Consistency of model averaging estimators

Xinyu Zhang

International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, China
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Highlights

• The local mis-specification is not assumed in this paper.
• I prove the root-\(n\) consistency of Smoothed AIC and Smoothed BIC estimators.
• I prove the root-\(n\) consistency of Mallows model averaging and jackknife model averaging estimators.

Abstract

Recently, there has been increasing interest in model averaging within frequentist paradigm. In the existing literature, for proving consistency of model averaging estimators, local mis-specification is assumed. In this paper, we show that under general fixed parameter setup, the model averaging estimators remain root-\(n\) consistent. This result provides a new theoretical basis for the use of model averaging estimators.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In the past two decades there has been a substantial deal of research devoted to model averaging. There are basically two major model averaging mechanisms: Bayesian model averaging and frequentist model averaging (FMA), and here we restrict our attention to FMA. Some FMA strategies have been developed and studied, in which, weighting strategy based on the AIC or BIC scores proposed by Buckland et al. (1997) is probably the most widely used method because of convenience. See, for example, Burnham and Anderson (2004) and Layton and Lee (2006). Recently, there has been increasing interest in asymptotically optimal model averaging including Mallows model averaging (Hansen, 2007), optimal mean squared error averaging (Liang et al., 2011), jackknife model averaging (JMA) (Hansen and Racine, 2012), heteroskedasticity-robust C\(p\) (Liu and Okui, 2013), and so on.

However, few authors have examined the consistency of model averaging estimators. Very recently, Hansen (2014) and Liu (2015) proved that the MMA and JMA estimators are root-\(n\) consistent. Both of them used local mis-specification assumption. Local mis-specification assumption means that some parameters are of order \(n^{-1/2}\), where \(n\) is sample size. Although this assumption provides a suitable framework for studying the asymptotic theories of FMA estimators, it also draws comments from Raftery and Zheng (2003) and Ishwaran and Rao (2003) because of its realism. Therefore, there is an immediate question to be answered: are these model averaging estimators still consistent under general fixed parameter setup? In the current paper, under general fixed parameter setup, we derive consistency of some commonly used FMA estimators, including estimators by weighting strategy based on the AIC or BIC scores, MMA, and JMA.

The plan of this paper is as follows. In Section 2 we introduce the model averaging estimators. Section 3 develops root-\(n\) consistency of these estimators. In Section 4, we conduct simulation study...
to check the root-\(n\) consistency. Section 5 concludes. Proofs are contained in Appendix.

2. Model averaging estimators

We are concerned with linear regression model:

\[
y_i = x_i^T \beta + z_i^T \gamma + e_i, \quad E(e_i | x_i, z_i) = 0, \\
E(e_i^2 | x_i, z_i) = \sigma^2(x_i, z_i), \quad i = 1, \ldots, n
\]

(1)

where \(y_i\) is a scaled dependent variable, \(x_i(1 \times 1)\) and \(z_i(2 \times 1)\) contain independent variables, and \(x_i\) and \(z_i(1 \times 1)\) are parameter vectors, \(e_i\) is an error term, and \((y_i, x_i, z_i)\) for \(i = 1, \ldots, n\) are independent. Here, \(x_i\) contains the variables that must be included in the model, while we are uncertain whether the variables in \(z_i\) are included. So we have 2\(^2\) sub-models of (1). To be more general, we combine \(M\) ones of these sub-models, where \(M\) can be smaller than 2\(^2\) because some sub-models may be excluded before model averaging. Model (1) is widely used (see, e.g., Leamer, 1978; Magnus and Durbin, 1999; Lian et al., 2011).

In matrix notation, model (1) can be rewritten as

\[
y = X\beta + Z\gamma + e.
\]

(2)

where \(y = (y_1, \ldots, y_n)^T\), \(X = (x_1, \ldots, x_n)^T\), \(Z = (z_1, \ldots, z_n)^T\), and \(e = (e_1, \ldots, e_n)^T\). Let \(q = q_1 + q_2\), \(\theta = (\theta_1, \ldots, \theta_q)^T = (\beta^T, \gamma^T)\) and \(H = (X, Z)\). We assume that \(H\) has full column rank \(q\). The \(m\)th sub-model (or candidate model) includes all variables in \(X\) and some (all or none) of the variables in \(Z\), denoted as \(Z_m\). Let \(H_m = (X, Z_m), k_m\) be the number of columns of \(H_m\), and \(\Pi_m\) be a selection matrix so that \(\Pi_m = (I_{k_m}, 0_{k_m \times (q - k_m)})\) or a column permutation thereof and thus \(H_m = H_m^T\), where \(I_{k_m}\) is a \(k_m \times k_m\) identity matrix. Under the \(m\)th candidate model, \(\theta_m = \Pi_m^T (H_m H_m)^{-1} H_m^T y\).

