Model averaging estimation for high-dimensional covariance matrices with a network structure

RONG ZHU[†], XINYU ZHANG[‡], YANYUAN MA[§] AND GUOHUA ZOU^{||}

[†]*School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK.*

Email: zhu_rong@amss.ac.cn.

[‡]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.*

Email: xinyu@amss.ac.cn

[§]*Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.*

Email: yzm63@psu.edu

^{||}*School of Mathematical Sciences, Capital Normal University, Beijing 100048, China.*

Email: ghzou@amss.ac.cn

First version received: 31 October 2019; final version accepted: 11 June 2020.

Summary: In this paper, we develop a model averaging method to estimate a high-dimensional covariance matrix, where the candidate models are constructed by different orders of polynomial functions. We propose a Mallows-type model averaging criterion and select the weights by minimizing this criterion, which is an unbiased estimator of the expected in-sample squared error plus a constant. Then, we prove the asymptotic optimality of the resulting model average covariance estimators. Finally, we conduct numerical simulations and a case study on Chinese airport network structure data to demonstrate the usefulness of the proposed approaches.

Keywords: *Asymptotic optimality, consistency, covariance regression network model, Mallows criterion, model averaging.*

JEL codes: *C10, C52.*

1. INTRODUCTION

The covariance matrix is a familiar concept and has been widely used in many fields. For example, Markowitz (1952) illustrated geometrically the relationship between belief and choice of portfolio according to their covariance matrix. Campbell et al. (1998) and Jagannathan and Ma (2003) considered finance and risk management based on covariance matrices. Bilmes (2000) improved the parsimony of speech recognition systems by adjusting the type of covariance matrices. Chen and Conley (2001) presented a semiparametric model for high-dimensional vector time series: they analysed a semiparametric spatial model by using the positive-definiteness of a covariance function for panel data in the context of time series. Friedman et al. (2008) considered the problem of estimating sparse graphs by incorporating the lasso penalty into the inverse covariance matrix. And, motivated by the arbitrage pricing theory in finance, Fan

et al. (2008) used the multi-factor model to reduce dimensionality and estimate the covariance matrix.

Let $\Sigma \equiv \text{cov}(\mathbf{Y}) \in \mathbf{R}^{p \times p}$ be the p -dimensional covariance matrix of a p -dimensional random vector $\mathbf{Y} = (Y_1, \dots, Y_p)^T$. The most commonly used covariance matrix estimator of Σ is the classic sample covariance matrix estimator. However, the sample covariance matrix estimator does not perform well when $p \rightarrow \infty$ and the sample size n is fixed or grows at a slower rate than p , because the number of unknown parameters is much larger than the sample size in this case. Inevitably, additional assumptions need to be made to ensure that the estimation of Σ is feasible. Here, we consider the novel strategy proposed by Lan et al. (2018), which uses a covariance regression network model (CRNM) to estimate the high-dimensional covariance matrix. In this framework, the covariance matrix is regarded as a polynomial function of the symmetric adjacency matrix, representing a given network structure. In this way, the estimation of a high-dimensional covariance matrix is converted into the estimation of the low-dimensional coefficients of the CRNM. These authors further developed a Bayesian information criterion (BIC) to select the order of the polynomial function and proved the consistency of the BIC. Their treatment can thus be classed as a model selection approach.

Model averaging can be viewed as a smooth extension of model selection and it can substantially reduce the risk relative to model selection (Hansen, 2014). Furthermore, model averaging is often more stable than model selection, in which a small change of the data may lead to a significant change in the model selected; see Yuan and Yang (2005) and Leung and Barron (2006) for further discussion. There has been much research into model averaging for regression models. For example, Buckland et al. (1997) proposed the smoothed AIC and smoothed BIC method, in which weights are assigned based on the information criterion scores obtained from different models. Hansen (2007) and Wan et al. (2010) developed a Mallows model averaging method for linear regression models. Other methods include jackknife model averaging (Hansen and Racine, 2012; Zhang et al., 2013), heteroscedasticity-robust C_p model averaging (Liu and Okui, 2013), leave-subject-out cross-validation (Gao et al., 2016), and Mahalanobis Mallows model averaging (Zhu et al., 2018).

In this paper, in order to improve the estimation of a high-dimensional covariance matrix with a network structure, we propose a model averaging method based on the normalized version of CRNM (Lan et al., 2018). We select the averaging weights through minimizing a Mallows-type criterion, which is an unbiased estimator of the expected in-sample squared error plus a constant. We then establish the asymptotic optimality of the resulting model average covariance (MAC) estimator when none of the candidate models is correct. We also show that MAC will be consistent during the estimation of the covariance matrix Σ if at least one of the candidate models is in fact correct. Following Lan et al. (2018), for the sake of convenience, we consider $p \rightarrow \infty$ and $n = 1$ first, and then discuss the extension when n increases.

The remainder of this paper is organized as follows. In Section 2, we describe the estimation method in the normalized CRNM (nCRNM), introduce the Mallows-type weight choice criterion, and propose the MAC estimator based on this criterion. The asymptotic optimality of the model averaging estimator is established in Section 3, while the consistency of the method is shown in Section 4. We extend the theorems of asymptotic optimality and consistency to the case that the sample size n is larger than one in Section 5. We compare the finite-sample properties of the MAC estimator with several information criterion-based model selection and averaging estimators in Section 6. A real data example is considered in Section 7. Section 8 contains some concluding remarks. Technical proofs are in Online Appendices.

2. MODEL SET-UP AND ESTIMATION

2.1. Covariance regression network model

Consider a graph with p nodes, where the i th node is used to represent the i th observation, for example the i th individual. Let $\tilde{\mathbf{A}} = (\tilde{a}_{ij}) \in \mathbf{R}^{p \times p}$ denote the symmetric adjacency matrix of the graph, where $\tilde{a}_{ij} = 1$ if the two nodes i and j are directly related, for example if the i th and j th individuals know each other, and let $\tilde{a}_{ij} = 0$ otherwise. We set the diagonal elements $\tilde{a}_{ii} = 0, i = 1, \dots, p$. Let $\tilde{\mathbf{A}}^k$ be the k th power of $\tilde{\mathbf{A}}$, with the zeroth power defined as identity, i.e., $\tilde{\mathbf{A}}^0 = \mathbf{I}_p$. The (ij) th component of $\tilde{\mathbf{A}}^k$ is then $\sum_{s_1, \dots, s_{k-1}} \tilde{a}_{is_1} \tilde{a}_{s_1 s_2} \dots \tilde{a}_{s_{k-1} j}$ and it counts the number of ways to connect the i th node and the j th node through a path with exactly k steps. In other words, $\tilde{\mathbf{A}}^k$ measures the k -path relationships of the p nodes.

Let Y_i be a random variable that describes a certain property of the i th node, such as the i th person's social activities in a period of time. Let $\mathbf{Y} = (Y_1, \dots, Y_p)^T$, and let the covariance matrix of $\mathbf{Y}$ be Σ . Naturally, Σ can be linked to the adjacency matrix $\tilde{\mathbf{A}}$. Lan et al. (2018) introduced the CRNM: $\Sigma = \sum_{k=0}^{\tilde{d}} \tilde{\beta}_k \tilde{\mathbf{A}}^k = \tilde{\beta}_0 \mathbf{I}_p + \tilde{\beta}_1 \tilde{\mathbf{A}} + \tilde{\beta}_2 \tilde{\mathbf{A}}^2 + \dots + \tilde{\beta}_{\tilde{d}} \tilde{\mathbf{A}}^{\tilde{d}}$, where $\tilde{d}$ is a positive integer and $(\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_{\tilde{d}})^T \in \mathbf{R}^{\tilde{d}+1}$ is the vector of regression coefficients. To ensure that the parameters in the CRNM are comparable between models, we normalize the elements of $\tilde{\mathbf{A}}$, $\tilde{a}_{ij}$, by dividing $\{(\sum_i \tilde{a}_{ij})(\sum_j \tilde{a}_{ij})\}^{1/2}$. Hence, the normalized matrix is $\mathbf{A} = (a_{ij}) \in \mathbf{R}^{p \times p}$ with $a_{ij} = \tilde{a}_{ij} / \{(\sum_i \tilde{a}_{ij})(\sum_j \tilde{a}_{ij})\}^{1/2}$. The (ij) th component of $\mathbf{A}^k$ is then $\sum_{s_1, \dots, s_{k-1}} a_{is_1} a_{s_1 s_2} \dots a_{s_{k-1} j}$, which is the 'normalized' number of ways to connect the i th node and the j th node through a path with k steps. Rather than using the CRNM in Lan et al. (2018), we assume the following normalized CRNM (nCRNM):

$$\Sigma = \sum_{k=0}^d \beta_k \mathbf{A}^k = \beta_0 \mathbf{I}_p + \beta_1 \mathbf{A} + \beta_2 \mathbf{A}^2 + \dots + \beta_d \mathbf{A}^d, \quad (2.1)$$

where d is a positive integer that can increase to infinity with p , and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T \in \mathbf{R}^{d+1}$ is the vector of regression coefficients.

To ensure that Σ in (2.1) is positive-definite, constraints are imposed on $\boldsymbol{\beta}$. Because the adjacency matrix $\mathbf{A}$ is assumed known, an obvious benefit of the nCRNM (2.1) is that it reduces the number of parameters in Σ from $p(p+1)/2$ to $d+1$. Note that (2.1) is not necessarily the smallest model, since it is possible that $\beta_k = 0$ for any $k \in \{0, \dots, d\}$. In particular, we allow the 'wasteful' case where $\beta_k = 0$ for all $k > d_0$, for a certain value $d_0 < d$.

