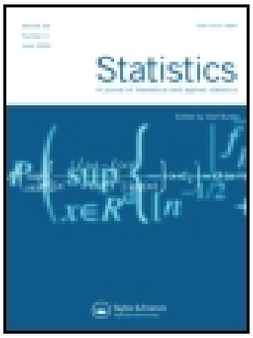

Statistics

A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: http://www.tandfonline.com/loi/gsta20

Model averaging for multivariate multiple regression models

Rong Zhu, Guohua Zou & Xinyu Zhang

To cite this article: Rong Zhu, Guohua Zou & Xinyu Zhang (2017): Model averaging for multivariate multiple regression models, Statistics, DOI: 10.1080/02331888.2017.1367794

To link to this article: http://dx.doi.org/10.1080/02331888.2017.1367794

Published online: 05 Sep 2017.

STATISTICS, 2017
https://doi.org/10.1080/02331888.2017.1367794

Taylor & Francis
Taylor & Francis Group

Model averaging for multivariate multiple regression models

Rong Zhu[a], Guohua Zou[b] and Xinyu Zhang[a]

[a]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, People's Republic of China; [b]School of Mathematical Science, Capital Normal University, Beijing, People's Republic of China

ABSTRACT
The theories and applications of model averaging have been developed comprehensively in the past two decades. In this paper, we consider model averaging for multivariate multiple regression models. In order to make use of the correlation information of the dependent variables sufficiently, we propose a model averaging method based on Mahalanobis distance which is related to the correlation of the dependent variables. We prove the asymptotic optimality of the resulting Mahalanobis Mallows model averaging (MMMA) estimators under certain assumptions. In the simulation study, we show that the proposed MMMA estimators compare favourably with model averaging estimators based on AIC and BIC weights and the Mallows model averaging estimators from the single dependent variable regression models. We further apply our method to the real data on urbanization rate and the proportion of non-agricultural population in ethnic minority areas of China.

ARTICLE HISTORY
Received 23 March 2017
Accepted 9 August 2017

KEYWORDS
Asymptotic optimality; Mahalanobis distance; Mallows criterion; model averaging; multivariate multiple regression model

1. Introduction

Model averaging method has been attracted more and more attention in the past two decades, which considers a series of candidate models and weights the results from these models. From the perspective of estimation and prediction, model averaging method can be regarded as an extension of model selection, which can avoid the instability on the model selection process and can also avoid selecting a poor model. There are two main research directions in model averaging: one is from the Bayesian point of view, and the other is from the frequentist point of view. See [1] for a comprehensive survey on Bayesian model averaging. However, in this paper, we focus on the frequentist model averaging. Buckland et al. [2] suggested smoothed-AIC and smoothed-BIC criteria, which weight all candidate model estimators and the negative exponents of AIC and BIC are used as the weights. Hansen [3] proposed the Mallows criterion for model averaging and selected the weights by minimizing this criterion, which is an estimator of the expected in-sample squared error. He constrained the model weights to a special discrete set and showed that the Mallows model averaging (MMA) estimator is asymptotically optimal in the sense of achieving the lowest possible squared error for a class of nested models. Wan et al. [4] extended Hansen's [3] study and showed that the asymptotic optimality of the Mallows criterion for model averaging holds for continuous model weights set and non-nested models. Subsequently, other model averaging criteria were also proposed. Examples include optimal mean squared error averaging by Liang et al. [5], Jackknife model averaging by Hansen and Racine [6] and Zhang et al. [7], heteroscedasticity-robust Cp model averaging by Liu and Okui [8], and leave-subject-out cross-validation by Gao et al. [9], among others. There are some other literatures focus on the asymptotic distribution theory of model averaging estimators, for example, [10–13].

CONTACT Guohua Zou ghzou@amss.ac.cn

All the proposed model averaging criteria mentioned above are for models with single dependent variable. To the best of our knowledge, there are no model averaging criteria for models with multiple dependent variables. The main purpose of this paper is to fill this gap. We will consider the multivariate multiple regression model (MMRM), which extends the dimension of dependent variables to be multidimensional and is a powerful tool to study the relationship between multiple independent variables and multiple dependent variables in the practical problems. These kinds of multiple-to-multiple problems are widely existent in economy, agriculture, environment, industry, biology and other fields. Here we provide a common example. To detect the atmospheric pollution in environmental science, we have the concentration of CO, SO, SO2, PM2.5, PM10, dust, smoke and methane as the multidimensional dependent variables. The concentration of atmospheric pollutants closely relates to the distribution of pollutant source, discharge value, topography, landforms, weather and other conditions. For this multiple-to-multiple problem, we need to consider using the MMRM to analyse the relationship between the variables. Other examples include that Izady et al. [14] used the MMRM to predict the groundwater fluctuations, Yanagihara and Satoh [15] applied the MMRM to study the Scottish election data, Jeong et al. [16] studied the daily precipitation with the MMRM, Dan et al. [17] employed the MMRM to investigate the relationship between the Vital Signs and the social characteristics of patients, and Shin et al. [18] analysed the mobile wireless networks by the MMRM. There are also lots of literatures for the theoretical study of MMRM, for example, [15,19–22].

In order to make use of the correlation information of the dependent variables sufficiently, a Mallows type model averaging criterion based on Mahalanobis distance [23] is proposed. We call this criterion as Mahalanobis Mallows criterion for model averaging. The weights are selected by minimizing the Mahalanobis Mallows criterion, which is an unbiased estimator of the expected in-sample Mahalanobis quadratic error plus a constant. We show that our Mahalanobis Mallows model averaging (MMMA) estimator is asymptotically optimal in the sense of achieving the lowest possible Mahalanobis quadratic error for continuous model weights set and non-nested models. The proving process of the asymptotic optimality is more complicated while comparing with the former work of Hansen [3] and Wan et al. [4] mentioned before, because we need to deal with the correlation between the dependent variables, and so need to provide relevant technical results.

The remainder of this paper begins with a description of the MMRM set-up and its parametric estimation in Section 2. Section 3 introduces the Mahalanobis Mallows criterion and the MMMA estimator. Section 4 discusses the asymptotic optimality of the MMMA estimator. Section 5 presents simulation evidence in support of the MMMA estimator. A real data example is studied in Sections 6 and 7 provides some concluding remarks. The technical proofs are relegated to the appendix.

2. Model set-up and parametric estimation

Our description of the model set-up follows Hansen's [3] notations. We consider the MMRM,

$$y_{ij} = \mu_{ij} + \epsilon_{ij} = \sum_{k=1}^{\infty} x_{ik}\theta_{kj} + \epsilon_{ij}, \quad i = 1, 2, \ldots, n, \;\; j = 1, 2, \ldots, p, \tag{1}$$

where y_{ij} is the jth component of the dependent variable $y_{(i)} = (y_{i1}, y_{i2}, \ldots, y_{ip})'$, which is a $p \times 1$ random vector; x_{ik} is the kth component of the independent variable $x_{(i)} = (x_{i1}, x_{i2}, \ldots)$, which is a countably infinite random vector; θ_{kj} is the unknown coefficient of x_{ik} corresponding to the jth dependent variable; and ϵ_{ij} is the jth component of the disturbance $\epsilon_{(i)} = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{ip})'$, which satisfies $E(\epsilon_{(i)} \mid x_{(i)}) = 0$, and $E(\epsilon_{(i)}\epsilon'_{(i)} \mid x_{(i)}) = \Sigma = (\sigma_{ij})_{p\times p} > 0$, which means Σ is a positive definite matrix. We assume p is fixed, $E\mu_{ij}^2 < \infty$ and μ_{ij} converges in mean square.

Equation (1) can also be written as the matrix form

$$Y = \mu + \Xi = X\theta + \Xi, \tag{2}$$

where $Y = (y_{(1)}, y_{(2)}, \ldots, y_{(n)})'$ is an $n \times p$ dependent variable matrix, $X = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})'$ is an $n \times \infty$ independent variable matrix, θ is an $\infty \times p$ parameter matrix, $\mu = X\theta$ is an $n \times p$ matrix function of X and $\Xi = (\epsilon_{(1)}, \epsilon_{(2)}, \ldots, \epsilon_{(n)})'$ is an $n \times p$ random error matrix.

We denote the p single dependent variable regression models as

$$y_j = \mu_j + \epsilon_j = X\theta_j + \epsilon_j, \quad j = 1, 2, \ldots, p, \tag{3}$$

where $y_j = (y_{1j}, y_{2j}, \ldots, y_{nj})'$ is the jth column of the dependent variable matrix Y, $\theta_j = (\theta_{1j}, \theta_{2j}, \ldots)'$ is the jth column of the parameter matrix θ, and $\epsilon_j = (\epsilon_{1j}, \epsilon_{2j}, \ldots, \epsilon_{nj})'$ is the jth column of the random error matrix Ξ.

We use M candidate MMRMs to approximate (2), where M is allowed to diverge to infinity as $n \to \infty$. The mth approximating (or candidate) MMRM can use any k_m regressors belonging to $x_{(i)}$. So the mth approximating model is

$$Y = \mu_{(m)} + \Xi = X_{(m)}\theta_{(m)} + \Xi, \quad m = 1, 2, \ldots, M,$$

where $X_{(m)}$ is an $n \times k_m$ matrix and has full column rank, and the columns of $X_{(m)}$ are constructed from the k_m columns of X, that is we only choose k_m independent variables in the mth candidate model, $\theta_{(m)}$ is a $k_m \times p$ parameter matrix.

The mth candidate model can be written as Equation (4), which is equivalent to Equation (5): we put the model in vector form by using the vectoring operation $\mathrm{Vec}(\cdot)$, which creates a column vector by stacking the column vectors of below one another, that is, we write

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = (I_p \otimes X_{(m)}) \begin{pmatrix} \theta_{1(m)} \\ \theta_{2(m)} \\ \vdots \\ \theta_{p(m)} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}, \tag{4}$$

where $\theta_{i(m)} = (\theta_{1i(m)}, \theta_{2i(m)}, \ldots, \theta_{k_m i(m)})'$ is the ith column of the parameter matrix $\theta_{(m)}$ in the mth candidate model. Denote

$$\mathrm{Vec}(Y) = D_{(m)}\mathrm{Vec}(\theta_{(m)}) + \mathrm{Vec}(\Xi), \tag{5}$$

where $D_{(m)} = I_p \otimes X_{(m)}$ is a $pn \times pk_m$ matrix.

We minimize the sum of squared errors and get the parameter estimators,

$$\widehat{\mathrm{Vec}(\theta_{(m)})} = (D'_{(m)}D_{(m)})^{-1}D'_{(m)}\mathrm{Vec}(Y) = \begin{pmatrix} (X'_{(m)}X_{(m)})^{-1}X'_{(m)}y_1 \\ (X'_{(m)}X_{(m)})^{-1}X'_{(m)}y_2 \\ \vdots \\ (X'_{(m)}X_{(m)})^{-1}X'_{(m)}y_p \end{pmatrix} = \begin{pmatrix} \hat{\theta}_{1(m)} \\ \hat{\theta}_{2(m)} \\ \vdots \\ \hat{\theta}_{p(m)} \end{pmatrix} = \mathrm{Vec}(\hat{\theta}_{(m)}),$$

where $\hat{\theta}_{j(m)} = (X'_{(m)}X_{(m)})^{-1}X'_{(m)}y_j$. Denote $P_{(m)} = X_{(m)}(X'_{(m)}X_{(m)})^{-1}X'_{(m)}$, which is an $n \times n$ projection matrix. Then under the mth candidate model, the estimator of μ is given by

$$\mathrm{Vec}(\hat{\mu}_{(m)}) = \mathrm{Vec}(X_{(m)}\hat{\theta}_{(m)}) = D_{(m)}\mathrm{Vec}(\hat{\theta}_{(m)}) = H_{(m)}\mathrm{Vec}(Y), \tag{6}$$

where $H_{(m)} = I_p \otimes P_{(m)}$. Equation (6) means $\mathrm{Vec}(\hat{\mu}_{(m)})$ is the linear function of $\mathrm{Vec}(Y)$.