Therefore, the model averaging estimator of \(\theta\) is

\[
\hat{\theta}(w) = \sum_{m=1}^M w_m \hat{\theta}_m,
\]

where \(w_m\) is weight corresponding to the \(m\)th candidate model and \(w = (w_1, \ldots, w_M)^T\), belonging to weight set \(w = [w \in [0, 1]^M : \sum_{m=1}^M w_m = 1]\). Next, we introduce some model averaging strategies.

Weighting strategy based on the AIC or BIC scores

In the \(m\)th candidate model, the AIC and BIC scores are

\[
AIC_m = n \log \hat{\sigma}_m^2 + 2k_m, \quad BIC_m = n \log \hat{\sigma}_m^2 + k_m \log n, \text{ respectively},
\]

where \(\hat{\sigma}_m^2 = \frac{\text{dim}1/2}{\text{dim}1/2 + \text{dim}1/2}\), \(\text{dim1/2}\) weights are

\[
\hat{\omega}_{AIC_m} = \exp(-\text{AIC}_m/2)/\sum_{m=1}^M \exp(-\text{AIC}_m/2), \quad m = 1, \ldots, M.
\]

(3)

where \(\hat{\omega}_{AIC_m}\) is the AIC or BIC score under the \(m\)th candidate model. The averaging estimators combined by weights in (3) are commonly called smoothed AIC (S-AIC) or smoothed BIC (S-BIC) estimators.

MMSE

Let \(\hat{\beta}_m = H_m(H_m^T H_m)^{-1} H_m^T y\), \(P_m = \sum_{m=1}^M w_m P_m\), and \(k = (k_1, \ldots, k_M)^T\). For homoscedastic error setting (that is \(\sigma^2(x_i, z_i) = \cdots = \sigma^2(x_i, z_i) = \sigma^2\)), Hansen (2007) proposed choosing weights by minimizing Mallows criterion \(C(w) = \|H \hat{\beta}_m - P_m y\|_2^2 + 2\hat{\sigma}_m^2 w_k\), where \(\|\cdot\|_2^2\) stands for the Euclidean norm, \(\hat{\sigma}_m^2 = (n - k_m)^{-1} \|H_m \hat{\beta}_m - P_m y\|_2^2\), and \(m^* = \arg \min_{m=1(1), M} \hat{\sigma}_m^2\). Denote \(\hat{\omega}_{MMA} = \arg \min_{w \in [0, 1]^M} C(w)\), so that the combined estimator \(\hat{\omega}(\hat{\omega}_{MMA})\) is the MMA estimator of \(\theta\). Although the MMA estimator is proposed under homoscedastic error setting, the following section shows that it is root-\(n\) consistent under model (1) which is conditionally heteroscedastic.

JMA

Write \(\hat{P}_m = D_m(\hat{P}_m - I_n) + I_n\), where \(D_m\) is a diagonal matrix with \((1 - \hat{p}_m^{-1})^{-1}\) being its \(i\)th diagonal element and \(\hat{p}_m^{-1}\) is the \(i\)th diagonal element of \(P_m\). Let \(\hat{P}_m(\hat{w}) = \sum_{m=1}^M w_m \hat{P}_m\). For heteroscedastic error setting, Hansen and Racine (2012) proposed choosing weights by minimizing cross-validation (or jackknife) criterion \(J(w) = \|H_m - P_m^{w}\|_2^2\). Let \(\hat{\omega}_{JMA} = \arg \min_{w \in [0, 1]^M} J(w)\), so that the combined estimator \(\hat{\omega}(\hat{\omega}_{JMA})\) is the JMA estimator of \(\theta\).

