Generalized C_p Model Averaging for Heteroskedastic Models

Qingfeng Liu[1]
Department of Economics, Otaru University of Commerce
5-21, Midori 3-chome, Otaru, Hokkaido, 047-8501, Japan

October 15, 2010

[1]Tel&Fax: +86 134 27 5312, E-mail: qliu@res.otaru-uc.ac.jp

Abstract

This paper proposed a model averaging method, which is called Generalized Mallows' C_p model averaging (GC). It works well for heteroskedastic models. Under some regularity conditions, we show that our GC has asymptotic optimality as a model averaging method, and also has asymptotic optimality as a model selection method as well for heteroskedastic model. Some Monte-Carlo studies are performed to investigate the small sample properties of GC. The simulation results show that our method works well, gives better performance than other alternative methods.

JET classification: C51 C52

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

1 Introduction

The methods of model selection help us to choose a single model as the optimal one from a set of candidate models. In the last two decades, model averaging methods are proposed as alternative to model selection. A model averaging estimator is obtained by taking weighted average of estimator from a set of candidate models. Comparing with model selection, model averaging avoids to select a very poor model, and improve estimate in the aspect of risk. Model averaging methods can be separated into two groups, Bayesian and frequentist (non-Bayesian) model averaging. Bayesian model averaging 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 Bayesian one. In the literature of frequentist model averaging, Buckland, Burnham, Burnham, and Augustin (1997) proposed a smoothed-AIC-based and smoothed-BIC-based methods, Hjort and Claeskens (2003) proposed a frequentist model averaging method and derived the inference for the estimate based on the likelihood function of models. Recently, Hansen (Hansen (2007), Hansen (2009), Hansen (2010)) proposed several model averaging methods, which works 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 average estimator (MMA) for models with homoskedastic error. The weights of models for MMA are determined by minimizing a criterion similar to Mallows' C_p. Our extension is a generalization of MMA. The new method denoeted as GC works for both homoskedastic and heteroskedastic errors. Under some regularity conditions, we show that GC has asymptotic optimality as a model averaging method, and also has asymptotic optimality as a model selection method as well.

This paper is organized as follows. In section 2, the GC model averaging is proposed, it's optimality is argued as well. In section 3, some simulation studies are performed to check the finite sample properties of GC. Conclusion remark is stated in section 4. The last section is an appendix on some technical proofs.

2 Generalized Mallows' C_p Model Averaging

Hansen (2007) proposed MMA. In his set up, the regressors are assumed to be ordered, and the candidate regression models are assumed to be nested.

Wan, Zhang, and Zou (2010) extend the results of Hansen (2007), by remove the ordered and nested assumption. Our setup follows Wan, Zhang, and Zou (2010). Following the notation of Wan, Zhang, and Zou (2010), we have model (1) as follows,

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

for $i = 1, \cdots, n$, where y_i is a real-valued scalar, $x_i = (x_{1i}, x_{2i}, \cdots)$ is an countably infinite real-valued vector, μ_i is assumed to be converging in mean square and $E\mu_i^2 < \infty$, and the error term e_i is assumed to be independent and heteroskedastic, that means $E\left(e_i^2 | x_i\right) = \sigma_i^2$. The matrix form of regressors is $X \equiv (x_1', x_2', \cdots)'$. The matrix form of eq.(1) is $y = \mu + e$. where $\mu = (\mu_1, \cdots, \mu_n)'$. Our concern is to propose a model averaging method to estimate μ_i with small risk (mean squared error, MSE).

