Economics Letters 145 (2016) 214–217

Contents lists available at ScienceDirect

Economics Letters

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

Model averaging with averaging covariance matrix

Shangwei Zhao [a], Xinyu Zhang [b,*], Yichen Gao [c]

[a] College of Science, Minzu University of China, China
[b] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
[c] ISEM, Capital University of Economics and Business, China

HIGHLIGHTS

- We use an average estimator which depends on all candidate models to estimate the covariance matrix.
- We choose weight vectors in the model average estimators of coefficients and covariance matrix simultaneously by minimizing the weight choice criterion.
- We prove the asymptotic optimality.
- Simulation experiments show that the proposed model averaging method is superior to its competitors.

ARTICLE INFO

Article history:
Received 6 March 2016
Received in revised form
21 May 2016
Accepted 8 June 2016
Available online 16 June 2016

JEL classification:
C2
C13

Keywords:
Asymptotic optimality
Heteroscedasticity
Model averaging

ABSTRACT

This article studies optimal model averaging for linear models with heteroscedasticity. We choose weights by minimizing Mallows-type criterion. Because the covariance matrix of random error in the criterion is unknown, an averaging estimator of covariance matrix is plugged into the criterion. The resulting model averaging estimator is proved to be asymptotically optimal under some regularity conditions. Simulation experiments show that the proposed model averaging method is superior to its competitors.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, asymptotically optimal least square model averaging methods have been actively developed. Hansen (2007), Liang et al. (2011), Zhang et al. (2015) and Xie (2015) respectively proposed Mallows model averaging, "OPT" model averaging, model averaging based on Kullback–Leibler distance, and prediction model averaging for combining least squared estimators with homoskedasticity. Hansen and Racine (2012) and Liu and Okui (2013) respectively proposed Jackknife model averaging (JMA) and Heteroskedasticity-Robust Cp (HRCp) model averaging for combining least squared estimators with heteroskedasticity. All these methods are asymptotically optimal in the sense that they minimize predictive squared error in large sample case.

This article contributes literature on least squared model averaging with heteroskedasticity. Similar to Mallows criterion of Hansen (2007), we use a weight choice criterion which is an unbiased estimator of expected predictive squared error up to a constant. But, in the weight choice criterion, the covariance matrix of random errors is unknown. One may estimate the covariance matrix based on a single candidate model, but this estimation is somehow arbitrary, because we have many candidate models. In the current paper, we use an average estimator which depends on all candidate models to estimate the covariance matrix. Then, we choose weight vectors in the model average estimators of coefficients and covariance matrix simultaneously by minimizing the weight choice criterion. The resulting weight vector is proved to be asymptotically optimal under some regularity conditions. We term this method as Model Averaging with Averaging Covariance Matrix (MAACM). This idea is similar to those in Xie (2015) and Gao et al. (submitted for publication) where the estimators of error variance also have average forms, but the models are different from ours. The existing optimal model averaging methods for heteroscedasticity include JMA and HRCp. In JMA, it is not necessary to estimate the unknown covariance matrix, but Liu and Okui (2013) show that JMA has a worse finite sample performance than HRCp. In HRCp, an

* Corresponding author.
E-mail address: xinyu@amss.ac.cn (X. Zhang).

http://dx.doi.org/10.1016/j.econlet.2016.06.011
0165-1765/© 2016 Elsevier B.V. All rights reserved.

estimate of the covariance matrix must be provided prior to choosing the weight vector, so the HRCp estimator is a two-step estimator. In contrast, the MAACM estimator is a continuous updating estimator requiring only one step of calculation. Finite sample simulation results show that MAACM outperforms JMA and HRCp.

The remainder of this paper is organized as follows. Section 2 introduces our model averaging estimation and presents the asymptotic optimality. Section 3 investigates the finite sample performance of the model average estimators. Technical proofs are contained in an Appendix.

