Generalized C_p Model Averaging for Heteroskedastic Models

Qingfeng Liu*†
Department of Economics, Otaru University of Commerce

Ryo Okui
Institute of Economic Research, Kyoto University

Revised Version

June 5, 2011

Abstract

This paper proposes a model averaging method, the generalized Mallows' C_p (GC) method, which works well for heteroskedastic models. Under some regularity conditions, we provide a feasible form of the GC method and show that the GC method has asymptotic optimality not only as a model averaging method but also as a model selection method for heteroskedastic models. We perform some Monte Carlo studies to investigate the small sample properties of the GC method. The simulation results show that our method works well and performs better than alternative methods.

JEL classification: C51 C52

Keywords: Model Averaging, Model Selection, Asymptotic Optimality, Mallows' C_p, Heteroskedastic error.

*Corresponding author: Qingfeng Liu, Department of Economics, Otaru University of Commerce, 5-21, Midori 3-chome, Otaru, Hokkaido 047-8501, Japan. (Tel & Fax: +86 134 27 5312, E-mail: qliu@res.otaru-uc.ac.jp).

†The authors would like to thank K. Akashi, Y. Feng, K. Hitomi, Y. Kawasaki, E. Kurozumi, K. Morimune, Y. Nishiyama, and N. Sueishi, for helpful discussions and comments.

1 Introduction

Model selection helps us to choose a single optimal model from a set of candidate models. In the last two decades, model averaging has been proposed as an alternative to model selection. A model averaging estimator is obtained by taking the weighted average of the estimators obtained from candidate models. As compared to model selection, model averaging seeks to avoid selecting a very poor model and to improve the estimate with regard to risk. Model averaging methods can be separated into two groups: Bayesian model averaging methods and frequentist (non-Bayesian) model averaging methods. Bayesian model averaging methods have been advocated by many researchers (see Draper (1995), Hoeting, Madigan, Raftery, and Volinsky (1999), and Clyde and George (2004)). On the other hand, frequentist model averaging methods have a shorter history than their Bayesian counterparts. In the literature on frequentist model averaging methods, Buckland, Burnham, Burnham, and Augustin (1997) proposed a smoothed-AIC (SAIC) based method and a smoothed-BIC (SBIC) based method, and Hjort and Claeskens (2003) proposed a frequentist model averaging method and derived the inference for the estimate based on the likelihood function of the model. Recently, Hansen (Hansen (2007), Hansen (2009), and Hansen (2010)) proposed several model averaging methods, which work for linear models, models based on series expansion, models with structural break, and models with a near unit root.

This paper extends Hansen (2007), which proposed a Mallows model averaging (MMA) estimator for models with homoskedastic errors. The

weights of the models for the MMA estimator are determined by minimizing a criterion similar to Mallows' C_p (MC). Our extension is a generalization of the MMA method. The GC method works for both homoskedastic and heteroskedastic errors not only as a model averaging method but also as a model selection method. For heteroskedastic situations, Andrews (1991) showed asymptotic optimality for a model selection criterion based on MC. However, Andrews (1991) did not provide a feasible form of this criterion, because of the difficulty associated with the consistent estimation of the covariance matrix. We provide a way to avoid the estimate of the covariance matrix, and are thus able to propose a feasible form of the GC method. Under some regularity conditions, we show that the GC method has asymptotic optimality not only as a model averaging method but also as a model selection method for models with heteroskedastic errors.

The rest of this paper is organized as follows. In section 2, the GC method and its feasible form for model averaging and model selection are proposed, and the optimality of the GC method is discussed. In section 3, some simulation studies are performed to check the finite sample properties of the GC method. Section 4 contains some concluding remarks. The appendix contains some technical proofs.

