Annals of the Institute of Statistical Mathematics (2020) 72:65–90
https://doi.org/10.1007/s10463-018-0686-5

A two-stage sequential conditional selection approach to sparse high-dimensional multivariate regression models

Zehua Chen[1] · **Yiwei Jiang**[1]

Received: 30 October 2017 / Revised: 13 July 2018 / Published online: 23 August 2018

Abstract
In this article, we deal with sparse high-dimensional multivariate regression models. The models distinguish themselves from ordinary multivariate regression models in two aspects: (1) the dimension of the response vector and the number of covariates diverge to infinity; (2) the nonzero entries of the coefficient matrix and the precision matrix are sparse. We develop a two-stage sequential conditional selection (TSCS) approach to the identification and estimation of the nonzeros of the coefficient matrix and the precision matrix. It is established that the TSCS is selection consistent for the identification of the nonzeros of both the coefficient matrix and the precision matrix. Simulation studies are carried out to compare TSCS with the existing state-of-the-art methods, which demonstrates that the TSCS approach outperforms the existing methods. As an illustration, the TSCS approach is also applied to a real dataset.

Keywords Conditional models · Multivariate regression · Precision matrix · Selection consistency · Sequential procedure · Sparse high-dimensional model

1 Introduction

Consider the following model:

$$Y = X\mathcal{B} + E, \tag{1}$$

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0686-5) contains supplementary material, which is available to authorized users.

✉ Zehua Chen
stachenz@nus.edu.sg

Yiwei Jiang
yiweijiang@u.nus.edu

[1] Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

where Y is a $n \times q$ matrix of n independent observations on a q-dimensional response vector, X is a $n \times p$ matrix of observations on p covariates, $\mathcal{B}$ is a $p \times q$ matrix of unknown regression coefficients and E is a $n \times q$ matrix of random variables whose rows are independent identically distributed as a q-variate normal distribution $N_q(\mathbf{0}, \Sigma)$. The inverse matrix $\Omega = \Sigma^{-1}$, which is of more practical interest, is usually referred to as the precision matrix. It is assumed that p and q are large; that in the asymptotic setting, they are allowed to diverge to infinity in a certain order of the sample size n; and that $\mathcal{B}$ and Ω are sparse, that is, only a few of their entries are nonzero. The nature of high dimensionality and sparsity distinguish model (1) from traditional multivariate regression models. We refer to model (1) as a sparse high-dimensional multivariate (SHM) regression model.

The SHM regression model arises in many important scientific fields such as genetics, medicine and econometrics. A few examples follow. In genetic studies, experiments have now been routinely carried out to obtain both genetic marker data and gene expression data on the same subjects. Both the number of markers and the number of genes are much larger than the number of subjects. The geneticists are interested in identifying the markers which regulate the gene expression levels, which is referred to as eQTL mapping in genetic studies, as well as identifying the conditional dependency among the genes. The data can be well modeled by the SHM regression model. The gene expression levels are treated as multi-response values and the genotypes of the markers are treated as covariates. The precision matrix describes the conditional relationship among the genes. Two genes are conditionally dependent if and only if the corresponding entry in the precision matrix is nonzero. Since the markers which regulate a particular gene are few and the number of genes which are conditionally dependent with a particular gene is also small, the matrices $\mathcal{B}$ and Ω in the regression model describing the data are sparse. In cancer research, the medical scientists are interested in investigating the influence of DNA copy numbers on RNA transcript levels to identify biomarkers for clinical purpose. The data of DNA copy number and RNA transcript level have the same structure and nature as in the above example. In financial econometrics, the future returns of stocks are predicted from the basis of their historical performance. The data are usually analyzed by a vector autoregressive model. Given the nature of the stock data, the model is a special case of the SHM regression model. Because of the wide range of its application, the SHM regression model has attracted the attention of many researchers. For recent developments of the research on the SHM regression model, see, to name but a few, Turlach et al. (2005), Yuan et al. (2007), Peng et al. (2010), Rothman et al. (2010), Obozinski et al. (2011), Yin and Li (2011), Chen and Huang (2012), Lee and Liu (2012), Wang (2015), etc.

In practical problems, there are two major purposes of using the SHM regression model: (i) to identify variables which are causal factors of the variation in the response variables, this amounts to identifying and estimating the nonzero entries of the coefficient matrix $\mathcal{B}$; (ii) to detect the inter-relation mechanism among the response variables, this amounts to identifying and estimating the nonzero entries of the precision matrix. In a particular problem, the emphasis is either one of the two purposes or both. The methods available in the literature for dealing with the SHM regression model can be roughly classified into two categories. The methods in the first category concentrate on the inference of $\mathcal{B}$, ignoring Ω. The methods in the second category

deal with simultaneously the inference of $\mathcal{B}$ and Ω. In the following, we give a brief review on these methods.

In the first category, a naive approach is to apply well-developed regularized methods for univariate regression models to each marginal model of (1) such as the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) and its variants, the smoothly clipped absolute derivation (SCAD) proposed by Fan and Li (2001) and so on. More sophisticated approaches are the multivariate version of the regularized methods. Essentially, the multivariate regularized methods aim to minimize $\mathrm{Tr}[(Y - X\mathcal{B})^\top (Y - X\mathcal{B})]$ by imposing some constraints on $\mathcal{B}$ such as $C(\mathcal{B}) \le c$ for some constant c, where $C(\mathcal{B})$ is a function of $\mathcal{B}$. In the dimension reduction method proposed in Yuan et al. (2007), $C(\mathcal{B}) = \sum_{j=1}^{\min(p,q)} \sigma_j(B)$ where $\sigma_j(B)$ is the jth singular value of $\mathcal{B}$. In the reduced rank method proposed in Chen and Huang (2012), $C(\mathcal{B}) = \mathrm{RANK}(\mathcal{B})$. Turlach et al. (2005) considers a penalized least squares approach, which is equivalent to taking $C(\mathcal{B}) = \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_\infty$, where $\boldsymbol{\beta}_j$ is the jth row of $\mathcal{B}$. Obozinski et al. (2011) considers a version of grouped LASSO which they referred to as support union recovery, which is equivalent to taking $C(\mathcal{B}) = \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2$. Peng et al. (2010) uses the sparse group LASSO penalty $\lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_1 + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2$, which is equivalent to taking $C(\mathcal{B}) = (\sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_1, \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2)$ and $c = (c_1, c_2)$. The methods in the first category have an advantage that the assumption of normality is not required. However, since these methods do not make use of the information contained in the correlation among the response variables, they lose a certain efficiency by the principle of sufficiency.

The methods in the second category are based on the assumption of multivariate normal distribution of the response variables. A penalized likelihood method is investigated by Rothman et al. (2010) and Yin and Li (2011). The method minimizes

$$-2\log L(\mathcal{B}, \Omega) + \lambda_1 \sum_{j \neq k} |\omega_{jk}| + \lambda_2 \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}|, \quad \text{w.r.t. } \mathcal{B} \text{ and } \Omega, \tag{2}$$

where $L(\mathcal{B}, \Omega)$ is the joint likelihood function of $\mathcal{B}$ and Ω. The method is referred to as multivariate regression with covariance estimation (MRCE) in Rothman et al. (2010) and as conditional Gaussian graphical model (cGGM) in Yin and Li (2011). In the graphical model literature, the SHM model with $\mathcal{B} = 0$ is called a Gaussian graphical model (GGM) where Ω represents a graph with the response variables as nods. The name cGGM reflects the fact that Ω represents the conditional graph given the covariates. Lee and Liu (2012) considered an extension of the penalized likelihood method by imposing certain weights on the L_1 penalties in (2). They considered three versions of the extension: plug-in weighted Lasso (PWL), plug-in weighted graphical Lasso (PWGL) and doubly penalized maximum likelihood (DML). In the first two versions, a given estimator of one of $\mathcal{B}$ and Ω is plugged-in and (2) is minimized with respect to the other. In the third version, (2) is minimized simultaneously with respect to $\mathcal{B}$ and Ω. Note that, without the penalty on $\mathcal{B}$, the estimate of $\mathcal{B}$ is the ordinary least squares estimate which does not depend on Ω. By imposing a penalty on $\mathcal{B}$, it effects a shrinkage of the ordinary least squares estimate. The Ω enters the scene in a role to affect the shrinkage, see formula (2.2) in Rothman et al. (2010). However, how this

effect on shrinkage improves the inference on $\mathcal{B}$ is not theoretically nor intuitively clear. Wang (2015) treated the simultaneous estimation of $\mathcal{B}$ and Ω in a conditional framework and proposed a method called aMCR (multivariate conditional regression with adaptive Lasso). For each j, the aMCR method estimates simultaneously the jth column $\boldsymbol{\beta}_j$ of $\mathcal{B}$ and the jth column $\boldsymbol{\xi}_j$ of a matrix Ξ which has a one-to-one correspondence with Ω by minimizing

$$\|\boldsymbol{y}_j - X\boldsymbol{\beta}_j - (Y_{j^-} - X\mathcal{B}_{j^-})\boldsymbol{\xi}_j\|_2^2 + \lambda_1 \sum_{k=1}^{p} u_{jk}|\beta_{kj}| + \lambda_2 \sum_{k\neq j} v_{jk}|\xi_{kj}|, \qquad (3)$$

where $\boldsymbol{y}_j$ is the jth column of Y, Y_{j^-} and $\mathcal{B}_{j^-}$ are matrices obtained from Y and $\mathcal{B}$, respectively, by omitting the jth column of the original matrix, and u_{jk} and v_{jk} are certain weights. Ideally, the information on Ω should be fully used when $\mathcal{B}$ is estimated and vice versa. However, the above approach does not fully use the information of Ω for the estimation of $\mathcal{B}$ and does not fully use the information of $\mathcal{B}$ for the estimation of Ω either. Further discussion on this point will be given later.