3. Root-\(n\) consistency

Before presenting asymptotic results, we define some special models. A model including all regressors with nonzero coefficients and only these regressors is named true model. Any candidate model omitting regressors with nonzero coefficients is called under-fitted model, contained by set \(U\). Typically, if \(q_2\) is large, then there are a large number of sub-models of (2) and thus practitioner generally screening out some of them as candidate models. It is possible that the true model is not one of the candidate models. We do not impose the assumption that the true model must be one of candidate models. We need that at least one candidate model is not under-fitted.

We list some regularity assumptions required for asymptotic results, where all limiting processes here and throughout the text are with respect to \(n \to \infty\).

Condition (C.1). \(\psi_n \equiv n^{-1/2} H^T e \to \psi\), where \(\psi\) is a positive definite matrix, and \(n^{-1/2} H^T e = O_p(1)\).

Condition (C.2). \(p^* = O(n^{-1})\), where \(p^* = \max_{1 \leq i \leq M} \max_{1 \leq j \leq n} P_m^i\).

Conditions (C.1) is very commonly used in literature such as Liu (2015). Condition (C.2) is quite mild. Instead, Andrews (1991) assumed that \(P_m^i \leq c k_m^{-1}\) for some constant \(c < \infty\), which is equivalent to Condition (C.2) under our model (2). The following theorems show that the S-AIC, S-BIC, MMA and JMA estimators are root-\(n\) consistent under these regularity conditions.

Theorem 1. Under Condition (C.1), \(\sqrt{n} \hat{\theta}(\hat{w}_{AIC}) - \theta = O_p(1)\) and \(\sqrt{n} \hat{\theta}(\hat{w}_{BIC}) - \theta = O_p(1)\).

Theorem 2. Under Condition (C.1), \(\sqrt{n} \hat{\theta}(\hat{w}_{MMA}) - \theta = O_p(1)\). If Condition (C.2) is further satisfied, then \(\sqrt{n} \hat{\theta}(\hat{w}_{JMA}) - \theta = O_p(1)\).

4. Simulation study

In this section we conducted a simulated example to check the consistency of the model averaging estimators. We generated data from the model (1) in which we set \(\beta = (1, 1)^T\), \(y = (0.5, 0.7, 0, -0.3)^T\), \(x_1 = (1, x_2)^T\), \(z_1 = (z_1, \ldots, z_4, x_2, z_1, \ldots, z_4)^T \sim \mathcal{N}(0, I_{16})\), and \(e_1 \sim \mathcal{N}(0, \Sigma_{16})\). We combined all \(2^4 = 16\) sub-models. We let sample size \(n\) vary in \(\{40, 60, 80, 100, 200, 300, 400, 600, 800, 1000\}\), and calculated mean squared errors (MSEs) of the model averaging estimators on the basis of 1000 replications.

To show the root \(n\) consistency of the model averaging estimators, we presented \(n^{0.95} \times\) MSE in Table 1. It is seen that when \(n\) increases, \(n^{0.95} \times\) MSE decreases, which accords to the root \(n\) consistency.
5. Conclusion

Without using the local mis-specification assumption, this paper has derived the root-$n$ consistency of the S-AIC, S-BIC, MMA, and JMA estimators. This result builds a theoretical basis for the use of model averaging estimators under fixed parameter setup. We conjecture that the consistency of other model averaging estimators can be developed using the proving technique of this paper.

Acknowledgments

The author thanks the referee for very constructive comments. This research was supported by National Natural Science Foundation of China (Grant nos. 11471324 and 11271355).

Appendix

We first define some notations. Denote $H_{m'}$ as a matrix consisting of columns of $H$ not contained in $H_m$ and $\Pi_{m'}$ as a selection matrix such that $H_{m'} = H\Pi_{m'}$. Let $\theta_m = \Pi_m \theta$ and $\theta_{m'} = \Pi_{m'} \theta$.

