R_{I_n}(W)} \right| \to_p 0, \tag{14}$$

and

$$\sup_W \left| \frac{tr\left(\sum_{m=1}^M w_m X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m'\right) - tr\left(\sum_{m=1}^M w_m X_m (X_m' \Omega^{-1} X_m)^{-1} X_m'\right)}{R_{I_n}(W)} \right| \to_p 0. \quad (15)$$

We first consider condition (14). We observe:

$$\begin{aligned}
& \|\hat{\mu}^F(W) - \hat{\mu}(W)\|^2 \\
= & \sum_{m=1}^M \sum_{l=1}^M w_m w_l \left| n(\hat{\Theta}_m^F - \hat{\Theta}_m)'(n^{-1} X_m' X_l)(\hat{\Theta}_l^F - \hat{\Theta}_l) \right| \\
\leq & \sum_{m=1}^M \sum_{l=1}^M w_m w_l n \left((\hat{\Theta}_m^F - \hat{\Theta}_m)' \left(\frac{1}{n} X_m' X_l\right) \left(\frac{1}{n} X_l' X_m\right) (\hat{\Theta}_m^F - \hat{\Theta}_m) \right)^{\frac{1}{2}} \\
& \times \left((\hat{\Theta}_l^F - \hat{\Theta}_l)'(\hat{\Theta}_l^F - \hat{\Theta}_l) \right)^{\frac{1}{2}} \\
\leq & \sum_{m=1}^M \sum_{l=1}^M w_m w_l \lambda_{\max}^{1/2} \left[\frac{X_m' X_l}{n} \frac{X_l' X_m}{n} \right] n \|\hat{\Theta}_m^F - \hat{\Theta}_m\| \|\hat{\Theta}_l^F - \hat{\Theta}_l\|,
\end{aligned}$$

where the first inequality follows by the Cauchy–Schwarz inequality and the second inequality follows by the property of the largest eigenvalue and the definition of the norm. Thus, by Lemma 4, we have:

$$\sup_W \|\hat{\mu}^F(W) - \hat{\mu}(W)\|^2 = O_p(k_M).$$

Condition (14) thus holds by the assumption of the theorem.

Next, we examine condition (13). As $Y = \mu + e$, we have:

$$\begin{aligned}
& \left| \frac{(Y - \hat{\mu}(W))'(\hat{\mu}^F(W) - \hat{\mu}(W))}{R_{I_n}(W)} \right| \\
\leq & \left| \frac{(\mu - \hat{\mu}(W))'(\hat{\mu}^F(W) - \hat{\mu}(W))}{R_{I_n}(W)} \right| + \left| \frac{e'(\hat{\mu}^F(W) - \hat{\mu}(W))}{R_{I_n}(W)} \right|.
\end{aligned}$$

By the Cauchy–Schwarz inequality, it follows that:

$$\begin{aligned}
& \left| \frac{(\mu - \hat{\mu}(W))'(\hat{\mu}^F(W) - \hat{\mu}(W))}{R_{I_n}(W)} \right| \\
\leq & \left(\frac{\|\mu - \hat{\mu}(W)\|^2}{R_{I_n}(W)} \right)^{1/2} \left(\frac{\|\hat{\mu}^F(W) - \hat{\mu}(W)\|^2}{R_{I_n}(W)} \right)^{1/2}.
\end{aligned}$$

As in the proof of Theorem 1, we have:

$$\left(\frac{\|\mu - \hat{\mu}(W)\|^2}{R_{I_n}(W)} \right) = O(1).$$

Therefore, because we have already shown condition (14), it holds that:

$$\sup_W \left| \frac{(\mu - \hat{\mu}(W))'(\hat{\mu}^F(W) - \hat{\mu}(W))}{R_{I_n}(W)} \right| \to^p 0.$$

We see that:

$$\begin{aligned}
|e'(\hat{\mu}^F(W) - \hat{\mu}(W))| &= \left| \sum_{m=1}^{M} w_m (\hat{\Theta}_m^F - \hat{\Theta})' X_m' e \right| \\
&= \left| \sum_{m=1}^{M} w_m n (\hat{\Theta}_m^F - \hat{\Theta})' \frac{1}{n} \sum_{i=1}^{n} X_{mi}' e_i \right| \\
&\leq \sum_{m=1}^{M} w_m n \left\| \hat{\Theta}_m^F - \hat{\Theta} \right\| \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} X_{mi}' e_i \right\|.
\end{aligned}$$