In this article, we propose a two-stage alternative sequential conditional selection (TSCS) procedure. The main consideration of the procedure is to make use of the correlation information fully to enhance the efficiency for the identification and estimation of the nonzeros of $\mathcal{B}$ and Ω. In the first stage, a sequential Lasso (SLasso) approach developed in Luo and Chen (2014b) is applied to each marginal model of (1) to yield the set of nonzeros of $\mathcal{B}$. The nonzero set is used to fit a regression model to the response matrix Y and to obtain a residual matrix. The residual matrix is treated as the response matrix of a Gaussian graphical model, and a GGM approach is applied to obtain an initial estimate of Ω. In the second stage, the correlation information is incorporated into the procedure by using the initial estimates obtained in the first stage for the conditional models on $\mathcal{B}$, the SLasso is applied to the conditional models to produce a updated estimate of $\mathcal{B}$, and the second step of the first stage is repeated with the updated estimate of $\mathcal{B}$ to give an updated estimate of Ω. We carry out theoretical and simulation studies to investigate whether or not the proposed approach can achieve selection consistency for both $\mathcal{B}$ and Ω and whether or not it can perform better than correlation-unadjusted approaches. The selection consistency for both $\mathcal{B}$ and Ω is rigorously established. A theoretical result suggesting the efficiency of the TSCS procedure over correlation-unadjusted approaches is derived. In a comprehensive simulation study, the TSCS is compared with the state-of-the-art methods in the literature, which demonstrates that the TSCS outperforms those existing methods.

The selection consistency of SLasso for an univariate high-dimensional regression model is established in Luo and Chen (2014b). In the TSCS procedure, to establish the selection consistency for Ω, the uniform selection consistency for a group of univariate models with group size diverging to infinity is required. In order to establish this uniform selection consistency, convergence rates must be determined for each single model, which is not trivial. Though the estimation of the precision matrix is well studied in the field of Gaussian graphical models, satisfactory methods for the case that the response variables depend on regression means are yet to be developed. The major contribution of this article is threefold: (i) the use of the conditional mechanism of multivariate normal distribution to enhance the efficiency of the estimation of $\mathcal{B}$, (ii) the

justification of using the residual matrix to estimate Ω in the framework of a Gaussian graphical model and (iii) the establishment of the theoretical results mentioned in the last paragraph.

The rest of the article is arranged as follows. The details of the development of the TSCS procedure are given in Sect. 2. The theoretical properties of TSCS are established in Sect. 3. Simulation studies are reported in Sect. 4. A real data example is provided in Sect. 5. Technical details are provided in a supplementary document.

2 The two-stage alternative sequential conditional selection procedure

We first give the notations to facilitate our discussion. Capital letters are used to denote matrices of random variables or covariates, e.g., Y, X. Bold lowercase letters are used to denote vectors, e.g., $\boldsymbol{y}$, $\boldsymbol{x}$. Scripted or roman letters are used to denote parameters, e.g., $\mathcal{B}$, Σ, β. A matrix with its jth column deleted is denoted by the notation of that matrix with a subscript j^-, e.g., Y_{j^-}, $\mathcal{B}_{j^-}$. An index set consists of a single j which is simply denoted by j. An index set consists of all indices but j is simply denoted by j^-. Let s be a general index set; the submatrix consists of the columns of a matrix with indices in s which is denoted by the notation of that matrix followed by (s), e.g., $X(s)$, $R(s)$. Let s and t be two index sets; the submatrix consists of the rows with indices in s and columns with indices in t of a matrix which is denoted by the notation of that matrix subscripted by st, e.g., Σ_{jj^-}. The projection matrix formed by the columns of a matrix with indices in s is denoted by $H(s)$ with the notation of that matrix as its superscript, e.g., $H^X(s) = X(s)[X(s)^\top X(s)]^{-1}X(s)^\top$. Let Y, X, $\mathcal{B}$ and E be as given in (1). Denote by $\boldsymbol{y}_j$, $\boldsymbol{x}_j$, $\boldsymbol{\beta}_j$ and $\boldsymbol{e}_j$, respectively, the jth column of Y, X, $\mathcal{B}$ and E. Let s_{0j} be the index set of the nonzero components of $\boldsymbol{\beta}_j$, i.e., $s_{0j} = \{k : 1 \le k \le p, \beta_{jk} \ne 0\}$. Denote the size of a set s by $|s|$. Let $p_{0j} = |s_{0j}|$.

The two-stage alternative sequential conditional selection (TSCS) procedure is motivated by the following consideration. We mentioned in the previous section that the naive approach for identifying and estimating the nonzeros of $\mathcal{B}$ is to apply the methods for univariate models to the marginal models of (1) given as follows:

$$\boldsymbol{y}_j = X\boldsymbol{\beta}_j + \boldsymbol{e}_j,\ \boldsymbol{e}_j \sim N(0, \sigma_j^2 I), \quad j = 1, \ldots, q, \tag{4}$$

where σ_j^2 is the jth diagonal entry of Σ. The naive approach is not efficient because it does not make use of the correlation information among $\boldsymbol{y}_j$'s. The variation of the error term in a marginal model can be attributed to two sources: the pure random errors and the correlation of the response vector with the other response vectors. If the variation caused by the correlation with the other response vectors is eliminated, the error variance will be reduced, and a better identification and estimation of the nonzeros of $\mathcal{B}$ can be achieved. It is then natural to consider, for a fixed j, the conditional model of $\boldsymbol{y}_j$ given the other response vectors. By the theory of multivariate normal distributions, $\boldsymbol{y}_j$ has a conditional normal distribution with mean $X\boldsymbol{\beta}_j + (Y_{j^-} - X\mathcal{B}_{j^-})\Sigma_{j^-j^-}^{-1}\Sigma_{j^-j}$ and variance $\tau_j^2 I$ where $\tau_j^2 = \sigma_j^2 - \Sigma_{jj^-}\Sigma_{j^-j^-}^{-1}\Sigma_{j^-j}$. Let $\tilde{\boldsymbol{y}}_j = \boldsymbol{y}_j - (Y_{j^-} - X\mathcal{B}_{j^-})\Sigma_{j^-j^-}^{-1}\Sigma_{j^-j}$. Then, we have the following models:

$$\tilde{\boldsymbol{y}}_j = X\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j,\ \boldsymbol{\epsilon}_j \sim N(0, \tau_j^2 I), \quad j = 1, \ldots, q. \tag{5}$$

Obviously, $\tau_j^2 \leq \sigma_j^2$, the equality holds if and only if the response variables are independent. If we have initial estimates of Ω (hence of Σ) and $\mathcal{B}$, we can substitute the initial estimates of Ω and $\mathcal{B}$ into $\tilde{\boldsymbol{y}}_j$, a better identification and estimation of $\boldsymbol{\beta}_j$ based on (5) can be expected.

For the inference on Ω, if $Y - X\mathcal{B}$ is observable, i.e., $\mathcal{B}$ is known, a naive estimate of Σ is given by the sample covariance matrix $S_n = \frac{1}{n}(Y - X\mathcal{B})^\top [I - \mathbf{1}\mathbf{1}^\top/n](Y - X\mathcal{B})$, where $\mathbf{1}$ is a vector of all elements 1, and an estimate of Ω might be given by S_n^{-1}. However, in the high-dimensional case where q is larger than n, S_n is non-invertible. The estimation of Ω becomes a challenging problem. The estimation of Ω in this context constitutes the analysis of the so-called Gaussian graphical models (GGM). There are roughly two major methodologies for GGM in the literature. The first one, which was initiated in Meinshausen and Bühlmann (2006), is called neighborhood detection. This methodology transfers the inference on Ω to the inference on the coefficients of q conditional univariate regression models. Various methods for univariate regression models including LASSO, scaled LASSO, Dantzig selector and sequential Lasso have been applied in this context, see Meinshausen and Bühlmann (2006), Yuan (2010), Sun and Zhang (2012), Peng et al. (2009) and Luo and Chen (2014a). The second methodology is to maximize a penalized profile likelihood function of Ω with various penalty functions, which results in the methods called GLasso (Friedman et al. 2008), G-Scad (Fan et al. 2009) and adaptive GLasso (Zhou et al. 2009), etc. In a GGM, the response matrix is assumed to have mean zero or a constant mean. In order to enable these methods to be used in our current context, the response matrix Y must be adjusted by its regression mean $X\mathcal{B}$. If Y is properly adjusted by taking into account its regression mean, then the GGM methods can be applied to the adjusted matrix.

The TSCS procedure consists of two stages. In the first stage, a naive approach is applied to the marginal (unconditional) models (4) to identify the nonzeros of $\mathcal{B}$, the nonzero set is used to fit a regression model to the response matrix Y and to obtain a residual matrix, and then a GGM method is applied to the residual matrix to obtain an initial estimate of Ω; in the second stage, the initial estimates are substituted into the $\tilde{\boldsymbol{y}}_j$'s in the conditional models (5) and an more efficient estimate of $\mathcal{B}$ is obtained from the conditional models; further, the procedure for estimating Ω in the first stage is repeated with an updated residual matrix to obtain a better estimate of Ω. For the identification and estimation of the nonzeros of $\mathcal{B}$, we adopt the approach of SLasso proposed in Luo and Chen (2014b). For the estimation of Ω, we adopt the neighborhood detection methodology and apply a pairwise version of SLasso dubbed as SSPS considered in Jiang and Chen (2016). In the following, we give the details of the TSCS procedure. For the sake of convenience, the estimation of $\mathcal{B}$ and Ω in both stages is referred to as, respectively, a $\mathcal{B}$-step and a Ω-step.

The method for $\mathcal{B}$-step. For each j, let $\hat{\boldsymbol{y}}_j$ denote the response vector. In the first stage, $\hat{\boldsymbol{y}}_j = \boldsymbol{y}_j$, in the second stage, $\hat{\boldsymbol{y}}_j = \boldsymbol{y}_j - (Y_{j^-} - X\hat{\mathcal{B}}_{j^-})\hat{\Sigma}_{j^-j^-}^{-1}\hat{\Sigma}_{j^-j} \equiv \hat{\boldsymbol{y}}_j(\hat{\mathcal{B}}_{j^-}, \hat{\boldsymbol{\xi}}_j)$, where $\hat{\mathcal{B}}_{j^-}$ and $\hat{\boldsymbol{\xi}}_j = \hat{\Sigma}_{j^-j^-}^{-1}\hat{\Sigma}_{j^-j}$ are initial estimates. The SLasso is used for each of the following models separately:

$$\hat{\boldsymbol{y}}_j = X\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, q.$$

The SLasso is a sequential procedure for univariate high-dimensional linear models which is equivalent to the procedure as follows. At each step, a current residual vector is obtained by fitting a linear model to the covariates which have already been selected, among the remaining covariates, the one having the largest correlation with the current residual is taken for the consideration of selection and is evaluated by EBIC (Chen and Chen 2008). For details of SLasso, the reader is referred to Luo and Chen (2014b). In the following, we describe the algorithm of the SLasso for the $\mathcal{B}$-step. Let the columns of X be standardized to have mean zero and squared norm n. Let $S = \{1, \ldots, p\}$ and s be any subset of S. Denote the algorithm by $\mathcal{B}(\hat{Y}, X)$ where $\hat{Y}$ and X are its inputs. The algorithm is as follows.