2 GC Method

Hansen (2007) proposed an MMA estimator. In his setup, the regressors are assumed to be ordered, and the candidate regression models are assumed to be nested. Wan, Zhang, and Zou (2010) extended the results of Hansen

(2007), by removing these assumptions. Our setup is similar to Wan, Zhang, and Zou (2010). The following is our model:

$$\begin{aligned} y_i &= \mu_i + e_i, \\ \mu_i &= \sum_{j=1}^{\infty} \theta_j x_{ij}, \\ E\left(e_i | x_i\right) &= 0, \end{aligned} \tag{1}$$

for $i = 1, \cdots, n$, where y_i is a real-valued scalar, $x_i = (x_{i1}, x_{i2}, \cdots)$ is a countably infinite real-valued vector, μ_i is assumed to be converging in mean square, and $E\mu_i^2 < \infty$. Our results almost all are conditional on x_i, for simplicity, we omit the conditional expression in some cases hereafter.

The most important difference between our setup and that of Hansen (2007) and Wan, Zhang, and Zou (2010) is that in their setup, the error term e_i is assumed to be homoskedastic and not heteroskedastic as in our setup. We assume that e_i is independent over i and $E\left(e_i^2 | x_i\right) = \sigma_i^2$. The matrix form of the regressors is $X \equiv (x_1', x_2', \cdots)'$. The matrix form of eq.(1) is $y = \mu + e$, where $y = (y_1, \cdots, y_n)'$, $\mu = (\mu_1, \cdots, \mu_n)'$, and $e = (e_1, \cdots, e_n)'$. We propose the GC method to estimate μ_i with a small risk (mean squared error, MSE).

The set of candidate models contains M models. The mth model has $k_m > 0$ regressors that can be any variables in x_i. Note that we do not restrict $k_1 < k_2 < \cdots < k_M$, as is the case with the nested models assumed

in Hansen (2007). The mth approximating model of model (1) is

$$y_i = \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)} + b_{i(m)} + e_i, \tag{2}$$

for $m = 1, 2, \cdots M$, where $x_{ij(m)}$ for $j = 1, \cdots, k_m$ denotes the regressors in the mth model, and $\theta_{j(m)}$ denotes the coefficients. We thus have a matrix form of eq.(2):

$$Y = X_{(m)}\Theta_{(m)} + b_{(m)} + e, \tag{3}$$

where $Y = (y_1, \cdots, y_n)'$, $X_{(m)}$ is an $n \times k_m$ matrix of the regressors with ij element $x_{ij(m)}$ and with full column rank, $\Theta_{(m)} = \left(\theta_{1(m)}, \cdots, \theta_{km(m)}\right)'$, $b_{(m)} = \left(b_{1(m)}, \cdots, b_{n(m)}\right)'$, and $e = (e_1, \cdots, e_n)'$. The LS estimator of $\Theta_{(m)}$ as derived from the mth model is $\hat{\Theta}_{(m)} = \left(X'_{(m)}X_{(m)}\right)^{-1} X'_{(m)}Y$. The estimator of μ is

$$\hat{\mu}_{(m)} = X_{(m)}\left(X'_{(m)}X_{(m)}\right)^{-1} X'_{(m)}Y \equiv P_{(m)}Y \tag{4}$$

and the residual is $\hat{e}_{(m)} = Y - \hat{\mu}_{(m)}$. The model averaging estimator of μ is defined as

$$\hat{\mu}(W) = \sum_{m=1}^{M} \omega_{(m)} P_{(m)} Y \equiv P(W) Y, \tag{5}$$

where $W = \left(\omega_{(1)}, \cdots, \omega_{(M)}\right)'$ is a weight vector in

$$\mathcal{H}_n = \left\{ W \in [0,1]^M : \sum_{m=1}^{M} \omega_{(m)} = 1 \right\}. \tag{6}$$

The setup of the weight vector is different from that in Hansen (2007) who

restricts the elements of the weight vector to be a/n, where a is some non-negative integers less than n, for the optimality of MMA.

Hansen's MMA was designed for models with homoskedastic errors. Although it is hoped that it can also be applied to models with heteroskedastic errors, there does not exist any theoretical support for optimality and good performance in the heteroskedastic case. In this section, we propose the GC method for the heteroskedastic error case. We will show the optimality of this method and check its small sample performance in the next section.