2. Model averaging estimation and its asymptotic optimality

Following Hansen and Racine (2012) and Liu and Okui (2013), we consider linear model

$$y_i = \mu_i + \epsilon_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \quad i = 1\ldots, n, \tag{1}$$

where $\mathbf{X}_i = (x_{i1}, \ldots, x_{i\infty})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_\infty)^{\mathrm{T}}$, $\mu_i = E(y_i|\mathbf{X}_i)$, and $\epsilon_1, \ldots, \epsilon_n$ are independent random errors with $E(\epsilon_i|\mathbf{X}_i) = 0$ and $E(\epsilon_i^2|\mathbf{X}_i) = \sigma_i^2$. So heteroscedasticity is allowed here. Let $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$.

Suppose that we have a total of S_n candidate models. For the sth candidate model, we use an $n \times p_s$ matrix $\mathbf{X}_{(s)}$, a subset of $\mathbf{X}$, as the covariate matrix. The total number S_n can be related to the sample size n. Using least squared estimation and assuming $\mathbf{X}_{(s)}$ to be of full column rank, the coefficient estimate under model s is $(\mathbf{X}_{(s)}{}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}\mathbf{y}$, and then the associate estimator of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{X}_{(s)}(\mathbf{X}_{(s)}{}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}\mathbf{y} = \mathbf{P}_{(s)}\mathbf{y}$, where $\mathbf{P}_{(s)} = \mathbf{X}_{(s)}(\mathbf{X}_{(s)}{}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}{}^{\mathrm{T}}$. Let weight vector $\mathbf{w} = (w_1, \ldots, w_{S_n})^{\mathrm{T}}$, belonging to the set $\mathcal{W} = \{\mathbf{w} \in [0, 1]^{S_n} : \sum_{s=1}^{S_n} w_s = 1\}$ and $\mathbf{P}(\mathbf{w}) = \sum_{s=1}^{S_n}\mathbf{P}_{(s)}$. Then, the model average estimator of $\boldsymbol{\mu}$ can be expressed by $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{s=1}^{S_n} w_s\widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{P}(\mathbf{w})\mathbf{y}$.

Let $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. The predictive squared loss of $\widehat{\boldsymbol{\mu}}(\mathbf{w})$ is $L_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$ and the expected loss is

$$R_n(\mathbf{w}) = E\{L_n(\mathbf{w})|\mathbf{X}\} = \|\mathbf{P}(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\mathbf{P}^{\mathrm{T}}(\mathbf{w})\}. \tag{2}$$

Define

$$C_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2 + 2\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}. \tag{3}$$

It is straightforward to show that $R_n(\mathbf{w}) = E\{C_n(\mathbf{w})|\mathbf{X}\} - E\{\mathrm{trace}(\boldsymbol{\Omega})|\mathbf{X}\}$, which indicates that for the selection of $\mathbf{w}$, we can ignore the offset $E\{\mathrm{trace}(\boldsymbol{\Omega})|\mathbf{X}\}$, which does not involve $\mathbf{w}$, and use $C_n(\mathbf{w})$ as if it were $R_n(\mathbf{w})$.

The weight choice criterion $C_n(\mathbf{w})$ still involves the unknown $\boldsymbol{\Omega}$. We estimate $\boldsymbol{\Omega}$ using residues from model averaging: $\widehat{\boldsymbol{\epsilon}}(\mathbf{w}) \equiv \{\widehat{\epsilon}_1(\mathbf{w}), \ldots, \widehat{\epsilon}_n(\mathbf{w})\}^{\mathrm{T}} = \mathbf{y} - \widehat{\boldsymbol{\mu}}(\mathbf{w})$. Specifically, the estimator of $\boldsymbol{\Omega}$ is

$$\widehat{\boldsymbol{\Omega}}(\mathbf{w}) = \mathrm{diag}\{\widehat{\epsilon}_1^2(\mathbf{w}), \ldots, \widehat{\epsilon}_n^2(\mathbf{w})\}. \tag{4}$$

