Generalized Cp Model Averaging for Heteroskedastic Models

Qingfeng Liu[1]
Otaru University of Commerce
Ryo Okui
Institute of Economic Research, Kyoto University

January 9, 2012

[1]Email: qliu@res.otaru-uc.ac.jp

The outline of this presentation

- A brief review of model averaging.
- Generalized Mallows' C_p Model Averaging for heteroscedastic models.
- Monte-Carlo studies.
- The conclusion remarks.

Model averaging in our daily life

- Doctors hold a consultation to determine an optimal treatment plan. Each doctor has one plan. Optimal plan=weighted averaged plan. The risk of misdiagnosis can be reduced.

Model averaging in our daily life

- Economists have many candidate models to explain economic phenomenon. Each model is reasonable to a certain extent.
- Using an averaged model (model averaging) instead of a particular model (model selection), the loss arising from misspecification can be reduced.

What is the model averaging in econometrics

- DGP

$$y = \mu(x) + e. \tag{1}$$

with $\mu(\cdot)$ is unknown. The target is to estimate μ at <u>low statistical risk</u>.

- We have a set of candidate models for $\mu(\cdot)$, to which K models belong

$$\mathcal{M} = \{M_1, M_2, \cdots, M_K\}.$$

- Based on model M_k, we can get $\hat{\mu}_{M_k}$ a estimator of μ.
- With a weight function $W(\cdot)$ (or a vector $W = (\omega_{(1)}, \cdots, \omega_{(K)})'$), model averaging estimator can be expressed as

$$\hat{\mu} = \sum_{M_k \in \mathcal{M}} W(M_k)\, \hat{\mu}_{M_k}. \tag{2}$$

Why do we use model averaging?

- Model averaging can reduce the loss and risk of estimation.
- Loss function and risk function for estimator with certain weight W

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

$$R_n(W) = E(L_n(W)|X), \tag{3}$$

- Optimality: we say a weight $\hat{W}$ or the estimator $\hat{\mu}(\hat{W})$ is optimal if

$$\frac{L_n(\hat{W})}{\inf_{W \in \mathcal{H}_n} L_n(W)} \underset{n\to\infty}{\overset{p}{\to}} 1, \tag{4}$$

$$\frac{R_n(\hat{W})}{\inf_{W \in \mathcal{H}_n} R_n(W)} \underset{n\to\infty}{\overset{p}{\to}} 1. \tag{5}$$

- In order to get an estimator of μ which achieves the infimum of the loss and risk, the task in the field of model averaging is to construct a model averaging criterion, by which one can find an optimal weight $\hat{W}$ and get the optimal estimator $\hat{\mu}(\hat{W})$.

The relationship between model averaging and model selection

- Model averaging is superior to model selection.
- A model selection method can be regarded as a model averaging with a special weight, $\mathrm{I}\left(M_k = M_{AIC}\right)$, where $\mathrm{I}(\cdot)$ is an indicator function.

$$\hat{\mu}_{AIC} = \sum_{M_k \in \mathcal{M}} \mathrm{I}\left(M_k = M_{AIC}\right) \hat{\mu}_{M_k}. \tag{6}$$

- Hence, with a optimal weight model averaging estimator can achieve lower risk than model selection estimator.

Existing Researches on model averaging method

- Bayesian model averaging estimators (For review see Hoeting (1999)).
- Weighted-average least squares (WALS) (Magnus etal., 2010, Magnus etal., 2011).
- Smoothed BIC, AIC (Buckland etal., 1997).
- Hansen's MMA for homoscedastic models (Hansen, 2007).
- JMA for homoscedastic models (Hansen and Racine, 2010).
- This paper extends Hansen's MMA, and propose a model averaging method for heteroskedatic case.

Bayesian model averaging

- Take $P\left(M_k\right)$ as the prior probability of model M_k, and $\pi\left(\theta_k|M_k\right)$ as the prior density of θ_k conditional on model M_k.
- Bayesian model averaging estimator

$$\hat{\mu} = E\left(\mu|y\right) = \sum_{k=1}^{K} P\left(M_k|y\right) E\left(\mu|M_k, y\right) \tag{7}$$

- Posterior density

$$\pi\left(\mu|y\right) = \sum_{k=1}^{K} \pi\left(\mu|M_k, y\right) P\left(M_k|y\right)$$

- Posterior density of M_k

$$P\left(M_k|y\right) = \frac{P\left(M_k\right)\lambda_k}{\sum_{k=1}^{K} P\left(M_k\right)\lambda_k} \tag{8}$$

- λ_k is the integrated likelihood of M_k

$$\lambda_k = \int L\left(y|M_k, \theta_k\right)\pi\left(\theta_k|M_k\right) d\theta_k \tag{9}$$