Algorithm $\mathcal{B}(\hat{Y}, X)$

For $j = 1, \ldots, q$, do

Step 1: Compute $\boldsymbol{x}_k^\top \hat{\boldsymbol{y}}_j$ for $k \in S$ and identify $s_{\text{TEMP}} = \{k : |\boldsymbol{x}_k^\top \hat{\boldsymbol{y}}_j| = \max_{l \in S} |\boldsymbol{x}_l^\top \hat{\boldsymbol{y}}_j|\}$. Let $s_{*1} = s_{\text{TEMP}}$ and compute EBIC(s_{*1}).

Step m $(m \geq 2)$**:** Compute $\boldsymbol{x}_k^\top \hat{\boldsymbol{\epsilon}}$ for $k \in s_{*m-1}^c$, where $\hat{\boldsymbol{\epsilon}} = [I - H^X(s_{*m-1})]\hat{\boldsymbol{y}}_j$, and identify $s_{\text{TEMP}} = \{k : |\boldsymbol{x}_k^\top \hat{\boldsymbol{\epsilon}}| = \max_{l \in s_{*m-1}^c} |\boldsymbol{x}_l^\top \hat{\boldsymbol{\epsilon}}|\}$. Let $s_{*m} = s_{*m-1} \cup s_{\text{TEMP}}$, and compute EBIC$(s_{*m})$. If EBIC$(s_{*m})$ > EBIC(s_{*m-1}), stop and set $\hat{s}_{0j} = s_{*m-1}$; otherwise, continue.

Output $\hat{s}_{0j}, \quad j = 1, \ldots, q.$

The form of the EBIC in the above algorithm will be given and discussed later.

The method for the Ω*-step*. Let $\tilde{Z} = Y - X\mathcal{B}$ and denote by $\tilde{\boldsymbol{z}}_j$ the jth column of $\tilde{Z}$. By the theory of multivariate normal distribution, $\tilde{\boldsymbol{z}}_j$ follows the conditional model below:

$$\tilde{\boldsymbol{z}}_j = \tilde{Z}_{j^-}\boldsymbol{\xi}_j + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, q, \tag{6}$$

where $\boldsymbol{\xi}_j = \Sigma_{j^-j^-}^{-1}\Sigma_{j^-j}$. Let $\Xi = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_q)$. Denote by ω_{jk} and ξ_{jk}, respectively, the (j, k)th entry of Ω and Ξ. It is well-known that

$$\xi_{jk} = -\omega_{jk}\tau_j^2, \quad \text{or} \quad \omega_{jk} = -\xi_{jk}/\tau_j^2. \tag{7}$$

The above relationship implies that the identification and estimation of the nonzeros of Ω are equivalent to the identification and estimation of the nonzeros of Ξ in model (6). The inference on Ω through the inference on Ξ is referred to as the methodology of neighborhood detection.

In the Ω-step, we replace $\tilde{\boldsymbol{z}}_j$ by $\boldsymbol{z}_j = \boldsymbol{y}_j - X(\hat{s}_{0j})\hat{\boldsymbol{\beta}}_j(\hat{s}_{0j})$. Denote by Z the matrix consisting of the columns $\boldsymbol{z}_j$'s. For any pair (j, k), the relationship between $\boldsymbol{z}_j$ and $\boldsymbol{z}_k$ mimics that between $\tilde{\boldsymbol{z}}_j$ and $\tilde{\boldsymbol{z}}_k$. In particular, $\hat{\sigma}_{jk} = \frac{1}{n}E(\boldsymbol{z}_j^\top \boldsymbol{z}_k) \to \frac{1}{n}E(\tilde{\boldsymbol{z}}_j^\top \tilde{\boldsymbol{z}}_k) = \sigma_{jk}$ uniformly, and $\hat{\sigma}_{jk} = 0$ if and only if $\sigma_{jk} = 0$. Thus, we can approximate model (6) by replacing $\tilde{Z}$ with Z and identify the nonzeros of Ω using the approximated model. The rigorous justification is delayed to the next section.

There is an intrinsic symmetry in the entries of Ξ, that is, $\text{sign}(\xi_{jk}) = \text{sign}(\xi_{kj})$. The SSPS is a procedure which takes this symmetry into account for the identification of the nonzeros of Ξ. The procedure is as follows. First, each z_j is scaled with its estimated conditional variance obtained by using a scaled Lasso algorithm proposed in Sun and Zhang (2013). The q models in (6) are combined into a single model with design matrix $\mathcal{Z} = \text{Diag}(Z_{1^-}, \ldots, Z_{q^-})$ and response vector $(z_1^\top/\hat{\tau}_1, \ldots, z_q^\top/\hat{\tau}_q)^\top$, where $\hat{\tau}_j$ is the square root of the estimated variance for the jth model in (6). The column of Z_{j^-} corresponding to ξ_{jk} is paired off with the column of Z_{k^-} corresponding to ξ_{kj} and the pairs are sequentially selected. At each step of the procedure, the response vector is fitted to the columns of $\mathcal{Z}$ which have already been selected to obtain a current residual vector, and the residual vector is projected onto the space spanned by each pair of the remaining column pairs, the pair which results in the largest L_2-norm of the projection is selected next. The EBIC is used as the stopping rule of the procedure. For more details, the reader is referred to Jiang and Chen (2016). Let $\mathcal{T}$ be a subset of $\{(j,k) : k \neq j, 1 \leq j, k \leq q\}$ and $\mathcal{T}^c$ its complement. Suppose $\mathcal{T}$ is symmetric in the sense that if $(j,k) \in \mathcal{T}$ then $(k,j) \in \mathcal{T}$. Let $t_j = \{k : (j,k) \in \mathcal{T}\} = \{k : (k,j) \in \mathcal{T}\}$. For any pair $(j,k) \in \mathcal{T}^c$, define

$$r_{jk}^2(\mathcal{T}) = \frac{[z_k^\top[I - H^Z(t_j)]z_j]^2}{\hat{\tau}_j^2 z_k^\top z_k} + \frac{[z_j^\top[I - H^Z(t_k)]z_k]^2}{\hat{\tau}_k^2 z_j^\top z_j}. \tag{8}$$

The $r_{jk}^2(\mathcal{T})$ defined above is in fact the squared L_2-norm of the projection of the residual vector determined by $\mathcal{T}$ onto the space spanned by the columns corresponding to the jth column in Z_{k^-} and the kth column in Z_{j^-}. Denote the algorithm for implementing the above procedure by $\Omega(Z)$ where Z is its input. The algorithm is given as follows.

Algorithm $\Omega(Z)$

Initial step: Set $\mathcal{T} = \emptyset$.
Selection step: For $(j,k) \in \mathcal{T}^c$, compute $r_{jk}^2(\mathcal{T})$ and identify

$$\mathcal{T}_{\text{TEMP}} = \left\{(j,k) : r_{jk}^2(\mathcal{T}) = \max_{(l,m)\in\mathcal{T}^c} r_{lm}^2(\mathcal{T})\right\}$$

Let $\mathcal{T}_{\text{NEW}} = \mathcal{T} \cup \mathcal{T}_{\text{TEMP}}$. Compare $\text{EBIC}(\mathcal{T}_{\text{NEW}})$ with $\text{EBIC}(\mathcal{T})$. If $\text{EBIC}(\mathcal{T}_{\text{NEW}}) < \text{EBIC}(\mathcal{T})$, go to the updating step; otherwise, output $\mathcal{T}$.
Updating step: Update $\mathcal{T}$ to $\mathcal{T} = \mathcal{T}_{\text{NEW}}$, go to the selection step.
Final step: Compute $\hat{\Omega} = \text{argmax}_{\omega_{jk}=0:(j,k)\in\mathcal{T}^c} L(Y, \Omega)$, where $L(Y, \Omega)$ is the profile likelihood of Ω while $\mathcal{B}$ is confined to the identified nonzeros.

The general form of the EBIC for a particular model M developed in Chen and Chen (2008) is given by

$$\text{EBIC}(M) = -2\log L_n(\hat{M}) + |M|\ln n + 2\gamma \ln \tau(\mathcal{S}_M),$$

where $L_n(\hat{M})$ is the maximum likelihood of the model, $|M|$ is the number of parameters of the model and $\tau(\mathcal{S}_M)$ is the size of the class of models containing M. For the

EBIC(s) in the $\mathcal{B}$-step, $M = s$ and $\tau(\mathcal{S}_M) = \binom{p}{|s|}$. For small $|s|$ (relative to p), $\binom{p}{|s|}$ is approximated by $p^{|s|}$. This gives the form of EBIC(s) as

$$\text{EBIC}(s) = n \ln(\|[I - H^X(s)]\hat{\mathbf{y}}_j\|_2^2) + |s| \ln n + 2|s|\gamma \ln p. \tag{9}$$

For the EBIC($\mathcal{T}$) in the Ω-step, $M = \mathcal{T}$ and $\tau(\mathcal{S}_M) = \binom{q(q-1)/2}{|\mathcal{T}|/2}$ which is approximated by $q^{|\mathcal{T}|}$. This yields the of form of EBIC($\mathcal{T}$) as

$$\text{EBIC}(\mathcal{T}) = n \sum_{j=1}^{q} \ln(\|[I - H^R(t_j)]\mathbf{z}_j\|_2^2/\hat{\tau}_j^2) + |\mathcal{T}| \ln n + 2\gamma|\mathcal{T}| \ln q. \tag{10}$$

In the theory of EBIC, there is a range of γ so that the EBIC is selection consistent. In the context of (9), the lower bound of the range is $1 - \frac{\ln n}{2 \ln p}$. In the context of (10), the lower bound of the range is $1 - \frac{\ln n}{2 \ln[q(q-1)/2]} \approx 1 - \frac{\ln n}{4 \ln q}$. It is recommended in Luo and Chen (2014b) to choose the value of γ slightly bigger than its lower bound. For the rationale of the recommendation, see page 1234 of Luo and Chen (2014b). But in a finite sample problem, it is as good to choose the lower bound as to choose a slightly larger value. Therefore, in our algorithms above, we simply take γ to be its lower bound.