In the existing model averaging methods such as MMA of Hansen (2007) and HRCp of Liu and Okui (2013), the variance of random error are generally estimated depending on a single model, which places too much confidence the single model. Here, we use the model average estimator $\widehat{\boldsymbol{\Omega}}(\mathbf{w})$. Replacing $\boldsymbol{\Omega}$ by $\widehat{\boldsymbol{\Omega}}(\mathbf{w})$ in (3), $C_n(\mathbf{w})$ becomes

$$\widehat{C}_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2 + 2\mathrm{trace}\{\mathbf{P}(\mathbf{w})\widehat{\boldsymbol{\Omega}}(\mathbf{w})\}.$$

Now, our weights are expressed by

$$\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\widehat{C}_n(\mathbf{w}). \tag{5}$$

The weight choice criterion $\widehat{C}_n(\mathbf{w})$ is a cubic function of $\mathbf{w}$. Numerous software packages are available for obtaining the solution to this problem (e.g., Fmincon of Matlab), and they generally work effectively and efficiently even when S_n is very large.

Let $\widetilde{p} = \max_s p_s$, $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}} R_n(\mathbf{w})$, $\mathbf{w}_s^o$ be a weight vector with the sth element taking on the value of unity and other elements zeros, $\rho_{ii}^{(s)}$ be the ith diagonal element of $\mathbf{P}_{(s)}$, $\max_i$ and $\max_s$ indicate the maximization over $i \in \{1, \ldots, n\}$ and $s \in \{1, \ldots, S_n\}$, respectively, and $\min_i$ indicates the minimization over $i \in \{1, \ldots, n\}$. We now list the regularity conditions required for the asymptotic optimality of the weights in (5), where all the limiting properties here and throughout the text hold under $n \to \infty$.

Condition (C.1). *For constant $\kappa > 0$, constant $\bar{\sigma}^2$, and some integral $G \geq 1$,*

$$\max_i E(\epsilon_i^{4G}|\mathbf{X}_i) \leq \kappa < \infty, \qquad S_n\xi_n^{-2G}\sum_{s=1}^{S_n}\{R_n(\mathbf{w}_s^o)\}^G \to 0,$$

$$\min_i \sigma_i^2 \geq \bar{\sigma}^2 \text{ almost surely.}$$

Condition (C.2). *There exists a constant c such that $|\rho_{ii}^{(s)}| \leq cn^{-1}p_s$ almost surely for all $s = \{1, \ldots, S\}$ and $i = \{1, \ldots, n\}$.*

Condition (C.3). $n^{-1}\widetilde{p}^2 = O(1)$.

The first two parts of Condition (C.1) are widely used in literature on model averaging; see, for example, the conditions (7) and (8) of Wan et al. (2010) and the assumptions 2.2 and 2.3 of Liu and Okui (2013). The third part of Condition (C.1) requires that the covariance matrix Ω does not degenerate as $n \to \infty$. Condition (C.2) is commonly used in the studies of asymptotic optimality of cross-validation methods (e.g., Andrews, 1991; Hansen and Racine, 2012). Condition (C.3), which is the same as the condition (12) of Wan et al. (2010), restricts the increasing rates of p_s's as $n \to \infty$, but it still allows p_s's increase with n. The following theorem builds the asymptotic optimality of model average estimator using weights $\widehat{\mathbf{w}}$.

Theorem 1. *Under Conditions (C.1), (C.2) and (C.3),*

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})} \to 1 \tag{6}$$

in probability as $n \to \infty$.

Theorem 1 shows that the model averaging procedure using weights $\widehat{\mathbf{w}}$ is asymptotically optimal in the sense that its squared loss is asymptotically identical to that of the infeasible best possible model average estimator. The proof of Theorem 1 is in the Appendix.