In estimating the covariance matrix Σ , it is common to assume that $Y_j, j = 1, \dots, p$, have the same mean, which can be consistently estimated by the sample average $\bar{Y} \equiv (\sum_{j=1}^p Y_j)/p$. One can further centre the data by subtracting the sample mean and work directly with $Y_j - \bar{Y}$. Hence, without loss of generality, we assume that $E(Y_j) = 0$ in the rest of the paper. For convenience, we also perform a standard eigenvalue decomposition on $\mathbf{A}$ to obtain

$$\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{U}^T, \quad (2.2)$$

where $\mathbf{U} \in \mathbf{R}^{p \times p}$ is an orthogonal matrix, $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ is a diagonal matrix, and λ_j is the j th-largest eigenvalue of $\mathbf{A}$. These preparations allow us to convert the model in (2.1) to a simple linear regression model, which facilitates straightforward estimation. Specifically, let $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_p)^T = \mathbf{U}^T \mathbf{Y} \in \mathbf{R}^p$, where $\mathbf{U}$ is defined in (2.2). Then $\tilde{\mathbf{Y}}$ has mean $\mathbf{0}$ and covariance matrix $\sum_{k=0}^d \beta_k \mathbf{D}^k$. Extracting the j th diagonal element of these relations, we obtain

$E(\tilde{Y}_j^2) = \sum_{k=0}^d \beta_k \lambda_j^k$, which is a multiple linear regression model if we view $\tilde{Y}_j^2$ as the response variable, λ_j^k as the k th regressor, and $\boldsymbol{\beta}$ as the regression coefficients. To fix the notation, we define $\mathbf{X}_j = (1, \lambda_j, \dots, \lambda_j^d)^T \in \mathbf{R}^{d+1}$, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)^T \in \mathbf{R}^{p \times (d+1)}$ and $\mathbf{V} = (\tilde{Y}_1^2, \dots, \tilde{Y}_p^2)^T$. Then, the nCRNM (2.1) is equivalently written as

$$\mathbf{V} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.3)$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Let $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $\boldsymbol{\Omega} = \text{var}(\boldsymbol{\epsilon})$, and $\text{diag}(\boldsymbol{\mu})$ be the diagonal matrix with the i th diagonal entry being the i th element of $\boldsymbol{\mu}$. With this notation,

$$\boldsymbol{\Sigma} = \mathbf{U} \left(\sum_{k=0}^d \beta_k \mathbf{D}^k \right) \mathbf{U}^T = \mathbf{U} \text{diag}(\boldsymbol{\mu}) \mathbf{U}^T. \quad (2.4)$$

Thus, for nCRNM (2.1) with a pre-determined d , one can use the ordinary least squares method when $\mathbf{X}$ is of full column rank to obtain the estimator of $\boldsymbol{\beta}$ (i.e., $\hat{\boldsymbol{\beta}}$) in (2.3) and use (2.4) to obtain $\hat{\boldsymbol{\Sigma}}$.

2.2. Candidate models and estimation

Each fixed polynomial order d represents a different nCRNM (2.1). In practice, which d value is suitable is often unclear. We thus consider many possible values of d , which corresponds to a series of possible models. We index these models by $m = 1, \dots, M$, where the total number of models M can be related to the total number of nodes p . To further increase flexibility, we allow the m th candidate nCRNM to contain k_m arbitrary monomials of $\mathbf{A}$, which do not have to be $\mathbf{A}^0$ to $\mathbf{A}^{k_m-1}$ but must include the intercept term $\mathbf{A}^0$. Similar to (2.3), the resulting m th nCRNM is equivalently written as

$$\mathbf{V} = \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)} + \boldsymbol{\epsilon}_{(m)}, \quad (2.5)$$

where $\mathbf{X}_{(m)}$ is a $p \times k_m$ submatrix of $\mathbf{X}$, $\boldsymbol{\beta}_{(m)}$ is the coefficient vector, and $\boldsymbol{\epsilon}_{(m)}$ is the error. Here we can assume that $\mathbf{X}$ is a $p \times (d+1)$ matrix with d sufficiently large so that $(d+1) \geq k_m$ for all $m = 1, \dots, M$ and $\mathbf{X}_{(m)}$ is of full column rank. Although we consider M models, we do not assume that any of these models are correct. Thus, we allow some or all of the M models to be misspecified.

It is easy to see that the ordinary least square estimator of $\boldsymbol{\beta}_{(m)}$ is $\hat{\boldsymbol{\beta}}_{(m)} = (\mathbf{X}_{(m)}^T \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \mathbf{V}$, and the estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}_{(m)} = \mathbf{P}_{(m)} \mathbf{V}$, where $\mathbf{P}_{(m)} \equiv \mathbf{X}_{(m)} (\mathbf{X}_{(m)}^T \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T$ is the projection matrix. By (2.4), the estimator of $\boldsymbol{\Sigma}$ based on the m th candidate model is

$$\hat{\boldsymbol{\Sigma}}_{(m)} = \mathbf{U} \text{diag}(\hat{\boldsymbol{\mu}}_{(m)}) \mathbf{U}^T.$$

We thus have M estimators of $\boldsymbol{\Sigma}$ based on the M candidate models.

2.3. Model averaging and weight choice criterion

Our purpose is now to combine the M estimators of $\boldsymbol{\Sigma}$ obtained from the M candidate models to achieve an optimal estimator of $\boldsymbol{\Sigma}$ through a weighted average of $\hat{\boldsymbol{\Sigma}}_{(m)}$ s. The optimality is reflected in the fact that the resulting estimator minimizes the distance to the true $\boldsymbol{\Sigma}$, while it simultaneously ensures that the estimated covariance matrix is positive-definite. Note that the positive-definiteness property is not guaranteed by the estimators described in Subsections 2.1 and 2.2. Of course, to obtain a positive-definite matrix by forming a weighted average of several

candidate matrices, we have to assume that at least one of these candidate matrices is positive-definite. This is reasonable and is certainly true when the set of candidates contains the trivial model, which contains only the intercept term $\mathbf{A}^0$.

Let $\mathbf{w} = (w_1, w_2, \dots, w_M)^T$ be a weight vector. A model average estimator of Σ is of the form

$$\widehat{\Sigma}(\mathbf{w}) = \sum_{m=1}^M w_m \widehat{\Sigma}_{(m)} = \mathbf{U} \text{diag}\{\widehat{\boldsymbol{\mu}}(\mathbf{w})\} \mathbf{U}^T,$$

where $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{m=1}^M w_m \widehat{\boldsymbol{\mu}}_{(m)}$. Similarly, define $\mathbf{P}(\mathbf{w}) \equiv \sum_{m=1}^M w_m \mathbf{P}_{(m)}$. We then have

$$\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{m=1}^M w_m \widehat{\boldsymbol{\mu}}_{(m)} = \sum_{m=1}^M w_m \mathbf{P}_{(m)} \mathbf{V} = \mathbf{P}(\mathbf{w}) \mathbf{V}.$$

We restrict the weight vector in the set $\mathcal{H}_p$, where

$$\mathcal{H}_p = \{\mathbf{w} \in [0, 1]^M : \sum_{m=1}^M w_m = 1 \text{ and } \widehat{\Sigma}(\mathbf{w}) \text{ is positive-definite}\}. \quad (2.6)$$

Obviously, $\mathcal{H}_p$ is not empty under our assumption that at least one candidate model is positive-definite. We measure the distance between two matrices using the Frobenius norm of the difference, and hence the Frobenius norm loss of $\widehat{\Sigma}(\mathbf{w})$ is naturally defined as

$$L_p(\mathbf{w}) = \|\widehat{\Sigma}(\mathbf{w}) - \Sigma\|_F^2 = \text{trace}\{(\widehat{\Sigma}(\mathbf{w}) - \Sigma)^T(\widehat{\Sigma}(\mathbf{w}) - \Sigma)\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Our purpose is to devise a weight choice criterion to minimize the expected loss $L_p(\mathbf{w})$.

We first note that

$$\begin{aligned} L_p(\mathbf{w}) &= \text{trace}\{(\widehat{\Sigma}(\mathbf{w}) - \Sigma)^T(\widehat{\Sigma}(\mathbf{w}) - \Sigma)\} \\ &= \text{trace}\{[\mathbf{U}\{\text{diag}(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})\}\mathbf{U}^T]^T[\mathbf{U}\{\text{diag}(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})\}\mathbf{U}^T]\} \\ &= \text{trace}\{\text{diag}^2(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})\} \\ &= \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2, \end{aligned} \quad (2.7)$$

where $\|\cdot\|$ denotes the L_2 norm. This means that the Frobenius norm loss of $\widehat{\Sigma}(\mathbf{w})$ is the same as the squared error loss of $\widehat{\boldsymbol{\mu}}(\mathbf{w})$. Then the corresponding risk function, defined as the expected loss function, can be calculated as

$$\begin{aligned} R_p(\mathbf{w}) &= \mathbb{E}\{L_p(\mathbf{w})\} \\ &= \mathbb{E}\|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2 \\ &= \mathbb{E}(\|(\mathbf{P}(\mathbf{w}) - \mathbf{I}_p)\boldsymbol{\mu} + \mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2) \\ &= \|(\mathbf{P}(\mathbf{w}) - \mathbf{I}_p)\boldsymbol{\mu}\|^2 + \text{trace}\{\mathbf{P}^T(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}. \end{aligned} \quad (2.8)$$

Of course, neither $L_p(\mathbf{w})$ nor $R_p(\mathbf{w})$ can be directly minimized because they depend on $\boldsymbol{\mu}$, which is unknown. We thus work around the difficulty by first replacing $\boldsymbol{\mu}$ with $\mathbf{V}$, and then adjusting for the offset. This leads us to construct an estimator of the risk function,

$$C_p(\mathbf{w}) = \|\mathbf{V} - \widehat{\boldsymbol{\mu}}(\mathbf{w})\|^2 + 2\text{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}. \quad (2.9)$$

It can be readily shown that

$$E\{C_p(\mathbf{w})\} = R_p(\mathbf{w}) + \text{trace}(\mathbf{\Omega}). \quad (2.10)$$

This implies that $C_p(\mathbf{w})$ is an unbiased estimator of the risk function $R_p(\mathbf{w})$ plus a constant irrelevant to the weight $\mathbf{w}$. Therefore, by minimizing $C_p(\mathbf{w})$, we expect that $L_p(\mathbf{w})$ and $R_p(\mathbf{w})$ are also minimized. This property in (2.10) is similar to the Mallows criterion proposed by Hansen (2007). We thus proceed to use the Mallows-type model averaging criterion. Alternatively, the jackknife model averaging (Hansen and Racine, 2012; Zhang et al., 2013) criterion can also be used for weight choice. Liu and Okui (2013) have shown that the Mallows-type and jackknife methods have a similar performance.