The model averaging criterion is defined as follows:

$$GC_n = \|Y - P(W)Y\|^2 + 2tr[\Omega P(W)], \tag{7}$$

where Ω is an $n \times n$ diagonal matrix with $\Omega_{ii} = \sigma_i^2$. Then, the estimator of the optimal weight vector is denoted as

$$\hat{W}_{GC} = \arg \min_{W \in \mathcal{H}_n} GC_n. \tag{8}$$

Our aim is to show the optimality of $\hat{W}_{GC}$ under some regularity conditions. We define the loss function and the risk function as

$$L_n(W) = \|\hat{\mu}(W) - \mu\|^2 \tag{9}$$

and

$$R_n(W) = E(L_n(W)|X), \tag{10}$$

respectively. Then, optimality implies

$$\frac{L_n\left(\hat{W}_{GC_n}\right)}{\inf_{W\in\mathcal{H}_n} L_n\left(W\right)} \to_p 1. \tag{11}$$

It can be easily seen that the expectation of GC_n is the sum of the risk function and a constant. Hence, GC_n can be regarded as an unbiased estimator of the risk function plus a constant.

Lemma 1 $E\left(GC_n\left(W\right)\right) = R_n\left(W\right) + \sum_{i=1}^{n}\sigma_i^2$.

The following theorem on the optimality of $\hat{W}_{GC}$ is an application of theorem 2.1* of Andrews (1991) and theorem 1' of Wan, Zhang, and Zou (2010).

Theorem 2 *For* $\xi_n \equiv \inf_{W\in\mathcal{H}_n} R_n\left(W\right)$ *and some integer* $1 \leq G < \infty$*, if*

$$E\left(e_i^{4G}|x_i\right) \leq \kappa < \infty, \tag{12}$$

$$M\xi_n^{-2G}\sum_{m=1}^{M}\left(R_n\left(W_m^0\right)\right)^G \to 0, \tag{13}$$

and $0 < \inf_i \sigma_i^2 \leq \sup_i \sigma_i^2 < \infty$*, then* $\frac{L_n\left(\hat{W}_{GC_n}\right)}{\inf_{W\in\mathcal{H}_n} L_n(W)} \xrightarrow{p} 1$*, where* W_m^0 *is a vector whose* m*th element is one and all other elements are zeros.*

Andrews (1991) showed the asymptotic optimality for a model selection method based on MC, but he did not propose a feasible criterion. The difficulty in providing a feasible criterion arises from the fact that without additional restrictions, one cannot expect to obtain a consistent estimator of the covariance matrix Ω; since for heteroskedastic errors, Ω has at least

n parameters and we have only n observations. To solve this problem, our idea is to estimate not Ω but the scalar $tr\left[\Omega P\left(W\right)\right]$. Using this approach, we propose the following feasible criterion for model averaging:

$$\widehat{GC}_n \equiv \left\|Y - P\left(W\right)Y\right\|^2 + 2\frac{n}{n-K}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}\left(W\right), \tag{14}$$

where $\hat{e}_i$ is the residual from the largest model with the most number of regressors, K is the number of the regressors of the largest model and $p_{ii}\left(W\right)$ is the ith diagonal element of $P\left(W\right)$. Let $\bar{e} \equiv \left(\tilde{e}_1, \cdots, \tilde{e}_M\right)$ with $\tilde{e}_m$ to be the $n \times 1$ residual vector from the mth model, let $\hat{\Omega} \equiv diag\left(\hat{e}_i^2, \cdots, \hat{e}_n^2\right)$, and let $\Xi \equiv \left(tr(\hat{\Omega}P_{(1)}), \cdots, tr(\hat{\Omega}P_{(M)})\right)'$. Then we have the following expression

$$\widehat{GC}_n = W'\bar{e}'\bar{e}W + 2\frac{n}{n-K}\Xi'W \tag{15}$$

The corresponding estimator of the optimal weight vector is

$$\hat{W}_{\widehat{GC}_n} \equiv \arg\min_{W \in \mathcal{H}_n} \widehat{GC}_n. \tag{16}$$

The following theorem shows that under some regularity conditions, if one replaces the term $tr\left[\Omega P\left(W\right)\right]$ in GC_n with $\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}\left(W\right)$ as in eq.(14), Theorem 2 still holds as the following theorem claims.