Our notations are almost identical to those in Wan, Zhang, and Zou (2010). The set of candidate models has M models. The mth model has $k_m > 0$ regressors which could be any variables in x_i. Notice that, we do not restrict $k_1 < k_2 < \cdots < k_M$, which means 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 $j = 1, \cdots, k_m$ is regressors in the mth model, and $\theta_{j(m)}$ are coefficients. We have a matrix form of eq.(2)

$$Y = X_{(m)}\Theta_{(m)} + b_{(m)} + e.$$

where $Y = (y_1, \cdots, y_n)'$, $X_{(m)}$ is the $n \times k_m$ matrix of regressors with ijth entry $x_{ij(m)}$, it is required to has 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 $e = (e_1,\ s, e_n)'$. The LS estimator of $\Theta_{(m)}$ from the mth model is $\hat{\Theta}_{(m)} = \left(X_{(m)}' X_{(m)}\right)^{-1} X_{(m)}' Y$. The estimator of μ is $\hat{\mu}_{(m)} = X_{(m)} \left(X_{(m)}' X_{(m)}'\right)^{-1} X_{(m)}' Y \equiv P_{(m)} Y$ and the residuals is $\hat{e}_{(m)} =$

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

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

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

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

The setup of the weight vector is different from that in Hansen (2007), he restricts the entries of the weight vector to be nonnegative integers time $1/n$ for the optimality of MMA.

Hansen's MMA can be applied to models with homoskedastic errors. Although it is hoped to be able to applied to models with heteroskedastic errors as well, but there is not any theory support for the optimality and no guaranty for good performance in the heteroskedastic case. In this section, we propose a Generalized Mallows C_p model averaging method (GC), which can be applied to models with heteroskedastic errors. We show the optimality of GC and will check it's small sample performance in the next section.

The model averaging criterion is defined as follows,

$$GC_n = n^{-1} \|Y - P(W) Y\|^2 + 2n^{-1} tr\left[\Omega P(W)\right],$$

where Ω is a $n \times n$ diagonal matrix which ii entry is σ_i^2. Then the optimal weight vector is derived as

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

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

$$L_n(W) = \|\hat{\mu}(W) - \mu\|^2, \tag{3}$$

and

$$R_n(W) = E\left(L_n(W) | X\right).$$

Then optimality means

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

It is easy to see that the expectation of GC_n is the risk function plus a constant. Hence GC_n can be regard as an unbiased estimator of the risk function plus a constant.

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

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

Theorem 2 *Under the assumption of Lemma 1, for* $\xi_n \equiv \inf_{W\in\mathcal{H}_n} R_n\left(W\right)$, *if* $E\left(e_i^{4G}|x_i\right) \leq \kappa < \infty$, $M\xi_n^{-2G}\sum_{m=1}^{M}\left(R_n\left(W_m^0\right)\right)^G \to 0$, $0 < \inf_i \sigma_i^2 \leq \sup_i \sigma_i^2 < \infty$, *and* $\inf_{W\in\mathcal{H}_n} L_n\left(W\right) = o\left(n\right)$, *then* $\frac{L_n\left(\hat{W}_{GC_n}\right)}{\inf_{W\in\mathcal{H}_n} L_n(W)} \to_p 1$.

The following theorem show that under some regularity conditions, if one replaces the term $n^{-1}tr\left[\Omega P\left(W\right)\right]$ in GC_n by $n^{-1}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}\left(W\right)$, where $\hat{e}_i$ is the residual from the model with all the regressors, and $p_{ii}\left(W\right)$ is the ii entry of $P\left(W\right)$, the above theorem will keep to be valid as the following theorem claims.

Theorem 3 *Using* $n^{-1}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}\left(W\right)$ *instead of* $n^{-1}tr\left[\Omega P\left(W\right)\right]$, *Theorem 1. is valid if* $\lim n^{-1}\sum_{i=1}^{n}\sigma_i^2 = \overline{\sigma^2} > 0$ *exists,* $\mu'\mu/n = O\left(1\right)$, $n\sup_{W\in\mathcal{H}_n}\max_{ii}\left[p_{ii}\left(W\right)\right]^2 \leq C_2 < \infty$, *and* $\sup_i\left[\left(\hat{\mu}_i\left(W\right) - \mu_i\right)^2|x\right] \leq n^{-1}C_3 R_n\left(W\right)$, *where* C_1, C_2 *and* C_3 *are positive constants.*

It is esay to understand that if we restrct the weight vector to be $W \in \{e_1, e_2, \cdots, e_M\}$, where e_i is a vectro with 1 as the ith entry and 0 as others, then GC works as a model selection procedure which select a single model. Sicne the above two Theorem work well for this model selection procedure, this model selection procedure has optimality as well.