Lemma 4 implies that $\sup_m \left\| \hat{\Theta}_m^F - \hat{\Theta} \right\| = O_p(\sqrt{k_M/n})$. Furthermore, it is easy to see that:

$$\sup_m \left\| \frac{1}{n} \sum_{i=1}^{n} X_{mi}' e_i \right\| = O_p\left(\sqrt{\frac{k_M}{n}} \right).$$

It therefore holds that:

$$\sup_W |e'(\hat{\mu}^F(W) - \hat{\mu}(W))| = O_p\left(k_M\right).$$

Thus, condition (13) holds by the assumption of the theorem.

Lastly, we consider condition (15). By Lemma 3, we have:

$$\begin{aligned}
&\sup_W \left| tr\left(\sum_{m=1}^{M} w_m X_m (X_m' \hat{\Omega}^{-1} X_m)^{-1} X_m^{-1} \right) - tr\left(\sum_{m=1}^{M} w_m X_m (X_m' \Omega^{-1} X_m)^{-1} X_m^{-1} \right) \right| \\
= \; & O_p\left(\frac{k_M^2}{\sqrt{n}} \right).
\end{aligned}$$

Therefore, condition (15) holds by the assumption of the theorem. □

A.4 Proof of Theorem 3

Proof. We consider a slightly modified version of the criterion:

$$\bar{C}_\Omega^F(W) = C_\Omega^F(W) + e'(\Omega^{-1} - \hat{\Omega}^{-1})e.$$

As $\bar{C}_\Omega^F(W)$ differs from $C_\Omega^F(W)$ only by the term that does not depend on W, the weights obtained by minimizing $\bar{C}_\Omega^F(W)$ are the same as that from $C_\Omega^F(W)$. Thus, it is sufficient to show the optimality of $\bar{C}_\Omega^F(W)$.

Similarly to the proof of Theorem 2, we need to show that:

$$\sup_W \frac{\left| \bar{C}_\Omega^F(W) - C_\Omega(W) \right|}{R_\Omega(W)} \to_p 0.$$

First, we observe that:

$$
\begin{aligned}
&\bar{C}_{\Omega}^{F}(W)-C_{\Omega}(W)\\
&=\|Y-\hat{\mu}(W)\|_{\Omega}^{2}-\|Y-\hat{\mu}_{F}(W)\|_{\Omega}^{2}+\|Y-\hat{\mu}_{F}(W)\|_{\Omega}^{2}-\|Y-\hat{\mu}_{F}(W)\|_{\hat{\Omega}}^{2}\\
&=-2Y'\Omega^{-1}\hat{\mu}\left(W\right)+\hat{\mu}'\left(W\right)\Omega^{-1}\hat{\mu}\left(W\right)-\left(-2Y'\Omega^{-1}\hat{\mu}_{F}\left(W\right)+\hat{\mu}_{F}'\left(W\right)\Omega^{-1}\hat{\mu}_{F}\left(W\right)\right)\\
&+\left(Y-\hat{\mu}_{F}(W)\right)'\left(\Omega^{-1}-\hat{\Omega}^{-1}\right)\left(Y-\hat{\mu}_{F}(W)\right)\\
&=2Y'\Omega^{-1}\left(\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right)+\left(\hat{\mu}\left(W\right)-\hat{\mu}_{F}\left(W\right)\right)'\Omega^{-1}\hat{\mu}_{F}\left(W\right)+\hat{\mu}'\left(W\right)\Omega^{-1}\left(\hat{\mu}\left(W\right)-\hat{\mu}_{F}\left(W\right)\right)\\
&+\left(Y-\hat{\mu}_{F}(W)\right)'\left(\Omega^{-1}-\hat{\Omega}^{-1}\right)\left(Y-\hat{\mu}_{F}(W)\right)\\
&=\left(Y-\hat{\mu}_{F}\left(W\right)\right)'\Omega^{-1}\left(\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right)+\left(Y'-\hat{\mu}\left(W\right)\right)\Omega^{-1}\left(\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right)\\
&+\left(Y-\hat{\mu}_{F}(W)\right)'\left(\Omega^{-1}-\hat{\Omega}^{-1}\right)\left(Y-\hat{\mu}_{F}(W)\right)\\
&=\left(\mu-\hat{\mu}_{F}\left(W\right)\right)'\Omega^{-1}\left(\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right)\\
&+\left(\mu-\hat{\mu}\left(W\right)\right)'\Omega^{-1}\left(\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right)+2e'\Omega^{-1}\left(\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right)\\
&+\left(\mu-\hat{\mu}_{F}(W)\right)'\left(\Omega^{-1}-\hat{\Omega}^{-1}\right)\left(\mu-\hat{\mu}_{F}(W)\right)+2e'\left(\Omega^{-1}-\hat{\Omega}^{-1}\right)\left(\mu-\hat{\mu}_{F}(W)\right).
\end{aligned}
$$