3. Simulation study

This simulation design is based on the setting of Hansen (2007), except that the error term is heteroscedastic. Specifically, we generated data from model (1) with $\mu_i = \sum_{j=1}^{\infty} x_{ij}\beta_j$ and normal errors $\epsilon_i \sim \mathrm{Normal}(0, x_{i2}^2)$. We set $x_{i1} = 1$ and observations of all other x_{ij}'s are generated from the Normal(0, 1) distribution and are independent. The coefficients $\theta_j = c\sqrt{2}\alpha j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 0.5$. We control c such that $R^2 = c^2/(1+c^2)$ vary in $\{0.1, \ldots, 0.9\}$. The sample size varies at 100, 300, 600 and 900. The number of approximating models is determined by $S_n = INT(3n^{1/3})$, where the function $INT(A)$ returns the smallest integer that exceeds A. The sth candidate model contains the first s observed covariates. We compare MAACM, HRCp and JMA methods based on the average predictive squared loss $L_n(\mathbf{w})$ in 1000 replications. For each

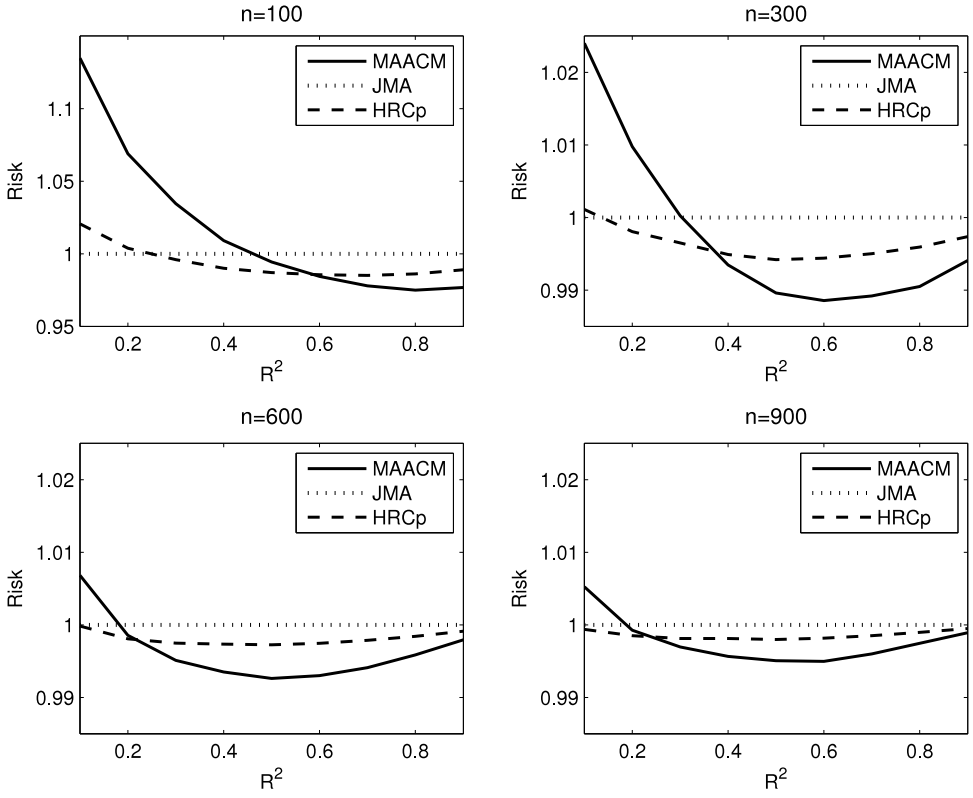

Fig. 1. Simulation results of Jackknife model averaging (JMA), Heteroskedasticity-Robust Cp (HRCp) model averaging, and model averaging with averaging covariance matrix (MAACM).

parameterization, we normalize the average loss by dividing by the average loss of JMA.

The simulation results are presented in Fig. 1. The R^2 is presented on the horizontal axis the relative losses is displayed on the vertical axis. We see that when $n = 100$, MAACM outperforms JMA and HRCp when $R^2 > 0.6$. The advantage of MAACM becomes more prominent as sample size increases. When $n = 600$ or $n = 900$, MAACM yields the smallest losses in the majority cases of R^2.

Acknowledgments

The authors thank the referee for very constructive comments. Zhang's work was partially supported by the National Natural Science Foundation of China (Grant nos. 71522004, 11471324 and 11271355) and a special talent grant from AMSS, CAS. Gao's work was partially supported by the National Natural Science Foundation of China (Grant no. 71501133).

