Economics Letters 130 (2015) 120–123

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

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.

ARTICLE INFO

Article history:
Received 17 October 2014
Received in revised form
25 February 2015
Accepted 14 March 2015
Available online 20 March 2015

JEL classification:
C51
C52

Keywords:
Consistency
Jackknife model averaging
Mallows model averaging
Smoothed AIC
Smoothed BIC

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.

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 (MMA) (Hansen, 2007), optimal mean squared error averaging (Liang et al., 2011), jackknife model averaging (JMA) (Hansen and Racine, 2012), heteroskedasticity-robust Cp (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

* Correspondence to: International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, China.
E-mail address: xinyu@amss.ac.cn.

http://dx.doi.org/10.1016/j.econlet.2015.03.017
0165-1765/

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:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + e_i, \qquad E(e_i|\mathbf{x}_i, \mathbf{z}_i) = 0,$$
$$E(e_i^2|\mathbf{x}_i, \mathbf{z}_i) = \sigma^2(\mathbf{x}_i, \mathbf{z}_i), \quad i = 1, \ldots, n \tag{1}$$

where y_i is a scaled dependent variable, $\mathbf{x}_i(q_1 \times 1)$ and $\mathbf{z}_i(q_2 \times 1)$ contain independent variables, $\boldsymbol{\beta}(q_1 \times 1)$ and $\boldsymbol{\gamma}(q_2 \times 1)$ are parameter vectors, e_i is an error term, and $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ for $i = 1, \ldots, n$ are independent. Here, $\mathbf{x}_i$ contains the variables that must be included in the model, while we are uncertain whether the variables in $\mathbf{z}_i$ are included. So we have 2^{q_2} sub-models of (1). To be more general, we combine M ones of these sub-models, where M can be smaller than 2^{q_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; Liang et al., 2011).

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^{\mathrm{T}}$, and $\mathbf{e} = (e_1, \ldots, e_n)^{\mathrm{T}}$. Let $q = q_1 + q_2$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^{\mathrm{T}} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{H} = (\mathbf{X}, \mathbf{Z})$. We assume that $\mathbf{H}$ has full column rank q. The mth sub-model (or candidate model) includes all variables in $\mathbf{X}$ and some (all or none) of variables in $\mathbf{Z}$, denoted as $\mathbf{Z}_m$. Let $\mathbf{H}_m = (\mathbf{X}, \mathbf{Z}_m)$, k_m be the number of columns of $\mathbf{H}_m$, and $\boldsymbol{\Pi}_m$ be a selection matrix so that $\boldsymbol{\Pi}_m = (\mathbf{I}_{k_m}, \mathbf{0}_{k_m \times (q-k_m)})$ or a column permutation thereof and thus $\mathbf{H}_m = \mathbf{H}\boldsymbol{\Pi}_m^{\mathrm{T}}$, where $\mathbf{I}_{k_m}$ is a $k_m \times k_m$ identity matrix. Under the mth candidate model, $\widehat{\boldsymbol{\theta}}_m = \boldsymbol{\Pi}_m^{\mathrm{T}}(\mathbf{H}_m^{\mathrm{T}}\mathbf{H}_m)^{-1}\mathbf{H}_m^{\mathrm{T}}\mathbf{y}$. Therefore, the model averaging estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m \widehat{\boldsymbol{\theta}}_m,$$

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

Weighting strategy based on the AIC or BIC scores

In the mth candidate model, the AIC and BIC scores are $\mathrm{AIC}_m = n \log \widehat{\sigma}_m^2 + 2k_m$ and $\mathrm{BIC}_m = n \log \widehat{\sigma}_m^2 + k_m \log n$, respectively, where $\widehat{\sigma}_m^2 = n^{-1}\|\mathbf{y} - \mathbf{H}\widehat{\boldsymbol{\theta}}_m\|^2$. Weights are

$$\widehat{\mathbf{w}}_{\mathrm{xIC},m} = \exp(-\mathrm{xIC}_m/2) / \sum_{m=1}^{M} \exp(-\mathrm{xIC}_m/2), \quad m = 1, \ldots, M, \tag{3}$$

where xIC_m is the AIC or BIC score under the mth candidate model. The averaging estimators combined by weights in (3) are commonly called smoothed AIC (S-AIC) or smoothed BIC (S-BIC) estimators.