In practice, the covariance matrix $\mathbf{\Omega}$ in $C_p(\mathbf{w})$ is unknown and needs to be estimated. Following Hansen (2007), Lan et al. (2018) and Zhang and Wang (2019), we estimate $\mathbf{\Omega}$ based on a candidate model containing the largest number of covariates, indexed by

$$M^* = \arg \max_{1 \leq m \leq M} k_m. \quad (2.11)$$

When several candidates have the same maximum number of covariates, we simply pick any one of such candidate models as the M^* model. This leads to the estimator

$$\widehat{\mathbf{\Omega}} = \text{diag}(\widehat{\epsilon}_{(M^*)1}^2, \dots, \widehat{\epsilon}_{(M^*)p}^2), \quad (2.12)$$

where $\widehat{\epsilon}_{(M^*)j} = \widetilde{Y}_j^2 - \widehat{\mu}_{(M^*)j}$, and $\widehat{\mu}_{(M^*)j}$ is the j th component of $\widehat{\boldsymbol{\mu}}_{(M^*)}$. Here, we have restricted $\widehat{\mathbf{\Omega}}$ to be diagonal, while $\mathbf{\Omega}$ itself may not be diagonal. We find that despite this seemingly crude practice, the resulting model averaging estimator is always optimal in that it minimizes the expected Frobenius loss $L_p(\mathbf{w})$ regardless of whether $\mathbf{\Omega}$ is diagonal or not.

When $\mathbf{\Omega}$ is replaced by $\widehat{\mathbf{\Omega}}$, $C_p(\mathbf{w})$ changes to

$$\widehat{C}_p(\mathbf{w}) \equiv \|\mathbf{V} - \widehat{\boldsymbol{\mu}}(\mathbf{w})\|^2 + 2\text{trace}\{\mathbf{P}(\mathbf{w})\widehat{\mathbf{\Omega}}\}, \quad (2.13)$$

in which all quantities are known except $\mathbf{w}$. Minimizing $\widehat{C}_p(\mathbf{w})$ with respect to $\mathbf{w}$ leads to

$$\widehat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathcal{H}_p} \widehat{C}_p(\mathbf{w}). \quad (2.14)$$

Substituting $\widehat{\mathbf{w}}$ in $\widehat{\mathbf{\Sigma}}(\mathbf{w})$ yields the model average estimator $\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{w}})$, which we name the model average covariance (MAC) estimator. We point out that the minimization of (2.13) with respect to $\mathbf{w}$ is a constrained quadratic programming problem, and hence the computation of the optimal weight vector is straightforward. For example, quadratic programming can be performed using the quadprog package in R, the quadprog command in MATLAB, or the qp command in SAS. Next, we present the asymptotic optimality of the MAC estimator $\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{w}})$.

3. ASYMPTOTIC OPTIMALITY

As we have pointed out, in order to obtain a final covariance matrix estimator based on several candidate estimators, at least one of these candidate estimators needs to be valid, i.e., positive-definite. We state this formally as Condition (C.1). We then proceed to establish the optimality property of our procedure under scenarios of both independent and dependent error components.

CONDITION (C.1) *There exists an $m \in \{1, \dots, M\}$ such that $\widehat{\mathbf{\Sigma}}_{(m)}$ is positive-definite.*

We first define some notation. Let $\widetilde{\mathcal{H}}_p = \{\mathbf{w} \in [0, 1]^M : \sum_{m=1}^M w_m = 1\}$, which is a larger space than $\mathcal{H}_p$. Write $\xi_p \equiv \inf_{\mathbf{w} \in \widetilde{\mathcal{H}}_p} R_p(\mathbf{w})$. Let $\bar{\lambda}(\cdot)$ and $\underline{\lambda}(\cdot)$ denote respectively the maximum and minimum eigenvalues of a matrix. Let $\mathbf{w}_m^0$ be an $M \times 1$ vector where the m th element is one and all the other elements are zero. We use $\rho_{jj}^{(m)}$ to denote the j th diagonal element of $\mathbf{P}_{(m)}$ and let $\rho = \max_{1 \leq m \leq M} \max_{1 \leq j \leq p} \rho_{jj}^{(m)}$. All limiting processes correspond to $p \rightarrow \infty$ unless stated otherwise.

3.1. Asymptotic optimality with independence

We first consider the situation where the elements of ϵ are independent of each other. This arises when, for example, $\mathbf{Y}$ is normally distributed. In this case, $\mathbf{Y} \sim N(\mathbf{0}, \Sigma)$, $\tilde{\mathbf{Y}} \sim N(\mathbf{0}, \sum_{k=0}^d \beta_k \mathbf{D}^k)$. This implies that the elements of $\tilde{\mathbf{Y}}$ are independent of each other; as a consequence, the elements of $\epsilon = \mathbf{V} - \mathbf{X}\beta$ are also independent of each other. We assume the following regularity conditions.

CONDITION (C.2) *There exist a fixed integer G and a constant κ so that $E(\epsilon_j^{4G}) \leq \kappa < \infty$ for $j = 1, \dots, p$.*

CONDITION (C.3) *$M \xi_p^{-2G} \sum_{m=1}^M \{R_p(\mathbf{w}_m^0)\}^G = o(1)$, for the same constant G as given in Condition (C.2).*

CONDITION (C.4) *There exists a constant c such that $\rho \leq ck_{M^*}/p$, where, recall, k_{M^*} is the number of columns of $\mathbf{X}$ in the largest model.*

CONDITION (C.5) *$\boldsymbol{\mu}^T \boldsymbol{\mu} / p = O(1)$, a.s.*

CONDITION (C.6) *$k_{M^*}^2 / p \leq \varphi < \infty$, where φ is a constant.*

REMARK 3.1 *Condition (C.2) places a moment restriction on the error term. Condition (C.3) imposes some relation between the best model risk and the total risks from all the models, in that the best model should not be too much better than all other models. Specifically, let $\underline{\xi}_p = \min_{1 \leq m \leq M} R_p(\mathbf{w}_m^0)$ and $\bar{\xi}_p = \max_{1 \leq m \leq M} R_p(\mathbf{w}_m^0)$, then, given $M < \infty$, a sufficient condition of Condition (C.3) is $\bar{\xi}_p \asymp \underline{\xi}_p^{1+\alpha}$ with $0 \leq \alpha < 1$, which means that the risk of the best model can increase more slowly than that of the worst model, but not too much more slowly. In addition, a consequence of Condition (C.3) is that $\xi_p \rightarrow \infty$, which implies that there is no correctly specified candidate model with a finite dimension. This can be seen from an argument of contradiction. If the m th candidate model is correctly specified with a finite k_m , then*

$$\begin{aligned} \xi_p &= \inf_{\mathbf{w} \in \widetilde{\mathcal{H}}_p} R_p(\mathbf{w}) \\ &\leq R_p(\mathbf{w}_m^0) \\ &= \|(\mathbf{P}_{(m)} - \mathbf{I}_p)\boldsymbol{\mu}\|^2 + \text{trace}\{\mathbf{P}_{(m)}^T \mathbf{P}_{(m)} \boldsymbol{\Omega}\} \\ &= \|(\mathbf{P}_{(m)} - \mathbf{I}_p)\mathbf{X}\beta\|^2 + \text{trace}\{\mathbf{P}_{(m)} \boldsymbol{\Omega} \mathbf{P}_{(m)}\} \\ &= \text{trace}\{\mathbf{P}_{(m)} \boldsymbol{\Omega} \mathbf{P}_{(m)}\} \\ &\leq \bar{\lambda}(\boldsymbol{\Omega})k_m, \end{aligned}$$

which is finite if $\bar{\lambda}(\boldsymbol{\Omega})$ is bounded and hence is a contradiction. Similar conditions are used in Wan et al. (2010), Liu and Okui (2013) and Ando and Li (2014). Condition (C.4) is commonly

used to ensure the asymptotic optimality of cross-validation; see, for example, Andrews (1991) and Hansen and Racine (2012). Condition (C.5) is about the sum of the squares of the elements of $\boldsymbol{\mu}$ and is commonly used in the context of linear regression; see, for example, Wan et al. (2010) and Liang et al. (2011). Condition (C.6) means that $k_{M^*}^2$ has order the same as or smaller than p . A similar requirement $k_{M^*}^2/n = O(1)$, where n is the sample size, has been used in Wan et al. (2010) and Liu et al. (2016).

THEOREM 3.1 Assume that Conditions (C.1)–(C.6) hold. Then as $p \rightarrow \infty$,

$$\frac{L_p(\widehat{\mathbf{w}})}{\inf_{\mathbf{w} \in \mathcal{H}_p} L_p(\mathbf{w})} \xrightarrow{p} 1. \quad (3.1)$$

Theorem 3.1 shows that the MAC estimator is asymptotically optimal in that it leads to a squared error loss that is asymptotically identical to that of the infeasible best possible model average estimator. The proof of Theorem 3.1 is given in Online Appendix A.1.