We now give the complete algorithm for the TSCS procedure below.

TSCS algorithm

Stage I:

(Ia) Call algorithm $\mathcal{B}(Y, X)$ and extract the output $\hat{s}_{0j}, \quad j = 1, \ldots, q$.

(Ib) For $j = 1, \ldots, q$, compute $\hat{\boldsymbol{\beta}}_j = (\hat{\boldsymbol{\beta}}_j(\hat{s}_{0j})^\top, \mathbf{0}^\top)^\top$, where $\hat{\boldsymbol{\beta}}_j(\hat{s}_{0j}) = [X(\hat{s}_{0j})^\top X(\hat{s}_{0j})]^{-1} X(\hat{s}_{0j})^\top \mathbf{y}_j$, and compute the residual matrix Z.

(Ic) Call algorithm $\Omega(Z)$ and extract the output $\hat{\Omega}$.

Stage II:

(IIa) Compute $\hat{\boldsymbol{\xi}}_j$'s from $\hat{\Omega}$ and $\hat{Y} = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_q)$ where $\hat{\mathbf{y}}_j = \hat{\mathbf{y}}(\hat{\mathcal{B}}_{j^-}, \hat{\boldsymbol{\xi}}_j)$. Call algorithm $\mathcal{B}(\hat{Y}, X)$ and extract the output $\hat{s}_{0j}, \quad j = 1, \ldots, q$.

(IIb) Repeat (Ib).

(IIc) Repeat (Ic).

We make some remarks to end this section. (i) It seems intuitively that if the second stage of the TSCS algorithm is further iterated then better estimates of $\mathcal{B}$ and Ω could be obtained. However, this is not the case. The second stage improves the first stage because the variances of the response variables are reduced from σ_j^2 to τ_j^2. But further iteration of the second stage will not do any better from a theoretical point of view. In fact, in our original simulation studies, we compared the two-stage algorithm with the version that further iterates the second stage and found that the further iterations, which result in similar results to the two-stage algorithm, do not really help. (ii) The TSCS algorithm and the aMCR method of Wang (2015) are common in using the conditional framework. However, there is an important difference. For estimating $\mathcal{B}$

or Ω, the TSCS algorithm makes a full adjustment for the effect of one on the other, while the aMCR only makes a partial adjustment. For example, for the estimation of Ω, the responses are adjusted for all the nonzero components of $\mathcal{B}$ in TSCS. But, in aMCR, they are adjusted for only a part of the nonzero components of $\mathcal{B}$, because, at fixed values of the penalty parameters, the active set of the penalized likelihood does not contain all the nonzero components of $\mathcal{B}$, the responses are only adjusted for those nonzero components which are in the active set. The lack of a full adjustment for the effect of one on the other makes aMCR inferior to TSCS, which is demonstrated in the simulation studies reported in Sect. 4.

3 Theoretical properties of TSCS

In the theoretical setting, we allow p, the total number of covariates, and p_{0j}, the number of nonzero components of $\boldsymbol{\beta}_j$, as well as q, the dimension of the response vector, and q_{0j}, the number of nonzero entries in the jth row of Ω, diverge to infinity in certain orders of n. In the analysis of the SHM regression models, it is desirable to establish the selection consistency for the identification of the nonzeros of $\mathcal{B}$ and Ω. We first show that, in the TSCS procedure, the identification of the nonzeros of $\mathcal{B}$ is uniformly selection consistent in q. Then by using this uniform selection consistency, we establish the selection consistency for the identification of the nonzeros of Ω under usual conditions for GGM models. Finally, we provide a proposition which suggests that the second stage of the TSCS procedure is potentially more efficient than its first stage from a theoretical viewpoint.

We start with the properties of the TSCS procedure for the inference on $\mathcal{B}$. Let s_j by any subset of s_{0j}. Define

$$\gamma_j\left(k, s_j\right) = \frac{1}{n}\boldsymbol{x}_k^\top\left[I_n - H^{\mathrm{X}}\left(s_j\right)\right]X\boldsymbol{\beta}_j.$$

The following conditions are assumed.

A1 $\ln p = O(n^\kappa)$, where $0 < \kappa < 1/3$; $\max_j p_{0j} = O(n^c)$, for some $0 < c < 1/6$;

A2 For any $s_j \subset s_{0j}$ but $s_j \neq s_{0j}$, $\max_{k\in s_{0j}\setminus s_j} |\gamma_j(k, s_j)| > \max_{k\notin s_{0j}} |\gamma_j(k, s_j)|$;

A3 $\min_{1\le j\le q}\left\{\lambda_{min}[\frac{1}{n}X^\top(s_{0j})X(s_{0j})]\min_{k\in s_{0j}}|\beta_{jk}|\right\} \ge Cn^{-1/6+\delta}$, where δ is an arbitrary small positive number.

A4 $\lim_{n\to\infty}\min_{1\le j\le q}\min_{s_j: s_{0j}\not\subset s_j, |s_j|\le kp_{0j}}\frac{\Delta(s_j,\boldsymbol{\mu}_j)}{p_{0j}\ln p} \to \infty$, where $\Delta(s_j, \boldsymbol{\mu}_j) = \boldsymbol{\mu}_j^\top[I - H^{\mathrm{X}}(s_j)]\boldsymbol{\mu}_j$, $\boldsymbol{\mu}_j = X\boldsymbol{\beta}_j$, and $k > 1$ is a fixed constant.

Condition A1 is simply a quantification of the high dimensionality and sparsity. Condition A2 is a natural requirement. Note that $[I - H^X(s_j)]X\boldsymbol{\beta}_j$ is the residual of the regression mean which are not explained by the covariates in s_j. Condition A2 requires that the maximum of the correlations of the remaining relevant covariates with the residual is larger than the maximum of the correlations of the remaining irrelevant covariates. This condition is actually weaker than the well-known *irrepresentability condition* required of the Lasso for selection consistency. For the argument of this and some examples, the reader is referred to Luo and Chen (2014b). Condition

A3 is imposed to guard against a fast increase in collinearity of the columns of the design matrix of the relevant covariates and a fast decay of the size of the corresponding regression coefficients when the number of relevant covariates diverges with the sample size. This condition is weaker than conditions (A2) and (A3) in Wang (2015) and condition (C) in Yin and Li (2011). Condition A4 is required for the selection consistency property of the EBIC. This condition is weaker than the so-called *sparse Riesz* condition assumed by other authors, e.g., Wainwright (2009) and Yin and Li (2011), see Chen and Chen (2008).

Let $\hat{s}_{0j}^{\mathrm{M}}$ and $\hat{s}_{0j}^{\mathrm{C}}$ be the sets of nonzeros for the jth column of $\mathcal{B}$ obtained, respectively, in the first and second stage of TSCS. Let $\mathcal{T}_0 = \{(j,k) : \omega_{jk} \neq 0\}$. Denote by $\hat{\mathcal{T}}_0^{\mathrm{M}}$ the estimate of $\mathcal{T}_0$ obtained in the first stage.

Theorem 1 *We have*

(i) *Assume conditions* A1–A4*,*

$$P\left(\hat{s}_{0j}^{\mathrm{M}} = s_{0j},\, j = 1, \ldots, q\right) \to 1$$

as $n \to \infty$.

(ii) *In addition, suppose* $P(\hat{\mathcal{T}}_0^{\mathrm{M}} = \mathcal{T}_0) \to 1$. *Then*

$$P\left(\hat{s}_{0j}^{\mathrm{C}} = s_{0j},\, j = 1, \ldots, q\right) \to 1$$

as $n \to \infty$.

Parts (i) and (ii) of the theorem state, respectively, the uniform selection consistency of TSCS for identifying the nonzeros of $\mathcal{B}$ in the first and second stage. The condition $P(\hat{\mathcal{T}}_0^{\mathrm{M}} = \mathcal{T}_0) \to 1$ is ensured by (i) and conditions **B1**–**B4** to be stated later. The theorem does not reveal whether or not the identification in the second stage is more efficient than in the first stage. A rigorous theoretical proof for the efficiency of the second stage over the first stage is difficult. However, we will provide at the end of this section a result which suggests the efficiency of the second stage from a theoretical viewpoint. The actual efficiency of the second stage over the first stage will be demonstrated in the simulation studies reported in Sect. 4.

The outline of the proof is as follows. Let $s_{j1}^{*\mathrm{M}} \subset \cdots \subset s_{jk}^{*\mathrm{M}} \subset \cdots$ be the sequence of the nonzero sets for the jth marginal model of (1) obtained in the first stage of TSCS. We first show that s_{0j}, the true nonzero set for the jth model, is a member of the sequence having a probability with an uniform lower bound converging to 1. Second, we show that the EBIC sequence, $\mathrm{EBIC}(s_{j1}^{*\mathrm{M}})$, $\mathrm{EBIC}(s_{j2}^{*\mathrm{M}})$, …, is decreasing until it reaches the true nonzero set having a probability with an uniform lower bound converging to 1. Third, we show that the EBIC is uniformly selection consistent in q with a lower bound of the convergence rate. Part (i) is implied by these results. For the sequence obtained in the second stage, the proof is similar.