A.1. Proof of Theorem 1

Since $\Psi$ is a positive definite matrix, it follows from Rao (1973) that

$$|\Psi| = |\Psi_{11}| |\Psi_{12}| = |\Psi_{11}| |\Psi_{22} - \Psi_{21} \Psi_{11}^{-1} \Psi_{12}|,$$

by which we have $|\Psi_{22} - \Psi_{21} \Psi_{11}^{-1} \Psi_{12}| > 0$. Thus by Condition (C.1), for any $m$, there exists a positive definite matrix $G_m$ such that

$$n^{-1}H_m^T(I_n - P_m)H_m = n^{-1}H_m^T H_{m'} H_{m'}^T H_m - n^{-1}H_m^T H_{m'} H_{m'}^T n^{-1}G_m \rightarrow G_m. \tag{A.1}$$

Let $g_m = \theta_m^T G_m \theta_m$. From (A.1) and Condition (C.1), we have for any $m \in U$,

$$n^{-1}||y - P_my||^2 = n^{-1}(e + H_m \theta_{m'})(I_n - P_m)(e + H_m \theta_{m'})$$

$$= n^{-1}e^T H_m^T (I_n - P_m) H_m \theta_{m'}$$

$$+ n^{-1}||e||^2 + 2n^{-1}e^T (I_n - P_m) H_m \theta_{m'} - n^{-1}e^T P_m e$$

$$= g_m + n^{-1}||e||^2 + O_p(1), \tag{A.2}$$

and for any $m \not\in U$,

$$n^{-1}||y - P_my||^2 = n^{-1}e^T (I_n - P_m) e = n^{-1}||e||^2 + O_p(1). \tag{A.3}$$

Let $m'$ be a model not belonging to $U$. So from above results and $n^{-1}||e||^2 = O_p(1)$ we obtain that when $m \in U$,

$$\hat{\omega}_{\text{MMA,m}}^{-1} \omega_{\text{MMA,m}} = \exp(\text{MMA}/2 - \text{MMA}/2)$$

$$= (||y - P_my||^2 - ||y - P_my||^2)^{n/2}$$

$$\times \exp(k_m - k_m) = O_p(c^n), \tag{A.4}$$

where $c$ is a generic constant belonging to $(0, 1)$. Since $\hat{\omega}_{\text{MMA}} \in W$, (A.4) implies $\hat{\omega}_{\text{MMA,m}} = O_p(c^n)$. From Condition (C.1), it is straightforward to show that for any $m \not\in U$,

$$\sqrt{n} (\theta_m - \theta) = O_p(1)$$

and for any $m \in U$,

$$\hat{\theta}_m = \Pi_m^T (H_m^T H_m)^{-1} H_m^T y = \Pi_m^T (H_m^T H_m)^{-1} H_m^T \theta$$

$$+ \Pi_m^T (H_m^T H_m)^{-1} H_m^T e = O_p(1). \tag{A.5}$$

Hence,

$$\sqrt{n} (\hat{\theta}_m - \theta) = O_p(1)$$

and for any $m \not\in U$,

$$\hat{\omega}_{\text{MMA,m}}^{-1} \omega_{\text{MMA,m}} = \exp(\text{MMA}/2 - \text{MMA}/2)$$

$$= (||y - P_my||^2 - ||y - P_my||^2)^{n/2} \exp(k_m - k_m)/2$$

$$= O_p(c^n n^{\eta_n/2}),$$

and thus $\hat{\omega}_{\text{MMA,m}} = O_p(c^n n^{\eta_n/2})$. Using the steps of (A.6), we have

$$\sqrt{n} (\hat{\theta}_m - \theta) = O_p(1). \tag{A.6}$$

A.2. Proof of Theorem 2

It is seen that

$$c(\omega) = ||(I_n - P_m)(y)||^2 + 2\sigma^2 w_k^T$$

$$= ||H\theta(\omega) - \mu - e||^2 + 2\sigma^2 w_k^T$$

$$= ||e||^2 + \theta(\omega - \theta)^T H^T H \theta(\omega - \theta)$$

$$- 2e^T H \hat{\theta}(\omega - \theta) + 2\sigma^2 w_k^T. \tag{A.7}$$

It follows from (A.2)–(A.3) that

$$\sigma^2 = (n - k_m^{-1})||y - H\hat{\theta}_m||^2 = O_p(1). \tag{A.8}$$

Let $j$ be a model not belonging to $U$. When $u_j = 1$, from (A.8) and Condition (C.1), we have $c(\omega) = ||e||^2 + \eta_n$ with $\eta_n = O_p(1)$. Hence, $c(\omega_{\text{MMA}}) \leq ||e||^2 + \eta_n$, which, together with (A.7), implies

$$\eta_n \geq \theta(\omega_{\text{MMA}} - \theta)^T H^T H \theta(\omega_{\text{MMA}} - \theta)$$

$$- 2e^T H \hat{\theta}(\omega_{\text{MMA}} - \theta) + 2\sigma^2 w_{k_m}^T. \tag{A.9}$$