Appendix. Proof of Theorem 1

Although $\mathbf{X}$ is stochastic in this article, we take it non-stochastic in this proof, because for both stochastic and non-stochastic cases, the proof steps are almost the same; see for example the proof of Theorem 2.1 of Zhang et al. (2013). Denote the largest singular values of a matrix $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$. From the first part of Condition (C.1), we have

$$\lambda_{\max}(\boldsymbol{\Omega}) = O(1). \tag{A.1}$$

Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$. Using (A.1) and transformation $\boldsymbol{\epsilon}^* = \boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon}$, from the proof of Theorem 1' of Wan et al. (2010), we have

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathcal{W}} \frac{\left|C_n(\mathbf{w}) - L_n(\mathbf{w}) - \|\boldsymbol{\epsilon}\|^2\right|}{R_n(\mathbf{w})} = o_p(1) \quad \text{and} \\ &\sup_{\mathbf{w}\in\mathcal{W}} \frac{|R_n(\mathbf{w}) - L_n(\mathbf{w})|}{R_n(\mathbf{w})} = o_p(1). \end{aligned} \tag{A.2}$$

It is seen that

$$\widehat{C}_n(\mathbf{w}) = C_n(\mathbf{w}) + \operatorname{trace}\{\mathbf{P}(\mathbf{w})\widehat{\boldsymbol{\Omega}}(\mathbf{w})\} - \operatorname{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}.$$

Hence, in order to prove (6), we need only to verify that

$$\sup_{\mathbf{w}\in\mathcal{W}}[|\operatorname{trace}\{\mathbf{P}(\mathbf{w})\widehat{\boldsymbol{\Omega}}(\mathbf{w})\} - \operatorname{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] = o_p(1). \tag{A.3}$$

Since $\mathbf{P}_{(s)}$ is symmetric and idempotent, we have

$$\max_s\{\lambda_{\max}(\mathbf{P}_s)\} = O(1) \quad \text{and} \quad \max_s\{\lambda_{\max}(\mathbf{P}_{(s)}{\mathbf{P}_{(s)}}^{\mathrm{T}})\} = O(1). \tag{A.4}$$

Let $\mathbf{Q}_{(s)} = \operatorname{diag}(\rho_{11}^{(s)}, \ldots, \rho_{nn}^{(s)})$ and $\mathbf{Q}(\mathbf{w}) = \sum_{s=1}^{S_n} w_s\mathbf{Q}_{(s)}$. Then, from (4), we have

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathcal{W}}[|\operatorname{trace}\{\mathbf{P}(\mathbf{w})\widehat{\boldsymbol{\Omega}}(\mathbf{w})\} - \operatorname{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\ &= \sup_{\mathbf{w}\in\mathcal{W}}[|\{\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\{\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\} \\ &\quad - \operatorname{trace}\{\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\ &= \sup_{\mathbf{w}\in\mathcal{W}}[|\{\boldsymbol{\epsilon} + \boldsymbol{\mu} - \mathbf{P}(\mathbf{w})\mathbf{y}\}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\{\boldsymbol{\epsilon} + \boldsymbol{\mu} - \mathbf{P}(\mathbf{w})\mathbf{y}\} \\ &\quad - \operatorname{trace}\{\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\ &\le \sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\boldsymbol{\epsilon} - \operatorname{trace}\{\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\ &\quad + 2\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\{\mathbf{P}(\mathbf{w})\mathbf{y} - \boldsymbol{\mu}\}|/R_n(\mathbf{w})] \\ &\quad + \sup_{\mathbf{w}\in\mathcal{W}}[|\{\mathbf{P}(\mathbf{w})\mathbf{y} - \boldsymbol{\mu}\}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\{\mathbf{P}(\mathbf{w})\mathbf{y} - \boldsymbol{\mu}\}|/R_n(\mathbf{w})] \\ &\le \sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\boldsymbol{\epsilon} - \operatorname{trace}\{\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\ &\quad + 2\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\{\mathbf{P}(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu}\}|/R_n(\mathbf{w})] \\ &\quad + 2\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon} - \operatorname{trace}\{\mathbf{Q}(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \end{aligned}$$