3.2. Asymptotic optimality without independence

We now consider the situation where the elements of $\boldsymbol{\epsilon}$ are dependent. Recall that $\boldsymbol{\Omega} = \text{var}(\boldsymbol{\epsilon})$, and thus we allow $\boldsymbol{\Omega}$ to be nondiagonal. Let $\mathcal{X} \equiv [\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \dots, \mathbf{X}_{(M)}]$, $r \equiv \text{rank}(\mathcal{X})$, and $\mathcal{P} \equiv \mathcal{X}(\mathcal{X}^T \mathcal{X})^- \mathcal{X}^T$, where $(\mathcal{X}^T \mathcal{X})^-$ is the Moore–Penrose generalized inverse matrix of $\mathcal{X}^T \mathcal{X}$. We can still establish the asymptotic optimality described in Theorem 3.1 with the following additional conditions.

CONDITION (C.7) $\xi_p^{-2} \sum_{m=1}^M R_p(\mathbf{w}_m^0) = o(1)$.

CONDITION (C.8) $r \xi_p^{-1} \rightarrow 0$.

CONDITION (C.9) $0 < c' \leq \underline{\lambda}(\boldsymbol{\Omega}) \leq \bar{\lambda}(\boldsymbol{\Omega}) \leq c < \infty$, where c' and c are positive constants.

REMARK 3.2 Condition (C.7) is similar to Condition (C.3) but is weaker. It is implied by Condition (C.3) with $G = 1$. Condition (C.8) is similar to condition (22) in Zhang et al. (2013). When all candidate models are nested within the largest candidate model M^* defined in (2.11), $r \leq k_{M^*}$, so Condition (C.8) is implied by $k_{M^*} \xi_p^{-1} \rightarrow 0$, which is again a restriction on k_{M^*} and is similar to Condition (C.6). Condition (C.9) is a commonly used boundedness requirement on the minimum and maximum eigenvalues of $\boldsymbol{\Omega}$.

THEOREM 3.2 Assume that Conditions (C.1), (C.4), (C.5), (C.7)–(C.9) hold. Then as $p \rightarrow \infty$, the asymptotic optimality in (3.1) still holds.

Theorem 3.2 shows that Theorem 3.1 remains valid when $\boldsymbol{\Omega}$ is a general matrix instead of a diagonal matrix. The proof of Theorem 3.2 is given in Online Appendix A.2. The reason why the optimality property is retained in Theorem 3.2 lies in Condition (C.8). This condition restricts the effect of the second term in the criterion $\widehat{C}_p(\mathbf{w})$ to be ignorable on the final weight choice compared with the first term, and thus the estimate of $\boldsymbol{\Omega}$ has an ignorable effect.

4. CONSISTENCY WITH CORRECTLY SPECIFIED CANDIDATE MODELS

As discussed in Remark 3.1, the optimality shown above essentially excludes the situation where at least one of the candidate models is correctly specified. Here, we show that when there is at least one correctly specified candidate model, the model averaging method described above can achieve the same convergence rate of these correctly specified candidate models. Let $\Pi_{(m)}$ be a $k_m \times (d+1)$ selection matrix so that $\mathbf{X}_{(m)} = \mathbf{X}\Pi_{(m)}^T$. Assume that the m^* th model is a correctly specified model, i.e.,

$$\Pi_{(m^*)}^T \beta_{(m^*)} = \beta. \quad (4.1)$$

Thus,

$$\mathbf{X}_{(m^*)} \beta_{(m^*)} = \mathbf{X} \beta. \quad (4.2)$$

Because the M^* th model has the largest number of covariates, for convenience we assume that the m^* th model is nested inside the M^* th model.

Under the m th candidate model, $\hat{\beta}_{(m)} = (\mathbf{X}_{(m)}^T \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \mathbf{V}$, and thus the estimator of β is $\Pi_{(m)}^T \hat{\beta}_{(m)}$. Then, the model averaging estimator of β is

$$\hat{\beta}(\hat{\mathbf{w}}) = \sum_{m=1}^M w_m \Pi_{(m)}^T \hat{\beta}_{(m)}. \quad (4.3)$$

To build the consistency of $\hat{\beta}(\hat{\mathbf{w}})$, we assume the following regularity condition. Let $\Psi_p = \mathbf{X}^T \mathbf{X} / p$.

CONDITION (C.10) $0 < c_1 \leq \underline{\lambda}(\Psi_p) \leq \bar{\lambda}(\Psi_p) \leq c_2 < \infty$, where c_1 and c_2 are constants.

REMARK 4.1 Condition (C.10) requires the additional components in bigger models to contribute sufficiently different structures. Condition (C.10) is the same as the condition (A1) of Zou and Zhang (2009). In their paper, c_i and c'_i are constants, and we sometimes use the same c_i to denote different constants.

LEMMA 4.1 If Conditions (C.9) and (C.10) are satisfied, then

$$\left\| \sqrt{p/(d+1)} (\Pi_{(m^*)}^T \hat{\beta}_{(m^*)} - \beta) \right\| = O_p(1). \quad (4.4)$$

REMARK 4.2 Lemma 4.1 shows the $\sqrt{p/(d+1)}$ -consistency of the correctly specified model coefficient $\hat{\beta}_{(m^*)}$, which is a very common convergence result when the dimension of coefficients diverges; see, for example, He and Shao (2000) and Fan and Peng (2004). The proof of Lemma 4.1 is given in Online Appendix A.3.

THEOREM 4.1 Assume that Conditions (C.4), (C.9) and (C.10) hold. Then $\|\sqrt{p/(d+1)}\{\hat{\beta}(\hat{\mathbf{w}}) - \beta\}\| = O_p(1)$.

Theorem 4.1 shows the $\sqrt{p/(d+1)}$ -consistency of the optimal model averaging coefficient $\hat{\beta}(\hat{\mathbf{w}})$. The results in Theorem 4.1 may appear counter-intuitive at first glance, since they indicate that the large size of the unknown matrix Σ is beneficial to us. This is a direct consequence of the assumption that the adjacency matrix $\mathbf{A}$ is known, and hence a larger p under such a setting does indeed represent more information. On the other hand, $d+1$ is the number of parameters, and hence $\sqrt{p/(d+1)}$ indeed resembles the usual diverging parametric convergence rate. The proof of Theorem 4.1 is given in Online Appendix A.4.

5. EXTENSIONS WITH $n \geq 2$

In this section, we extend the asymptotic optimality and consistency properties in Sections 3 and 4 to the case of $n \geq 2$. Suppose we have a sample $\{\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}\} \in \mathbf{R}^p$, which consists of independent and identically distributed copies of $\mathbf{Y}$. Denote $\tilde{\mathbf{Y}}^{(i)} = (\tilde{Y}_1^{(i)}, \dots, \tilde{Y}_p^{(i)})^T = \mathbf{U}^T \mathbf{Y}^{(i)}$, $\mathbf{V}^{(i)} = (\tilde{Y}_1^{(i)2}, \dots, \tilde{Y}_p^{(i)2})^T$, $\boldsymbol{\epsilon}^{(i)} = \mathbf{V}^{(i)} - \mathbf{X}\boldsymbol{\beta} = (\epsilon_1^{(i)}, \dots, \epsilon_p^{(i)})^T$ and $\boldsymbol{\epsilon}_{(m)}^{(i)} = \mathbf{V}^{(i)} - \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)}$ for $i = 1, 2, \dots, n$. Let $\bar{\mathbf{V}} = \sum_{i=1}^n \mathbf{V}^{(i)}/n$, $\bar{V}_i$ be the i th component of $\bar{\mathbf{V}}$, $\bar{\boldsymbol{\epsilon}} = \sum_{i=1}^n \boldsymbol{\epsilon}^{(i)}/n = (\bar{\epsilon}_1, \dots, \bar{\epsilon}_p)^T$ and $\bar{\boldsymbol{\epsilon}}_{(m)} = \sum_{i=1}^n \boldsymbol{\epsilon}_{(m)}^{(i)}/n$. Then, we know $E(\bar{\boldsymbol{\epsilon}}) = \mathbf{0}$ and $\text{var}(\bar{\boldsymbol{\epsilon}}) = \boldsymbol{\Omega}/n$. By model (2.3) and the m th candidate model (2.5), we have that

$$\bar{\mathbf{V}} = \boldsymbol{\mu} + \bar{\boldsymbol{\epsilon}} = \mathbf{X}\boldsymbol{\beta} + \bar{\boldsymbol{\epsilon}} \quad \text{and} \quad \bar{\mathbf{V}} = \mathbf{X}_{(m)}\boldsymbol{\beta}_{(m)} + \bar{\boldsymbol{\epsilon}}_{(m)}. \quad (5.1)$$

Then, the ordinary least squares estimator of $\boldsymbol{\beta}_{(m)}$ is $\hat{\boldsymbol{\beta}}_{(m)} = (\mathbf{X}_{(m)}^T \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^T \bar{\mathbf{V}}$, and the estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}_{(m)} = \mathbf{P}_{(m)} \bar{\mathbf{V}}$. For simplicity, we do not explicitly redefine $\hat{\boldsymbol{\Sigma}}_{(m)}$, $\hat{\boldsymbol{\Sigma}}(\mathbf{w})$, $\hat{\boldsymbol{\mu}}_{(m)}$, $\hat{\boldsymbol{\mu}}(\mathbf{w})$ and $L_p(\mathbf{w})$, other than pointing out that they are the same as defined in Section 2 except with the $\mathbf{V}$ in their corresponding expressions replaced by $\bar{\mathbf{V}}$.