As

$$
\begin{aligned}
\frac{R_{I_n}(W)}{R_{\Omega}(W)}&=\frac{E\left\|\mu-\hat{\mu}\left(W\right)\right\|^{2}}{E\left\|\mu-\hat{\mu}\left(W\right)\right\|_{\Omega}^{2}}\\
&=\frac{E\left\|\mu-\hat{\mu}\left(W\right)\right\|^{2}}{E\left(\mu-\hat{\mu}\left(W\right)\right)'\Omega^{-1}\left(\mu-\hat{\mu}\left(W\right)\right)}\\
&\leq\frac{E\left\|\mu-\hat{\mu}\left(W\right)\right\|^{2}}{\inf_{i}\left(\sigma_{i}^{-2}\right)E\left\|\mu-\hat{\mu}\left(W\right)\right\|^{2}}\\
&=O\left(1\right),
\end{aligned}
$$

and because we have shown in the proof of Theorem 2 that:

$$
\begin{aligned}
\sup_{W}\frac{\left\|\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right\|^{2}}{R_{I_n}(W)}&\rightarrow_{p}0,\\
\sup_{W}\frac{\left\|\mu-\hat{\mu}\left(W\right)\right\|^{2}}{R_{I_n}(W)}&=O_{p}(1),
\end{aligned}
$$

we have:

$$
\begin{aligned}
&\sup_{W}\frac{\left|\left(\mu-\hat{\mu}_{F}\left(W\right)\right)'\Omega^{-1}\left(\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right)\right|}{R_{\Omega}(W)}\\
&\leq\sup_{W}\frac{\sup_{i}\left(\sigma_{i}^{-2}\right)\left\|\mu-\hat{\mu}_{F}\left(W\right)\right\|\left\|\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right\|}{R_{\Omega}(W)}\\
&=\sup_{W}\frac{\sup_{i}\left(\sigma_{i}^{-2}\right)\left\|\mu-\hat{\mu}_{F}\left(W\right)\right\|\left\|\hat{\mu}_{F}\left(W\right)-\hat{\mu}\left(W\right)\right\|}{R_{I_n}(W)}\frac{R_{I_n}(W)}{R_{\Omega}(W)}\\
&=o_{p}\left(1\right).
\end{aligned}
$$

Moreover, we have:

$$\begin{aligned}
&\sup_W \frac{\left|\left(\mu - \hat{\mu}_F(W)\right)' \left(\Omega^{-1} - \hat{\Omega}^{-1}\right) \left(\mu - \hat{\mu}_F(W)\right)\right|}{R_{\Omega_n}(W)} \\
&\leq \sup_W \frac{\sup_i \left|\sigma_i^{-2} - \hat{\sigma}_i^{-2}\right| \left\|\mu - \hat{\mu}_F(W)\right\|^2}{R_{I_n}(W)} \frac{R_{I_n}(W)}{R_\Omega(W)} \\
&= o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
&\sup_W \frac{\left|e' \left(\Omega^{-1} - \hat{\Omega}^{-1}\right) \left(\mu - \hat{\mu}_F(W)\right)\right|}{R_\Omega(W)} \\
&\leq \sup_W \frac{\sup_i \left|\sigma_i^{-2} - \hat{\sigma}_i^{-2}\right| \|e\| \left\|\mu - \hat{\mu}_F(W)\right\|}{R_{I_n}(W)} \frac{R_{I_n}(W)}{R_\Omega(W)} \\
&= \sup_W \frac{O_p\left(n^{-1/2}\right) \|e\| \left\|\mu - \hat{\mu}_F(W)\right\|}{R_{I_n}(W)} \frac{R_{I_n}(W)}{R_\Omega(W)} \\
&= O_p\left(n^{-1/2}\right) O_p\left(n^{1/2}\right) o_p(1) O(1) = o_p(1).
\end{aligned}$$