MMA

Let $\mathbf{P}_m = \mathbf{H}_m(\mathbf{H}_m^{\mathrm{T}}\mathbf{H}_m)^{-1}\mathbf{H}_m^{\mathrm{T}}$, $\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{P}_m$, and $\mathbf{k} = (k_1, \ldots, k_M)^{\mathrm{T}}$. For homoscedastic error setting (that is $\sigma^2(\mathbf{x}_1, \mathbf{z}_1) = \cdots = \sigma^2(\mathbf{x}_n, \mathbf{z}_n) = \sigma^2$), Hansen (2007) proposed choosing weights by minimizing Mallows criterion $\mathcal{C}(\mathbf{w}) = \|\{\mathbf{I}_n - \mathbf{P}(\mathbf{w})\}\mathbf{y}\|^2 + 2\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{k}$, where $\|\cdot\|^2$ stands for the Euclidean norm, $\widehat{\sigma}^2 = (n - k_{m^*})^{-1}\|\mathbf{y} - \mathbf{H}\widehat{\boldsymbol{\theta}}_{m^*}\|^2$, and $m^* = \arg\max_{m \in \{1,\ldots,M\}} k_m$. Denote $\widehat{\mathbf{w}}_{\mathrm{MMA}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \mathcal{C}(\mathbf{w})$, so that the combined estimator $\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}})$ is the MMA estimator of $\boldsymbol{\theta}$. 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 $\mathbf{P}_m^{\mathrm{JMA}} = \mathbf{D}_m(\mathbf{P}_m - \mathbf{I}_n) + \mathbf{I}_n$, where $\mathbf{D}_m$ is a diagonal matrix with $(1 - P_{ii}^m)^{-1}$ being its ith diagonal element and P_{ii}^m is the ith diagonal element of $\mathbf{P}_m$. Let $\mathbf{P}^{\mathrm{JMA}}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{P}_m^{\mathrm{JMA}}$. For heteroscedastic error setting, Hansen and Racine (2012) proposed choosing weights by minimizing cross-validation (or jackknife) criterion $\mathcal{J}(\mathbf{w}) = \|\{\mathbf{I}_n - \mathbf{P}^{\mathrm{JMA}}(\mathbf{w})\}\mathbf{y}\|^2$. Let $\widehat{\mathbf{w}}_{\mathrm{JMA}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \mathcal{J}(\mathbf{w})$, so that the combined estimator $\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{JMA}})$ is the JMA estimator of $\boldsymbol{\theta}$.

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 $\mathcal{U}$. 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). $\boldsymbol{\Psi}_n \equiv n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} \to \boldsymbol{\Psi}$ *where* $\boldsymbol{\Psi}$ *is a positive definite matrix, and* $n^{-1/2}\mathbf{H}^{\mathrm{T}}\mathbf{e} = O_p(1)$.

Condition (C.2). $p^* = O(n^{-1})$, *where* $p^* = \max_{1 \le m \le M} \max_{1 \le i \le n} P_{ii}^m$.

Conditions (C.1) is very commonly used in literature such as Liu (2015). Condition (C.2) is quite mild. Instead, Andrews (1991) assumed that $P_{ii}^m \le c k_m n^{-1}$ 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),* $\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{AIC}}) - \boldsymbol{\theta}\} = O_p(1)$ *and* $\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{BIC}}) - \boldsymbol{\theta}\} = O_p(1)$.

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

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 $\boldsymbol{\beta} = (1, 1)'$, $\boldsymbol{\gamma} = (0.5, 0.7, 0, -0.3)'$, $\mathbf{x}_i = (1, x_{i,2})'$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{i,4})$, $(x_{i,2}, z_{i1}, \ldots, z_{i,4}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_5)$, and $e_i \sim \mathcal{N}(0, x_{i,2}^2)$ We combined all $2^4 = 16$ sub-modes. 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.