The loss function is $L_p(\mathbf{w}) = \|\hat{\boldsymbol{\Sigma}}(\mathbf{w}) - \boldsymbol{\Sigma}\|_F^2 = \|\hat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2 = \|\mathbf{P}(\mathbf{w})\bar{\mathbf{V}} - \boldsymbol{\mu}\|^2$, and the corresponding risk function is

$$\begin{aligned} R_p(\mathbf{w}) &= E\{L_p(\mathbf{w})\} \\ &= E\|\hat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2 \\ &= E(\|(\mathbf{P}(\mathbf{w}) - \mathbf{I}_p)\boldsymbol{\mu} + \mathbf{P}(\mathbf{w})\bar{\boldsymbol{\epsilon}}\|^2) \\ &= \|(\mathbf{P}(\mathbf{w}) - \mathbf{I}_p)\boldsymbol{\mu}\|^2 + \text{trace}\{\mathbf{P}^T(\mathbf{w})\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}/n. \end{aligned} \quad (5.2)$$

The estimator of the risk function is

$$C_p(\mathbf{w}) = \|\bar{\mathbf{V}} - \hat{\boldsymbol{\mu}}(\mathbf{w})\|^2 + 2\text{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}/n,$$

which is an unbiased estimator of the risk $R_p(\mathbf{w})$ up to a constant, i.e.,

$$E\{C_p(\mathbf{w})\} = R_p(\mathbf{w}) + \text{trace}(\boldsymbol{\Omega})/n.$$

In practice, we estimate $\boldsymbol{\Omega}$ by

$$\tilde{\boldsymbol{\Omega}} = \text{diag}(\tilde{\epsilon}_{(M^*)1}^2, \dots, \tilde{\epsilon}_{(M^*)p}^2), \quad (5.3)$$

where $\tilde{\epsilon}_{(M^*)j}^2 = n(\bar{V}_j - \hat{\mu}_{(M^*)j})^2$. Then, $C_p(\mathbf{w})$ is changed to

$$\hat{C}_p(\mathbf{w}) \equiv \|\bar{\mathbf{V}} - \hat{\boldsymbol{\mu}}(\mathbf{w})\|^2 + 2\text{trace}\{\mathbf{P}(\mathbf{w})\tilde{\boldsymbol{\Omega}}\}/n, \quad (5.4)$$

and the optimal weight vector is

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathcal{H}_p} \hat{C}_p(\mathbf{w}). \quad (5.5)$$

Next, we consider the asymptotic optimality of the model averaging estimator $\boldsymbol{\Sigma}(\hat{\mathbf{w}})$ as $n \geq 2$ and $p \rightarrow \infty$ when no correct model is contained in the candidate model set.

COROLLARY 5.1 *When $\boldsymbol{\Omega}$ is diagonal, if Conditions (C.1)–(C.6) hold, then as $n \geq 2$ and $p \rightarrow \infty$, the asymptotic optimality in (3.1) still holds.*

Table 1. The MFL under a normal distribution for Design 1 ($\times 100$).

r	p	n	MAC	SAIC	SBIC	AIC	BIC
5	200	1	5.780	6.467	6.992	7.383	8.923
		$\lfloor p^{1/5} \rfloor$	3.571	3.991	4.072	4.639	4.736
		$\lfloor p^{1/3} \rfloor$	1.463	1.638	1.775	1.908	2.144
	400	1	3.086	3.456	3.637	3.966	4.271
		$\lfloor p^{1/5} \rfloor$	1.243	1.518	1.745	1.723	2.081
		$\lfloor p^{1/3} \rfloor$	0.551	0.631	0.798	0.640	0.916
10	200	1	7.720	8.305	8.638	9.181	9.959
		$\lfloor p^{1/5} \rfloor$	4.819	5.143	5.344	5.460	5.913
		$\lfloor p^{1/3} \rfloor$	1.733	1.790	1.805	1.962	1.953
	400	1	3.878	4.372	4.608	4.660	5.265
		$\lfloor p^{1/5} \rfloor$	1.331	1.429	1.417	1.572	1.566
		$\lfloor p^{1/3} \rfloor$	0.637	0.726	0.821	0.787	0.946

Notes: The smallest, second smallest and largest MFLs in each row are highlighted in **bold**, **bold italic** and *italic*, respectively.

COROLLARY 5.2 When Ω is nondiagonal, if Conditions (C.1), (C.4), (C.5), (C.7)–(C.9) hold, then as $n \geq 2$ and $p \rightarrow \infty$, the asymptotic optimality in (3.1) still holds.

Corollaries 5.1 and 5.2 are the generalized versions of Theorems 3.1 and 3.2, where the sample size n is larger than 1. The proofs of Corollaries 5.1 and 5.2 are given in Online Appendices A.5 and A.6. From these proofs, it is straightforward to see that when $n \rightarrow \infty$, these corollaries still hold.

Similarly, we also establish the consistency of the methods when at least one candidate model is correct. Since n can go to infinity, the convergence order in Lemma 4.1 will depend on n . Specifically, we have the following result.

LEMMA 5.1 If Conditions (C.9) and (C.10) are satisfied, then

$$\left\| \sqrt{np/(d+1)}(\Pi_{(m^*)}^T \hat{\beta}_{(m^*)} - \beta) \right\| = O_p(1). \quad (5.6)$$

REMARK 5.1 Lemma 5.1 is a generalization of Lemma 4.1 when the sample size n is allowed to diverge. The proof of Lemma 5.1 is given in Online Appendix A.7.

COROLLARY 5.3 Assume that Conditions (C.4), (C.9) and (C.10) hold. Then $\|\hat{\beta}(\hat{\mathbf{w}}) - \beta\| = O_p\{\sqrt{(d+1)/np} + \sqrt{(d+1)/p^{3/2}}\}$.

From Corollary 5.3, if $n = o(\sqrt{p})$, then $\|\hat{\beta}(\hat{\mathbf{w}}) - \beta\| = O_p\{\sqrt{(d+1)/(np)}\}$ under those conditions. Otherwise, $\|\hat{\beta}(\hat{\mathbf{w}}) - \beta\| = O_p\{\sqrt{(d+1)/p^{3/2}}\}$. Thus, as far as the convergence rate is concerned, there is no need to increase n to be much larger than $\sqrt{p}$ because no more gain can be obtained. The proof of Corollary 5.3 is given in Online Appendix A.8.

6. SIMULATION STUDY

This section is devoted to a comparison of the finite-sample performance of the MAC estimator with the AIC- and BIC-based model selection and averaging estimators. The AIC and BIC scores for the m th candidate model are $\text{AIC}_m = p \log(\hat{\sigma}_m^2) + 2k_m$ and $\text{BIC}_m = p \log(\hat{\sigma}_m^2) + k_m \log(p)$, respectively, where $\hat{\sigma}_m^2 = \|\mathbf{V} - \hat{\mu}_{(m)}\|^2/p$. (When $n \geq 2$, we use $\bar{\mathbf{V}}$ instead of $\mathbf{V}$ and set

Table 2. The MFL under an exponential distribution for Design 1 ($\times 100$).

r	p	n	MAC	SAIC	SBIC	AIC	BIC
5	200	1	9.828	<i>10.631</i>	11.169	11.581	<i>13.106</i>
		$\lfloor p^{1/5} \rfloor$	5.036	<i>5.245</i>	5.316	5.736	<i>5.817</i>
		$\lfloor p^{1/3} \rfloor$	2.512	<i>2.784</i>	2.937	3.068	<i>3.374</i>
	400	1	5.120	<i>5.542</i>	5.726	5.997	<i>6.258</i>
		$\lfloor p^{1/5} \rfloor$	1.997	<i>2.212</i>	2.381	2.437	<i>2.752</i>
		$\lfloor p^{1/3} \rfloor$	0.833	<i>0.919</i>	1.068	0.941	<i>1.189</i>
10	200	1	11.093	<i>11.650</i>	11.997	12.494	<i>13.450</i>
		$\lfloor p^{1/5} \rfloor$	5.464	<i>5.576</i>	5.813	5.789	<i>6.366</i>
		$\lfloor p^{1/3} \rfloor$	2.507	<i>2.593</i>	2.619	2.808	<i>2.814</i>
	400	1	5.321	<i>5.866</i>	6.205	6.142	<i>6.909</i>
		$\lfloor p^{1/5} \rfloor$	2.028	2.148	2.135	2.338	<i>2.295</i>
		$\lfloor p^{1/3} \rfloor$	0.856	<i>0.948</i>	1.043	1.015	<i>1.150</i>

Notes: The smallest, second smallest and largest MFLs in each row are highlighted in **bold**, ***bold italic*** and *italic*, respectively.

Table 3. The MFL under a mixture distribution for Design 1 ($\times 100$).

r	p	n	MAC	SAIC	SBIC	AIC	BIC
5	200	1	9.801	<i>10.476</i>	10.954	11.563	<i>12.515</i>
		$\lfloor p^{1/5} \rfloor$	5.952	<i>7.443</i>	7.460	8.038	8.036
		$\lfloor p^{1/3} \rfloor$	2.230	<i>2.481</i>	2.601	2.792	<i>3.040</i>
	400	1	5.233	<i>5.671</i>	5.803	6.255	<i>6.430</i>
		$\lfloor p^{1/5} \rfloor$	1.861	<i>2.328</i>	2.548	2.500	<i>2.891</i>
		$\lfloor p^{1/3} \rfloor$	0.769	<i>0.864</i>	1.023	0.885	<i>1.143</i>
10	200	1	10.421	<i>10.994</i>	11.285	11.908	<i>12.716</i>
		$\lfloor p^{1/5} \rfloor$	5.737	<i>5.915</i>	6.124	6.135	<i>6.754</i>
		$\lfloor p^{1/3} \rfloor$	2.556	<i>2.698</i>	2.741	2.916	<i>2.922</i>
	400	1	5.543	<i>5.921</i>	6.274	6.148	<i>7.025</i>
		$\lfloor p^{1/5} \rfloor$	1.869	1.977	1.946	2.150	<i>2.085</i>
		$\lfloor p^{1/3} \rfloor$	0.825	<i>0.919</i>	1.010	0.988	<i>1.143</i>

Notes: The smallest, second smallest and largest MFLs in each row are highlighted in **bold**, ***bold italic*** and *italic*, respectively.