Theorem 3 *When $\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}\left(W\right)$ is used instead of $tr\left[\Omega P\left(W\right)\right]$, Theorem 2 is valid if*

$$0 < \lim n^{-1}\sum_{i=1}^{n}\sigma_i^2 = \overline{\sigma^2} < \infty, \tag{17}$$

$$\mu'\mu/n = O\left(1\right), \tag{18}$$

$$\max_{1\leq m\leq M}\max_{1\leq i\leq n} p_{m,ii} = O\left(n^{-1/2}\right), \tag{19}$$

$$\frac{\tilde{p}e'e}{\xi_n} \xrightarrow{p} 0, \tag{20}$$

$$\lim \lambda_{\max}(n) = \infty, \tag{21}$$

$$\log\left(\lambda_{\max}(n)\right) = O\left(n^{1/2}\right), \tag{22}$$

where $\tilde{p} \equiv \sup_{W\in\mathcal{H}_n} \max_{1\leq i\leq n}\left(p_{ii}(W)\right)$, $\lambda_{\max}(n)$ *is the maximum eigenvalue of* $\tilde{X}'\tilde{X}$ *with* $\tilde{X}$ *denoting the matrix of the regressors of the largest model, and* $p_{m,ii}$ *is the* i*th diagonal element of* $P_{(m)}$.

In Li (1987), Andrews (1991), and Hansen and Racine (2010), there are some restrictions similar to (19), such as $\max_{1\leq m\leq M}\max_{1\leq i\leq n} p_{m,ii} \to 0$. Using some properties of $P_{(m)}$, which is an idempotent matrix, one can show that such restrictions are reasonable, they exclude only extremely unbalanced models. If such restrictions do not hold, the variances of some $\hat{\mu}_i$s, i.e., some elements of the estimate $\hat{\mu}$ based on an unbalanced single model, will be extremely large, and as such, $\hat{\mu}_i$s will be much less accurate.

It can be easily seen that if we restrict the weight vector to be $W \in \{i_1, i_2, \cdots, i_M\}$, where i_i is a vector whose ith element is one and other elements are zeros, then the GC method works as a model selection procedure to select a single model. The above two theorems are valid for this model selection procedure; further, this model selection procedure has optimality. The criterion for model selection can be expressed as follows:

$$\widehat{GC}_n(m) \equiv \|Y - P_m Y\|^2 + 2\frac{n}{n-K}\sum_{i=1}^{n}\hat{e}_i^2 p_{m,ii}. \tag{23}$$

The estimator of the indicator of the optimal model can be obtained as follows:

$$\hat{m} \equiv \arg \min_{1 \leq m < M} GC_n(m). \tag{24}$$

3 Monte Carlo Studies

To investigate the finite sample performance of our method, we conduct two Monte Carlo simulations. The number of replications is 1000 for both simulations. For comparison, not only the results of the GC method but also the results of a model averaging method based on generalized cross-validation (GCV, Craven and Wahba (1978)), MMA (Hansen (2007)), SAIC and SBIC (Buckland, Burnham, Burnham, and Augustin (1997)), and AIC (Akaike (1973)) methods are shown. The model averaging method based on GCV has been introduced by Liu (2010) in an unpublished working paper, the criterion is defined as