Table 1
$n^{0.95} \times$ MSE of the model averaging estimators in simulation study.

n	40	60	80	100	200	300	400	600	800	1000
MMA	1.227	1.148	1.089	1.073	1.067	0.932	0.928	0.910	0.885	0.860
JMA	1.297	1.203	1.139	1.113	1.092	0.946	0.935	0.916	0.889	0.863
S-AIC	1.737	1.577	1.450	1.370	1.164	0.940	0.883	0.861	0.841	0.816
S-BIC	1.339	1.245	1.171	1.141	1.066	0.903	0.882	0.867	0.845	0.820

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 $\mathbf{H}_{m^c}$ as a matrix consisting of columns of $\mathbf{H}$ not contained in $\mathbf{H}_m$ and $\mathbf{\Pi}_{m^c}$ as a selection matrix such that $\mathbf{H}_{m^c} = \mathbf{H}\mathbf{\Pi}_{m^c}^{\mathrm{T}}$. Let $\boldsymbol{\theta}_m = \mathbf{\Pi}_m\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{m^c} = \mathbf{\Pi}_{m^c}\boldsymbol{\theta}$.

A.1. Proof of Theorem 1

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

$$|\mathbf{\Psi}| = \begin{vmatrix} \mathbf{\Psi}_{11} & \mathbf{\Psi}_{12} \\ \mathbf{\Psi}_{21} & \mathbf{\Psi}_{22} \end{vmatrix} = |\mathbf{\Psi}_{11}||\mathbf{\Psi}_{22} - \mathbf{\Psi}_{21}\mathbf{\Psi}_{11}^{-1}\mathbf{\Psi}_{12}|,$$

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

$$\begin{aligned} & n^{-1}\mathbf{H}_{m^c}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{H}_{m^c} \\ & = n^{-1}\mathbf{H}_{m^c}^{\mathrm{T}}\mathbf{H}_{m^c} - n^{-1}\mathbf{H}_{m^c}^{\mathrm{T}}\mathbf{H}_m(\mathbf{H}_m^{\mathrm{T}}\mathbf{H}_m)^{-1}\mathbf{H}_m^{\mathrm{T}}\mathbf{H}_{m^c} \to \mathbf{G}_m. \end{aligned} \tag{A.1}$$

Let $g_m = \boldsymbol{\theta}_{m^c}^{\mathrm{T}}\mathbf{G}_m\boldsymbol{\theta}_{m^c}$. From (A.1) and Condition (C.1), we have for any $m \in \mathcal{U}$,

$$\begin{aligned} n^{-1}\|\mathbf{y} - \mathbf{P}_m\mathbf{y}\|^2 &= n^{-1}(\mathbf{e} + \mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{e} + \mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c}) \\ &= n^{-1}\boldsymbol{\theta}_{m^c}^{\mathrm{T}}\mathbf{H}_{m^c}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c} \\ &\quad + n^{-1}\|\mathbf{e}\|^2 + 2n^{-1}\mathbf{e}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c} \\ &\quad - n^{-1}\mathbf{e}^{\mathrm{T}}\mathbf{P}_m\mathbf{e} \\ &= g_m + n^{-1}\|\mathbf{e}\|^2 + o_p(1), \end{aligned} \tag{A.2}$$

and for any $m \notin \mathcal{U}$,

$$n^{-1}\|\mathbf{y} - \mathbf{P}_m\mathbf{y}\|^2 = n^{-1}\mathbf{e}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{e} = n^{-1}\|\mathbf{e}\|^2 + o_p(1). \tag{A.3}$$