$\hat{\mu}_{(m)} = \mathbf{P}_{(m)} \bar{\mathbf{V}}.$) Buckland et al. (1997) suggested smoothed AIC (SAIC) and smoothed BIC (SBIC) model averaging methods, where the weight for the m th model is simply set as $\exp(-\text{AIC}_m/2)/\sum_{l=1}^M \exp(-\text{AIC}_l/2)$, and similarly for BIC. Owing to their ease of use, the SAIC and SBIC weight choice methods have been used extensively in the literature (Wan and Zhang, 2009; Millar et al., 2014).

In the following, we consider two kinds of experimental design. In the first design, all candidate models are misspecified, while in the second one, some candidate models are correctly specified.

6.1. Experimental designs

Design 1 (All candidate models are misspecified). We set the true covariance matrix Σ as $\Sigma = \sum_{k=0}^d \beta_k \mathbf{A}^k$, where $\beta_k = (2/3)^k$ and the individual elements of $\mathbf{A}$, a_{ij} , are calculated by

Table 4. The MFL under a normal distribution for Design 2 ($\times 100$).

β	r	p	n	MAC	SAIC	SBIC	AIC	BIC
$\beta^{(1)}$	5	200	1	21.577	24.045	24.885	26.990	29.942
			$\lfloor p^{1/5} \rfloor$	12.983	14.752	14.289	16.624	16.140
			$\lfloor p^{1/3} \rfloor$	5.287	5.470	5.536	6.071	6.313
		400	1	11.103	12.604	12.678	14.173	14.393
			$\lfloor p^{1/5} \rfloor$	4.422	5.161	5.212	5.670	5.857
			$\lfloor p^{1/3} \rfloor$	1.719	1.824	1.796	1.983	1.958
	10	200	1	24.202	25.178	25.367	28.237	28.729
			$\lfloor p^{1/5} \rfloor$	15.434	16.109	16.480	17.396	18.228
			$\lfloor p^{1/3} \rfloor$	5.180	5.289	5.417	5.617	6.059
		400	1	13.505	13.953	14.699	15.180	16.458
			$\lfloor p^{1/5} \rfloor$	4.382	4.447	4.481	4.739	4.905
			$\lfloor p^{1/3} \rfloor$	1.878	1.903	1.905	1.969	2.037
$\beta^{(2)}$	5	200	1	17.819	19.836	20.807	22.336	24.616
			$\lfloor p^{1/5} \rfloor$	10.138	11.543	12.006	12.506	13.587
			$\lfloor p^{1/3} \rfloor$	3.987	4.110	4.131	4.321	4.381
		400	1	8.332	9.742	10.687	10.623	12.495
			$\lfloor p^{1/5} \rfloor$	3.521	4.295	4.157	4.545	4.407
			$\lfloor p^{1/3} \rfloor$	1.238	1.359	1.287	1.477	1.392
	10	200	1	21.224	21.724	21.570	23.124	23.196
			$\lfloor p^{1/5} \rfloor$	13.374	13.885	14.068	14.667	15.207
			$\lfloor p^{1/3} \rfloor$	4.486	4.647	4.868	4.814	5.248
		400	1	11.434	11.830	12.125	12.550	13.279
			$\lfloor p^{1/5} \rfloor$	3.746	3.838	4.057	4.009	4.407
			$\lfloor p^{1/3} \rfloor$	1.626	1.671	1.710	1.717	1.793
$\beta^{(3)}$	5	200	1	15.920	18.239	18.245	20.264	20.769
			$\lfloor p^{1/5} \rfloor$	9.256	10.970	10.732	11.693	11.854
			$\lfloor p^{1/3} \rfloor$	3.820	4.003	3.922	4.358	4.287
		400	1	7.563	9.210	9.155	10.037	10.318
			$\lfloor p^{1/5} \rfloor$	3.295	4.121	3.923	4.458	4.335
			$\lfloor p^{1/3} \rfloor$	1.262	1.363	1.398	1.490	1.586
	10	200	1	20.512	21.196	20.628	22.084	21.468
			$\lfloor p^{1/5} \rfloor$	12.942	13.595	13.606	14.150	14.329
			$\lfloor p^{1/3} \rfloor$	4.314	4.560	4.682	4.775	4.983
		400	1	10.950	11.458	11.306	12.039	11.805
			$\lfloor p^{1/5} \rfloor$	3.454	3.652	3.748	3.810	3.982
			$\lfloor p^{1/3} \rfloor$	1.535	1.630	1.675	1.687	1.777

Notes: The smallest, second smallest and largest MFLs in each row are highlighted in **bold**, **bold italic** and *italic*, respectively.

$\tilde{a}_{ij}$, which are independently generated from a binary distribution with probability $\pi = r/p$ with $r = 5$ or 10 for any $i \neq j$, while $\tilde{a}_{jj} = 0$. We set $p \in \{200, 400\}$ and $n \in \{1, \lfloor p^{1/5} \rfloor, \lfloor p^{1/3} \rfloor\}$, where the function $\lfloor x \rfloor$ returns the largest integers not greater than x .

The r and p values together control the dimension and the sparsity of the network structure. We let $d = \lfloor 2p^{1/5} \rfloor$ be the maximum order for $\mathbf{A}$. Here we have chosen very small coefficients to ensure the positive-definiteness of Σ . In addition, the response vector is $\mathbf{Y} = \Sigma^{1/2}\mathbf{Z}$, where each component of $\mathbf{Z}$ is independently and identically simulated from either a

Table 5. The MFL under an exponential distribution for Design 2 ($\times 100$).

β	r	p	n	MAC	SAIC	SBIC	AIC	BIC
$\beta^{(1)}$	5	200	1	31.665	34.181	35.037	37.668	39.586
			$\lfloor p^{1/5} \rfloor$	15.732	16.454	16.369	18.079	18.267
			$\lfloor p^{1/3} \rfloor$	7.938	8.351	8.498	8.989	9.468
		400	1	15.544	17.232	17.525	18.749	19.361
			$\lfloor p^{1/5} \rfloor$	6.021	6.392	6.580	6.872	7.198
			$\lfloor p^{1/3} \rfloor$	2.397	2.506	2.453	2.653	2.609
	10	200	1	29.631	30.776	30.796	33.824	33.871
			$\lfloor p^{1/5} \rfloor$	14.371	14.837	15.341	16.025	17.127
			$\lfloor p^{1/3} \rfloor$	6.630	6.789	6.967	7.218	7.547
		400	1	15.451	15.858	16.652	17.158	18.424
			$\lfloor p^{1/5} \rfloor$	5.627	5.718	5.850	6.017	6.307
			$\lfloor p^{1/3} \rfloor$	2.242	2.266	2.290	2.327	2.427
$\beta^{(2)}$	5	200	1	22.577	24.649	25.606	27.176	29.480
			$\lfloor p^{1/5} \rfloor$	11.560	12.278	13.001	13.189	14.650
			$\lfloor p^{1/3} \rfloor$	5.390	5.704	5.729	5.994	6.129
		400	1	10.597	12.065	13.314	12.916	15.328
			$\lfloor p^{1/5} \rfloor$	4.035	4.387	4.381	4.600	4.570
			$\lfloor p^{1/3} \rfloor$	1.631	1.751	1.698	1.866	1.804
	10	200	1	24.457	25.007	24.816	26.269	26.431
			$\lfloor p^{1/5} \rfloor$	11.745	12.177	12.482	12.967	13.621
			$\lfloor p^{1/3} \rfloor$	5.404	5.608	5.821	5.838	6.230
		400	1	12.708	13.020	13.339	13.780	14.430
			$\lfloor p^{1/5} \rfloor$	4.613	4.734	4.922	4.882	5.271
			$\lfloor p^{1/3} \rfloor$	1.873	1.920	1.967	1.975	2.038
$\beta^{(3)}$	5	200	1	20.949	23.550	23.422	25.406	26.054
			$\lfloor p^{1/5} \rfloor$	10.631	11.572	11.508	12.419	12.514
			$\lfloor p^{1/3} \rfloor$	5.136	5.518	5.399	5.958	5.812
		400	1	9.642	11.283	11.661	12.048	13.004
			$\lfloor p^{1/5} \rfloor$	3.916	4.299	4.249	4.602	4.578
			$\lfloor p^{1/3} \rfloor$	1.661	1.754	1.811	1.880	1.997
	10	200	1	23.945	24.523	23.774	25.440	24.640
			$\lfloor p^{1/5} \rfloor$	11.259	11.799	11.802	12.341	12.515
			$\lfloor p^{1/3} \rfloor$	5.168	5.475	5.571	5.738	5.920
		400	1	12.174	12.623	12.469	13.153	12.994
			$\lfloor p^{1/5} \rfloor$	4.365	4.602	4.697	4.777	4.898
			$\lfloor p^{1/3} \rfloor$	1.797	1.882	1.912	1.968	2.006

Notes: The smallest, second smallest and largest MFLs in each row are highlighted in **bold**, **bold italic** and *italic*, respectively.

standard normal distribution (norm), a standardized exponential distribution (exp), or a mixture of two normal distributions (mix). For the 'mix' case, the first distribution has mean zero and variance 5/9, and the second one has mean zero and variance 5, with the mixture coefficient itself generated from a binomial distribution with probability 0.9, i.e., $Z_j \sim UN(0, 5/9) + (1 - U)N(0, 5)$, where $U \sim \text{Binomial}(0.9)$. We repeat the simulation 500 times under each simulation setting.