$$GCV_n(W) = \frac{\|Y - \hat{\mu}(W)\|^2}{(n - trP(W))^2}. \tag{25}$$

The optimal weight vector selected by the GCV method is defined as

$$\hat{W}_{GCV} = \arg \min_{W \in \mathcal{H}_n} GCV_n(W). \tag{26}$$

We have the DGP as

$$y_i = \sum_{j=1}^{\infty} \theta_j x_{ij} + e_i. \tag{27}$$

We cut off the infinite order at $j = 30$. The parameters are determined as in Hansen (2007): $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$. We set the values as $c = 0.2, 0.4, 0.6, \cdots, 2$ and $\alpha = 0.5$. The parameter c affects the population R^2 of eq.(27): R^2 increases with c. The sample size is $n = 150$, the number of models is $M = 10$, and the biggest model has 10 regressors. For simplicity, in the simulations, we employ a nested setting: the $(k+1)$th model is nested in the kth model. x_{ij}s are independent over j, $j = 1, \cdots, m$, and set to be i.i.d. $N(0,1)$ over i. The first simulation is with homoskedastic errors; we set e_i to be i.i.d. $N\left(0, \sigma^2\right)$, where $\sigma = 1$. In the second simulation study, we set e_i to be independent and heteroskedastic $N\left(0, \sigma_i^2\right)$, where $\sigma_i = x_{i2}^2$. Since the arguments in the above sections are restricted in the situation conditional on X, we first generate X and then fix the data of X through all the replications. We define the sample MSE as $MSE = 1/1000 \sum_{i=1}^{1000} (\hat{\mu} - \mu)^2$, and calculate the MSE ratios (the ratios of the MSEs of the aforementioned methods and the MSE of the GC method). The MSE ratios are plotted in Figures I and II for homoskedastic and heteroskedastic errors, respectively.

We can see that the AIC method is dominated by the SAIC method for almost all values of c (R^2) in both simulations. The performance of the SBIC method is the worst in the homoskedastic case, but is better than some other methods for small values of c in the heteroskedastic case. The AIC method and the MC method perform moderately, and the GCV method and the MMA method are better than them in both cases.

The most important results are on the comparisons between the GC method and the GCV method and between the GC method and the MMA method. The GCV method is a perfect alternative to the MMA method;

both have almost the same MSE ratios. In the homoskedastic case, these three methods have similar MSEs: the GC method performs slightly worse than the GCV method and the MMA method when $c < 0.4$ but slightly better when c (R^2) is bigger. In the heteroskedastic case, the situation is much different. The GC method performs the best, and is particularly better than the GCV method and the MMA method when c is small. From these results, we get that the GC method works well for models with heteroskedastic errors.

4 Conclusion

We proposed a model averaging method for heteroskedastic models. We argued the optimality of this method, and performed Monte Carlo simulations to investigate its small sample properties. The results of these simulations show that the proposed method works well, particularly for models with heteroskedastic errors.

5 Appendix

Proof of Theorem 2. The proof of optimality in Wan, Zhang, and Zou (2010) is an application of Theorem 2 of Whittle (1960). Since Theorem 2 of Whittle (1960) holds even with heteroskedastic errors, when $\sigma^2 trP(W)$ is replaced with $tr\Omega P(W)$ and $\sigma^2 trP^2(W)$ is replaced with $tr\Omega P^2(W)$, the proof of Theorem 2 is almost the same as the proof of Theorem 1' in Wan, Zhang, and Zou (2010). ■

Proof of Theorem 3. We denote the projection matrix of the largest

model as P^*, and the ith diagonal element of P^* as p_{ii}^*. We define $\bar{P}(W)$ as a diagonal matrix which ith diagonal element is $p_{ii}(W)$.

Condition (19) implies that $\tilde{p} = O\left(n^{-1/2}\right)$ and the number of regressors in the largest model $K = O\left(n^{1/2}\right)$; condition (13) implies that $\xi_n \to \infty$. From the properties of an idempotent matrix, we have $tr\bar{P}^2\left(W_m^0\right) \leq trP^2\left(W_m^0\right)$ and $0 \leq p_{ii}(W) \leq 1$. We use C to denote some constant which could take different values in the following proof.