3. Model averaging estimation and weight choice criterion

Let $w = (w_1, w_2, \ldots, w_M)'$ be a weight vector in the unit simplex of R^M:

$$\mathscr{H}_n = \left\{ w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}.$$

The model averaging estimator of $\mathrm{Vec}(\mu)$ is

$$\mathrm{Vec}(\hat{\mu}(w)) = \sum_{m=1}^{M} w_m \mathrm{Vec}(\hat{\mu}_{(m)}) = \sum_{m=1}^{M} w_m H_{(m)} \mathrm{Vec}(Y) = H(w)\mathrm{Vec}(Y),$$

where $P(w) = \sum_{m=1}^{M} w_m P_{(m)}$ and $H(w) = \sum_{m=1}^{M} w_m H_{(m)} = I_p \otimes P(w)$.

In order to make use of the correlation information of the dependent variables sufficiently, we adopt the Mahalanobis distance [23] to define the quadratic loss function,

$$L_n(w) = \{\mathrm{Vec}(\hat{\mu}(w)) - \mathrm{Vec}(\mu)\}' \Sigma_0^{-1} \{\mathrm{Vec}(\hat{\mu}(w)) - \mathrm{Vec}(\mu)\},$$

where $\Sigma_0 = \Sigma \otimes I_n$ and $\otimes$ means the Kronecker product. The corresponding risk function is

$$\begin{aligned} R_n(w) &= E(L_n(w) \mid X) \\ &= \mathrm{Vec}(\mu)'(I - H(w))' \Sigma_0^{-1} (I - H(w))\mathrm{Vec}(\mu) + p\mathrm{tr}\{P'(w)P(w)\}. \end{aligned} \tag{7}$$

Let

$$C_n(w) = \{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}' \Sigma_0^{-1} \{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\} + 2p\mathrm{tr}(P(w)). \tag{8}$$

By simple calculation, we obtain

$$E(C_n(w)) = E(L_n(w)) + np,$$

which means $C_n(w)$ is an unbiased estimator of the expected in-sample Mahalanobis quadratic error plus a constant. Therefore, we use $C_n(w)$ as the weight choice criterion and call this criterion as Mahalanobis Mallows criterion for model averaging. The optimal weight vector can be obtained by minimizing the Mahalanobis Mallows criterion over the weight set $\mathscr{H}_n$, that is,

$$\hat{w} = \arg\min_{w \in \mathscr{H}_n} C_n(w). \tag{9}$$

The corresponding estimator $\hat{\mu}(\hat{w})$ is referenced as the MMMA estimator.

Considering a special case with the covariance matrix Σ being a diagonal matrix, the Mahalanobis Mallows criterion defined in Equation (8) can be written as

$$C_n(w) = \sum_{i=1}^{p} C_{ni}(w)/\sigma_{ii}, \tag{10}$$

where $C_{ni}(w) = y_i'(I - P(w))'(I - P(w))y_i + 2\sigma_{ii}\mathrm{tr}(P(w))$ is the Mallows criterion in the ith single dependent variable regression model introduced by Hansen [3]. Equation (10) means that when the covariance matrix is diagonal, the weight choice criterion is equivalent to the sum of Mallows criteria for all single dependent variable regression models scaled by σ_{ii}.

In addition, if instead of $L_n(w)$, we use the sum of squares loss functions

$$L_{0,n}(w) = \{\mathrm{Vec}(\hat{\mu}(w)) - \mathrm{Vec}(\mu)\}'\{\mathrm{Vec}(\hat{\mu}(w)) - \mathrm{Vec}(\mu)\},$$

which does not contain Σ_0^{-1}, then we will get a weight choice criterion

$$C_{0,n}(w) = \{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\} + 2\mathrm{tr}(\Sigma)\mathrm{tr}(P(w)).$$

It is easy to show that $E(C_{0,n}(w)) = E(L_{0,n}(w)) + n\mathrm{tr}(\Sigma)$, where the second term has nothing to do with w. By some calculations, we have

$$C_{0,n}(w) = \sum_{i=1}^{p} C_{ni}(w),$$

that means that when using the sum of squares loss functions $L_{0,n}(w)$, the weight choice criterion for MMRMs is equivalent to the sum of Mallows criteria for all single dependent variable regression models.

4. Asymptotic optimality of MMMA estimator

Let $\xi_n = \inf_{w\in\mathscr{H}_n} R_n(w)$ and w_m^0 be an $M \times 1$ vector in which the mth element is one and the others are zeros. We consider the asymptotic optimality of the proposed MMMA estimator, which is given in the next theorem.

Theorem 4.1: *As $n \to \infty$, if, for some fixed integer G, $1 \le G < \infty$, and constant $\kappa < \infty$,*

$$E(\epsilon_{ij}^{4G}|x_{(i)}) \le \kappa < \infty, \quad \text{for all } i = 1, \ldots, n, \quad j = 1, 2, \ldots, p, \tag{11}$$

and

$$M\xi_n^{-2G} \sum_{m=1}^{M} \{R_n(w_m^0)\}^G \to 0, \tag{12}$$

then

$$\frac{L_n(\hat{w})}{\inf_{w\in\mathscr{H}_n} L_n(w)} \xrightarrow{p} 1. \tag{13}$$

Theorem 4.1 shows that the model averaging procedure using $\hat{w}$ is asymptotically optimal in the sense that the resulting Mahalanobis quadratic loss is asymptotically identical to that of the infeasible best possible model averaging estimator. The proof of Theorem 4.1 is provided in Appendix A.2.

Remark 4.1: Condition (11) is a counterpart to condition (7) of Wan et al. [4], and places a bound on the conditional moments of random errors. The convergence condition given in Equation (12) is similar to condition (8) of Wan et al. [4]. $\xi_n \to \infty$ is obviously a necessary condition for (12) to hold. As Hansen [3] remarked, it simply means there is no finite approximating model for which the bias is zero. Condition (12) also requires $M\sum_{m=1}^{M}\{R_n(w_m^0)\}^G \to \infty$ at a rate slower than $\xi_n^{-2G} \to \infty$.

So far we have assumed that the covariance matrix Σ is known. This is, of course, not the case in practice. We can use the largest approximating model to estimate Σ, which is also suggested by

Hansen [3]. The largest approximating model is the model such that $k_{M^*} = \max\{k_1, k_2, \ldots, k_M\}$. Then the estimator of Σ based on the largest approximating model M^* is given by

$$\hat{\Sigma} = \frac{1}{n - k_{M^*}}(Y - \hat{\mu}_{(M^*)})'(Y - \hat{\mu}_{(M^*)}). \tag{14}$$

Let $\bar{\lambda}(\cdot)$ be the maximum singular value of a matrix. The following theorem shows that Theorem 4.1 remains valid if Σ is unknown and replaced by $\hat{\Sigma}$, and provided that some mild conditions are satisfied.

Theorem 4.2: *When Σ is unknown and replaced by $\hat{\Sigma}$, the asymptotic optimality of MMMA estimator in Equation* (13) *remains valid if*

$$\bar{\lambda}(\hat{\Sigma} - \Sigma)\xi_n^{-1} n \xrightarrow{p} 0, \tag{15}$$

$$\frac{k_{M^*}}{n} \to 0, \tag{16}$$

$$k_{M^*} \to \infty, \tag{17}$$

and

$$\mathrm{Vec}(\mu)'\mathrm{Vec}(\mu)/n = O(1), \tag{18}$$

as $n \to \infty$.

Theorem 4.2 presents the asymptotic optimality of $\hat{\mu}(\hat{w})$ when the covariance matrix Σ is unknown. The proof of Theorem 4.2 is given in Appendix A.3.

Remark 4.2: Condition (15) requires that the convergence rate of $\hat{\Sigma} - \Sigma$ be faster than that of $(\xi_n^{-1} n)^{-1}$. When all the $X_{(m)}$ are of full column ranks, Conditions (16) and (17) place constraints on the number of regressors in the largest approximating model. Conditions (16) and (17) are similar to those of Theorem 2 in Hansen [3]. Condition (18) concerns the average of the squares of the conditional means μ_{ij}'s. This condition is similar to Condition (11) in Wan et al. [4].

5. Simulation study

5.1. Simulation setup

Our setting is similar to the infinite-order regression by Hansen [3]. We generate the data by $Y = \mu + \Xi = X\theta + \Xi$, where Y is an $n \times 2$ dependent variable matrix and X is an $n \times \infty$ independent variable matrix. For practical calculations, instead of ∞, we take X to be an $n \times M_0$ independent variable matrix, where M_0 is a large number and is set to be 200 in this section. The component of the first column of X is 1 for intercept. The remaining x_{ij} are i.i.d. $N(0, 1)$. Each row of the random error matrix Ξ is $\epsilon'_{(i)}$ and $\epsilon_{(i)} \overset{\text{i.i.d}}{\sim} N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & r_0 \\ r_0 & 1 \end{pmatrix}$, and is independent of $x_{(i)}$. We consider r_0 varying between 0, 0.3, 0.6 and 0.9. The greater the r_0 is, the greater the correlation of random errors is. The coefficient matrix θ is an $M_0 \times 2$ matrix, and the element of the jth row is $(\theta_{j1}, \theta_{j2})$. We set $\theta_{j1} = c\sqrt{2\alpha}j^{-(1/2)\alpha - 1/2}$ and $\theta_{j2} = c\sqrt{\alpha}j^{-\alpha - 1/10}$, which are similar to the setting of Hansen [3]. The coefficient c is selected to control the population R^2, which is set as $(\mathrm{var}(\mu_{i1}) + \mathrm{var}(\mu_{i2}))/(\mathrm{var}(y_{i1}) + \mathrm{var}(y_{i2})) \approx (\mathrm{var}(\mu_{i1}) + \mathrm{var}(\mu_{i2}))/(\mathrm{var}(\mu_{i1}) + \mathrm{var}(\mu_{i2}) + \mathrm{var}(\epsilon_{i1}) + \mathrm{var}(\epsilon_{i2}))$, and varied on a grid between 0.1 and 0.9. The parameter α is varied between 0.5, 1.0 and 1.5. The sample size is set to be n =50, 100 and

200. The number of candidate models M is determined by the rule $M = \text{round}(3n^{1/3})$, which is used by Hansen [3], that is, $M = 11$, 14 and 18, where round$(\cdot)$ means rounding. The M candidate models are the nested models, and the mth candidate model contains the first m independent variables.

We consider eight kinds of model selection and model averaging methods:

(1) AIC (Akaike information criterion) model selection method for MMRMs [24]: $\text{AIC}_m = n(\ln|\hat{\Sigma}_m| + p) + 2\{pk_m + p(p+1)/2\}$, $m = 1, 2, \ldots, M$, where $\hat{\Sigma}_m = (Y - \hat{\mu}_{(m)})'(Y - \hat{\mu}_{(m)})/n$ and $p = 2$. We select the model that minimizes AIC;
(2) AICc model selection method for MMRMs [24]: $\text{AICc}_m = n(\ln|\hat{\Sigma}_m| + p) + 2\{pk_m + p(p+1)/2\}\{n/(n - k_m - p - 1)\}$, where $\hat{\Sigma}_m$ and p are the same as those in AIC. We select the model that minimizes AICc;
(3) BIC (Bayesian information criterion) model selection method for MMRMs: $\text{BIC}_m = n(\ln|\hat{\Sigma}_m| + p) + \ln(n)\{pk_m + p(p+1)/2\}$, where $\hat{\Sigma}_m$ and p are the same as those in AIC. We select the model that minimizes BIC;
(4) S-AIC, that is, smoothed AIC model averaging method, which uses the negative exponent of AIC as the weight: $w_m = \exp(-\text{AIC}_m/2)/\sum_{l=1}^{M}\exp(-\text{AIC}_l/2)$ is the weight of the mth candidate model;
(5) S-AICc, that is, smoothed AICc model averaging method, which uses the negative exponent of AICc as the weight: $w_m = \exp(-\text{AICc}_m/2)/\sum_{l=1}^{M}\exp(-\text{AICc}_l/2)$ is the weight of the mth candidate model;
(6) S-BIC, that is, smoothed BIC model averaging method, which uses the negative exponent of BIC as the weight: $w_m = \exp(-\text{BIC}_m/2)/\sum_{l=1}^{M}\exp(-\text{BIC}_l/2)$ is the weight of the mth candidate model;
(7) MMA, which treats the p dimensional MMRM as p single dependent variable regression models and is used to obtain the weights for each single dependent variable regression model;
(8) MMMA, the Mahalanobis Mallows criterion for the model averaging estimator, which is the proposed criterion in this paper.