By Lemma 4, it follows that:

$$\begin{aligned}
&\sup_W \frac{\left|e'\Omega^{-1}\left(\hat{\mu}_F(W) - \hat{\mu}(W)\right)\right|}{R_\Omega(W)} \\
&= \sup_W \frac{\left|\sum_{m=1}^M w_m \left(\hat{\Theta}_m^F - \hat{\Theta}_m\right)' X_m' \Omega^{-1} e\right|}{R_\Omega(W)} \\
&= \sup_W \frac{\left|\sum_{m=1}^M w_m n \left(\hat{\Theta}_m^F - \hat{\Theta}_m\right)' \frac{1}{n}\Sigma_{i=1}^n X_{mi}' \sigma_i^{-2} e_i\right|}{R_\Omega(W)} \\
&\leq \sup_W \frac{\sum_{m=1}^M w_m n \left\|\hat{\Theta}_m^F - \hat{\Theta}_m\right\| \left\|\frac{1}{n}\Sigma_{i=1}^n X_{mi}' \sigma_i^{-2} e_i\right\|}{R_{I_n}(W)} \frac{R_{I_n}(W)}{R_\Omega(W)} \\
&= O(n) O_p\left(\sqrt{k_M/n}\right) O_p\left(\sqrt{k_M/n}\right) \left(\inf_{W \in \mathcal{H}_M(N)} R_{I_n}(W)\right)^{-1} \\
&= O_p(k_M) \left(\inf_{W \in \mathcal{H}_M(N)} R_{I_n}(W)\right)^{-1} = o_p(1),
\end{aligned}$$

and we have the desired result.

□

References

Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," in Petrov, B. and Csáki, F. eds, *Second International Symposium on Information Theory*, Akadémiai Kiadó, 267–281

Akaike, H. (1979): "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting," *Biometrika*, 66, 237–242.

Berk, K. N. (1974): "Consistent Autoregressive Spectral Estimates," *Annals of Statistics*, 2(3), 489–502.

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

Claeskens, G. and N. L. Hjort (2008): *Model Selection and Model Averaging*, Cambridge University Press.

Donald, S. G. and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–1191.

Draper, D. (1995): "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 45–97.

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

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

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

Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14(4), 382–417, with comments by M. Clyde, D. Draper and E. I. George, and a rejoinder by the authors.

Kuersteiner, G. and R. Okui (2010): "Constructing Optimal Instruments by First Stage Prediction Averaging," *Econometrica*, 78(2), 697–718.

Lewis, R. and G. C. Reinsel (1985): "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, 16, 393–411.

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

Liang, H., G. Zou, A. T. K. Wan and X. Zhang (2011): "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 106(495), 1053–1066.

Liu, C.-A. (2011): "A Plug-In Averaging Estimator for Regressions with Heteroskedastic Errors," working paper, University of Wisconsin, Madison.

Liu, Q. and R. Okui (2012): "Generalized Cp Model Averaging for Heteroskedastic Models," working paper, Otaru University of Commerce.

Magnus, J. R., O. Powell and P. Prüfer (2010): "A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics," *Journal of Econometrics*, 154, 139–153.

Magnus, J. R., A. K. Wan and X. Zhang (2011): "Weighted Average Least Squares Estimation with Nonspherical Disturbances and an Application to the Hong Kong Housing Market," *Computational Statistics and Data Analysis*, 55, 1331–1341.

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

Robinson, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55(4), 875–891

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

Wan, A. T., X. Zhang and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156(2), 277–283.

Whittle, P. (1960): "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and its Applications*, 5(3), 302–305.

Wiener, N. and P. Masani (1958): "The Prediction Theory of Multivariate Stochastic Processes, II. The Linear Predictor," *Acta Mathematica*, 99, 93–137.

Yuan, Z., and Y. Yang (2005): "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Society*, 100(472), 1202–1214.

Zhang, X., A. T. K. Wan and S. Z. Zhou (2012): "Focused Information Criteria, Model Selection, and Model Averaging in a Tobit Model with a Nonzero Threshold," *Journal of Business & Economic Statistics*, 30(1), 132–142.