We now turn to the properties of the TSCS for the inference on Ω. Recall that $\mathcal{T}_0 = \{(j,k) : \omega_{jk} \neq 0\}$. Let $t_{0j} = \{k : (j,k) \in \mathcal{T}_0\}$ and $q_{0j} = |t_{0j}|$. Denote by $\mathcal{T}$ any subset of $\mathcal{T}_0$ and let $t_j = \{k : (j,k) \in \mathcal{T}\}$. Note that $t_j \subset t_{0j}$. For any pair $(j,k) \in \mathcal{T}^c$, define

$$\rho_{jk}^2(\mathcal{T}) = \frac{\left[\left(\Sigma_{kj^-} - \Sigma_{kt_j}\Sigma_{t_jt_j}^{-1}\Sigma_{t_jj^-}\right)\boldsymbol{\xi}_j\right]^2}{\tau_j^2\sigma_k^2} + \frac{\left[\left(\Sigma_{jk^-} - \Sigma_{jt_k}\Sigma_{t_kt_k}^{-1}\Sigma_{t_kk^-}\right)\boldsymbol{\xi}_k\right]^2}{\tau_k^2\sigma_j^2} \tag{11}$$

Note that the $r_{jk}^2(\mathcal{T})$ defined in (8) for the Ω-step is the empirical form of (11). The following conditions are assumed.

B1 $q = o(\ln p)$; $\max_j q_{0j} = O(n^\delta)$, for some $\delta < 1/6$.

B2 For any $\mathcal{T} \subset \mathcal{T}_0$ but $\mathcal{T} \neq \mathcal{T}_0$, $\max_{(j,k)\in\mathcal{T}_0\setminus\mathcal{T}} \rho_{jk}^2(\mathcal{T}) > \max_{(j,k)\notin\mathcal{T}_0} \rho_{jk}^2(\mathcal{T})$;

B3 $\lim_{n\to\infty} \min_{1\le j\le q}\left\{\lambda_{min}(\Sigma_{t_{0j}t_{0j}}) \min_{k\in t_{0j}} |\xi_{jk}|\right\} \ge Cn^{-1/6+\delta}$, where δ is an arbitrary small positive number.

B4 $\lim_{n\to\infty} \min_{1\le j\le q} \min\left\{\frac{\Delta_n(t_j)-\Delta_n(t_j\cup\{l\})}{(\max_{1\le j\le q} q_{0i})^2 \ln q} : t_j \neq t_{0j}, l \in t_j^c \cap t_{0j}\right\} = \infty$, where, for $t \subset t_{0j}$, $\Delta_n(t) = \boldsymbol{\xi}_j^\top[\Sigma_{j^-j^-} - \Sigma_{j^-t}\Sigma_{tt}^{-1}\Sigma_{tj^-}]\boldsymbol{\xi}_j$.

The above conditions are required for the selection consistency of a Gaussian graphical model, see Jiang and Chen (2016), that is, if in the Ω-step, $\tilde{Z} = Y - X\mathcal{B}$ were observed and used, the above conditions ensure the selection consistency. But, instead of $\tilde{Z}$, what was used is Z whose jth column $\boldsymbol{z}_j$ equals $\boldsymbol{y}_j - X(\hat{s}_{0j})\hat{\boldsymbol{\beta}}_j(\hat{s}_{0j})$. To establish the selection consistency of the Ω-step, the sample variance–covariance matrix of Z must provide a good estimate of Σ in a certain sense. The uniform selection consistency in the $\mathcal{B}$-step endows the matrix Z with this desired property. We have the following lemma.

Lemma 1 *Assume that the correlations $\sigma_{jk}/(\sigma_j\sigma_k)$ are bounded by a constant less than 1, the variances σ_j^2 are bounded, and $P(\hat{s}_{0j} = s_{0j},\ j = 1,\ldots,q) \to 1$. Then,*

$$P\left(\max_{1\le j,k\le q} |\frac{1}{n}\boldsymbol{z}_j^\top\boldsymbol{z}_k - \sigma_{jk}| > n^{-1/3}c_0\right) \to 0,$$

where c_0 is a fixed constant.

With Lemma 1, conditions B2–B4 can be transferred into empirical versions in terms of Z similar to A2–A4. The following theorem can then be proved in exactly the same way as that for Gaussian graphical models.

Theorem 2 *Assume* B1–B4 *and the conditions for Lemma 1. Let $\hat{\mathcal{T}}_0$ be the index set of the identified nonzero entries of Ω in a Ω-step. Then,*

$$P\left(\hat{\mathcal{T}}_0 = \mathcal{T}_0\right) \to 1, \quad \text{as } n \to \infty.$$

We end this section by the following proposition. Suppose we could observe $\tilde{\boldsymbol{y}}_j = \boldsymbol{y}_j - (Y_{j^-} - X\mathcal{B}_{j^-})\Sigma_{j^-j^-}^{-1}\Sigma_{j^-j}$. By replacing $\boldsymbol{y}_j$'s with $\tilde{\boldsymbol{y}}_j$'s in the $\mathcal{B}$-step, we obtain a sequence, $\tilde{s}_{j1}^* \subset \cdots \subset \tilde{s}_{jk}^* \subset \cdots$. Then, we have

Proposition 1 *Let*

$$C_n = \min_{1\le j\le q}\left\{\frac{\sqrt{n}}{\ln p}\lambda_{\min}\left[\frac{1}{n}X^\top(s_{0j})X(s_{0j})\right]\min_{k\in s_{0j}}|\beta_{jk}|\right\}$$

Under conditions A1–A3, *as* $n \to \infty$,

(i) $P(s_{j1}^{*\mathrm{M}} \subset \cdots \subset s_{jk}^{*\mathrm{M}} \subset \cdots \subset \hat{s}_{jp_{0j}}^{*\mathrm{M}} = s_{0j}, j = 1, \ldots, q) > 1 - r_n$, *where*

$$r_n = \frac{2 \max_{1 \le j \le q} \sigma_j}{C_n^{1/2} \ln p} \exp\left\{ -\frac{(\ln p)^2}{2} + \ln p + \ln q + \ln(\max p_{0j}) \right\}.$$

(ii) $P(\tilde{s}_{j1}^{*} \subset \cdots \subset \tilde{s}_{jk}^{*} \subset \cdots \subset \tilde{s}_{jp_{0j}}^{*} = s_{0j}; \quad j = 1, \ldots, q) > 1 - \tilde{r}_n$, *where*

$$\tilde{r}_n = \frac{2 \max_{1 \le j \le q} \tau_j}{C_n^{1/2} \ln p} \exp\left\{ -\frac{(\ln p)^2}{2} + \ln p + \ln q + \ln(\max p_{0j}) \right\}.$$

Since $\tau_j < \sigma_j$, $j = 1, \ldots, q$, $1 - \tilde{r}_n > 1 - r_n$. Thus, the probability in (ii) has a higher lower bound than that in (i). This justifies partially the incorporation of the correlation information for identifying the nonzeros of $\mathcal{B}$ in the second stage of TSCS.

The detailed proofs for the results in this section are given in supplementary document.

4 Simulation studies

We conducted two simulation studies. The first simulation study is a comparison of TSCS with two representative existing methods, MRCE and aMCR, which deal with $\mathcal{B}$ and Ω simultaneously. The second simulation study is a comparison of TSCS with the naive approach which ignores the information of correlation in the identification and estimation of the nonzeros of $\mathcal{B}$.

Simulation study I

We set $n = 100, 200$, $(p_0, p) = ([4n^{0.16}], [5e^{n^{0.3}}])$ and $q = 50, 200$. We adopt a common practice in the literature of graphical models for the generation of the covariance matrix of the response variable, that is, a graph is used to generate Ω and it is then inverted to obtain Σ. The following four graphs are considered: *AR(1), ER, Tridiag* and *BA*, see Luo and Chen (2014a) for the description of these graphs. The design matrix X and the nonzero set s_{0j} are generated in four different types. The vectors $\boldsymbol{\beta}_j$ are generated to attain different signal-to-noise ratios. The details of the types of X and s_{0j} and the generation of the $\boldsymbol{\beta}_j$ are given in supplementary document.

For $q = 50$, we consider the settings that are the combinations of the graphs, $n = 100$ and 200, signal-to-noise ratio $h = 0.8$ and 0.6. For $q = 200$, we considered the combinations of the graphs with only $n = 100$ and $h = 0.8$, but, for each combination, two designs which we refer to as block design and noise design are considered for Ω. In the block design, Ω is a diagonal block matrix with four identical diagonal blocks which are precisely the precision matrix in the case of $q = 50$. In the noise design, Ω is a diagonal block matrix consisting of two diagonal blocks, the first block is the same as that in the block design and the second one is an identity matrix of dimension 150.

Under each setting, TSCS, MRCE and aMCR are applied to the same data, and the simulation for each setting is replicated 100 times. The performances of the methods

are evaluated by the average over the 100 replicates of each of the measures given in the following. On Ω, the measures are positive discovery rate (PDR), false discovery rate (FDR), the number of nonzero entries of $\hat{\Omega}$ ($|\hat{\Omega}|$), the spectral norm ($\|\hat{\Omega} - \Omega\|_S$), the matrix ℓ_1 norm ($\|\hat{\Omega} - \Omega\|_{\ell_1}$) and the Frobenius norm ($\|\hat{\Omega} - \Omega\|_F$). On $\mathcal{B}$, the measures are PDR, FDR and the predictive mean squares error (PMSE). The PDR, FDR and PMSE are defined as follows:

$$\text{PDR} = \frac{\big|\{(i,j) : c_{ij} \neq 0 \text{ and } \hat{c}_{ij} \neq 0\}\big|}{\big|\{(i,j) : c_{ij} \neq 0\}\big|}, \quad \text{FDR} = \frac{\big|\{(i,j) : c_{ij} = 0 \text{ and } \hat{c}_{ij} \neq 0\}\big|}{\big|\{(i,j) : \hat{c}_{ij} \neq 0\}\big|},$$

where $c_{ij} = \beta_{ij}$ or ω_{ij}.

$$\text{PMSE} = \frac{1}{n}\left\|Y^{'} - X\hat{B}\right\|_F^2,$$

where $Y^{'}$ is a new matrix of observation not used to obtain the estimate $\hat{B}$. For the method aMCR, since there is no explicit estimate of Ω given in Wang (2015), in the computation of the losses in terms of the matrix norms, we used the same method as in TSCS for the computation of the estimate of Ω, that is, the estimate is computed as $\hat{\Omega} = \text{argmax}_{\omega_{jk}=0:(j,k)\in\mathcal{T}^c} L(Y, \Omega)$, where $L(Y, \Omega)$ is the profile likelihood of Ω while $\mathcal{B}$ is confined to the identified nonzeros.