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

$$\begin{aligned} \widehat{w}_{\mathrm{AIC},m}\widehat{w}_{\mathrm{AIC},m_o}^{-1} &= \exp(\mathrm{AIC}_{m_o}/2 - \mathrm{AIC}_m/2) \\ &= (\|\mathbf{y} - \mathbf{P}_m\mathbf{y}\|^{-2}\|\mathbf{y} - \mathbf{P}_{m_o}\mathbf{y}\|^2)^{n/2} \\ &\quad \times \exp(k_{m_o} - k_m) = O_p(c^n), \end{aligned} \tag{A.4}$$

where c is a generic constant belonging to $(0, 1)$. Since $\widehat{\mathbf{w}}_{\mathrm{AIC}} \in \mathcal{W}$, (A.4) implies $\widehat{w}_{\mathrm{AIC},m} = O_p(c^n)$. From Condition (C.1), it is straightforward to show that for any $m \notin \mathcal{U}$,

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

and for any $m \in \mathcal{U}$,

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_m &= \mathbf{\Pi}_m^{\mathrm{T}}(\mathbf{H}_m^{\mathrm{T}}\mathbf{H}_m)^{-1}\mathbf{H}_m^{\mathrm{T}}\mathbf{y} = \mathbf{\Pi}_m^{\mathrm{T}}(\mathbf{H}_m^{\mathrm{T}}\mathbf{H}_m)^{-1}\mathbf{H}_m^{\mathrm{T}}\mathbf{H}\boldsymbol{\theta} \\ &\quad + \mathbf{\Pi}_m^{\mathrm{T}}(\mathbf{H}_m^{\mathrm{T}}\mathbf{H}_m)^{-1}\mathbf{H}_m^{\mathrm{T}}\mathbf{e} = O_p(1). \end{aligned} \tag{A.5}$$

Hence,

$$\begin{aligned} & \sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{AIC}}) - \boldsymbol{\theta}\} \\ & = \sum_{m=1}^{M}\widehat{w}_{\mathrm{AIC},m}\sqrt{n}(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}) \\ & = \sum_{m\in\mathcal{U}}\widehat{w}_{\mathrm{AIC},m}\sqrt{n}(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}) + \sum_{m\notin\mathcal{U}}\widehat{w}_{\mathrm{AIC},m}\sqrt{n}(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}) \\ & = \sum_{m\in\mathcal{U}}O_p(c^n)\sqrt{n}O_p(1) + \sum_{m\notin\mathcal{U}}\widehat{w}_{\mathrm{AIC},m}O_p(1) = O_p(1). \end{aligned} \tag{A.6}$$

Similar to argument of (A.4), we can prove that for any $m \in \mathcal{U}$,

$$\begin{aligned} \widehat{w}_{\mathrm{BIC},m}\widehat{w}_{\mathrm{BIC},m_o}^{-1} &= \exp(\mathrm{BIC}_{m_o}/2 - \mathrm{BIC}_m/2) \\ &= (\|\mathbf{y} - \mathbf{P}_m\mathbf{y}\|^{-2}\|\mathbf{y} - \mathbf{P}_{m_o}\mathbf{y}\|^2)^{n/2}n^{(k_{m_o} - k_m)/2} \\ &= O_p(c^n n^{q_2/2}), \end{aligned}$$

and thus $\widehat{w}_{\mathrm{BIC},m} = O_p(c^n n^{q_2/2})$. Using the steps of (A.6), we have $\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{BIC}}) - \boldsymbol{\theta}\} = O_P(1)$.

A.2. Proof of Theorem 2

It is seen that

$$\begin{aligned} \mathcal{C}(\mathbf{w}) &= \|\{\mathbf{I}_n - \mathbf{P}(\mathbf{w})\}\mathbf{y}\|^2 + 2\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{k} \\ &= \|\mathbf{H}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\mu} - \mathbf{e}\|^2 + 2\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{k} \\ &= \|\mathbf{e}\|^2 + \{\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}\}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\{\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}\} \\ &\quad - 2\mathbf{e}^{\mathrm{T}}\mathbf{H}\{\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}\} + 2\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{k}. \end{aligned} \tag{A.7}$$

It follows from (A.2)–(A.3) that

$$\widehat{\sigma}^2 = (n - k_{m^*})^{-1}\|\mathbf{y} - \mathbf{H}\widehat{\boldsymbol{\theta}}_{m^*}\|^2 = O_p(1). \tag{A.8}$$