To evaluate the estimators, we compute the averaged Mahalanobis quadratic loss

$$\text{Risk} = \frac{1}{D}\sum_{d=1}^{D}\{\text{Vec}(\hat{\mu}^{(d)}(w)) - \text{Vec}(\mu^{(d)})\}'\Sigma_0^{-1}\{\text{Vec}(\hat{\mu}^{(d)}(w)) - \text{Vec}(\mu^{(d)})\}.$$

The subscript d means the dth simulation run. We take $D = 500$. For each parameterization, the risk is normalized by dividing by the risk of the infeasible optimal least squares estimator.

5.2. Simulation results

The risk calculations are presented in Figures 1–3 for $\alpha = 1.5$ and $n = 50$, 100 and 200, respectively. To save space, we only report the results with $\alpha = 1.5$. The simulation results with $\alpha = 0.5$ and 1 are similar and available from the author upon request. In each figure, four different correlation coefficients $r_0 = 0$, 0.3, 0.6 and 0.9 are considered.

We find that all the model selection methods perform worse than the corresponding model averaging methods, that is the S-AIC method achieves a lower risk than the AIC method, the S-AICc method achieves a lower risk than the AICc method, and the S-BIC method achieves a lower risk than the BIC method, implying that the model averaging methods are superior to the model selection methods. For simplicity, we do not plot the AIC, AICc and BIC model selection methods in all figures. (The figures which contain the model selection methods are available on request from the authors.)

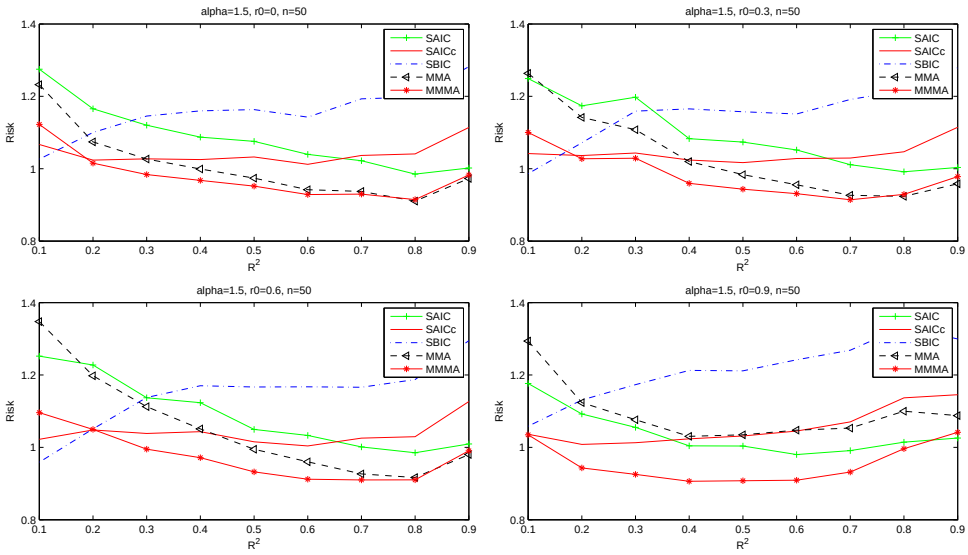

Figure 1. Risk comparison of five methods ($\alpha = 1.5$, $n = 50$).

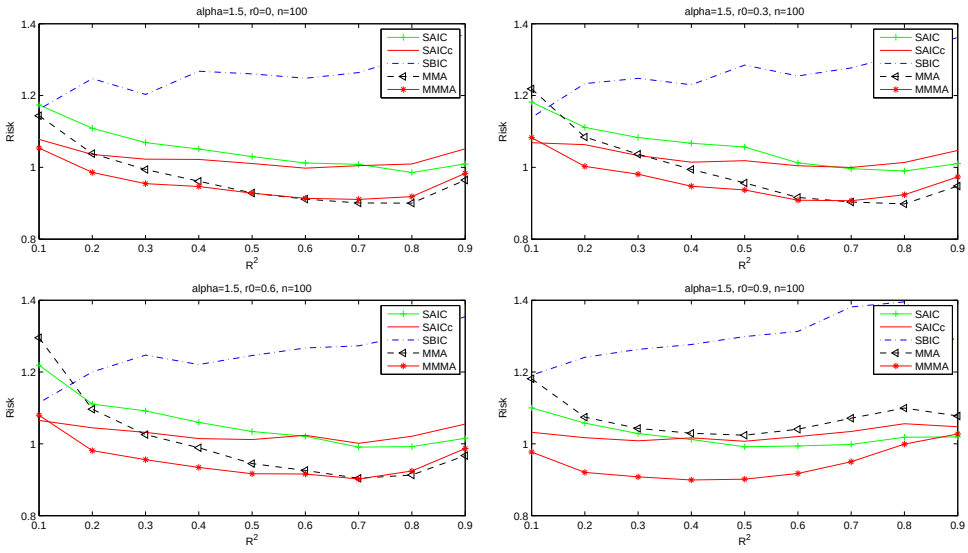

Figure 2. Risk comparison of five methods ($\alpha = 1.5$, $n = 100$).

It is seen from all the figures that, with the increasing of r_0, the risk gap between MMMA and other methods is becoming larger, which shows the obvious advantages of our Mahalanobis Mallows criterion. When r_0 is high, such as $r_0 = 0.9$, the MMMA method is uniformly superior to other methods. In many situations, the MMMA normalized risk is less than 1, which means that it is often smaller than that of the infeasible optimal model selection. The good performance of MMMA method owes to its consideration of making sufficient use of the correlation information of the dependent variables.

It is clear from the last two figures that when r_0 is low and moderate, the performance of the MMMA method depends on R^2. When R^2 is small, the MMMA method performs best in most of the situations. When R^2 is large, the MMA method can perform better than the MMMA method. This is

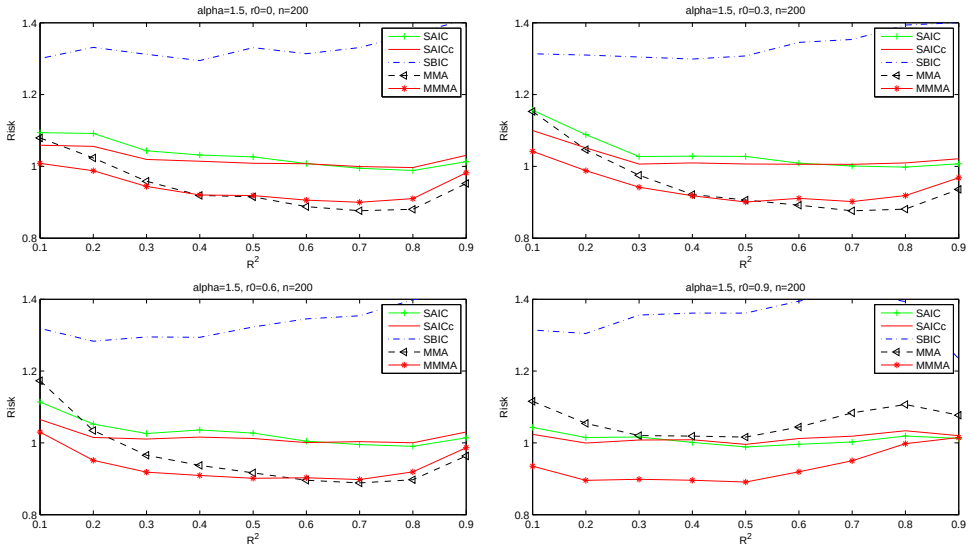

Figure 3. Risk comparison of five methods ($\alpha = 1.5$, $n = 200$).

not surprising, because when $r_0 = 0$, each single dependent variable regression model is uncorrelated, and in this situation, we have no correlation information which can be used.

We also simulate the different coefficient matrix θ and consider the different distribution of independent variables $x_{i,j}$, and the conclusions are similar.

6. Empirical application

We apply our method to analyse the data of 77 ethnic minority areas in China, collected comes from the sixth national population census in China in 2010. We treat the urbanization rate and the proportion of non-agricultural population as the two dimensinal dependent variables, and we consider six independent variables which may be related to the dependent variables: the proportion of the population of ethnic minorities, urban and rural income ratio, per capita investment in fixed assets, per capita GDP, the second industry output value ratio and the service output value ratio.

We consider all possible, totally 64, candidate models. To compare the performance of the eight methods mentioned in Section 5, we use any T ethnic minority areas data to do the regression process and get the coefficient estimates of the eight methods, where $T = 50$ and 60. Then we forecast the risk by calculating the averaged Mahalanobis quadratic loss with the remaining 77-T ethnic minority areas data. This process is done $N = 500$ times, and we calculate the mean and the median risks of the eight methods, whose calculation formula is the same as that in Section 5. We also calculate the optimal rate of each method, and here the optimal rate means the number of times that the risk of this method is the smallest among the eight methods, divided by the total number $N = 500$. The results are displayed in Table 1, where the optimal values are highlighted in bold.

It is observed from Table 1 that both the mean risk and the median risk of MMMA are lower than those of other methods, respectively, and the optimal rate of MMMA is the highest among all the methods, which showing that our proposed method has obvious advantages over other methods. From the perspective of the mean and the median, the best method is MMMA, followed by some model averaging methods, such as MMA, S-BIC, S-AICc and S-AIC. And the worst methods are some model selection methods, such as AIC and BIC. From the perspective of the optimal rate, the best is MMMA, followed by AIC, then MMA, BIC and S-BIC, and the worst are S-AIC, AICc and S-AICc. In most of the situations, the model selection methods perform worse than the corresponding model averaging methods.

Table 1. Comparison of the risk results of eight methods.

	Method	AIC	AICc	BIC	S-AIC	S-AICc	S-BIC	MMA	MMMA
$T = 50$	Mean	4.032	4.022	3.999	4.006	3.962	3.935	3.929	**3.910**
	Median	3.618	3.610	3.645	3.604	3.593	3.570	3.599	**3.569**
	Optimal rate	0.182	0.076	0.094	0.064	0.022	0.132	0.170	**0.260**
$T = 60$	Mean	6.222	6.191	6.233	6.193	6.142	6.140	6.132	**6.056**
	Median	5.385	5.300	5.310	5.307	5.269	5.336	5.251	**5.148**
	Optimal rate	0.182	0.038	0.122	0.062	0.022	0.106	0.154	**0.314**

Note: Bold values indicate the best results.

7. Concluding remarks

The MMRMs have become popular in many fields because such models consider the multiple-to-multiple relationship between the independent variables and the dependent variables. In this paper, we have investigated the model averaging approach for such models. In order to make sufficient use of the correlation information of the dependent variables, we have derived a weight choice criterion for model averaging estimators based on Mahalanobis distance. The asymptotic optimality of the resulting estimator has been proved, and both simulation study and real data analysis have shown its validity.

This is the first paper which considers the MMRM in the model averaging literature, and so many problems remain for further studies. At least three issues deserve future research. First, we have not discussed the asymptotic distribution of the proposed MMMA estimator in this paper. Since the optimal weight vector depends on the observations, more efforts are needed to derive the asymptotic distribution of the MMMA estimator. Second, this paper has extended MMA criterion from the models with single dependent variable to the models with multiple dependent variables. It is also interesting to study other model averaging methods for MMRMs, such as Jackknife model averaging by Hansen and Racine [6]. Third, we have considered only the simple data for multivariate multiple regression models. How to develop optimal model averaging methods for complex data, such as longitudinal data, high-dimensional data, time series data, missing data and measurement error data, for MMRMs is also an interesting problem.