For the sake of clarity, we only report the results in two settings: (i) $n = 100$, $(p_0, p) = ([4n^{0.16}], [5e^{n^{0.3}}])$, $q = 50$, $h = 0.8$; (ii) $n = 100$, $(p_0, p) = ([4n^{0.16}], [5e^{n^{0.3}}])$, $q = 200$, $h = 0.8$ and Ω is generated by the block design. The results are given, respectively, in Tables 1 and 2. The results under other settings convey similar messages which we are going to discuss in the following. The full results can be found in Jiang (2015).

Now, we discuss the findings under the settings with $q = 50$. First, it is interesting to notice from Table 1 that (a) the performances of each method for the estimation of $\mathcal{B}$ differ significantly in different covariance structures of X; however, with the same covariance structure of X, the performances are quite comparable across different response correlation structures; (b) the performances of each method for the estimation of Ω differ significantly in different response correlation structures and also differ across different covariance structures of X. (Particularly, the performance is worse when the corresponding estimation on $\mathcal{B}$ is worse.) It suggests that, for the methods considered, their performance on the estimation of $\mathcal{B}$ is affected by the covariance structure of X but is insensitive to the response correlation structure; on the other hand, their performance on the estimation of Ω is mainly affected by the response correlation structure and is affected by the covariance structure of X at a lesser extent.

Next, on the comparison of the three methods, it can be seen from Table 1 that (a) the performances of TSCS and aMCR are better than MRCE across all settings in terms of all the measures considered, i.e., higher PDR and lower FDR for both Ω and $\mathcal{B}$, smaller matrix norms in the estimation of Ω and smaller PMSE for the estimate $\hat{\mathcal{B}}$; (b) the overall performance of TSCS is better than that of aMCR, the former has comparable PDR and PMSE but much lower FDR than the latter, though aMCR has slightly higher PDR and slightly smaller PMSE than TSCS in a few settings, its FDR

Table 1 Average of $\hat{\Omega}$-related PDR, FDR, $|\hat{\Omega}|$ and matrix loss in three norms and $\hat{B}$-related PDR, FDR and PMSE when $q = 50$, $n = 100$ and $h = 0.8$

			$\hat{\Omega}$						$\hat{B}$		
Graphs	X	Method	PDR	FDR	$\lvert\hat{\Omega}\rvert$	$\lVert\cdot\rVert_S$	$\lVert\cdot\rVert_{\ell_1}$	$\lVert\cdot\rVert_F$	PDR	FDR	PMSE
AR(1)	I	MRCE	0.570	0.894	396	3.142	3.496	8.164	0.315	0.337	463.671
		aMCR	0.887	0.695	145	23.512	26.598	32.378	0.998	0.663	155.381
		TSCS	0.967	0.076	51	1.321	1.713	3.053	0.990	0.077	128.802
	II	MRCE	0.427	0.869	192	3.588	3.937	8.346	0.218	0.306	453.568
		aMCR	0.807	0.718	143	20.311	23.285	27.702	0.887	0.677	170.376
		TSCS	0.557	0.352	42	1.537	1.891	5.228	0.584	0.116	201.877
	III	MRCE	0.947	0.958	1109	14.683	17.495	18.430	0.115	0.373	518.764
		aMCR	0.850	0.714	147	22.348	25.925	30.440	0.960	0.673	162.134
		TSCS	0.851	0.153	49	1.290	1.713	3.460	0.903	0.089	146.621
	IV	MRCE	0.425	0.940	345	1.756	2.027	6.925	0.008	0.197	576.900
		aMCR	0.981	0.676	154	5.822	8.106	8.643	0.619	0.625	141.103
		TSCS	0.911	0.180	55	1.157	1.691	2.792	0.570	0.266	149.923
ER	I	MRCE	0.183	0.801	69	3.982	5.088	11.847	0.071	0.267	286.144
		aMCR	0.616	0.694	123	24.923	27.583	41.272	0.998	0.667	81.345
		TSCS	0.671	0.168	49	2.315	3.512	6.018	0.976	0.077	69.561
	II	MRCE	0.154	0.788	46	3.611	5.084	11.346	0.115	0.112	264.267
		aMCR	0.524	0.721	116	22.166	25.787	34.712	0.879	0.667	86.329
		TSCS	0.271	0.514	34	2.695	4.312	7.791	0.548	0.112	103.292
	III	MRCE	0.161	0.781	61	4.952	5.794	12.125	0.123	0.396	276.151
		aMCR	0.575	0.709	122	25.385	28.740	40.169	0.962	0.673	84.158
		TSCS	0.507	0.282	43	2.192	3.603	6.229	0.854	0.089	79.776

Table 1 continued

			$\hat{\Omega}$						$\hat{B}$		
Graphs	X	Method	PDR	FDR	$\lvert\hat{\Omega}\rvert$	$\lVert\cdot\rVert_S$	$\lVert\cdot\rVert_{\ell_1}$	$\lVert\cdot\rVert_F$	PDR	FDR	PMSE
	IV	MRCE	0.108	0.706	42	4.231	5.606	9.205	0.553	0.625	152.464
		aMCR	0.804	0.653	147	9.782	13.212	15.939	0.614	0.631	72.736
		TSCS	0.569	0.282	49	2.085	3.576	5.642	0.551	0.253	76.609
Tridiag	I	MRCE	0.049	0.803	20	6.243	7.232	23.330	0.113	0.175	239.109
		aMCR	0.626	0.668	139	26.172	30.178	42.712	0.998	0.627	66.626
		TSCS	0.647	0.066	50	3.771	4.882	8.454	0.988	0.085	58.003
	II	MRCE	0.030	0.813	13	6.167	7.173	22.953	0.208	0.108	202.526
		aMCR	0.597	0.701	147	23.085	27.381	35.929	0.886	0.653	74.018
		TSCS	0.428	0.361	49	5.178	6.462	15.745	0.614	0.117	87.634
	III	MRCE	0.374	0.697	247	10.498	12.705	25.806	0.529	0.455	158.372
		aMCR	0.614	0.680	140	25.112	29.026	39.679	0.961	0.649	70.220
		TSCS	0.601	0.146	51	4.012	5.272	9.438	0.928	0.096	63.365
	IV	MRCE	0.015	0.453	15	5.533	6.560	19.324	0.629	0.610	84.470
		aMCR	0.695	0.649	149	10.222	15.361	16.932	0.627	0.613	62.991
		TSCS	0.631	0.171	55	3.545	5.010	7.965	0.606	0.254	66.637
BA	I	MRCE	0.675	0.952	694	10.029	18.967	16.195	0.298	0.616	800.088
		aMCR	0.697	0.830	202	23.482	27.936	32.204	0.998	0.648	245.365
		TSCS	0.730	0.314	53	5.250	11.064	7.206	0.995	0.093	209.772
	II	MRCE	0.738	0.941	653	10.858	20.464	16.723	0.297	0.357	711.443
		aMCR	0.651	0.846	208	20.027	26.433	28.410	0.887	0.667	270.467
		TSCS	0.479	0.587	57	9.334	17.833	13.425	0.656	0.123	308.791

Table 1 continued

			$\hat{\Omega}$						$\hat{B}$		
Graphs	X	Method	PDR	FDR	$\lvert\hat{\Omega}\rvert$	$\lVert\cdot\rVert_S$	$\lVert\cdot\rVert_{\ell_1}$	$\lVert\cdot\rVert_F$	PDR	FDR	PMSE
	III	MRCE	0.879	0.959	1092	22.066	30.758	28.650	0.107	0.683	889.077
		aMCR	0.687	0.836	206	20.547	25.502	28.982	0.963	0.663	257.452
		TSCS	0.686	0.369	54	6.094	11.978	8.448	0.946	0.099	228.755
	IV	MRCE	0.581	0.938	455	10.222	18.999	15.929	0.002	0.000	966.933
		aMCR	0.796	0.778	181	6.407	12.759	10.261	0.641	0.625	232.100
		TSCS	0.683	0.402	57	6.271	12.097	8.747	0.626	0.266	243.749

Table 2 Average of $\hat{\Omega}$-related PDR, FDR, $|\hat{\Omega}|$ and matrix loss in three norms and $\hat{B}$-related PDR, FDR and PMSE for block precision matrix design when $q = 200$, $n = 100$ and $h = 0.8$

			$\hat{\Omega}$						$\hat{B}$		
Graphs	X	Method	PDR	FDR	$\lvert\hat{\Omega}\rvert$	$\lVert\cdot\rVert_S$	$\lVert\cdot\rVert_{\ell_1}$	$\lVert\cdot\rVert_F$	PDR	FDR	PMSE
AR(1)	I	MRCE	0.002	0.976	16	4.390	4.790	15.775	0.817	0.507	1232.233
		aMCR	0.837	0.849	1090	40.903	70.339	89.933	0.998	0.682	634.426
		TSCS	0.901	0.046	185	1.577	2.041	6.468	0.979	0.080	533.867
	II	MRCE	0.005	0.974	66	2.470	2.617	15.286	0.231	0.181	1812.271
		aMCR	0.763	0.870	1156	35.751	65.651	75.093	0.882	0.684	692.143
		TSCS	0.317	0.380	101	1.625	1.979	11.464	0.547	0.116	842.243
	III	MRCE	0.003	0.380	7	3.737	4.073	14.886	0.711	0.533	1158.763
		aMCR	0.801	0.857	1107	37.344	65.619	80.767	0.961	0.686	655.552
		TSCS	0.695	0.119	155	1.533	1.970	7.880	0.873	0.093	612.762
	IV	MRCE	0.169	0.983	1925	1.785	2.016	14.609	0.031	0.332	2280.273
		aMCR	0.980	0.862	1418	20.746	54.075	41.403	0.629	0.636	568.345
		TSCS	0.794	0.132	179	1.334	1.943	6.337	0.559	0.263	613.254
ER	I	MRCE	0.005	0.802	7	4.885	5.684	21.802	0.617	0.601	843.912
		aMCR	0.520	0.841	804	37.735	69.395	100.108	0.998	0.672	323.042
		TSCS	0.510	0.148	146	2.690	4.222	12.287	0.971	0.075	277.059
	II	MRCE	0.007	0.822	12	4.088	5.237	21.764	0.252	0.255	991.202
		aMCR	0.467	0.865	840	35.050	67.951	85.665	0.880	0.671	357.354
		TSCS	0.161	0.590	94	2.877	4.701	16.011	0.539	0.112	435.270
	III	MRCE	0.005	0.665	5	4.312	5.353	22.897	0.221	0.253	1014.938
		aMCR	0.484	0.849	796	36.261	63.058	94.668	0.959	0.675	334.296
		TSCS	0.338	0.290	116	2.635	4.431	13.458	0.836	0.087	320.785