Let j be a model not belonging to $\mathcal{U}$. When $w_j = 1$, from (A.8) and Condition (C.1), we have $\mathcal{C}(\mathbf{w}) = \|\mathbf{e}\|^2 + \eta_n$ with $\eta_n = O_p(1)$. Hence, $\mathcal{C}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) \le \|\mathbf{e}\|^2 + \eta_n$, which, together with (A.7), implies

$$\begin{aligned} \eta_n \ge\ & \{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\} \\ & - 2\mathbf{e}^{\mathrm{T}}\mathbf{H}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\} + 2\widehat{\sigma}^2\widehat{\mathbf{w}}_{\mathrm{MMA}}^{\mathrm{T}}\mathbf{k}. \end{aligned} \tag{A.9}$$

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

$$\begin{aligned} & \lambda_{\min}(\mathbf{\Psi}_n)\|\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\}\|^2 \\ & \le \{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\} \\ & \le \eta_n + 2\mathbf{e}^{\mathrm{T}}\mathbf{H}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\} - 2\widehat{\sigma}^2\widehat{\mathbf{w}}_{\mathrm{MMA}}^{\mathrm{T}}\mathbf{k} \\ & \le \eta_n + 2\|n^{-1/2}\mathbf{e}^{\mathrm{T}}\mathbf{H}\|\,\|\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\}\|, \end{aligned}$$

and thus

$$\begin{aligned} & \lambda_{\min}(\mathbf{\Psi}_n)[\|\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}) - \boldsymbol{\theta}\}\| - \{\lambda_{\min}(\mathbf{\Psi}_n)\}^{-1}\|n^{-1/2}\mathbf{e}^{\mathrm{T}}\mathbf{H}\|]^2 \\ & \le \eta_n + \{\lambda_{\min}(\mathbf{\Psi}_n)\}^{-1}\|n^{-1/2}\mathbf{e}^{\mathrm{T}}\mathbf{H}\|^2, \end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\|\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}})-\boldsymbol{\theta}\}\| \\
&\quad\in\big[-(\{\lambda_{\min}(\boldsymbol{\Psi}_n)\}^{-1}[\eta_n+\{\lambda_{\min}(\boldsymbol{\Psi}_n)\}^{-1}\|n^{-1/2}\mathbf{e}^{\mathrm{T}}\mathbf{H}\|^2])^{1/2} \\
&\qquad+\{\lambda_{\min}(\boldsymbol{\Psi}_n)\}^{-1}\|n^{-1/2}\mathbf{e}^{\mathrm{T}}\mathbf{H}\|, \\
&\qquad(\{\lambda_{\min}(\boldsymbol{\Psi}_n)\}^{-1}[\eta_n+\{\lambda_{\min}(\boldsymbol{\Psi}_n)\}^{-1}\|n^{-1/2}\mathbf{e}^{\mathrm{T}}\mathbf{H}\|^2])^{1/2} \\
&\qquad+\{\lambda_{\min}(\boldsymbol{\Psi}_n)\}^{-1}\|n^{-1/2}\mathbf{e}^{\mathrm{T}}\mathbf{H}\|\big],
\end{aligned}$$

which, together with $\eta_n = O_p(1)$ and Condition (C.1), implies $\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}})-\boldsymbol{\theta}\}=O_p(1)$.

For JMA, denote $\mathbf{Q}_m$ as an $n\times n$ diagonal matrix with the ith diagonal element $Q_{m,ii}=P_{ii}^m/(1-P_{ii}^m)$, so $\mathbf{D}=\mathbf{Q}+\mathbf{I}_n$. Let $\boldsymbol{\Xi}$ be an $M\times M$ matrix with the mjth element

$$\begin{aligned}
\Xi_{mj} &= (\mathbf{e}+\mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c})^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_m)(\mathbf{Q}_m+\mathbf{Q}_j+\mathbf{Q}_m\mathbf{Q}_j)(\mathbf{I}_n-\mathbf{P}_j) \\
&\quad\times(\mathbf{e}+\mathbf{H}_{j^c}\boldsymbol{\theta}_{j^c})-2k_m\widehat{\sigma}^2,
\end{aligned}$$

so that

$$\mathcal{J}(\mathbf{w})=\mathcal{C}(\mathbf{w})+\mathbf{w}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{w}. \tag{A.10}$$