Since

$$\begin{aligned}\widehat{GC} &\equiv (Y - \hat{\mu}(W))'(Y - \hat{\mu}(W)) + 2\frac{n}{n-K}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W) \\ &= GC + 2\left(\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W) - tr\left[\Omega P(W)\right]\right) + \frac{2K}{n-K}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W).\end{aligned} \tag{28}$$

to prove Theorem 3, we only need to show that

$$\sup_{W\in\mathcal{H}_n}\left\{\left|\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W) - tr\left[\Omega P(W)\right]\right| \Big/ R_n(W)\right\} \xrightarrow{p} 0. \tag{29}$$

$$\sup_{W\in\mathcal{H}_n}\left\{\frac{K}{n-K}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W) \Big/ R_n(W)\right\} \xrightarrow{p} 0. \tag{30}$$

Since $K = O\left(n^{1/2}\right)$, eq.(30) can be proved according to assumptions (20).

Furthermore, it can be easily seen that

$$\begin{aligned}
&\sup_{W\in\mathcal{H}_n}\left\{\left|\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W)-tr\left[\Omega P(W)\right]\right|\Big/ R_n(W)\right\}\\
&\leq \sup_{W\in\mathcal{H}_n}\left|\hat{e}'\bar{P}(W)\hat{e}-E\left(e'\bar{P}(W)e\right)\right|/\xi_n\\
&\leq \sup_{W\in\mathcal{H}_n}\left\{\left|\hat{e}'\bar{P}(W)\hat{e}-e'\bar{P}(W)e\right|+\left|e'\bar{P}(W)e-E\left(e'\bar{P}(W)e\right)\right|\right\}/\xi_n\\
&\leq \sup_{W\in\mathcal{H}_n}\left\{\left|\hat{e}'\bar{P}(W)\hat{e}\right|+\left|e'\bar{P}(W)e\right|+\left|e'\bar{P}(W)e-E\left(e'\bar{P}(W)e\right)\right|\right\}/\xi_n\\
&\leq \tilde{p}\left\{\hat{e}'\hat{e}+e'e\right\}/\xi_n+\sup_{W\in\mathcal{H}_n}\left|e'\bar{P}(W)e-E\left(e'\bar{P}(W)e\right)\right|/\xi_n\\
&= \tilde{p}\left\{(\mu+e)'(I-P^*)(\mu+e)+e'e\right\}/\xi_n\\
&+ \sup_{W\in\mathcal{H}_n}\left|e'\bar{P}(W)e-E\left(e'\bar{P}(W)e\right)\right|/\xi_n\\
&= \tilde{p}\left\{\mu'(I-P^*)\mu+2\mu'(I-P^*)e+e'(I-P^*)e+e'e\right\}/\xi_n\\
&+ \sup_{W\in\mathcal{H}_n}\left|e'\bar{P}(W)e-E\left(e'\bar{P}(W)e\right)\right|/\xi_n\\
&\leq \tilde{p}\left\{\mu'(I-P^*)\mu+2\left|\mu'(I-P^*)e\right|+e'P^*e+2e'e\right\}/\xi_n\\
&+ \sup_{W\in\mathcal{H}_n}\left|e'\bar{P}(W)e-E\left(e'\bar{P}(W)e\right)\right|/\xi_n. \qquad (31)
\end{aligned}$$

From conditions (13), (18), (19), and $R_n(W)\geq\mu'(I-P(W))\mu$, we have

$$\begin{aligned}
\tilde{p}\frac{\mu'(I-P^*)\mu}{\xi_n} &\leq \left(\tilde{p}^2\frac{\mu'(I-P^*)\mu}{\xi_n^2}\mu'\mu\right)^{1/2}\\
&\leq \left(\tilde{p}^2\frac{R_n(W_0^*)}{\xi_n^2}\mu'\mu\right)^{1/2}\\
&= \sqrt{O(n^{-1})\,o(1)\,O(n)}\to 0, \qquad (32)
\end{aligned}$$

where W_0^* is the weight vector giving weight 1 to the largest model.