Table 2 continued

			$\hat{\Omega}$						$\hat{B}$		
Graphs	X	Method	PDR	FDR	$\lvert\hat{\Omega}\rvert$	$\lVert\cdot\rVert_S$	$\lVert\cdot\rVert_{\ell_1}$	$\lVert\cdot\rVert_F$	PDR	FDR	PMSE
	IV	MRCE	0.269	0.966	1956	3.201	5.469	19.859	0.006	0.041	1173.614
		aMCR	0.775	0.829	1138	23.870	65.180	54.533	0.609	0.637	293.599
		TSCS	0.411	0.268	138	2.403	4.074	11.845	0.531	0.262	314.227
Tridiag	I	MRCE	0.000	0.005	0	6.229	7.434	43.279	0.795	0.562	535.729
		aMCR	0.599	0.811	928	36.639	59.644	99.018	0.998	0.639	269.958
		TSCS	0.629	0.047	192	4.671	6.134	17.254	0.985	0.087	234.255
	II	MRCE	0.000	0.108	0	6.242	7.256	45.017	0.299	0.138	747.138
		aMCR	0.558	0.848	1082	34.407	62.426	89.791	0.884	0.658	299.988
		TSCS	0.324	0.393	156	5.582	6.995	33.287	0.582	0.119	364.330
	III	MRCE	0.001	0.212	1	6.223	7.670	41.811	0.814	0.656	446.587
		aMCR	0.577	0.825	963	36.459	60.610	97.262	0.961	0.656	282.527
		TSCS	0.546	0.123	181	4.745	6.252	20.660	0.910	0.097	260.994
	IV	MRCE	0.135	0.629	1162	5.920	7.171	41.545	0.293	0.565	737.285
		aMCR	0.675	0.827	1155	22.748	55.741	59.008	0.627	0.626	248.413
		TSCS	0.582	0.138	197	4.288	6.090	17.176	0.586	0.267	267.354
BA	I	MRCE	0.094	0.966	539	11.271	21.040	33.155	0.182	0.500	3454.254
		aMCR	0.668	0.904	1376	39.767	64.596	82.219	0.999	0.665	981.557
		TSCS	0.659	0.251	173	7.344	14.589	16.497	0.984	0.097	856.464
	II	MRCE	0.073	0.958	344	10.973	20.360	32.009	0.368	0.569	2797.879
		aMCR	0.639	0.919	1553	35.740	71.026	73.182	0.889	0.676	1099.499
		TSCS	0.329	0.536	139	10.763	20.064	29.214	0.594	0.125	1341.996

Table 2 continued

			$\hat{\Omega}$						$\hat{B}$		
Graphs	X	Method	PDR	FDR	$\lvert\hat{\Omega}\rvert$	$\lVert\cdot\rVert_S$	$\lVert\cdot\rVert_{\ell_1}$	$\lVert\cdot\rVert_F$	PDR	FDR	PMSE
	III	MRCE	0.085	0.964	466	10.835	20.193	32.168	0.337	0.549	3092.597
		aMCR	0.655	0.908	1395	36.586	63.869	75.822	0.963	0.676	1033.160
		TSCS	0.587	0.309	167	8.905	16.923	20.215	0.919	0.104	954.123
	IV	MRCE	0.209	0.978	1825	11.458	21.108	34.826	0.009	0.159	3782.157
		aMCR	0.772	0.892	1421	17.810	47.038	36.437	0.639	0.639	923.662
		TSCS	0.615	0.323	179	8.888	16.838	19.850	0.606	0.278	980.495

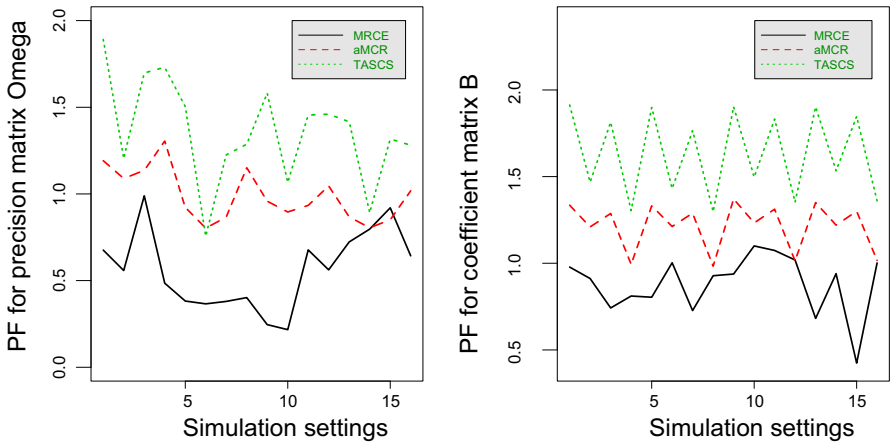

Fig. 1 The plot of PDR + (1− FDR) for the identification of nonzeros of Ω and $\mathcal{B}$ in the 16 settings when $q = 50$, $n = 100$ and $h = 0.8$

is too high to be acceptable when the identification of the nonzeros of $\mathcal{B}$ and Ω is of a major concern, for example, in eQTL mapping problems.

Now turn to the case that $q = 200$. Comparing Table 2 with Table 1, it can be found that both PDR and FDR for identifying the nonzeros of Ω become worse for all methods under all settings, that is, the increase in the dimension of the response vector increases the difficulty for the identification of the nonzeros of Ω and that the changes in PDR and FDR for identifying the nonzeros of $\mathcal{B}$ are insignificant for all methods under all settings, that is, the efficiency for identifying the nonzeros of $\mathcal{B}$ does not seem to be affected too much by the increase in the dimension of the response vector.

On the comparison of the three methods, the findings found in Table 1 remain in Table 2, that is, the relative performances of the three methods in the case of $q = 200$ are the same as those in the case of $q = 50$. But there is an additional point worthy of mention. For the identification of the nonzeros of Ω, although the PDRs of both TSCS and aMCR are reduced by about the same rate, the FDRs of TSCS remain almost the same (even smaller in certain cases) and the FDRs of aMCR have, however, soared to a large extent. (The minimum changed from 0.649 to 0.811; the maximum changed from 0.846 to 0.919.) This indicates a obvious advantage of TSCS over aMCR.

To further illustrate the simulation results of the comparison, we take PDR + (1 − FDR) as an overall measure of the performance, and this overall measure is plotted for each method at the 16 settings reported in Tables 1 and 2, respectively, in Figs. 1 and 2.

Simulation study II

In this simulation study, we compare TSCS with the naive approach: separate identification and estimation of the nonzeros of $\boldsymbol{\beta}_j$ in each of the marginal models by the SLasso method. The purpose of this simulation study is to demonstrate the efficiency gain by incorporating the information of correlation into the inference on $\mathcal{B}$.

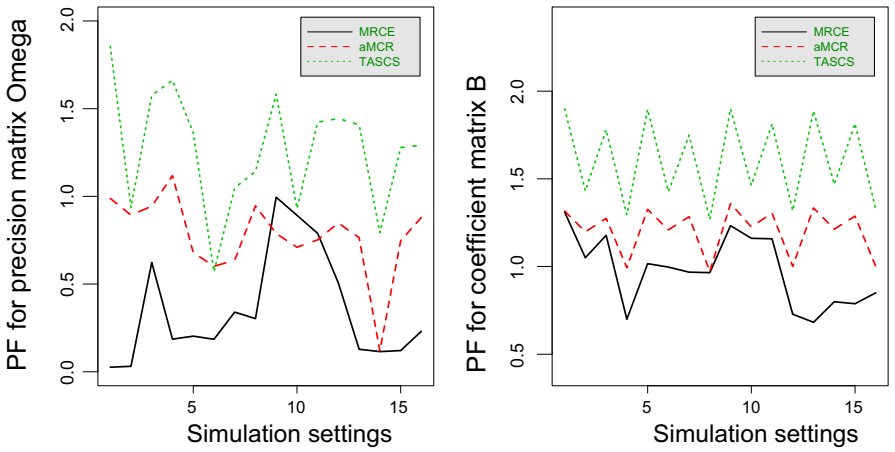

Fig. 2 The plot of PDR + (1− FDR) for the identification of nonzeros of Ω and $\mathcal{B}$ in the 16 settings when $q = 200$, $n = 100$ and $h = 0.8$

In the simulation settings, the observations on the p-dimensional covariate vector are generated as normal vectors with mean $\mathbf{0}$ and covariance matrix Σ_X. The Σ_X takes three forms — (i) Identity (I): $\Sigma_X = I_p$, (ii) Power Decay (PD): $\Sigma_X = (0.5^{|i-j|})_{p\times p}$ and (iii) Equal Correlation (EC): $\Sigma_X = (\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5$ if $i \neq j$, 1 if $i = j$.

The observations of the response vector are generated as follows. For each component of the response vector, p_0 covariates are randomly selected as its true features except in the case of power decay where the true features are taken in the same way as in the second type of simulation I. Note that, although the number p_0 is the same for each component, the p_0 true features are different from component to component. The $\boldsymbol{\beta}_j$'s are generated in the same way as in simulation study I but without scaling. The rows of the error matrix E are generated as i.i.d. samples from $N_q(\mathbf{0}, \Sigma)$, where Σ is determined as follows. First, its jth diagonal element is determined as

$$\sigma_j^2 = \frac{1-h}{h}\boldsymbol{\beta}_j^\top \Sigma_X \boldsymbol{\beta}_j, \quad j = 1, \ldots, q,$$

where h is a specified signal-to-noise ratio. Let $D = \text{diag}(\sigma_1, \ldots, \sigma_q)$ and $R = (\rho_{ij})_{q\times q}$ be a correlation matrix. Then, Σ is taken as DRD. Four forms of R are considered — (i) Identity (I): $R = I_p$, (ii) Band: $\rho_{i,i+1} = \rho_{i+1,i} = 0.5$, (iii) Power Decay (PD): $\rho_{ij} = (0.8^{|i-j|})$ and (iv) Equal Correlation (EC): $\rho_{ij} = 0.5$, $i \neq j$.