Smoothed BIC and AIC

- According to Claeskens and Hjort (2008) $BIC \approx -2\log(\lambda_k)$.
- Assuming $P(M_k)$ is $k-$homogeneous, from (8)

$$P(M_k|y) = \frac{P(M_k)\lambda_k}{\sum_{k=1}^{K} P(M_k)\lambda_k} \tag{10}$$

we have

$$P(M_k|y) \approx \frac{\exp(-BIC_k/2)}{\sum_{k=1}^{K}\exp(-BIC_k/2)}. \tag{11}$$

- Smoothed-BIC-Based estimator

$$\hat{\mu}_{MA-BIC} = \sum_{M_k \in \mathcal{M}} c_{BIC}(M_k)\,\hat{\mu}_{M_k}, \tag{12}$$

$$c_{BIC}(M_k) = \frac{\exp(-BIC_k/2)}{\sum_{k=1}^{K}\exp(-BIC_k/2)}. \tag{13}$$

- Smoothed-AIC has a similar form

$$c_{AIC}(M_k) = \frac{\exp(-AIC_k/2)}{\sum_{k=1}^{K}\exp(-AIC_k/2)}. \tag{14}$$

Asymptotic distribution of model averaging estimators under parametric setup

- Hjort and Claeskens (2003) take the following local misspecification setup, avoiding domination by bias

$$f_{true}\left(y\right) = f_n\left(y\right) = f\left(y, \theta_0, \gamma\right), \tag{15}$$

$$\gamma = \gamma_0 + \frac{1}{\sqrt{n}}\delta. \tag{16}$$

- The most narrow model is $f_{narr}\left(y, \theta\right) = f\left(y, \theta, \gamma_0\right)$, the full model is $f_{full}\left(y, \theta, \gamma\right)$ including all parameters in δ.
- Model averaging estimator follow non-normal distribution

$$\hat{\mu} = \sum_{j \in 2^K} W\left(M_{S_j}\right)\hat{\mu}_{S_j}. \tag{17}$$

Setup and Purpose

- DGP: infinite dimensional linear model

$$y_i = \mu_i + e_i, \tag{18}$$

$$\mu_i = \sum_{j=1}^{\infty} \theta_j x_{ij}, \tag{19}$$

$$E\left(e_i | x_i\right) = 0,$$

$$E\mu_i^2 < \infty$$

- Heteroskedasticity

$$E\left(e_i^2 | x_i\right) = \sigma_i^2,$$

- Propose a model averaging method for heteroskedatic case, estimate μ_i at low risk.

- Notice that we change the meaning of the notation M and K hereafter.
- M denotes the total number of candidate models in the candidate set. The mth model has $k_m > 0$ regressors which could be any variables in x_i.
- The mth approximating model

$$y_i = \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)} + b_{i(m)} + e_i \tag{20}$$

$$b_{i(m)} = \sum_{j=1}^{\infty} \theta_j x_{ij} - \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)} \tag{21}$$

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

- The LS estimator from the mth model

$$\hat{\Theta}_{(m)} = \left(X'_{(m)}X_{(m)}\right)^{-1} X'_{(m)}Y$$

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

- The model averaging estimator of μ

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

where

$$W=\left(\omega_{(1)},\cdots,\omega_{(M)}\right)'\in H_n\equiv\left\{W\in\left[0,1\right]^M:\sum_{m=1}^{M}\omega_{(m)}=1\right\}.$$

- In Hansen (2007)
$H_n\equiv\left\{W\in\left[0,1\right]^M:\sum_{m=1}^{M}\omega_{(m)}=1,\omega_{(m)}=c/n,c=1,\cdots,n.\right\}$

Hansen's MMA for homoscedastic models

- Hansen's MMA (Mallows' Cp Model Averaging): in order to obtain a optimal model averaging estimator, which can achieve the infimum of the loss and risk, Hansen proposed the following criterion to select optimal weight

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

- Optimal weight

$$\hat{W}_{\widehat{C_n}} = \arg\min_{W \in \mathcal{H}_n} \widehat{C}_n.$$

- Hansen's MMA has optimality for homoscedastic models but not for heteroscedastic models.

Our Generalized Cp for heteroscedastic models

- We propose a Generalized Cp model averaging method which has optimality for heteroscedastic models, $E\left(e_i^2|x_i\right)=\sigma_i^2$.
- Generalized Cp model averaging is an extension of Hansen's MMA and Andrews (1991).
- Generalized Cp model averaging criterion

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

where Ω is a $n\times n$ diagonal matrix which ii entry is σ_i^2.
- The expectation of GC is the risk function plus a constant.