Moreover, using similar techniques as in Wan, Zhang, and Zou (2010), i.e., by applying Chebyshev's inequality, Theorem 2 of Whittle (1960) and $R_n(W) \geq \mu'(I - P(W))\mu$, denoting the ith element of $\mu'(I - P^*)$ as η_i, for any $\delta > 0$, we have

$$\begin{aligned}
&P\left\{2\tilde{p}\frac{|\mu'(I-P^*)e|}{\xi_n} > \delta\right\} \\
&\leq \frac{E\left(\mu'(I-P^*)e\right)^2 4\tilde{p}^2}{\xi_n^2\delta^2} \\
&\leq C\left[\sum_{i=1}^{n}\eta_i^2\sigma_i^2\right]^{1/2}\frac{\tilde{p}^2}{\xi_n^2\delta^2} \\
&\leq C\frac{\tilde{p}^2}{\xi_n^2\delta^2}\sup_{1\leq i\leq n}\sigma_i^2\mu'(I-P^*)\mu \\
&\leq C\frac{\tilde{p}^2}{\xi_n^2\delta^2}R_n(W_0^*) \to 0, \qquad (33)
\end{aligned}$$

hence

$$2\tilde{p}\frac{|\mu'(I-P^*)e|}{\xi_n} \xrightarrow{p} 0. \qquad (34)$$

Moreover according to Lemma 1 of Lai and Wei (1982) and assumptions (21) and (22), we have

$$\begin{aligned}
&\tilde{p}\left|e'P^*e\right|/\xi_n \\
&= O\left(n^{-1/2}\right)O\left(\log\lambda_{\max}(n)\right)o(1) \quad a.s. \\
&= O\left(n^{-1/2}\right)O\left(n^{1/2}\right)o(1) \quad a.s. \qquad (35) \\
&\xrightarrow{a.s.} 0. \qquad (36)
\end{aligned}$$

Furthermore, using Chebyshev's inequality, Theorem 2 of Whittle (1960),

and C_m, $m = 1, \cdots, M$, denoting some constant, for any $\delta > 0$, we have,

$$
\begin{aligned}
& P\left\{ \sup_{W \in \mathcal{H}_n} \left| \left(e' \bar{P}(W) e\right) - E\left(e' \bar{P}(W) e\right) \right| / \xi_n > \delta \right\} \\
& \leq \sum_{m=1}^{M} P\left\{ \left| \left(e' \bar{P}\left(W_m^0\right) e\right) - E\left(e' \bar{P}\left(W_m^0\right) e\right) \right| > \delta \xi_n \right\} \\
& \leq \sum_{m=1}^{M} E\left\{ \frac{\left[\left(e' \bar{P}\left(W_m^0\right) e\right) - E\left(e' \bar{P}\left(W_m^0\right) e\right)\right]^{2G}}{\delta^{2G} \xi_n^{2G}} \right\} \\
& \leq \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^{M} C_m \left\{ \sum_{i=1}^{n} p_{ii}^2\left(W_m^0\right) \left[E\left(e_i^{4G}\right)\right]^{1/G} \right\}^G \\
& \leq \max_{1 \leq m \leq M}(C_m) \max_{1 \leq i \leq n} \left(\left[E\left(e_i^{4G}\right)\right]^{1/G} \right) \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^{M} \left\{ \sum_{i=1}^{n} p_{ii}^2\left(W_m^0\right) \right\}^G \\
& = C \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^{M} \left[tr \bar{P}^2\left(W_m^0\right) \right]^G \\
& \leq C \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^{M} \left[tr P^2\left(W_m^0\right) \right]^G \\
& = C \delta^{-2G} \xi_n^{-2G} \left(\inf_{1 \leq i \leq n} \sigma_i^2 \right)^{-G} \sum_{m=1}^{M} \left[tr \left\{ \inf_{1 \leq i \leq n} \sigma_i^2 I P^2\left(W_m^0\right) \right\} \right]^G \\
& \leq C \delta^{-2G} \xi_n^{-2G} \left(\inf_{1 \leq i \leq n} \sigma_i^2 \right)^{-G} \sum_{m=1}^{M} \left[tr \left\{ \Omega P^2\left(W_m^0\right) \right\} \right]^G \\
& \leq C \delta^{-2G} \xi_n^{-2G} \sum_{m=1}^{M} \left[R_n\left(W_m^0\right) \right]^G \to 0, \qquad (37)
\end{aligned}
$$