Acknowledgments

The authors thank the editor, the associate editor and the referee for helpful comments and suggestions.

Funding

Zou's work was partially supported by the National Natural Science Foundation of China [Grant nos. 11331011 and 11529101] and a grant from the Ministry of Science and Technology of China [Grant no. 2016YFB0502301]. Zhang's work was partially supported by the National Natural Science Foundation of China [Grant nos. 71522004 and 11471324] and a special talent grant from Academy of Mathematics and Systems Science , Chinese Academy of Sciences.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Hoeting JA, Madigan D, Raftery AE, et al. Bayesian model averaging: a tutorial. Statist Sci. 1999;14:382–417.
[2] Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. Biometrics. 1997;53:603–618.
[3] Hansen BE. Least squares model averaging. Econometrica. 2007;75:1175–1189.
[4] Wan ATK, Zhang X, Zou G. Least squares model averaging by Mallows criterion. J Econometrics. 2010;156:277–283.

[5] Liang H, Zou G, Wan ATK, et al. Optimal weight choice for frequentist model average estimators. J Amer Statist Assoc. 2011;106:1053–1066.
[6] Hansen BE, Racine JS. Jackknife model averaging. J Econometrics. 2012;167:38–46.
[7] Zhang X, Wan ATK, Zou G. Model averaging by jackknife criterion in models with dependent data. J Econometrics. 2013;174:82–94.
[8] Liu Q, Okui R. Heteroscedasticity-robust Cp model averaging. Econom J. 2013;16:463–472.
[9] Gao Y, Zhang X, Wang S, et al. Model averaging based on leave-subject-out cross-validation. J Econometrics. 2016;192:139–151.
[10] Claeskens G, Carroll RJ. An asymptotic theory for model selection inference in general semiparametric problems. Biometrika. 2007;94:249–265.
[11] Hjort NL, Claeskens G. Frequentist model average estimators. J Amer Statist Assoc. 2003;98:879–899.
[12] Liu C. Distribution theory of the least squares averaging estimator. J Econometrics. 2015;186:142–159.
[13] Wang H, Zou G, Wan ATK. Model averaging for varying-coefficient partially linear measurement error models. Electron J Stat. 2012;6:1017–1039.
[14] Izady A, Soheili M, Davari K, et al. Groundwater level forecasting using multivariate-multiple regression model. ICWR 2009 international conference on water resources; 2009.
[15] Yanagihara H, Satoh K. An unbiased Cp criterion for multivariate ridge regression. J Multivariate Anal. 2010;101:1226–1238.
[16] Jeong DI, St-Hilaire A, Ouarda TBMJ, et al. Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator. Clim Change. 2012;114:567–591.
[17] Dan ED, Jude O, Chukwukailo NA. Application of multivariate multiple linear regression model on vital signs and social characteristics of patients. Int J Eng Sci. 2013;2:40–56.
[18] Shin Y, Chae CB, Kim S. Multivariate multiple regression models for a big data-empowered SON framework in mobile wireless networks. Mobile Inf Syst. 2016;2016:1–10.
[19] Fujikoshi Y, Satoh K. Modified AIC and Cp in multivariate linear regression. Biometrika. 1997;84:707–716.
[20] Fujikoshi Y, Sakurai T, Yanagihara H. Consistency of high-dimensional AIC-type and Cp-type criteria in multivariate linear regression. J Multivariate Anal. 2014;123:184–200.
[21] Sparks RS, Coutsourides D, Troskie L. The multivariate Cp. Comm Statist Theory Methods. 1983;12:1775–1793.
[22] Yanagihara H, Wakaki H, Fujikoshi Y. A consistency property of AIC for multivariate linear model when the dimension and the sample size are large. Electron J Stat. 2015;43:231–245.
[23] Mahalanobis PC. On the generalised distance in statistics. Proc Natl Inst Sci India. 1936;2:49–55.
[24] Bedrick EJ, Tsai C-L. Model selection for multivariate regression in small samples. Biometrics. 1994;50:226–231.
[25] Whittle P. Bounds for the moments of linear and quadratic forms in independent variables. Theory Probab Appl. 1960;5:302–305.

Appendix

To prove Theorems 4.1 and 4.2, we first give some lemmas.

A.1 Useful preliminary results

Lemma A.1: *Assume that Conditions* (11) *hold, then for some fixed integer* $1 \le G < \infty$, *the* $4G$*th conditional moment of the absolute value of each element of* $\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)$ *is* $O(1)$.

Proof: Note that $\Sigma_0^{-1/2} = \Sigma^{-1/2} \otimes I_n$. Denote

$$\Sigma^{-1/2} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix}_{p\times p}.$$

Then for all $j = 1, 2, \ldots, p$ and $l = 1, 2, \ldots, n$, the $\{(j-1)n + l\}$th element of the vector $\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)$ can be expressed as $\sum_{i=1}^{p} a_{ji}\epsilon_{li}$. Using the Minkowski's inequality, we obtain

$$E\left(\left|\sum_{i=1}^{p} a_{ji}\epsilon_{li}\right|^{4G} |x_{(l)}\right) \le \left[\sum_{i=1}^{p}\{E(|a_{ji}\epsilon_{li}|^{4G}|x_{(l)})\}^{1/4G}\right]^{4G} = \left[\sum_{i=1}^{p}|a_{ji}|\{E(\epsilon_{li}^{4G}|x_{(l)})\}^{1/4G}\right]^{4G}.$$

From Condition (11), we obtain $E(\epsilon_{li}^{4G} \mid x_{(l)}) = O(1)$. Since Σ is positive definite and p is fixed, Σ^{-1} is also positive definite and a_{ji} is finite for all $i, j = 1, 2, \ldots, p$. For some fixed integer $1 \le G < \infty$, $E(|\sum_{i=1}^{p} a_{ji}\epsilon_{li}|^{4G}|x_{(l)}) = O(1)$. ■

Lemma A.2: *Suppose that Conditions* (11), (16) *and* (17) *hold, then*

$$\hat{\Sigma} - \Sigma = o_p(1), \tag{A1}$$

where $\hat{\Sigma} = (1/(n - k_{M^*}))(Y - \hat{\mu}_{(M^*)})'(Y - \hat{\mu}_{(M^*)})$ *is defined in Equation* (14), *which is an estimator of* Σ *based on the largest approximating model.*

Proof: We denote $\Psi_m \subseteq \{1, 2, \ldots\}$ as an index set of the independent variables in the mth candidate model and Ψ_m^c as the complement of Ψ_m. The size of Ψ_m is k_m. From Equation (1), we have

$$\mu_{ij} = \sum_{k=1}^{\infty} x_{ik}\theta_{kj} = \sum_{k\in\Psi_{M^*}} x_{ik}\theta_{kj} + \sum_{k\in\Psi_{M^*}^c} x_{ik}\theta_{kj}, \quad i = 1, 2, \ldots, n, \;\; j = 1, 2, \ldots, p. \tag{A2}$$

Let $\mu_{(M^*)ij} = \sum_{k\in\Psi_{M^*}} x_{ik}\theta_{kj}$ be the ijth component of $\mu_{(M^*)}$ and $b_{(M^*)ij} = \sum_{k\in\Psi_{M^*}^c} x_{ik}\theta_{kj}$ be the ijth component of $b_{(M^*)}$, where $b_{(M^*)} = \mu - \mu_{(M^*)}$ represents the approximation error in the M^*th model.

From Equation (14), it is easily seen that

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{n - k_{M^*}}(Y - \hat{\mu}_{(M^*)})'(Y - \hat{\mu}_{(M^*)}) \\
&= \frac{Y'(I_n - P_{(M^*)})Y}{n - k_{M^*}} \\
&= \frac{(b_{(M^*)} + \Xi)'(I_n - P_{(M^*)})(b_{(M^*)} + \Xi)}{n - k_{M^*}} \\
&= \frac{b'_{(M^*)}(I_n - P_{(M^*)})b_{(M^*)}}{n - k_{M^*}} + 2\frac{b'_{(M^*)}(I_n - P_{(M^*)})\Xi}{n - k_{M^*}} + \frac{\Xi'(I_n - P_{(M^*)})\Xi}{n - k_{M^*}}.
\end{aligned}$$

We need only to verify that

$$\frac{b'_{(M^*)}(I_n - P_{(M^*)})b_{(M^*)}}{n - k_{M^*}} = o_p(1), \tag{A3}$$

$$\frac{b'_{(M^*)}(I_n - P_{(M^*)})\Xi}{n - k_{M^*}} = o_p(1), \tag{A4}$$

and

$$\frac{\Xi'(I_n - P_{(M^*)})\Xi}{n - k_{M^*}} - \Sigma = o_p(1). \tag{A5}$$

First, we consider (A3). Since $P_{(M^*)}$ is an idempotent matrix, we obtain

$$\frac{b'_{(M^*)}(I_n - P_{(M^*)})b_{(M^*)}}{n - k_{M^*}} \le \frac{b'_{(M^*)}b_{(M^*)}}{n - k_{M^*}}. \tag{A6}$$

Denote $(b'_{(M^*)}b_{(M^*)}/(n - k_{M^*}))_{ij}$ as the ijth component of $b'_{(M^*)}b_{(M^*)}/(n - k_{M^*})$ and e_i as a $p \times 1$ vector, in which the ith element is one and the others are zeros. By Condition (17) and the model set-up that $E(\mu_{ij}^2) < \infty$ and μ_{ij} converges in mean square, we have $E(b_{(M^*)ij}^2) \to 0$. Further, it is seen that

$$\sum_{i=1}^{n} E(b_{(M^*)ij}^2) = o(n). \tag{A7}$$

Since $b_{(M^*)ij}^2$ is non-negative, we obtain $b_{(M^*)ij}^2 = o_p(1)$. This implies $\sum_{i=1}^{n}(\sum_{k\in\Psi_{M^*}^c} x_{ik}\theta_{kj})^2 = o_p(n)$. By Cauchy–Schwarz inequality and Condition (16), we have

$$\begin{aligned}
\left|\left(\frac{b'_{(M^*)}b_{(M^*)}}{n - k_{M^*}}\right)_{ij}\right| &= \left|\frac{e'_i b'_{(M^*)}b_{(M^*)}e_j}{n - k_{M^*}}\right| \\
&= \left|\frac{\sum_{l=1}^{n}(\sum_{k\in\Psi_{M^*}^c} x_{lk}\theta_{ki})(\sum_{k\in\Psi_{M^*}^c} x_{lk}\theta_{kj})}{n - k_{M^*}}\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\{\sum_{l=1}^{n}(\sum_{k\in\Psi_{M^*}^c} x_{lk}\theta_{ki})^2\}^{1/2}\{\sum_{l=1}^{n}(\sum_{k\in\Psi_{M^*}^c} x_{lk}\theta_{kj})^2\}^{1/2}}{n-k_{M^*}} \\
&= \frac{o_p(n)}{n-k_{M^*}} \\
&= o_p(1),
\end{aligned} \tag{A8}
$$

where

$$
b_{(M^*)} = \begin{pmatrix} \sum_{k\in\Psi_{M^*}^c} x_{1k}\theta_{k1} & \sum_{k\in\Psi_{M^*}^c} x_{1k}\theta_{k2} & \cdots & \sum_{k\in\Psi_{M^*}^c} x_{1k}\theta_{kp} \\ \sum_{k\in\Psi_{M^*}^c} x_{2k}\theta_{k1} & \sum_{k\in\Psi_{M^*}^c} x_{2k}\theta_{k2} & \cdots & \sum_{k\in\Psi_{M^*}^c} x_{2k}\theta_{kp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k\in\Psi_{M^*}^c} x_{nk}\theta_{k1} & \sum_{k\in\Psi_{M^*}^c} x_{nk}\theta_{k2} & \cdots & \sum_{k\in\Psi_{M^*}^c} x_{nk}\theta_{kp} \end{pmatrix}.
$$

From (A8), we see that each element of $b'_{(M^*)}b_{(M^*)}/(n-k_{M^*}) = o_p(1)$. Combining this and (A6), we obtain (A3).