3 Monte-Carlo Studies

In order to investigate the finite sample performance of our method, we carry out two Monte-Carlo simulations. The number of replications is 1000 for both simulations. For comparison, not only the results of GC but also the results of GCV (Liu 2010 Liu (2010)), MMA (Hansen 2007), Smoothed-AIC Buckland, Burnham, Burnham, and Augustin (1997), Smoothed-BIC Buckland, Burnham, Burnham, and Augustin (1997) and AIC (Akaike 1973)

Akaike (1973) are shown. GCV is a model averaging method proposed by Liu (2010) in an unpublished paper. The GCV is defined as

$$GCV_n(W) = \frac{n^{-1}\|Y - \hat{\mu}\|^2}{\left(1 - n^{-1}k(W)\right)^2}$$

and the optimal weight vector selected by GCV is defined as

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

Following the setting in Hansen (2007), we have DGP as

$$y_i = \sum_{j=1}^{\infty} \theta_j x_{ij} + e_i. \tag{4}$$

We cut off the infinite-order at $j = 30$. The parameters are determined by the same rule of Hansen (2007) $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$, we take the values $c = 0.2, 0.4, 0.6, \cdots, 2$, and $\alpha = 0.5$. The parameter c affects the population R^2 of eq.(4), bigger c brings bigger R^2. The sample size $n = 150$, the number of models $M = 10$ and the biggest model has 10 regressors. In the simulations just for simplicity, we employ nested setting, that means the $(k+1)$th model is nested in the kth. x_{ji} are independent over j, $j = 1, \cdots, m$, and set to be i.i.d. $N(0,1)$ with respect to i. The first one is a simulation with homoskedastic errors, by setting e_i to be $i.i.d.$ $N\left(0, \sigma^2\right)$ with $\sigma = 1$. In the second simulation study, we set e_i to be independent and heteroskedastic, follow $N\left(0, \sigma_i^2\right)$ with $\sigma_i = x_{2i}^2$. Since all the arguments in above sections are restricted in the situation conditional on X, we generate X once, then fixed the data of X through all replications. Defining MSE as $MSE = 1/1000\sum_{i=1}^{1000}(\hat{\mu} - \mu)^2$, after performing simulations, we calculate MSE ratios which are MSEs of all methods aforementioned divided by MSE of GC. The MSE ratios are plotted in Figures I and II for homoskedastic and heteroskedastic errors respectively.

We can see that, AIC is dominated by Smoothed-AIC(de noted as SAIC in the figures) with respect of almost all different values of c, different values of population R^2, in both simulations. The performance of SBIC is the poorest in the homoskedastic case, but is better than some others with small c in the heteroskedastic case. AIC and Mallows C_p (denoted as MC in the figures) have moderate performances, and GCV and MMA are better than them in both cases.

The most important result is on the comparison between GC and the pair, GCV and MMA. GCV is totally an alternative of MMA, they get

almost same MSE ratios. In the homoskedastic case, those three method get similar MSEs, our method GC works a little poorer than GCV and MMA, when $c < 0.4$, but a bit better than them with bigger c (bigger population R^2). In the heteroskedastic case, the situation is much different. Our GC has the best performance, particularly, GC works much better than GCV and MMA when c is small, and even for big values of c, GC is better than all the others. From these results, we know that our GC method works well for models with heteroskedastic errors.