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

Optimality of GC

Th. 2. As $n \to \infty$, and $M \to \infty$, for $\xi_n \equiv \inf_{W \in \mathcal{H}_n} R_n(W)$ and some integer $1 \leq G < \infty$, if

$$E\left(e_i^{4G} | x_i\right) \leq \kappa < \infty, \tag{22}$$

$$M\xi_n^{-2G} \sum_{m=1}^{M} \left(R_n\left(W_m^0\right)\right)^G \to 0, \tag{23}$$

$\mu'\mu/n = O(1)$, and $0 < \inf_i \sigma_i^2 \leq \sup_i \sigma_i^2 < \infty$, then

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

W_m^0 is a vector whose mth element is one and all other elements are zeros.

Feasible GC

- Replace $tr\left[\Omega P\left(W\right)\right]$ by

$$\frac{n}{n-K}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}\left(W\right)$$

$$\widehat{GC}_n \equiv \left\|Y - P\left(W\right)Y\right\|^2 + 2\frac{n}{n-K}\sum_{i=1}^{n}\hat{e}_i^2 p_{ii}\left(W\right), \qquad (24)$$

where $\hat{e}_i$ is the residual from the biggest model, and K is the number of regressors in the biggest model.

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

Optimality of feasible GC

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

$$0 < \lim n^{-1} \sum_{i=1}^{n} \sigma_i^2 = \overline{\sigma^2} < \infty, \tag{25}$$

$$\max_{1 \le m \le M} \max_{1 \le i \le n} p_{m,ii} = O\left(n^{-1/2}\right), \tag{26}$$

$$\frac{\tilde{p} e' e}{\xi_n} \xrightarrow{p} 0, \tag{27}$$

where $\tilde{p} \equiv \sup_{W \in \mathcal{H}_n} \max_{1 \le i \le n} \left(p_{ii}(W)\right)$, and $p_{m,ii}$ is the ith diagonal element of $P_{(m)}$.

- The proof of optimality under some regularity conditions is an extension of Wan et al. (2010).

GC works as a model selection criterion

- The criterion for model selection:

$$\widehat{GC}_n(m) \equiv \|Y - P_m Y\|^2 + 2\frac{n}{n-K}\sum_{i=1}^{n} \hat{e}_i^2 p_{m,ii}. \tag{28}$$

- The estimator of the indicator of the optimal model:

$$\hat{m} \equiv \arg \min_{1 \le m < M} \widehat{GC}_n(m). \tag{29}$$

Outline of the proof of Th.2.

- Since

$$GC_n = L_n(W) + \|e\|^2 + 2\langle e, (I - P(W))\mu\rangle \\ + 2\left(tr\left[\Omega P(W)\right] - \langle e, P(W)\mu\rangle\right)$$

- We just need to show

$$\sup_{W \in \mathcal{H}_n} \left|\langle e, (I - P(W))\mu\rangle\right| / R_n(W) \to_p 0$$

$$\sup_{W \in \mathcal{H}_n} \left|tr\left[\Omega P(W)\right] - \langle e, P(W)\mu\rangle\right| / R_n(W) \to_p 0$$

$$\sup_{W \in \mathcal{H}_n} \left|L_n(W) / R_n(W) - 1\right| \to_p 0$$

Outline of the proof of Th.3.

- $\tilde{p} \equiv \sup_{W\in\mathcal{H}_n} \max_{1\le i\le n} \left(p_{ii}\left(W\right)\right)$, P^* is the projection matrix of the model with all regressors, p_{ii}^* is the ith diagonal element of P^*, $\bar{p}^* \equiv n^{-1}\sum_{i=1}^n p_{ii}^*$.
- Condition (26) implies that $\tilde{p} = O\left(n^{-1/2}\right)$ and $K = O\left(n^{1/2}\right)$; condition (23) implies that $\xi_n \to \infty$.
- Since

$$\widehat{GC} = GC + 2\left(\sum_{i=1}^n \hat{e}_i^2 p_{ii}\left(W\right) - tr\left[\Omega P\left(W\right)\right]\right) + \frac{2K}{n-K}\sum_{i=1}^n \hat{e}_i^2 p_{ii}\left(W\right). \tag{30}$$

to prove Theorem 3, we only need to show that

$$\sup_{W\in\mathcal{H}_n}\left\{\left|\sum_{i=1}^n \hat{e}_i^2 p_{ii}\left(W\right) - tr\left[\Omega P\left(W\right)\right]\right| \Big/ R_n\left(W\right)\right\} \xrightarrow{p} 0. \tag{31}$$

$$\sup_{W\in\mathcal{H}_n}\left\{\frac{K}{n-K}\sum_{i=1}^n \hat{e}_i^2 p_{ii}\left(W\right) \Big/ R_n\left(W\right)\right\} \xrightarrow{p} 0. \tag{32}$$