$$+2\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}\{\mathbf{Q}(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})]$$
$$+\sup_{\mathbf{w}\in\mathcal{W}}[|\{\mathbf{P}(\mathbf{w})\mathbf{y}-\boldsymbol{\mu}\}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\{\mathbf{P}(\mathbf{w})\mathbf{y}-\boldsymbol{\mu}\}|/R_n(\mathbf{w})]$$
$$\equiv \Xi_1+\Xi_2+\Xi_3+\Xi_4+\Xi_5. \tag{A.5}$$

Define $\rho=\max_s\max_i\rho_{ii}^{(s)}$. From Condition (C.2), we have

$$\rho=O(n^{-1}\widetilde{p}). \tag{A.6}$$

By (2) and the third part of Condition (C.1), we have

$$R_n(\mathbf{w}_s^o)\geq \mathrm{trace}\{\mathbf{P}_{(s)}\boldsymbol{\Omega}\mathbf{P}_{(s)}\}\geq\bar{\sigma}^2\mathrm{trace}(\mathbf{P}_{(s)})=\bar{\sigma}^2p_s,$$

which together with the second part of Condition (C.1) imply that

$$\xi_n\to\infty\quad\text{and}\quad S_n\xi_n^{-2G}=o(1). \tag{A.7}$$

Using (A.1), (A.4), (A.6), Chebyshev's inequality and Theorem 2 of Whittle (1960), we obtain that, for any $\delta>0$,

$$\begin{aligned}\Pr(\Xi_1>\delta)&\leq\sum_{s=1}^{S_n}\Pr[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}_{(s)}\boldsymbol{\epsilon}-\mathrm{trace}(\mathbf{Q}_{(s)}\boldsymbol{\Omega})|>\delta\xi_n]\\&\leq\delta^{-2G}\xi_n^{-2G}\sum_{s=1}^{S_n}E\{\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}_{(s)}\boldsymbol{\epsilon}-\mathrm{trace}(\mathbf{Q}_{(s)}\boldsymbol{\Omega})\}^{2G}\\&\leq c_1\delta^{-2G}\xi_n^{-2G}\sum_{s=1}^{S_n}\mathrm{trace}^G\{\boldsymbol{\Omega}^{1/2}\mathbf{Q}_{(s)}\boldsymbol{\Omega}\mathbf{Q}_{(s)}\boldsymbol{\Omega}^{1/2}\}\\&\leq c_1\delta^{-2G}\xi_n^{-2G}\lambda_{\max}^{2G}(\boldsymbol{\Omega})n^G\rho^{2G}S_n\\&=\xi_n^{-2G}S_n\{O(n^{-1}\widetilde{p}^2)\}^G\end{aligned} \tag{A.8}$$

and

$$\begin{aligned}&\Pr(\Xi_3>\delta)\\&\leq\sum_{s=1}^{S_n}\Pr\{|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}_{(s)}\mathbf{P}_{(s)}\boldsymbol{\epsilon}-\mathrm{trace}(\mathbf{Q}_{(s)}\mathbf{P}_{(s)}\boldsymbol{\Omega})|>\delta\xi_n\}\\&\leq\delta^{-2G}\xi_n^{-2G}\sum_{s=1}^{S_n}E[\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{Q}_{(s)}\mathbf{P}_{(s)}\boldsymbol{\epsilon}-\mathrm{trace}(\mathbf{Q}_{(s)}\mathbf{P}_{(s)}\boldsymbol{\Omega})]^{2G}\\&\leq c_2\delta^{-2G}\xi_n^{-2G}\sum_{s=1}^{S_n}\mathrm{trace}^G\{\boldsymbol{\Omega}^{1/2}\mathbf{Q}_{(s)}\mathbf{P}_{(s)}\boldsymbol{\Omega}\mathbf{P}_{(s)}^{\mathrm{T}}\mathbf{Q}_{(s)}\boldsymbol{\Omega}^{1/2}\}\\&\leq c_2\delta^{-2G}\xi_n^{-2G}\lambda_{\max}^{2G}(\boldsymbol{\Omega})n^G\rho^{2G}S_n\max_s\lambda_{\max}^G(\mathbf{P}_{(s)}\mathbf{P}_{(s)}^{\mathrm{T}})\\&=\xi_n^{-2G}S_n\{O(n^{-1}\widetilde{p}^2)\}^G,\end{aligned} \tag{A.9}$$