Next, we prove (A4). We first consider the ijth element of $b'_{(M^*)}(I_n - P_{(M^*)})\Xi/(n-k_{M^*})$. Let

$$
D = (I_n - P_{(M^*)})b_{(M^*)}e_ie_i'b'_{(M^*)}(I_n - P_{(M^*)}) = (d_{ij})_{n\times n},
$$

where d_{ij} is the ijth element of D. By (A7), we have

$$
\begin{aligned}
E\{\mathrm{tr}(D)\} &= E[\mathrm{tr}\{e_i'b'_{(M^*)}(I_n - P_{(M^*)})(I_n - P_{(M^*)})b_{(M^*)}e_i\}] \\
&\leq E\{\bar{\lambda}(I_n - P_{(M^*)})\bar{\lambda}(I_n - P_{(M^*)})e_i'b'_{(M^*)}b_{(M^*)}e_i\} \\
&\leq E\{e_i'b'_{(M^*)}b_{(M^*)}e_i\} \\
&= \sum_{l=1}^{n} E\left\{\left(\sum_{k\in\Psi_{M^*}^c} x_{lk}\theta_{ki}\right)^2\right\} \\
&= o(n),
\end{aligned}
$$

and so

$$
\begin{aligned}
E\left(\frac{e_i'b'_{(M^*)}(I_n - P_{(M^*)})\Xi e_j}{n-k_{M^*}}\right)^2 &= E\left(\frac{e_i'b'_{(M^*)}(I_n - P_{(M^*)})\Xi e_je_j'\Xi'(I_n - P_{(M^*)})b_{(M^*)}e_i}{(n-k_{M^*})^2}\right) \\
&= E\left\{\mathrm{tr}\left(\frac{(I_n - P_{(M^*)})b_{(M^*)}e_ie_i'b'_{(M^*)}(I_n - P_{(M^*)})\Xi e_je_j'\Xi'}{(n-k_{M^*})^2}\right)\right\} \\
&= E\left\{\mathrm{tr}\left(\frac{D\Xi e_je_j'\Xi'}{(n-k_{M^*})^2}\right)\right\} \\
&= E\left(\frac{\sum_{l=1}^{n}\sum_{i=1}^{n} d_{li}\epsilon_{ij}\epsilon_{lj}}{(n-k_{M^*})^2}\right) \\
&= \frac{E\{\sum_{i=1}^{n} d_{ii}(\epsilon_{ij}^2)\}}{(n-k_{M^*})^2} \\
&= \frac{E(\sum_{i=1}^{n} d_{ii})\sigma_{jj}}{(n-k_{M^*})^2} \\
&= \frac{E\{\mathrm{tr}(D)\}\sigma_{jj}}{(n-k_{M^*})^2}
\end{aligned}
$$

$$= \frac{o(n)}{(n-k_{M^*})^2}$$
$$= o\left(\frac{1}{n}\right),$$

then the ijth element of $b'_{(M^*)}(I_n - P_{(M^*)})\Xi/(n-k_{M^*})$ is $o_p(1/\sqrt{n})$, which implies (A4).

In the following, we prove (A5). From the model set-up in Section 2, we see that the ith row of Ξ is $\epsilon_{(i)}$, which satisfies $E(\epsilon_{(i)} \mid x_{(i)}) = 0$ and $E(\epsilon_{(i)}\epsilon'_{(i)} \mid x_{(i)}) = \Sigma = (\sigma_{ij})_{p\times p}$; also, $\{\epsilon_{(i)},\ i = 1, 2, \ldots, n\}$ are independent and identically distributed. Let

$$\hat{\sigma}_{ij} = \frac{\epsilon'_i(I_n - P_{(M^*)})\epsilon_j}{n-k_{M^*}},$$

then we have

$$\begin{aligned} E(\hat{\sigma}_{ij}) &= E\left(\frac{\epsilon'_i(I_n - P_{(M^*)})\epsilon_j}{n-k_{M^*}}\right) \\ &= \frac{\operatorname{tr}\{(I_n - P_{(M^*)})E(\epsilon_j\epsilon'_i)\}}{n-k_{M^*}} \\ &= \frac{\operatorname{tr}\{(I_n - P_{(M^*)})\sigma_{ij}I_n\}}{n-k_{M^*}} \\ &= \sigma_{ij}. \end{aligned} \tag{A9}$$

We now consider the sum of $\hat{\sigma}_{ij}$. Denote

$$A = I_n - P_{(M^*)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

then $\epsilon'_i(I_n - P_{(M^*)})\epsilon_j$ can be written as $\sum_{l=1}^n \sum_{k=1}^n a_{lk}\epsilon_{li}\epsilon_{kj}$. By the fact that A is an idempotent matrix, we have

$$\operatorname{tr}(A'A) = n - k_{M^*}. \tag{A10}$$

By the property of the trace of a matrix, we obtain

$$\operatorname{tr}^2(A) = \left(\sum_{l=1}^n a_{ll}\right)^2 = \sum_{l=1}^n a_{ll}^2 + \sum_{l\neq k} a_{ll}a_{kk} \geq \sum_{l=1}^n a_{ll}^2 \quad \text{and} \quad \sum_{l\neq k} a_{ll}a_{kk}, \tag{A11}$$

and

$$\operatorname{tr}(A'A) = \sum_{l=1}^n \sum_{k=1}^n a_{lk}^2 = \sum_{l=1}^n a_{ll}^2 + \sum_{l\neq k} a_{lk}^2 \geq \sum_{l=1}^n a_{ll}^2 \quad \text{and} \quad \sum_{l\neq k} a_{lk}^2. \tag{A12}$$

It is easily seen that

$$\begin{aligned} E[\{\epsilon'_i(I_n - P_{(M^*)})\epsilon_j\}^2] &= E\left\{\left(\sum_{l=1}^n \sum_{k=1}^n a_{lk}\epsilon_{li}\epsilon_{kj}\right)^2\right\} \\ &= E\left(\sum_{l=1}^n \sum_{k=1}^n a_{lk}^2\epsilon_{li}^2\epsilon_{kj}^2\right) + E\left(\sum_{l=1}^n \sum_{k=1}^n \sum_{d=1,d\neq k}^n a_{lk}a_{ld}\epsilon_{li}^2\epsilon_{kj}\epsilon_{dj}\right) \\ &\quad + E\left(\sum_{l=1}^n \sum_{k=1}^n \sum_{c=1,c\neq l}^n a_{lk}a_{ck}\epsilon_{li}\epsilon_{ci}\epsilon_{kj}^2\right) \\ &\quad + E\left(\sum_{l=1}^n \sum_{k=1}^n \sum_{c=1,c\neq l}^n \sum_{d=1,d\neq k}^n a_{lk}a_{cd}\epsilon_{li}\epsilon_{kj}\epsilon_{ci}\epsilon_{dj}\right). \end{aligned} \tag{A13}$$

We compute the four parts of the above equation. Denote $\kappa' = \kappa^{1/G}$. By Cauchy–Schwarz inequality and Condition (11), we obtain $E(\epsilon_{ij}^4) \leq \{E(\epsilon_{ij}^{4G})\}^{1/G} \leq \kappa^{1/G} = \kappa'$, for all $i = 1, \ldots, n$ and $j = 1, 2, \ldots, p$. Further, by (A10)–(A12),

we have

$$
\begin{aligned}
E\left(\sum_{l=1}^{n}\sum_{k=1}^{n} a_{lk}^2\epsilon_{li}^2\epsilon_{kj}^2\right) &= E\left(\sum_{l=1}^{n} a_{ll}^2\epsilon_{li}^2\epsilon_{lj}^2\right) + E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}^2\epsilon_{li}^2\epsilon_{kj}^2\right)\\
&= \sum_{l=1}^{n} a_{ll}^2 E(\epsilon_{li}^2\epsilon_{lj}^2) + \sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}^2 E(\epsilon_{li}^2)E(\epsilon_{kj}^2)\\
&\leq \sum_{l=1}^{n} a_{ll}^2\{E(\epsilon_{li}^4)\}^{1/2}\{E(\epsilon_{lj}^4)\}^{1/2} + \sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}^2\sigma_{ii}\sigma_{jj}\\
&\leq \kappa'\sum_{l=1}^{n} a_{ll}^2 + \sigma_{ii}\sigma_{jj}\sum_{k=1}^{n}\sum_{l=1,l\neq k}^{n} a_{lk}^2\\
&\leq \kappa'\mathrm{tr}(A'A) + \sigma_{ii}\sigma_{jj}\mathrm{tr}(A'A)\\
&= (\kappa' + \sigma_{ii}\sigma_{jj})(n - k_{M^*}),
\end{aligned}
$$

$$
\begin{aligned}
E\left(\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{d=1,d\neq k}^{n} a_{lk}a_{ld}\epsilon_{li}^2\epsilon_{kj}\epsilon_{dj}\right) &= E\left(\sum_{l=1}^{n}\sum_{d=1,d\neq l}^{n} a_{ll}a_{ld}\epsilon_{li}^2\epsilon_{lj}\epsilon_{dj}\right) + E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}a_{ll}\epsilon_{li}^2\epsilon_{kj}\epsilon_{lj}\right)\\
&\quad + E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{d=1,d\neq k\neq l}^{n} a_{lk}a_{ld}\epsilon_{li}^2\epsilon_{kj}\epsilon_{dj}\right)\\
&= \sum_{l=1}^{n}\sum_{d=1,d\neq l}^{n} a_{ll}a_{ld}E(\epsilon_{li}^2\epsilon_{lj})E(\epsilon_{dj}) + \sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}a_{ll}E(\epsilon_{li}^2\epsilon_{lj})E(\epsilon_{kj})\\
&\quad + \sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{d=1,d\neq k\neq l}^{n} a_{lk}a_{ld}E(\epsilon_{li}^2)E(\epsilon_{kj})E(\epsilon_{dj})\\
&= 0,
\end{aligned}
$$

$$
\begin{aligned}
E\left(\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{c=1,c\neq l}^{n} a_{lk}a_{ck}\epsilon_{li}\epsilon_{ci}\epsilon_{kj}^2\right) &= E\left(\sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n} a_{ll}a_{cl}\epsilon_{li}\epsilon_{ci}\epsilon_{lj}^2\right) + E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}a_{kk}\epsilon_{li}\epsilon_{ki}\epsilon_{kj}^2\right)\\
&\quad + E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{c=1,c\neq k\neq l}^{n} a_{lk}a_{ck}\epsilon_{li}\epsilon_{ci}\epsilon_{kj}^2\right)\\
&= \sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n} a_{ll}a_{cl}E(\epsilon_{li}\epsilon_{lj}^2)E(\epsilon_{ci}) + \sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}a_{kk}E(\epsilon_{li})E(\epsilon_{ki}\epsilon_{kj}^2)\\
&\quad + \sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{c=1,c\neq k\neq l}^{n} a_{lk}a_{ck}E(\epsilon_{li})E(\epsilon_{ci})E(\epsilon_{kj}^2)\\
&= 0,
\end{aligned}
$$