where I is a $n \times n$ identity matrix. Therefore, we have $\sup_{W \in \mathcal{H}_n} | \left(e' \bar{P}(W) e\right) - E\left(e' \bar{P}(W) e\right) | \} / \xi_n \xrightarrow{p} 0$. From the above results and eq.(20) we have eq.(29). The proof of Theorem 3 is complete. ■

References

Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle," in *Proc. of the 2nd Int. Symp. on Information Theory*, ed. by P. B. N., and C. F., pp. 267–281.

Andrews, D. W. (1991): "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, 47, 359–377.

Buckland, S. T., C. Burnham, K. P. Burnham, and N. H. Augustin (1997): "Model selection: an integral part of inference," *Biometrics*, 53, 603–618.

Clyde, M., and E. I. George (2004): "Model Uncertainty," *Statistical Science*, 19(1), 81–94.

Craven, P., and G. Wahba (1978): "Smoothing noisy data with spline functions," Numerische Mathematik, 31, 377–403.

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.

——— (2009): "Averaging Estimators for Regressions with a Possible Structural Break," *Econometric Theory*, 35(6), 1498–1514.

——— (2010): "Averaging Estimators for Autoregressions with a Near Unit Root," *Journal of Econometrics*, 158(1), 142–155.

Hansen, B. E., and J. Racine (2010): "Jackknife Model Averaging," *Unpublished Working Paper*.

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, David Draper and E. I. George, and a rejoinder by the authors.

Lai, T. L., and C. Z. Wei (1982): "Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems," *Annals of Statistics*, 10(1), 154–166.

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

Liu, Q. (2010): "Generalized CV and Generalized Cp Model Averaging," *Unpublished Working Paper*.

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.

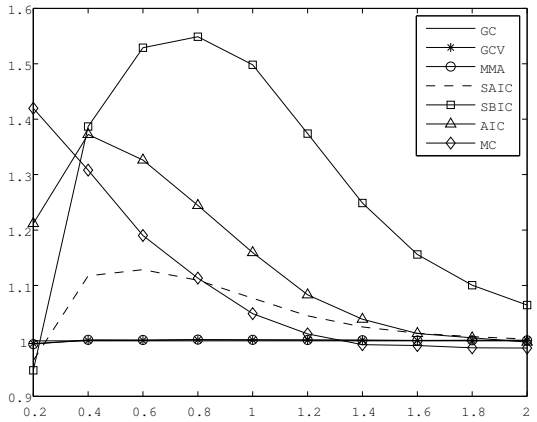

Figure 1. MSE ratios of models with homoskedastic errors.

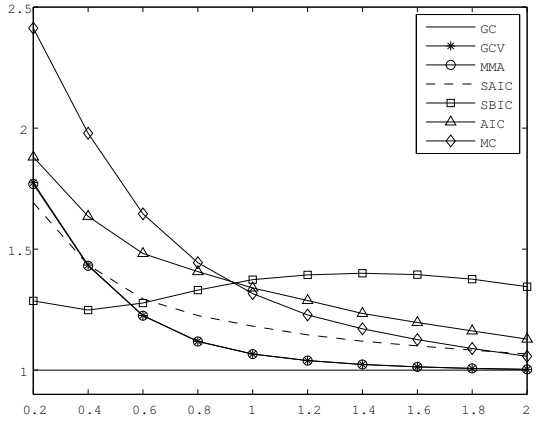

Figure 2. MSE ratios of models with heteroskedastic errors.