Let $\mathcal{S}(\mathbf{K})$ be the largest singular value of a matrix $\mathbf{K}$. From Li (1987), we know that any two $n\times n$ matrices $\mathbf{K}_1$ and $\mathbf{K}_2$, $\mathcal{S}(\mathbf{K}_1\mathbf{K}_2)\le\mathcal{S}(\mathbf{K}_1)\mathcal{S}(\mathbf{K}_2)$ and $\mathcal{S}(\mathbf{K}_1+\mathbf{K}_2)\le\mathcal{S}(\mathbf{K}_1)+\mathcal{S}(\mathbf{K}_2)$, by which and Conditions (C.1)–(C.2), we have

$$\begin{aligned}
&(\mathbf{e}+\mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c})^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_m)(\mathbf{Q}_m+\mathbf{Q}_j+\mathbf{Q}_m\mathbf{Q}_j) \\
&\qquad\times(\mathbf{I}_n-\mathbf{P}_j)(\mathbf{e}+\mathbf{H}_{j^c}\boldsymbol{\theta}_{j^c}) \\
&\quad\le\|\mathbf{e}+\mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c}\|\,\|\mathbf{e}+\mathbf{H}_{j^c}\boldsymbol{\theta}_{j^c}\| \\
&\qquad\times\mathcal{S}\{(\mathbf{I}_n-\mathbf{P}_m)(\mathbf{Q}_m+\mathbf{Q}_j+\mathbf{Q}_m\mathbf{Q}_j)(\mathbf{I}_n-\mathbf{P}_j)\} \\
&\quad\le\|\mathbf{e}+\mathbf{H}_{m^c}\boldsymbol{\theta}_{m^c}\|\,\|\mathbf{e}+\mathbf{H}_{j^c}\boldsymbol{\theta}_{j^c}\|\{2p^*+(p^*)^2\}=O_p(1),
\end{aligned} \tag{A.11}$$

so for any $\mathbf{w}\in\mathcal{W}$, $\mathbf{w}^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{w}=O_p(1)$. Now, using the arguments in proving root-n consistency of $\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}})$, we can obtain $\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}_{\mathrm{JMA}})-\boldsymbol{\theta}\}=O_p(1)$.

References

Andrews, D.W.K., 1991. Asymptotic optimality of generalized c_l, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. J. Econometrics 47, 359–377.

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

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference understanding aic and bic in model selection. Sociol. Methods Res. 33, 261–304.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hansen, B.E., 2014. Model averaging, asymptotic risk, and regressor groups. Quant. Econ. 5, 495–530.

Hansen, B.E., Racine, J., 2012. Jacknife model averaging. J. Econometrics 167, 38–46.

Ishwaran, R.H., Rao, J.S., 2003. Discussion. J. Amer. Statist. Assoc. 98, 922–925.

Layton, D.F., Lee, S.T., 2006. Embracing model uncertainty: strategies for response pooling and model averaging. Environ. Resour. Econ. 34, 51–85.

Leamer, E., 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York.

Li, K.-C., 1987. Asymptotic optimality for Cp, Cl, cross-validation and generalized cross-validation: discrete index set. Ann. Statist. 15, 958–975.

Liang, H., Zou, G., Wan, A.T.K., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. J. Amer. Statist. Assoc. 106, 1053–1066.

Liu, C.-A., 2015. Distribution theory of the least squares averaging estimator. J. Econometrics 186, 142–159.

Liu, Q., Okui, R., 2013. Heteroskedasticity-robust Cp model averaging. Econom. J. 16, 462–473.

Magnus, J., Durbin, J., 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. Econometrica 67, 639–643.

Raftery, A.E., Zheng, Y., 2003. Discussion: performance of Bayesian model averaging. J. Amer. Statist. Assoc. 98, 931–938.

Rao, C.R., 1973. Linear Statistical Inference and its Applications. Wiley.