and

$$
\begin{aligned}
&E\left(\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{c=1,c\neq l}^{n}\sum_{d=1,d\neq k}^{n} a_{lk}a_{cd}\epsilon_{li}\epsilon_{kj}\epsilon_{ci}\epsilon_{dj}\right)\\
&\quad = E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}a_{kl}\epsilon_{li}\epsilon_{kj}\epsilon_{ki}\epsilon_{lj}\right) + E\left(\sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n} a_{ll}a_{cc}\epsilon_{li}\epsilon_{lj}\epsilon_{ci}\epsilon_{cj}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+E\left(\sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n}\sum_{d=1,d\neq c\neq l}^{n} a_{ll}a_{cd}\epsilon_{li}\epsilon_{lj}\epsilon_{ci}\epsilon_{dj}\right)+E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{c=1,c\neq k\neq l}^{n} a_{lk}a_{cl}\epsilon_{li}\epsilon_{kj}\epsilon_{ci}\epsilon_{lj}\right)\\
&+E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{d=1,d\neq k\neq l}^{n} a_{lk}a_{kd}\epsilon_{li}\epsilon_{kj}\epsilon_{ki}\epsilon_{dj}\right)+E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{d=1,d\neq k\neq l}^{n} a_{lk}a_{dd}\epsilon_{li}\epsilon_{kj}\epsilon_{di}\epsilon_{dj}\right)\\
&+E\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{c=1,c\neq k\neq l}^{n}\sum_{d=1,d\neq c\neq k\neq l}^{n} a_{lk}a_{cd}\epsilon_{li}\epsilon_{kj}\epsilon_{ci}\epsilon_{dj}\right)\\
=&\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}a_{kl}E(\epsilon_{li}\epsilon_{lj})E(\epsilon_{kj}\epsilon_{ki})\\
&+\sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n} a_{ll}a_{cc}E(\epsilon_{li}\epsilon_{lj})E(\epsilon_{ci}\epsilon_{cj})\\
&+\sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n}\sum_{d=1,d\neq c\neq l}^{n} a_{ll}a_{cd}E(\epsilon_{li}\epsilon_{lj})E(\epsilon_{ci})E(\epsilon_{dj})\\
&+\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{c=1,c\neq k\neq l}^{n} a_{lk}a_{cl}E(\epsilon_{li}\epsilon_{lj})E(\epsilon_{kj})E(\epsilon_{ci})\\
&+\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{d=1,d\neq k\neq l}^{n} a_{lk}a_{kd}E(\epsilon_{li})E(\epsilon_{kj}\epsilon_{ki})E(\epsilon_{dj})\\
&+\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{d=1,d\neq k\neq l}^{n} a_{lk}a_{dd}E(\epsilon_{li})E(\epsilon_{kj})E(\epsilon_{di}\epsilon_{dj})\\
&+\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\sum_{c=1,c\neq k\neq l}^{n}\sum_{d=1,d\neq c\neq k\neq l}^{n} a_{lk}a_{cd}E(\epsilon_{li})E(\epsilon_{kj})E(\epsilon_{ci})E(\epsilon_{dj})\\
=&\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}^2\sigma_{ij}^2+\sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n} a_{ll}a_{cc}\sigma_{ij}^2\\
=&\,\sigma_{ij}^2\left(\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n} a_{lk}^2+\sum_{l=1}^{n}\sum_{c=1,c\neq l}^{n} a_{ll}a_{cc}\right)\\
\leq&\,\sigma_{ij}^2\{\mathrm{tr}(A'A)+\mathrm{tr}^2(A)\}\\
=&\,\sigma_{ij}^2\{(n-k_{M^*})+(n-k_{M^*})^2\}.
\end{aligned}
$$

Then (A13) can be written as

$$
\begin{aligned}
E[\{\epsilon_i'(I_n-P_{(M^*)})\epsilon_j\}^2] &\leq (\kappa'+\sigma_{ii}\sigma_{jj})(n-k_{M^*})+\sigma_{ij}^2\{(n-k_{M^*})+(n-k_{M^*})^2\}\\
&=(\kappa'+\sigma_{ii}\sigma_{jj}+\sigma_{ij}^2)(n-k_{M^*})+\sigma_{ij}^2(n-k_{M^*})^2. \qquad \text{(A14)}
\end{aligned}
$$

By (A9), (A14) and Condition (16), we obtain that the variance of $\hat{\sigma}_{ij}$ is

$$
\begin{aligned}
\mathrm{Var}(\hat{\sigma}_{ij}) &= \mathrm{Var}\left(\frac{\epsilon_i'(I_n-P_{(M^*)})\epsilon_j}{n-k_{M^*}}\right)\\
&= E\left\{\left(\frac{\epsilon_i'(I_n-P_{(M^*)})\epsilon_j}{n-k_{M^*}}\right)^2\right\}-\left\{E\left(\frac{\epsilon_i'(I_n-P_{(M^*)})\epsilon_j}{n-k_{M^*}}\right)\right\}^2\\
&\leq \frac{(\kappa'+\sigma_{ii}\sigma_{jj}+\sigma_{ij}^2)(n-k_{M^*})+\sigma_{ij}^2(n-k_{M^*})^2}{(n-k_{M^*})^2}-\sigma_{ij}^2
\end{aligned}
$$

$$= \frac{\kappa' + \sigma_{ii}\sigma_{jj} + \sigma_{ij}^2}{n - k_{M^*}}$$
$$= O\left(\frac{1}{n}\right).$$

That is to say

$$\hat{\sigma}_{ij} - \sigma_{ij} = O_p\left(\frac{1}{\sqrt{n}}\right), \quad \text{for all } i, j = 1, 2, \ldots, p,$$

so

$$\frac{\Xi'(I_n - P_{(M^*)})\Xi}{n - k_{M^*}} - \Sigma = o_p(1),$$

which implies (A5). ■

Lemma A.3: *Assume that* $\Sigma > 0$ *and Conditions* (11), (16) *and* (17) *hold, then*

$$\bar{\lambda}(\hat{\Sigma}^{-1}) = O_p(1). \tag{A15}$$

Proof: Since Σ is positive definite and p is fixed, Σ^{-1} is also positive definite and $\bar{\lambda}(\Sigma^{-1})$ is finite. From Lemma A.2, we see that $\hat{\Sigma} - \Sigma = o_p(1)$, which implies $\hat{\Sigma}^{-1} - \Sigma^{-1} = o_p(1)$, therefore,

$$\bar{\lambda}(\hat{\Sigma}^{-1} - \Sigma^{-1}) = o_p(1), \tag{A16}$$

then we have

$$\bar{\lambda}(\hat{\Sigma}^{-1}) \le \bar{\lambda}(\hat{\Sigma}^{-1} - \Sigma^{-1}) + \bar{\lambda}(\Sigma^{-1}) = o_p(1) + O(1) = O_p(1).$$

■

A.2 Proof of Theorem 4.1

The Mahalanobis Mallows criterion for model averaging is

$$\begin{aligned} C_n(w) &= \{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'\Sigma_0^{-1}\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\} + 2p\mathrm{tr}(P(w)) \\ &= L_n(w) + 2\,\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I - H(w))\mathrm{Vec}(\mu) \\ &\quad - 2\mathrm{Vec}(\Xi)'\Sigma_0^{-1}H(w)\mathrm{Vec}(\Xi) \\ &\quad + \mathrm{Vec}(\Xi)'\Sigma_0^{-1}\mathrm{Vec}(\Xi) + 2p\mathrm{tr}(P(w)), \end{aligned}$$

where $\mathrm{Vec}(\Xi)'\Sigma_0^{-1}\mathrm{Vec}(\Xi)$ is independent of w, so we need only to verify that

$$\sup_{w \in \mathscr{H}_n} \left| \frac{\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I - H(w))\mathrm{Vec}(\mu)}{R_n(w)} \right| \xrightarrow{p} 0, \tag{A17}$$

$$\sup_{w \in \mathscr{H}_n} \left| \frac{\mathrm{Vec}(\Xi)'\Sigma_0^{-1}H(w)\mathrm{Vec}(\Xi) - ptr(P(w))}{R_n(w)} \right| \xrightarrow{p} 0, \tag{A18}$$

and

$$\sup_{w \in \mathscr{H}_n} \left| \frac{L_n(w)}{R_n(w)} - 1 \right| \xrightarrow{p} 0. \tag{A19}$$

We observe, for any $\delta > 0$, that

$$\begin{aligned} &P\left(\sup_{w \in \mathscr{H}_n} \left| \frac{\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I - H(w))\mathrm{Vec}(\mu)}{R_n(w)} \right| > \delta \right) \\ &= P\left(\max_{1 \le m \le M} |\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I - H(w))\mathrm{Vec}(\mu)| > \delta\xi_n\right) \\ &= P(\{|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I - H(w_1^0))\mathrm{Vec}(\mu)| > \delta\xi_n\} \cup \{|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I - H(w_2^0))\mathrm{Vec}(\mu)| > \delta\xi_n\} \\ &\cup \cdots \cup \{|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I - H(w_M^0))\mathrm{Vec}(\mu)| > \delta\xi_n\}) \end{aligned}$$

$$
\begin{aligned}
&\le \sum_{m=1}^{M} P(\{|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I-H(w_m^0))\mathrm{Vec}(\mu)| > \delta\xi_n\}) \\
&\le \sum_{m=1}^{M} \frac{E|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}(I-H(w_m^0))\mathrm{Vec}(\mu)|^{2G}}{\delta^{2G}\xi_n^{2G}} \\
&\le \sum_{m=1}^{M} \frac{c_1\|\Sigma_0^{-1/2}(I-H(w_m^0))\mathrm{Vec}(\mu)\|^{2G}}{\delta^{2G}\xi_n^{2G}} \\
&\le \sum_{m=1}^{M} \frac{c_1\{R_n(w_m^0)\}^G}{\delta^{2G}\xi_n^{2G}} \\
&= \frac{c_1}{\delta^{2G}M}\cdot\frac{M\sum_{m=1}^{M}\{R_n(w_m^0)\}^G}{\xi_n^{2G}} \to 0,
\end{aligned}
$$

where c_1 is a positive constant and $c_2 - c_{13}$ in the following proofs denote different positive numbers, the first inequality follows from the Bonferroni's inequality, the second inequality is obtained by an extension of Chebyshev's inequality, the third inequality is ensured by Theorem 2 of Whittle [25], and the fourth inequality follows from Equation (7). Along with Condition (12), we see that (A17) is true.

Similar to the proving process of (A17), we have

$$
\begin{aligned}
&P\left(\sup_{w\in\mathscr{H}_n}\left|\frac{\mathrm{Vec}(\Xi)'\Sigma_0^{-1}H(w)\mathrm{Vec}(\Xi)-p\mathrm{tr}(P(w))}{R_n(w)}\right| > \delta\right) \\
&\le P\left(\sup_{w\in\mathscr{H}_n}\sum_{m=1}^{M} w_m|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}H(w)\mathrm{Vec}(\Xi)-p\mathrm{tr}(P(w))| > \delta\xi_n\right) \\
&= P\left(\max_{1\le m\le M}|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}H_{(m)}\mathrm{Vec}(\Xi)-p\mathrm{tr}(P_{(m)})| > \delta\xi_n\right) \\
&\le \sum_{m=1}^{M} P(|\mathrm{Vec}(\Xi)'\Sigma_0^{-1}H(w_m^0)\mathrm{Vec}(\Xi)-p\mathrm{tr}(P(w_m^0))| > \delta\xi_n) \\
&\le \sum_{m=1}^{M} \frac{E|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\Sigma_0^{-1/2}H(w_m^0)\Sigma_0^{1/2}\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}-p\mathrm{tr}(P(w_m^0))|^{2G}}{(\delta\xi_n)^{2G}} \\
&\le \sum_{m=1}^{M} \frac{c_2\mathrm{tr}^G\{\Sigma_0^{-1/2}H(w_m^0)\Sigma_0H'(w_m^0)\Sigma_0^{-1/2}\}}{(\delta\xi_n)^{2G}} \\
&\le \sum_{m=1}^{M} \frac{c_2[\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})(p\mathrm{tr}\{P(w_m^0)P'(w_m^0)\})]^G}{(\delta\xi_n)^{2G}} \\
&\le \sum_{m=1}^{M} \frac{c_2\{\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})R_n(w_m^0)\}^G}{(\delta\xi_n)^{2G}} \\
&= \frac{c_2\{\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})\}^G}{\delta^{2G}M}\cdot\frac{M\sum_{m=1}^{M}\{R_n(w_m^0)\}^G}{\xi_n^{2G}} \to 0,
\end{aligned}
$$

which, together with Condition (12), implies (A18).