Table 6. The MFL under a mixture distribution for Design 2 ($\times 100$).

β	r	p	n	MAC	SAIC	SBIC	AIC	BIC
$\beta^{(1)}$	5	200	1	31.494	34.076	34.723	37.717	39.548
			$\lfloor p^{1/5} \rfloor$	26.376	33.483	32.943	35.667	34.934
			$\lfloor p^{1/3} \rfloor$	7.194	7.459	7.611	8.072	8.545
		400	1	15.549	17.122	17.399	18.847	19.249
			$\lfloor p^{1/5} \rfloor$	7.677	9.692	9.613	10.261	10.414
			$\lfloor p^{1/3} \rfloor$	2.260	2.391	2.382	2.538	2.538
	10	200	1	28.938	29.761	29.665	32.988	32.564
			$\lfloor p^{1/5} \rfloor$	15.583	16.243	16.769	17.558	18.745
			$\lfloor p^{1/3} \rfloor$	6.903	7.142	7.372	7.525	7.956
		400	1	15.148	15.589	16.181	16.905	17.920
			$\lfloor p^{1/5} \rfloor$	5.323	5.427	5.542	5.735	6.076
			$\lfloor p^{1/3} \rfloor$	2.210	2.265	2.260	2.378	2.406
$\beta^{(2)}$	5	200	1	22.950	25.278	26.137	27.904	30.076
			$\lfloor p^{1/5} \rfloor$	21.712	29.250	29.787	30.200	31.647
			$\lfloor p^{1/3} \rfloor$	5.266	5.571	5.599	5.758	5.916
		400	1	10.800	12.232	13.236	13.167	15.212
			$\lfloor p^{1/5} \rfloor$	6.479	8.413	8.338	8.642	8.639
			$\lfloor p^{1/3} \rfloor$	1.557	1.681	1.607	1.803	1.713
	10	200	1	24.239	24.780	24.694	26.198	26.343
			$\lfloor p^{1/5} \rfloor$	12.803	13.285	13.529	14.025	14.709
			$\lfloor p^{1/3} \rfloor$	5.839	6.101	6.318	6.307	6.737
		400	1	12.337	12.662	12.935	13.370	14.056
			$\lfloor p^{1/5} \rfloor$	4.367	4.474	4.641	4.616	4.944
			$\lfloor p^{1/3} \rfloor$	1.864	1.914	1.952	1.970	2.045
$\beta^{(3)}$	5	200	1	21.398	24.199	24.175	26.096	27.105
			$\lfloor p^{1/5} \rfloor$	20.849	28.529	28.268	29.559	29.689
			$\lfloor p^{1/3} \rfloor$	4.998	5.417	5.327	5.758	5.717
		400	1	9.911	11.531	11.626	12.332	12.932
			$\lfloor p^{1/5} \rfloor$	6.340	8.301	8.151	8.602	8.592
			$\lfloor p^{1/3} \rfloor$	1.578	1.672	1.702	1.805	1.879
	10	200	1	23.860	24.410	23.873	25.182	24.778
			$\lfloor p^{1/5} \rfloor$	12.298	12.975	12.941	13.641	13.765
			$\lfloor p^{1/3} \rfloor$	5.570	5.935	6.026	6.129	6.233
		400	1	11.712	12.230	11.979	12.738	12.441
			$\lfloor p^{1/5} \rfloor$	4.071	4.337	4.412	4.542	4.673
			$\lfloor p^{1/3} \rfloor$	1.771	1.877	1.912	1.955	2.007

Notes: The smallest, second smallest and largest MFLs in each row are highlighted in **bold**, **bold italic** and *italic*, respectively.

In all candidate models, $\mathbf{A}^0$ is included and $\mathbf{A}^1$, $\mathbf{A}^2$, $\mathbf{A}^3$ may or may not be included. Thus, we consider a total of $M = 2^3 = 8$ candidate models. Note that $d = 5$ when $p = 200$, and $d = 6$ when $p = 400$, and hence all candidate models are misspecified.

Design 2 (Some candidate models are correctly specified). We set the covariance matrix Σ as $\Sigma = \sum_{k=0}^4 \beta_k \mathbf{A}^k$. We consider three settings of β , which are $\beta^{(1)} = (1, 0.5, 3, 0, 0)^T$, $\beta^{(2)} = (1, 0.5, 0, 0, 3)^T$, and $\beta^{(3)} = (1, 0, 0, 0.5, 3)^T$. Like in Design 1, we always include $\mathbf{I}_p$, while

Table 7. The MSE of the MAC estimator $\widehat{\beta}(\widehat{\mathbf{w}})$ under Design 2.

β	p	n	$r = 5$			$r = 10$		
			Norm	Exp	Mix	Norm	Exp	Mix
$\beta^{(1)}$	200	1	38.156	42.237	42.153	72.417	71.881	79.302
		$\lfloor p^{1/5} \rfloor$	19.555	23.064	30.482	51.202	38.125	41.443
		$\lfloor p^{1/3} \rfloor$	11.337	9.225	13.505	16.759	18.577	20.464
	400	1	27.714	26.913	25.331	73.968	69.991	61.756
		$\lfloor p^{1/5} \rfloor$	7.020	11.045	14.087	23.588	25.602	24.684
		$\lfloor p^{1/3} \rfloor$	4.684	5.221	5.467	10.584	11.597	11.795
$\beta^{(2)}$	200	1	40.017	38.634	40.506	60.281	59.040	64.643
		$\lfloor p^{1/5} \rfloor$	19.370	22.159	28.908	37.361	28.449	32.889
		$\lfloor p^{1/3} \rfloor$	10.276	8.056	12.130	12.684	13.218	15.312
	400	1	24.519	24.133	24.501	58.768	59.701	54.037
		$\lfloor p^{1/5} \rfloor$	6.793	10.555	15.713	20.278	21.214	20.150
		$\lfloor p^{1/3} \rfloor$	4.348	4.918	4.737	8.412	9.281	9.228
$\beta^{(3)}$	200	1	41.005	38.222	41.347	55.315	58.739	64.556
		$\lfloor p^{1/5} \rfloor$	20.543	22.352	28.944	37.371	26.602	31.185
		$\lfloor p^{1/3} \rfloor$	10.724	8.065	12.109	11.796	12.293	14.133
	400	1	26.025	23.938	24.792	58.897	59.446	52.941
		$\lfloor p^{1/5} \rfloor$	6.714	10.575	15.647	19.404	20.189	18.872
		$\lfloor p^{1/3} \rfloor$	4.396	4.902	4.765	8.020	8.883	8.598

Table 8. Model selection criterion values and model averaging weights in the analysis of Chinese airports in 2017.

Models	Model selection criterion values		Weights		
	AIC	BIC	SAIC	SBIC	MAC
(0)	1361.723	1365.148	0.000	0.000	0.000
(0, 1)	1357.781	1364.630	0.000	0.000	0.000
(0,2)	1195.696	1202.546	0.002	0.043	0.593
(0,3)	1266.882	1273.731	0.000	0.000	0.132
(0, 1, 2)	1198.360	1208.635	0.000	0.002	0.000
(0, 1, 3)	1242.146	1252.421	0.000	0.000	0.275
(0,2,3)	1190.737	1201.012	0.019	0.093	0.000
(0, 1, 2, 3)	1182.865	1196.565	0.979	0.861	0.000

Notes: Model (0,2,3) means the model including $\mathbf{A}^0$, $\mathbf{A}^2$ and $\mathbf{A}^3$. The minimum AIC and BIC values are highlighted in bold.

allowing all other components β_k to be zeros. Thus, we consider a total of $M = 2^4 = 16$ candidate models. All other aspects of Design 2 are the same as in Design 1.

We evaluate the performance of various estimators based on the following mean Frobenius loss (MFL):

$$\text{MFL} = \frac{1}{pD} \sum_{\tau=1}^D \|\widehat{\Sigma}^{(\tau)} - \Sigma\|_F^2,$$

(6.1)

Table 9. MFL in the analysis of Chinese airports in 2017.

Method		MAC	SAIC	SBIC	AIC	BIC
MAC	Mean	84.669	149.813	148.348	150.454	149.039
	Median	63.676	75.379	70.190	77.443	73.255
	Standard deviation	94.413	304.467	304.698	304.360	304.712
	Optimal rate	0.526	0.064	0.140	0.172	0.098
SAIC	Mean	137.931	354.814	353.701	355.525	354.004
	Median	94.863	102.324	98.094	105.817	97.684
	Standard deviation	235.541	1403.513	1403.687	1403.396	1403.677
	Optimal rate	0.528	0.070	0.118	0.202	0.082
SBIC	Mean	129.912	305.138	303.586	305.603	304.121
	Median	94.681	103.598	99.498	103.875	101.281
	Standard deviation	271.782	1512.634	1512.853	1512.588	1512.798
	Optimal rate	0.538	0.064	0.120	0.208	0.070
AIC/BIC	Mean	115.859	226.284	224.767	226.895	225.073
	Median	89.564	98.179	95.068	100.434	96.556
	Standard deviation	150.260	650.444	650.764	650.334	650.732
	Optimal rate	0.524	0.074	0.106	0.212	0.084

Notes: The methods listed in the first column are used to estimate the correlation matrix, and then the estimated correlation matrix is used as the truth to generate the corresponding data. The minimum values in every rows are highlighted in bold. Since AIC and BIC select the same model in this real data analysis, the last method is AIC/BIC.

where $\widehat{\Sigma}^{(\tau)}$ is the estimator of Σ obtained by a given method in the τ th trial, and $D = 500$ is the number of replications. In Design 2, where some candidate models are correctly specified, in order to illustrate the consistency of the MAC method shown in Corollary 5.3 we also calculate the mean squared error (MSE) :

$$\text{MSE} = \frac{1}{D} \sum_{\tau=1}^D \|\widehat{\beta}^{(\tau)}(\widehat{\mathbf{w}}^{(\tau)}) - \beta\|^2, \quad (6.2)$$

where $\mathbf{w}^{(\tau)}$ and $\widehat{\beta}^{(\tau)}(\widehat{\mathbf{w}}^{(\tau)})$ are the chosen weight vector and the estimator of β obtained in the τ th trial.