4 Conclusion

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

5 Appendix

Proof of Theorem 2. After replace $\sigma^2 trP(W)$ by $tr\Omega P(W)$ and $\sigma^2 trP^2(W)$ by $tr\Omega P^2(W)$, the proof of Theorem 1 is almost the same as Wan, Zhang, and Zou (2010) proof of Theorem 1'. ■

Proof of Theorem 3. It is easy to see that

$$\begin{aligned}
&\sup_{W\in\mathcal{H}_n}\left\{\left|\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W) - tr\left[\Omega P(W)\right]\right| \Big/ R_n(W)\right\} \\
&\leq \sup_{W\in\mathcal{H}_n}\left|\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}(W) - \sum_{i=1}^{n}\sigma_i^2 p_{ii}(W)\right| \Big/ \xi_n \\
&\leq \sup_{W\in\mathcal{H}_n}\max_{ii}\left(p_{ii}(W)\right)\left|\sum_{i=1}^{n}\left(\hat{e}_i^2 - \sigma_i^2\right)\right| \Big/ \xi_n \\
&= \sup_{W\in\mathcal{H}_n}\max_{ii}\left(p_{ii}(W)\right)\left|\sum_{i=1}^{n}\left(\hat{e}_i^2 - \overline{\sigma_n^2}\right)\right| \Big/ \xi_n.
\end{aligned}$$

Then the rest of the proof is straightforward using the technique in the proof of the Theorem 2 of Wan, Zhang, and Zou (2010). ■

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.

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.

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.

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.

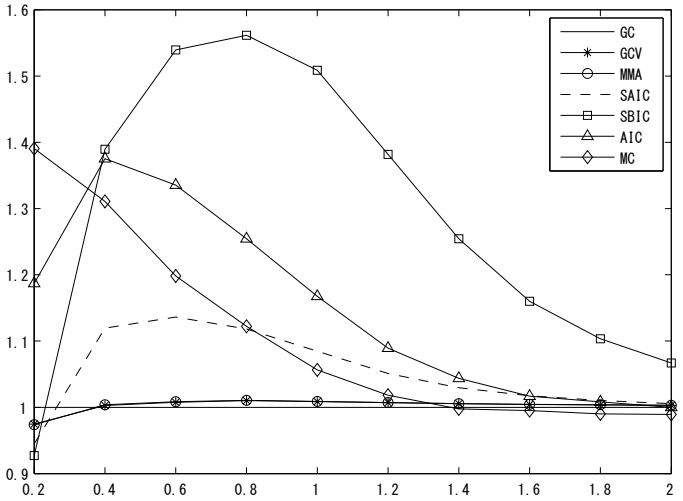

Figure 1. MSE ratios of models with homoscedastic errors.

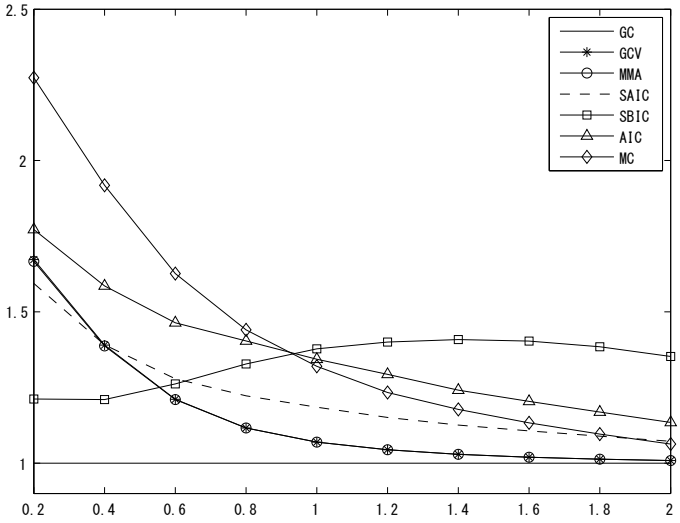

Figure 2. MSE ratios of models with heteroscedastic errors.