Note that

$$
\begin{aligned}
&\sup_{w\in\mathscr{H}_n}\left|\frac{L_n(w)}{R_n(w)}-1\right| \\
&= \sup_{w\in\mathscr{H}_n}\left|\frac{-2\,\mathrm{Vec}(\Xi)'H'(w)\Sigma_0^{-1}(I-H(w))\mathrm{Vec}(\mu)+\mathrm{Vec}(\Xi)'H'(w)\Sigma_0^{-1}H(w)\mathrm{Vec}(\Xi)-p\mathrm{tr}\{P(w)P'(w)\}}{R_n(w)}\right|.
\end{aligned}
$$

In order to prove (A19), it is sufficient to verify that

$$\sup_{w\in\mathscr{H}_n} \frac{|\mathrm{Vec}(\Xi)'H'(w)\Sigma_0^{-1}(I-H(w))\mathrm{Vec}(\mu)|}{R_n(w)} \xrightarrow{p} 0, \tag{A20}$$

and

$$\sup_{w\in\mathscr{H}_n} \frac{|\mathrm{Vec}(\Xi)'H'(w)\Sigma_0^{-1}H(w)\mathrm{Vec}(\Xi)-p\mathrm{tr}\{P(w)P'(w)\}|}{R_n(w)} \xrightarrow{p} 0. \tag{A21}$$

Recognizing that for any weight vectors w and $w_*\in H_n$,

$$\mathrm{tr}\{P^2(w_*)P^2(w)\}\le \bar{\lambda}^2\{P(w_*)\}\mathrm{tr}\{P^2(w)\}\le \mathrm{tr}\{P^2(w)\},$$

we obtain

$$\begin{aligned}
&P\left(\sup_{w\in\mathscr{H}_n} \frac{|\mathrm{Vec}(\Xi)'H'(w)\Sigma_0^{-1}(I-H(w))\mathrm{Vec}(\mu)|}{R_n(w)} > \delta\right)\\
&\le P\left(\sup_{w\in\mathscr{H}_n}\sum_{t=1}^{M}\sum_{m=1}^{M} w_t w_m|\mathrm{Vec}(\Xi)'H'_{(t)}\Sigma_0^{-1}(I-H_{(m)})\mathrm{Vec}(\mu)| > \delta\xi_n\right)\\
&\le P\left(\max_{1\le t\le M}\max_{1\le m\le M}|\mathrm{Vec}(\Xi)'H'_{(t)}\Sigma_0^{-1}(I-H_{(m)})\mathrm{Vec}(\mu)| > \delta\xi_n\right)\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} P(|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\Sigma_0^{1/2}H'(w_t^0)\Sigma_0^{-1}(I-H(w_m^0))\mathrm{Vec}(\mu)| > \delta\xi_n)\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{E|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\Sigma_0^{1/2}H'(w_t^0)\Sigma_0^{-1}(I-H(w_m^0))\mathrm{Vec}(\mu)|^{2G}}{(\delta\xi_n)^{2G}}\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{c_3\|\Sigma_0^{1/2}H'(w_t^0)\Sigma_0^{-1}(I-H(w_m^0))\mathrm{Vec}(\mu)\|^{2G}}{(\delta\xi_n)^{2G}}\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{c_3[\bar{\lambda}(\Sigma_0)\bar{\lambda}(\Sigma_0^{-1})\bar{\lambda}\{H'(w_t^0)H(w_t^0)\}]^G\|\Sigma_0^{-1/2}(I-H(w_m^0))\mathrm{Vec}(\mu)\|^{2G}}{(\delta\xi_n)^{2G}}\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{c_3\{\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})R_n(w_m^0)\}^G}{(\delta\xi_n)^{2G}}\\
&= \frac{c_3\{\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})\}^G}{\delta^{2G}}\cdot\frac{M\sum_{m=1}^{M}\{R_n(w_m^0)\}^G}{\xi_n^{2G}} \to 0,
\end{aligned}$$

which implies (A20). Furthermore, we have

$$\begin{aligned}
&P\left(\sup_{w\in\mathscr{H}_n} \frac{|\mathrm{Vec}(\Xi)'H'(w)\Sigma_0^{-1}H(w)\mathrm{Vec}(\Xi)-p\mathrm{tr}\{P(w)P'(w)\}|}{R_n(w)} > \delta\right)\\
&\le P\left(\sup_{w\in\mathscr{H}_n}\sum_{t=1}^{M}\sum_{m=1}^{M} w_t w_m|\mathrm{Vec}(\Xi)'H'_{(t)}\Sigma_0^{-1}H_{(m)}\mathrm{Vec}(\Xi)-p\mathrm{tr}\{P_{(t)}P'_{(m)}\}| > \delta\xi_n\right)\\
&\le P\left(\max_{1\le t\le M}\max_{1\le m\le M}|\mathrm{Vec}(\Xi)'H'_{(t)}\Sigma_0^{-1}H_{(m)}\mathrm{Vec}(\Xi)-p\mathrm{tr}\{P_{(t)}P'_{(m)}\}| > \delta\xi_n\right)\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} P(|\mathrm{Vec}(\Xi)'H'(w_t^0)\Sigma_0^{-1}H(w_m^0)\mathrm{Vec}(\Xi)-p\mathrm{tr}\{P(w_t^0)P'(w_m^0)\}| > \delta\xi_n)\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{E|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\Sigma_0^{1/2}H'(w_t^0)\Sigma_0^{-1}H(w_m^0)\Sigma_0^{1/2}\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}-p\mathrm{tr}\{P(w_t^0)P'(w_m^0)\}|^{2G}}{(\delta\xi_n)^{2G}}\\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{c_4\mathrm{tr}^G\{\Sigma_0^{1/2}H'(w_t^0)\Sigma_0^{-1}H(w_m^0)\Sigma_0H'(w_m^0)\Sigma_0^{-1}H(w_t^0)\Sigma_0^{1/2}\}}{(\delta\xi_n)^{2G}}
\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{c_4[\bar{\lambda}^2(\Sigma_0)\bar{\lambda}^2(\Sigma_0^{-1})\bar{\lambda}\{H(w_m^0)H'(w_m^0)\}\mathrm{tr}\{H'(w_t^0)H(w_t^0)\}]^G}{(\delta\xi_n)^{2G}} \\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{c_4[\{\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})\}^2 p\mathrm{tr}\{P(w_m^0)P'(w_m^0)\}]^G}{(\delta\xi_n)^{2G}} \\
&\le \sum_{t=1}^{M}\sum_{m=1}^{M} \frac{c_4\{\{\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})\}^2 R_n(w_m^0)\}^G}{(\delta\xi_n)^{2G}} \\
&= \frac{c_4\{\bar{\lambda}(\Sigma)\bar{\lambda}(\Sigma^{-1})\}^{2G}}{\delta^{2G}} \cdot \frac{M\sum_{m=1}^{M}\{R_n(w_m^0)\}^G}{\xi_n^{2G}} \to 0,
\end{aligned}
$$

which implies (A21).

A.3 Proof of Theorem 4.2

When Σ is replaced by $\hat{\Sigma}$, the Mahalanobis Mallows criterion can be written as

$$
\begin{aligned}
\hat{C}_n(w) &= \{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'\hat{\Sigma}_0^{-1}\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\} + 2p\mathrm{tr}(P(w)) \\
&= C_n(w) + \{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'(\hat{\Sigma}_0^{-1} - \Sigma_0^{-1})\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\},
\end{aligned}
$$

where $\hat{\Sigma}_0^{-1} = \hat{\Sigma}^{-1} \otimes I_n$. Hence, from the result of Theorem 4.1, it suffices to prove that, as $n \to \infty$,

$$
\begin{aligned}
&\sup_{w\in\mathscr{H}_n} |\hat{C}_n(w) - C_n(w)|/R_n(w) \\
&\quad = \sup_{w\in\mathscr{H}_n} |\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'(\hat{\Sigma}_0^{-1} - \Sigma_0^{-1})\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}|/R_n(w) \\
&\quad \le \bar{\lambda}(\hat{\Sigma}_0^{-1} - \Sigma_0^{-1}) \sup_{w\in\mathscr{H}_n} |\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}|/R_n(w) \\
&\quad \xrightarrow{p} 0,
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{\lambda}(\hat{\Sigma}_0^{-1} - \Sigma_0^{-1}) &= \bar{\lambda}\{(\hat{\Sigma}^{-1} - \Sigma^{-1}) \otimes I_n\} \\
&= \bar{\lambda}(\hat{\Sigma}^{-1} - \Sigma^{-1}) \\
&= \bar{\lambda}\{\Sigma^{-1}(\Sigma - \hat{\Sigma})\hat{\Sigma}^{-1}\} \\
&\le \bar{\lambda}(\Sigma^{-1})\bar{\lambda}(\Sigma - \hat{\Sigma})\bar{\lambda}(\hat{\Sigma}^{-1})
\end{aligned}
$$

and

$$
\begin{aligned}
&\sup_{w\in\mathscr{H}_n} |\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}|/R_n(w) \\
&\quad \le \sup_{w\in\mathscr{H}_n} |\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}'\{\mathrm{Vec}(Y) - \mathrm{Vec}(\hat{\mu}(w))\}|/\xi_n \\
&\quad = \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\mu)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\mu)|/\xi_n \\
&\qquad + \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\Xi) - \mathrm{tr}\{(I_n - H(w))'(I_n - H(w))\Sigma_0\}|/\xi_n \\
&\qquad + 2\sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\mu)|/\xi_n \\
&\qquad + \sup_{w\in\mathscr{H}_n} |\mathrm{tr}\{(I_n - H(w))'(I_n - H(w))\Sigma_0\}|/\xi_n \\
&\quad = \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\mu)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\mu)|/\xi_n \\
&\qquad + \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'\mathrm{Vec}(\Xi) - \mathrm{tr}\{\Sigma_0\}|/\xi_n
\end{aligned}
$$

$$
\begin{aligned}
&- 2 \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'H(w)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H(w)\Sigma_0\}|/\xi_n \\
&+ \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'H'(w)H(w)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H'(w)H(w)\Sigma_0\}|/\xi_n \\
&+ 2 \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\mu)|/\xi_n \\
&+ \sup_{w\in\mathscr{H}_n} |\mathrm{tr}\{(I_n - H(w))'(I_n - H(w))\Sigma_0\}|/\xi_n.
\end{aligned}
$$

Since $\bar{\lambda}(\Sigma^{-1})$ is finite and $\bar{\lambda}(\hat{\Sigma}^{-1}) = O_p(1)$, we need to verify that

$$\bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\mu)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\mu)|/\xi_n \xrightarrow{p} 0, \tag{A22}$$

$$\bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'\mathrm{Vec}(\Xi) - \mathrm{tr}(\Sigma_0)|/\xi_n \xrightarrow{p} 0, \tag{A23}$$

$$\bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'H(w)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H(w)\Sigma_0\}|/\xi_n \xrightarrow{p} 0, \tag{A24}$$

$$\bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'H'(w)H(w)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H'(w)H(w)\Sigma_0\}|/\xi_n \xrightarrow{p} 0, \tag{A25}$$

$$\bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\mu)|/\xi_n \xrightarrow{p} 0, \tag{A26}$$

and

$$\bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} |\mathrm{tr}\{(I_n - H(w))'(I_n - H(w))\Sigma_0\}|/\xi_n \xrightarrow{p} 0. \tag{A27}$$