where c_1 and c_2 are positive constants. It follows from (A.7)–(A.9) and Condition (C.3) that $\Xi_1+\Xi_3=o_p(1)$. From (2), (A.1), (A.4), (A.6), Condition (C.1), and the second part of (A.2), we have

$$\begin{aligned}\Xi_2&\leq\sup_{\mathbf{w}\in\mathcal{W}}\{\|\boldsymbol{\epsilon}\|^2\rho^2\|\mathbf{P}(\mathbf{w})\boldsymbol{\mu}-\boldsymbol{\mu}\|^2/R_n^2(\mathbf{w})\}^{1/2}\\&\leq\|\boldsymbol{\epsilon}\|\rho\xi^{-1/2}=\xi_n^{-1/2}O(n^{-1/2}\widetilde{p}),\\\Xi_4&\leq\xi_n^{-1}\rho\lambda_{\max}(\boldsymbol{\Omega})\sup_{\mathbf{w}\in\mathcal{W}}[\mathrm{trace}\{\mathbf{P}(\mathbf{w})\}]\\&\leq\xi_n^{-1}\rho\lambda_{\max}(\boldsymbol{\Omega})\max_s\{\mathrm{trace}(\mathbf{P}_s)\}\\&\leq\xi_n^{-1}\rho\lambda_{\max}(\boldsymbol{\Omega})\max_s\{\lambda_{\max}(\mathbf{P}_s)\}\max_s\{\mathrm{rank}(\mathbf{P}_s)\}\\&=\xi_n^{-1}O(n^{-1}\widetilde{p}^2)=\xi_n^{-1}O(n^{-1}\widetilde{p}^2)\end{aligned}$$

and

$$\begin{aligned}\Xi_5&\leq\rho\sup_{\mathbf{w}\in\mathcal{W}}[\{\mathbf{P}(\mathbf{w})\mathbf{y}-\boldsymbol{\mu}\}^{\mathrm{T}}\{\mathbf{P}(\mathbf{w})\mathbf{y}-\boldsymbol{\mu}\}/R_n(\mathbf{w})]\\&=\rho\sup_{\mathbf{w}\in\mathcal{W}}[L_n(\mathbf{w})/R_n(\mathbf{w})]=O(n^{-1}\widetilde{p}).\end{aligned}$$

Thus, by (A.7) and Condition (C.3), we see that $\Xi_2+\Xi_4+\Xi_5=o_p(1)$. This completes the proof.

References

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

Gao, Y., Zhang, X., Wang, S., Chong, T., Zou, G., 2016. Frequentist model averaging for threshold models. (submitted for publication).

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

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

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, Q., Okui, R., 2013. Heteroskedasticity-robust Cp model averaging. Econom. J. 16, 463–472.

Wan, A.T.K., Zhang, X., Zou, G., 2010. Least squares model averaging by Mallows criterion. J. Econometrics 156, 277–283.

Whittle, P., 1960. Bounds for the moments of linear and quadratic forms in independent variables. Theory Probab. Appl. 5, 302–305.

Xie, T., 2015. Prediction model averaging estimator. Econom. Lett. 131, 5–8.

Zhang, X., Wan, A.T.K., Zou, G., 2013. Model averaging by jackknife criterion in models with dependent data. J. Econometrics 174, 82–94.

Zhang, X., Zou, G., Carroll, R., 2015. Model averaging based on kullback-leibler distance. Statist. Sinica 25, 1583–1598.