Let $\lambda_{\text{min}}(K)$ be the smallest eigenvalue of a matrix $K$. From (A.9), we have

$$\lambda_{\text{min}}(\Psi_n) \|\sqrt{n}(\hat{\theta}(\omega_{\text{MMA}}) - \theta)\|$$

$$\leq \|\hat{\theta}(\omega_{\text{MMA}} - \theta)\| H^T H \theta(\omega_{\text{MMA}} - \theta)$$

$$\leq \eta_n + 2e^T H \hat{\theta}(\omega_{\text{MMA}} - \theta) - 2\sigma^2 w_{k_m}^T$$

$$\leq \eta_n + \eta_n \|H\theta(\omega_{\text{MMA}} - \theta)\| \leq \eta_n + \lambda_{\text{min}}(\Psi_n) - \eta_n \|H\theta(\omega_{\text{MMA}} - \theta)\|$$

and thus

$$\lambda_{\text{min}}(\Psi_n) \|\sqrt{n}(\hat{\theta}(\omega_{\text{MMA}}) - \theta)\| - \{\lambda_{\text{min}}(\Psi_n)\}^{-1} \|n^{-1/2} e^T H\|$$

$$\leq \eta_n + \{\lambda_{\text{min}}(\Psi_n)\}^{-1} \|n^{-1/2} e^T H\|^2.$$
which is equivalent to
\[
\sqrt{n}|\hat{\vartheta}(\hat{\theta}_{\text{MMA}}) - \theta| \leq (\lambda_{\min}(\Psi_n))^{-1}\|\eta_n + (\lambda_{\min}(\Psi_n))^{-1}n^{-1/2}e^TH\|^2)^{1/2}
\]\[
+ (\lambda_{\min}(\Psi_n))^{-1}\|\eta_n + (\lambda_{\min}(\Psi_n))^{-1}n^{-1/2}e^TH\|^2)^{1/2}
\]
which, together with \(\eta_n = O_p(1)\) and \text{Condition (C.1)}, implies
\[
\sqrt{n}|\hat{\vartheta}(\hat{\theta}_{\text{MMA}}) - \theta| = O_p(1).
\]
For JMA, denote \(Q_m\) as an \(n \times n\) diagonal matrix with the \(i\)th diagonal element \(Q_{m,ii} = p_{m} / (1 - p_{m})\), so \(D = Q + I_n\). Let \(\Xi\) be an \(M \times M\) matrix with the \(j\)th element
\[
\Xi_{mj} = (e + H_m\theta_m)\top(I_n - P_m)(Q_m + Q_j + Q_mQ_j)(I_n - P_j)
\]
so that
\[
\mathcal{f}(w) = C(w) + w^\top \Xi w. \tag{A.10}
\]
Let \(\delta(K)\) be the largest singular value of a matrix \(K\). From \cite{Li (1987)}, we know that any two \(n \times n\) matrices \(K_1\) and \(K_2\), \(\delta(K_1K_2) \leq \delta(K_1)\delta(K_2)\) and \(\delta(K_1 + K_2) \leq \delta(K_1) + \delta(K_2)\), by which and \text{Conditions (C.1)-(C.2)}, we have
\[
(e + H_m\theta_m)\top(I_n - P_m)(Q_m + Q_j + Q_mQ_j)(I_n - P_j)
\]
\[
\leq \|e + H_m\theta_m\| \|e + H_m\theta_j\|
\]
\[
\leq \|e + H_m\theta_m\| \|e + H_m\theta_j\|^{2p^* + (p^*)^2} = O_p(1), \tag{A.11}
\]
so for any \(w \in \mathcal{W}, w^\top \Xi w = O_p(1)\). Now, using the arguments in proving root-\(n\) consistency of \(\hat{\vartheta}(\hat{\theta}_{\text{MMA}})\), we can obtain \(\sqrt{n}|\hat{\vartheta}(\hat{\theta}_{\text{MMA}}) - \theta| = O_p(1)\).

References