6.2. Results

The MFLs of Design 1 are presented in Tables 1–3 under normal, exponential and mixture distributions, respectively. The MFLs of Design 2 are presented in Tables 4–6, under normal, exponential and mixture distributions, respectively. To facilitate comparisons, we flag the best, second best and worst estimators in each case in bold, bold italic and italic respectively. The MSEs of Design 2 are shown in Table 7.

First, different distributions of $\mathbf{Z}$ have little quantitative effect on the performance of each method. Second, in most cases, evaluated by MFL, MAC is the best, and SAIC/SBIC outperforms AIC/BIC, which is as expected. Tables A1–A6 in the online Supporting Information show the standard errors for the MFLs, for which, we can see that our method performs the best stable in all cases. Finally, we can see from Table 7 that as n or p increases, the MSE of $\widehat{\beta}(\widehat{\mathbf{w}})$ decreases, which reflects the consistency of $\widehat{\beta}(\widehat{\mathbf{w}})$.

7. EMPIRICAL APPLICATION

To illustrate the usefulness of the proposed method, we now apply the proposed MAC method to analyse the data on passenger traffic volume in airports. The dataset consists of yearly passenger traffic volumes at $p = 227$ airports in Mainland China in 2017,¹ obtained from the Civil Aviation Administration of China². The response variable $\mathbf{Y}$ contains the centralized values of the logarithm of passenger traffic volumes at these airports. It is intuitive to see that the number of nonstop flights at an airport affects the passenger traffic volume at the airport. So, we use the matrix of nonstop flights as the adjacency matrix $\mathbf{A}$. Specifically, the (i, j) entry of $\mathbf{A}$ is $1/\sqrt{\#_i\#_j}$ if there is at least one nonstop flight between airports i and j , and 0 otherwise, where $\#_i$ is the number of airports that have a nonstop flight with airport i . We set the largest order of $\mathbf{A}$ to be $d = 3$, and then the numbers of candidate models to be considered is $M = 2^d = 8$.

We consider the five methods studied in Section 6. The values of AIC, BIC and the weights of SAIC, SBIC and MAC are reported in Table 8. We can see that the AIC and BIC methods select the same model $(0, 1, 2, 3)$, which means that $\mathbf{A}^0, \mathbf{A}^1, \mathbf{A}^2$ and $\mathbf{A}^3$ are in the model, and the SAIC and SBIC methods also put almost all the weights on these models. However, the MAC method puts weights on the models $(0, 2)$, $(0, 3)$ and $(0, 1, 3)$.

To compare the different methods, we use the estimated covariance matrix $\hat{\Sigma}_{\text{method}}$ from each method as the true covariance matrix, i.e., we let $\Sigma = \hat{\Sigma}_{\text{method}}$. Then we create $D = 500$ replications of $\mathbf{Y}$ randomly generated from $N(\mathbf{0}, \Sigma)$. Subsequently, we estimate Σ based on the five methods used above and calculate their mean and median of the MFLs across the 500 replications. Specifically,

$$\text{MFL}_{\text{mean}} = \frac{1}{D} \sum_{\tau=1}^D (\|\hat{\Sigma}^{(\tau)} - \Sigma\|_F^2 / p), \quad (7.1)$$

and

$$\text{MFL}_{\text{median}} = \text{median}_{\tau=1,2,\dots,D} (\|\hat{\Sigma}^{(\tau)} - \Sigma\|_F^2 / p), \quad (7.2)$$

where $\hat{\Sigma}^{(\tau)}$ is the estimator of Σ obtained by a given method in the τ th trial. The results are shown in Table 9, which also reports the standard deviation of each method and the optimal rate of each method, defined as the proportion of times in which the method results in the smallest MFL across the D replication trials.

The results show that the MAC method consistently outperforms all other methods, regardless of how the true Σ is selected and which performance evaluation criterion is used. We find this quite remarkable. In terms of the mean and median of the MFLs, the MAC method performs the best among all five estimators. In particular, the mean MFL of the MAC method is about half of that by any other method in all cases. The standard deviation of the MFL of MAC is much smaller than those of others in all cases, which means that the MAC performance is the most stable. In terms of the percentage of times where a method shows optimal performance (optimal rate), the MAC estimator always attains the highest score among the five methods, often with a value of more than 50%, indicating that in over half of the 500 trials, MAC is the champion. Depending on the means and medians, SAIC/SBIC performs slightly better than AIC/BIC.

¹ In 2017, there were 229 civil airports in Mainland China, 227 of which had scheduled flights.

² <http://news.carnoc.com/list/438/438561.html?from=timeline>.

8. CONCLUDING REMARKS

In this article, we used the covariance regression network model, which treats the covariance as a polynomial function of the symmetric adjacency matrix. The model averaging method was used to estimate the high-dimensional covariance matrix, and the candidate models were constructed through different orders of a polynomial function. The optimal weights were obtained by minimizing the newly proposed Mallows-type model averaging criterion. We proved the asymptotic optimality and consistency of the resulting MAC estimators for different situations. Both numerical simulations and a case study on Chinese airport network structure data were conducted to demonstrate the validity of the proposed approach.

It is worth noting that our method combines polynomials of $\mathbf{A}$. If the true covariance matrix is far from the polynomial form, the MAC method may not yield an accurate estimate. In addition, when conjecturing that Σ is a combination of banded and Toeplitz-type matrices, these matrices should be included in candidate models, but then there is no single orthogonal matrix to diagonalize them simultaneously. In such a case, MAC will not be applicable. How to develop a model averaging method for estimating this kind of covariance matrix warrants further study.

ACKNOWLEDGEMENTS

We thank the referee, the associate editor, the co-editor Victor Chernozhukov and Prof. Hansheng Wang for many constructive comments and suggestions. We thank Prof. Wei Lan for providing his codes. Zhang, the corresponding author, was supported by the National Key R&D Program of China (2020AAA0105200), the National Natural Science Foundation of China (grant nos. 71925007, 11688101 and 71631008), the Beijing Academy of Artificial Intelligence, and the Youth Innovation Promotion Association of the Chinese Academy of Sciences. Ma was supported by grants from the National Science Foundation and the National Institutes of Health. Zou was supported by the National Natural Science Foundation of China (grant nos. 11971323 and 12031016). All errors remain the authors.

REFERENCES

- Ando, T. and K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* 109, 254–65.
- Andrews, D. W. K. (1991). Asymptotic optimality of generalized c_l , cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* 47, 359–77.
- Bilmes, J. A. (2000). Factored sparse inverse covariance matrices. *IEEE International Conference* 2, 1009–12.
- Buckland, S. T., K. P. Burnham and N. H. Augustin (1997). Model selection: an integral part of inference. *Biometrics* 53, 603–18.
- Campbell, J. Y., A. W. Lo, A. C. Mackinlay and R. F. Whitelaw (1998). The econometrics of financial markets. *Macroeconomic Dynamics* 2, 559–62.
- Chen, X. H. and T. G. Conley (2001). A new semi-parametric spatial model for panel time series. *Journal of Econometrics* 105, 59–83.
- Fan, J., Y. Fan and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* 147, 186–97.

- Fan, J. and H. Peng (2004). On nonconcave penalized likelihood with diverging number of parameters. *Annals of Statistics* 32, 928–61.
- Friedman, J. H., T. Hastie and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–41.
- Gao, Y., X. Zhang, S. Wang and G. Zou (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics* 192, 139–51.
- Hansen, B. E. (2007). Least squares model averaging. *Econometrica* 75, 1175–89.
- Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics* 5, 495–530.
- Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics* 167, 38–46.
- He, X. and Q. M. Shao (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis* 73, 120–35.
- Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: why imposing the wrong constraints helps. *Journal of Finance* 58, 1651–84.
- Lan, W., Z. Fang, H. Wang and C. Tsai (2018). Covariance matrix estimation via network structure. *Journal of Business and Economic Statistics* 36, 359–69.
- Leung, G. and A. Barron (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on information theory* 52, 3396–410.
- Liang, H., G. Zou, A. T. K. Wan and X. Zhang (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106, 1053–66.
- Liu, Q. and R. Okui (2013). Heteroskedasticity-robust C_p model averaging. *Econometrics Journal* 16, 463–72.
- Liu, Q., R. Okui and A. Yoshimura (2016). Generalized least squares model averaging. *Econometric Reviews* 35, 1692–752.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance* 7, 77–91.
- Millar, C. P., E. Jardim, F. Scott, G. C. Osio, I. Mosqueira and N. Alzoriz (2014). Model averaging to streamline the stock assessment process. *ICES Journal of Marine Science* 72, 93–98.
- Wan, A. T. K. and X. Zhang (2009). On the use of model averaging in tourism research. *Annals of Tourism Research* 36, 525–32.
- Wan, A. T. K., X. Zhang and G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156, 277–83.
- Whittle, P., (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications* 5, 331–5.
- Yuan, Z. and Y. Yang (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association* 100, 1202–14.
- Zhang, X., (2010). Model averaging and its applications. Ph.D. Thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences.
- Zhang, X., A. T. K. Wan and G. Zou (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* 174, 82–94.
- Zhang, X. and W. Wang (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica* 29, 693–718.
- Zhu, R., G. Zou and X. Zhang (2018). Model averaging for multivariate multiple regression models. *Statistics* 52, 205–27.
- Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* 37, 1733–51.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Victor Chernozhukov handled this manuscript.