By Conditions (15) and (18), we have

$$
\begin{aligned}
&\bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\mu)'(I_n - H(w))'(I_n - H(w))\mathrm{Vec}(\mu)|/\xi_n \\
&\le \bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} \sum_{m=1}^{M}\sum_{t=1}^{M} w_m w_t |\mathrm{Vec}(\mu)'(I_n - H_{(m)})'(I_n - H_{(t)})\mathrm{Vec}(\mu)|/\xi_n \\
&\le \bar{\lambda}(\Sigma - \hat{\Sigma}) \sup_{w\in\mathscr{H}_n} \max_{1\le m\le M} \max_{1\le t\le M} |\mathrm{Vec}(\mu)'(I_n - H(w_m^0))'(I_n - H(w_t^0))\mathrm{Vec}(\mu)|/\xi_n \\
&\le \bar{\lambda}(\Sigma - \hat{\Sigma}) \max_{1\le m\le M} \max_{1\le t\le M} \bar{\lambda}(I_n - H(w_m^0))\bar{\lambda}(I_n - H(w_t^0))|\mathrm{Vec}(\mu)'\mathrm{Vec}(\mu)|/\xi_n \\
&= \bar{\lambda}(\Sigma - \hat{\Sigma})O(\xi_n^{-1}n) \xrightarrow{p} 0,
\end{aligned}
$$

which implies (A22). Denote (A23) as

$$\bar{\lambda}(\Sigma - \hat{\Sigma})\xi_n^{-1}n \sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'\mathrm{Vec}(\Xi) - \mathrm{tr}(\Sigma_0)|/n = \bar{\lambda}(\Sigma - \hat{\Sigma})\xi_n^{-1}n \cdot |\mathrm{Vec}(\Xi)'\mathrm{Vec}(\Xi) - \mathrm{tr}(\Sigma_0)|/n. \tag{A28}$$

For any $\delta > 0$,

$$
\begin{aligned}
P(|\mathrm{Vec}(\Xi)'\mathrm{Vec}(\Xi) - \mathrm{tr}(\Sigma_0)|/n > \delta) &= P(|\mathrm{Vec}(\Xi)'\mathrm{Vec}(\Xi) - \mathrm{tr}(\Sigma_0)| > \delta n) \\
&\le \frac{E|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\Sigma_0\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\} - \mathrm{tr}(\Sigma_0)|^2}{\delta^2 n^2} \\
&\le \frac{c_5\mathrm{tr}(\Sigma_0^2)}{\delta^2 n^2} \\
&\le \frac{c_5\bar{\lambda}^2(\Sigma_0)\mathrm{tr}(I_{np})}{\delta^2 n^2} \\
&\le \frac{c_6 p}{\delta^2 n} \to 0,
\end{aligned}
$$

which, together with Condition (15) and (A28), implies (A23). For proving (A24), by $\bar{\lambda}(\hat{\Sigma} - \Sigma) = o_p(1)$, one has

$$
\begin{aligned}
&P\left(\sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'H'(w)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H(w)\Sigma_0\}|/\xi_n > \delta\right)\\
&\le P\left(\sup_{w\in\mathscr{H}_n} \sum_{m=1}^{M} w_m|\mathrm{Vec}(\Xi)'H'_{(m)}\mathrm{Vec}(\Xi) - \mathrm{tr}\{H_{(m)}\Sigma_0\}| > \delta\xi_n\right)\\
&= P\left(\max_{1\le m\le M} |\mathrm{Vec}(\Xi)'H(w_m^0)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H(w_m^0)\Sigma_0\}| > \delta\xi_n\right)\\
&\le \sum_{m=1}^{M} P(|\mathrm{Vec}(\Xi)'H(w_m^0)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H(w_m^0)\Sigma_0\}| > \delta\xi_n)\\
&\le \sum_{m=1}^{M} \frac{E|\mathrm{Vec}(\Xi)'H(w_m^0)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H(w_m^0)\Sigma_0\}|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\le \sum_{m=1}^{M} \frac{E|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\{\Sigma_0^{1/2}H(w_m^0)\Sigma_0^{1/2}\}\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\} - \mathrm{tr}\{H(w_m^0)\Sigma_0\}|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\le \sum_{m=1}^{M} \frac{c_7\mathrm{tr}^G\{\Sigma_0^{1/2}H(w_m^0)\Sigma_0H'(w_m^0)\Sigma_0^{1/2}\}}{\delta^{2G}\xi_n^{2G}}\\
&\le \sum_{m=1}^{M} \frac{c_7\{\bar{\lambda}^2(\Sigma_0)\}^G\mathrm{tr}^G\{H'(w_m^0)\Sigma_0^{-1}H(w_m^0)\Sigma_0\}}{\delta^{2G}\xi_n^{2G}}\\
&\le \sum_{m=1}^{M} \frac{c_8\{R_n(w_m^0)\}^G}{\delta^{2G}\xi_n^{2G}} \to 0.
\end{aligned}
$$

Now, we prove (A25). For any $\delta > 0$,

$$
\begin{aligned}
&P\left(\sup_{w\in\mathscr{H}_n} |\mathrm{Vec}(\Xi)'H'(w)H(w)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H'(w)H(w)\Sigma_0\}|/\xi_n > \delta\right)\\
&\le P\left(\sup_{w\in\mathscr{H}_n} \sum_{m=1}^{M}\sum_{t=1}^{M} w_mw_t|\mathrm{Vec}(\Xi)'H'_{(t)}H_{(m)}\mathrm{Vec}(\Xi) - \mathrm{tr}\{H'_{(t)}H_{(m)}\Sigma_0\}| > \delta\xi_n\right)\\
&\le P\left(\max_{1\le m\le M}\max_{1\le t\le M} |\mathrm{Vec}(\Xi)'H'(w_t^0)H(w_m^0)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H'(w_t^0)H(w_m^0)\Sigma_0\}| > \delta\xi_n\right)\\
&\le \sum_{m=1}^{M}\sum_{t=1}^{M} P(|\mathrm{Vec}(\Xi)'H'(w_t^0)H(w_m^0)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H'(w_t^0)H(w_m^0)\Sigma_0\}| > \delta\xi_n)\\
&\le \sum_{m=1}^{M}\sum_{t=1}^{M} \frac{E|\mathrm{Vec}(\Xi)'H'(w_t^0)H(w_m^0)\mathrm{Vec}(\Xi) - \mathrm{tr}\{H'(w_t^0)H(w_m^0)\Sigma_0\}|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\le \sum_{m=1}^{M}\sum_{t=1}^{M} \frac{E|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\Sigma_0^{1/2}H'(w_t^0)H(w_m^0)\Sigma_0^{1/2}\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\} - \mathrm{tr}\{H'(w_t^0)H(w_m^0)\Sigma_0\}|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\le \sum_{m=1}^{M}\sum_{t=1}^{M} \frac{c_9\mathrm{tr}^G\{\Sigma_0^{1/2}H'(w_t^0)H(w_m^0)\Sigma_0H'(w_m^0)H(w_t^0)\Sigma_0^{1/2}\}}{\delta^{2G}\xi_n^{2G}}\\
&\le \sum_{m=1}^{M}\sum_{t=1}^{M} \frac{c_9\{\bar{\lambda}^2(\Sigma_0)\bar{\lambda}^2(H(w_m^0))\}^G\mathrm{tr}^G\{H'(w_t^0)\Sigma_0^{-1}H(w_t^0)\Sigma_0\}}{\delta^{2G}\xi_n^{2G}}
\end{aligned}
$$

$$\leq \sum_{m=1}^{M}\sum_{t=1}^{M}\frac{c_{10}\{R_n(w_t^0)\}^G}{\delta^{2G}\xi_n^{2G}}$$
$$= \frac{c_{10}M\sum_{t=1}^{M}\{R_n(w_t^0)\}^G}{\delta^{2G}\xi_n^{2G}} \to 0,$$

then we obtains (A25) by $\bar{\lambda}(\hat{\Sigma}-\Sigma)=o_p(1)$. Similarly,

$$\begin{aligned}
&P\left(\sup_{w\in\mathscr{H}_n}|\mathrm{Vec}(\Xi)'(I_n-H(w))'(I_n-H(w))\mathrm{Vec}(\mu)|/\xi_n>\delta\right)\\
&\leq P\left(\sup_{w\in\mathscr{H}_n}\sum_{m=1}^{M}\sum_{t=1}^{M}w_mw_t|\mathrm{Vec}(\Xi)'(I_n-H_{(t)})'(I_n-H_{(m)})\mathrm{Vec}(\mu)|>\delta\xi_n\right)\\
&\leq P\left(\max_{1\leq m\leq M}\max_{1\leq t\leq M}\left|\mathrm{Vec}(\Xi)'(I_n-H(w_t^0))'(I_n-H(w_m^0))\mathrm{Vec}(\mu)\right|>\delta\xi_n\right)\\
&\leq \sum_{m=1}^{M}\sum_{t=1}^{M}P(|\mathrm{Vec}(\Xi)'(I_n-H(w_t^0))'(I_n-H(w_m^0))\mathrm{Vec}(\mu)|>\delta\xi_n)\\
&\leq \sum_{m=1}^{M}\sum_{t=1}^{M}\frac{E|\mathrm{Vec}(\Xi)'(I_n-H(w_t^0))'(I_n-H(w_m^0))\mathrm{Vec}(\mu)|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\leq \sum_{m=1}^{M}\sum_{t=1}^{M}\frac{E|\{\Sigma_0^{-1/2}\mathrm{Vec}(\Xi)\}'\Sigma_0^{1/2}(I-H(w_t^0))'(I_n-H(w_m^0))\mathrm{Vec}(\mu)|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\leq \sum_{m=1}^{M}\sum_{t=1}^{M}\frac{c_{11}\|\Sigma_0^{1/2}(I_n-H(w_t^0))'(I_n-H(w_m^0))\mathrm{Vec}(\mu)\|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\leq \sum_{m=1}^{M}\sum_{t=1}^{M}\frac{c_{11}\{\bar{\lambda}^2(\Sigma_0)\bar{\lambda}^2(I_n-H(w_t^0))\}^G\|\Sigma_0^{-1/2}(I_n-H(w_m^0))\mathrm{Vec}(\mu)\|^{2G}}{\delta^{2G}\xi_n^{2G}}\\
&\leq \sum_{m=1}^{M}\sum_{t=1}^{M}\frac{c_{12}\{R_n(w_m^0)\}^G}{\delta^{2G}\xi_n^{2G}}\\
&= \frac{c_{12}M\sum_{m=1}^{M}\{R_n(w_m^0)\}^G}{\delta^{2G}\xi_n^{2G}}\to 0,
\end{aligned}$$

which, together with $\bar{\lambda}(\hat{\Sigma}-\Sigma)=o_p(1)$, implies (A26). Finally, we prove (A27).

$$\begin{aligned}
&\bar{\lambda}(\Sigma-\hat{\Sigma})\sup_{w\in\mathscr{H}_n}|\mathrm{tr}\{(I_n-H(w))'(I_n-H(w))\Sigma_0\}|/\xi_n\\
&\leq \bar{\lambda}(\hat{\Sigma}-\Sigma)\sup_{w\in\mathscr{H}_n}\sum_{m=1}^{M}\sum_{t=1}^{M}w_mw_t|\mathrm{tr}\{(I_n-H(w_t^0))'(I_n-H(w_m^0))\Sigma_0\}|/\xi_n\\
&\leq \bar{\lambda}(\hat{\Sigma}-\Sigma)\sup_{w\in\mathscr{H}_n}\max_{1\leq m\leq M}\max_{1\leq t\leq M}|\mathrm{tr}\{(I_n-H(w_t^0))'(I_n-H(w_m^0))\Sigma_0\}|/\xi_n\\
&\leq \bar{\lambda}(\hat{\Sigma}-\Sigma)\sup_{w\in\mathscr{H}_n}\max_{1\leq m\leq M}\max_{1\leq t\leq M}\bar{\lambda}(I_n-H(w_t^0))\bar{\lambda}(I_n-H(w_m^0))\bar{\lambda}(\Sigma_0)\mathrm{tr}(I_{np})/\xi_n\\
&\leq \bar{\lambda}(\hat{\Sigma}-\Sigma)c_{13}np/\xi_n,
\end{aligned}$$

which implies (A27) by Condition (15). This completes the proof of Theorem 4.2.