Monte-Carlo Studies

- The data generating process is:

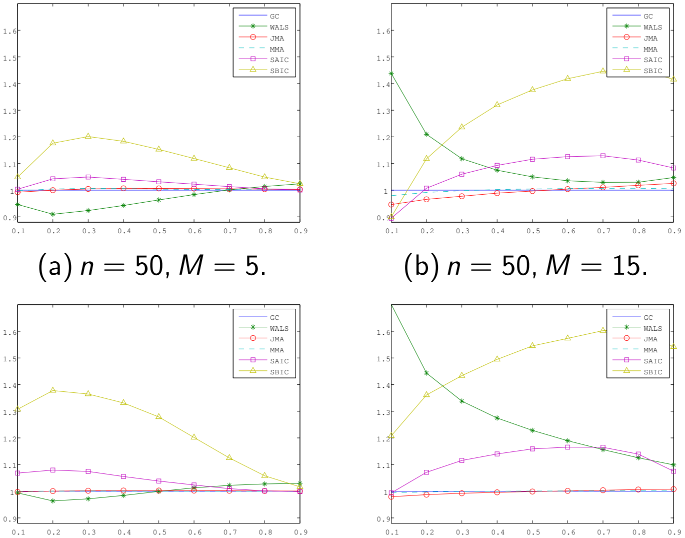
$$y_i = \sum_{j=1}^{10000} \theta_j x_{ij} + e_i.$$

- Draw a random sample of $\{x_i, e_i\}$ for each replication such that $x_{i1} = 1$ and other x_{ij} are i.i.d. $N(0, 1)$.
- $e_i \sim N(0, \sigma_i^2)$ is independent of x_{ij}.
- $\sigma_i^2 = 1$ (homoskedastic), and $\sigma_i^2 = x_{2i}^4 + 0.01$ (heteroskedastic).
- $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, where the parameter $\alpha = 0.5$, which determines how quickly the magnitude of θ_j decays as j increases, and we vary the values of c so that the population R^2 increases with c from 0.1 to 0.9

- The sample size is $n = 50$ and $n = 150$.
- The number of observable regressors K is 5 and 15 when $n = 50$, and 10 and 30 when $n = 150$.
- We consider K different models so that $M = K$. The kth model includes the first k regressors and the $(k+1)$th model is nested in the kth model.
- The number of replications is 1000.

Remarks

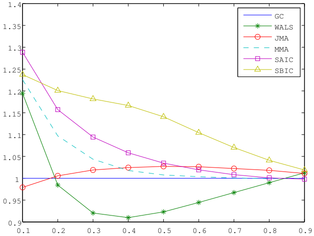
- WALS for heteroskedastic models, proposed by Magnus etal. 2011, is a Bayesian combination of frequentist estimators. It has bounded risk, and it's computational effort is negligible.
- JMA is propose by Hansen and Racine (2010) based on Jackknife for heteroskedastic models.



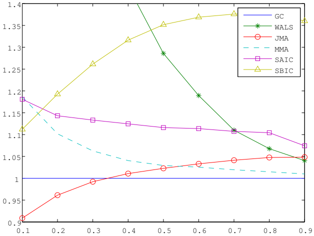
(a) $n = 50$, $M = 5$. (b) $n = 50$, $M = 15$.

(c) $n = 150$, $M = 10$. (d) $n = 150$, $M = 30$.

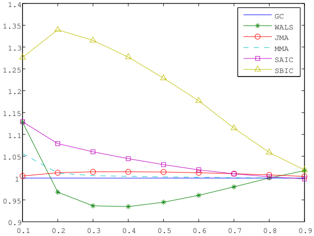
Figure: Homoskedastic Cases



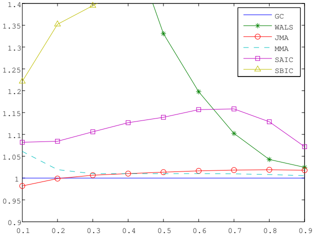
(a) $n = 50$, $M = 5$.



(b) $n = 50$, $M = 15$.



(c) $n = 150$, $M = 10$.



(d) $n = 150$, $M = 30$.

Figure: Heteroskedastic Cases

Conclusion remark

- We proposed a model averaging methods for heteroscedastic models.
- Our Gp model averaging method optimality of this method.
- The results of Monte-Carlo studies showed that our method works well.

Thank you very much and welcome to Otaru city!