In this simulation study, we considered $n = 100$, $p = [5e^{n^{0.3}}]$ and 500, $p_0 = [4n^{0.16}]$, $q = 10$ and 20, $h = 0.75$ and 0.8. TSCS and the naive approach are applied to each set of the simulated data. The PDR, FDR and PMSE of the two methods are averaged over 200 replicated simulations. Since the results are similar across the simulation settings, for the sake of clarity, we only report the results for the setting that $n = 100$, $p = [5e^{n^{0.3}}] = 267$, $p_0 = [4n^{0.16}] = 8$, $q = 10$ and $h = 0.8$, which is given in Table 3. The following two points manifest themselves in Table 3. (i) In the case of $R = I$, i.e., the response variables are indeed independent, the performances

Table 3 Average (SD) of PDR, FDR and PMSE over 200 replications of the naive approach and the TSCS in the simulation setting that $n = 100$, $p = [5e^{n^{0.3}}]$, $q = 10$ and $h = 0.8$

Σ_X	R	Method	PDR	FDR	PMSE
I	I	Naive	0.962 (0.048)	0.062 (0.029)	105.198 (11.679)
		TSCS	0.970 (0.041)	0.077 (0.028)	105.140 (10.118)
	Band	Naive	0.971 (0.039)	0.056 (0.027)	103.184 (10.824)
		TSCS	0.996 (0.018)	0.121 (0.042)	95.081 (7.138)
	PD	Naive	0.965 (0.044)	0.057 (0.031)	104.223 (13.328)
		TSCS	0.999 (0.006)	0.093 (0.038)	90.645 (8.580)
	EC	Naive	0.957 (0.054)	0.064 (0.032)	107.127 (13.788)
		TSCS	0.992 (0.024)	0.078 (0.034)	95.810 (8.668)
PD	I	Naive	0.518 (0.085)	0.101 (0.048)	163.178 (19.150)
		TSCS	0.525 (0.080)	0.128 (0.050)	163.742 (18.971)
	Band	Naive	0.520 (0.071)	0.095 (0.042)	165.151 (17.935)
		TSCS	0.626 (0.100)	0.146 (0.048)	148.930 (19.042)
	PD	Naive	0.519 (0.072)	0.099 (0.051)	165.782 (21.468)
		TSCS	0.812 (0.093)	0.127 (0.044)	121.127 (19.909)
	EC	Naive	0.525 (0.073)	0.100 (0.046)	164.442 (21.419)
		TSCS	0.670 (0.079)	0.123 (0.046)	141.705 (20.325)
EC	I	Naive	0.545 (0.109)	0.136 (0.061)	173.987 (34.855)
		TSCS	0.549 (0.113)	0.158 (0.064)	175.433 (34.924)
	Band	Naive	0.538 (0.110)	0.135 (0.058)	177.883 (33.385)
		TSCS	0.697 (0.137)	0.159 (0.055)	152.361 (33.528)
	PD	Naive	0.546 (0.118)	0.135 (0.065)	174.281 (41.824)
		TSCS	0.864 (0.104)	0.123 (0.053)	123.276 (35.208)
	EC	Naive	0.537 (0.116)	0.135 (0.060)	175.933 (37.885)
		TSCS	0.714 (0.112)	0.126 (0.051)	146.438 (34.545)

of TSCS and the naive approach are comparable. This is expected since, in this case, the conditional univariate models reduce to the unconditional marginal models. (ii) In the cases, when $R \neq I$, i.e., the response variables are correlated, however, TSCS performs much better than the naive approach. The former has a much higher PDR, a much smaller PMSE than and a comparable (though slightly higher) FDR with the latter. This simulation study demonstrates the efficiency gain of TSCS for the identification and estimation of $\mathcal{B}$. TSCS does not cause any adverse effect even if the response variables are indeed independent.

We make an analysis of the computation time to conclude this section. We recorded the time of each method under the following settings: (i) $n = 100, q = 50, h = 0.8$; (ii) $n = 200, q = 50, h = 0.8$; (iii) $n = 100, q = 200, h = 0.8$, with block precision matrix; (iv) $n = 100, q = 200, h = 0.8$, with noise precision matrix. In all the settings, $(p_0, p) = ([4n^{0.16}], [5e^{n^{0.3}}])$. The computation time for each method in each setting is averaged over the four graphs and the four X-structures with ten replicates. The average times are given as follows:

Simulation	Time (in seconds) MRCE	aMCR	TSCS
(i)	106	85	16
(ii)	314	178	45
(iii)	701	389	729
(iv)	1083	391	425

When $q = 50$, the computation time required by TSCS is much less than the other two methods. When $q = 200$, TSCS requires more time than aMCR and less time than MRCE on average. That TSCS requires more time when $q = 200$ is because that more time is needed for obtaining $\hat{\tau}_j$'s by using the scaled Lasso algorithm in the Ω-step.

5 A real example

The TSCS approach is applied to the Glioblastoma multiforme (GBM) cancer data which are available at https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp. The data consist of 11861 gene expression levels and 534 microRNA values from 202 subjects. It is of interest to investigate how the microRNA values are affected by the gene expression levels and how they are related to each other. The data were previously analyzed by the Cancer Genome Atlas (TCGA) Research Network (McLendon et al. 2008; Verhaak et al. 2010), Lee and Liu (2012) and Wang (2015).

In order to make a comparison with the previous analyses, we follow the same analysis process as in Wang (2015) and the references therein. Six subjects with missing microRNA values are excluded. A pre-screening procedure is carried out on the genes and the microRNAs based on their median absolute deviations (MAD). Five hundred genes with the top largest MADs of expression levels and 20 microRNAs with the top largest MADs of the microRNA values are extracted from the original data for the analysis. Thus, the final data consist of 500 gene expression levels and 20 microRNA values of 196 subjects. The microRNA values and the gene expression levels are modeled as follows:

$$\boldsymbol{y}^\top = \boldsymbol{x}^\top \mathcal{B} + \boldsymbol{e}^\top, \tag{12}$$

where $\boldsymbol{y}$ is a vector of 20 microRNA values, $\boldsymbol{x}$ is a vector of 500 gene expression levels and $\boldsymbol{e}$ is a multivariate normal vector distributed as $N_{20}(\mathbf{0}, \Omega^{-1})$. The final data are randomly divided into a training dataset with 120 subjects and a testing dataset with the remaining 76 subjects. The training dataset is used to estimate the coefficient matrix B and the precision matrix Ω. The testing dataset is used to calculate the predictive squared error (PSE) which is given by

$$\text{PSE} = \frac{1}{76 \times 20} \sum_i \| \boldsymbol{y}_i - \hat{B}^\top \boldsymbol{x}_i \|_2^2,$$

Table 4 Average (standard deviation) of PSE and number of involved genes resulted from TSCS, CW, PWL, DML and aMCR in the analysis of the Glioblastoma Multiforme Cancer Data

Method	TSCS	CW	PWL	DML	aMCR
PSE	1.193	1.298	1.248	1.229	1.190
	(0.009)	(0.038)	(0.032)	(0.032)	(0.012)
Num. Genes	43	500	17	78	65
	(0.624)	(0.000)	(13.565)	(32.151)	(1.750)

where the sum is taken over the testing dataset. As in Wang (2015), we repeated the above procedure 50 times and computed the average and standard deviation of the PSE and the number of genes actually involved in model (12).

In Lee and Liu (2012), three different methods: Curds and Whey method (CW) (Breiman and Friedman 1997), PWL and DML (Lee and Liu 2012), are applied to analyze the data in a slightly different procedure and the average and standard deviation are taken over ten replications. The PWL and DML methods are similar to MRCE but impose the adaptive Lasso penalty on both $\mathcal{B}$ and Ω. For comparison, the average and standard deviation of the PSE and the number of genes resulted from TSCS, CW, PWL, DML and aMCR are reported together in Table 4. The values for CW, PWL and DML are copied from Lee and Liu (2012), and those for aMCR are copied from Wang (2015). As given in Table 4, the PSE of TSCS is almost the same as that of aMCR which is the smallest. The difference of PSE between TSCS and aMCR is indeed negligible (which is less than a third of their pooled standard deviation). However, TSCS results in a much more parsimonious model than aMCR. The TSCS needs only 43 genes to achieve about the same PSE as aMCR that needs around 65 genes. The network graph of the 20 microRNAs detected by TSCS is given in supplementary document.

References

Breiman, L., Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *59*(1), 3–54.

Chen, J., Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Chen, L., Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, *107*(500), 1533–1545.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., Feng, Y., Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, *3*(2), 521.

Friedman, J., Hastie, T., Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphica Lasso. *Biostatistics*, *9*(3), 432–441.

Jiang, Y. (2015). Sequential approaches in graphical models and multivariate response regression models. In PhD thesis, Department of Statistics & Applied Probability, National University of Singapore.

Jiang, Y., Chen, Z. (2016). A sequential scaled pairwise selection approach to edge detection in nonparanormal graphical models. *Canadian Journal of Statistics*, *44*(1), 25–43.

Lee, W., Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis*, *111*, 241–255.

Luo, S., Chen, Z. (2014a). Edge detection in sparse Gaussian graphical models. *Computational Statistics & Data Analysis*, *70*, 138–152.

Luo, S., Chen, Z. (2014b). Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, *109*(507), 1229–1240.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068.

Meinshausen, N., Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

Obozinski, G., Wainwright, M. J., Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, *39*(1), 1–47.

Peng, J., Wang, P., Zhou, N., Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, *104*(486), 735–746.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., et al. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, *4*(1), 53.

Rothman, A. J., Levina, E., Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, *19*(4), 947–962.

Sun, T., Zhang, C. (2012). Scaled sparse linear regression. *Biometrika*, *99*(4), 879–898.

Sun, T., Zhang, C. H. (2013). Sparse matrix inversion with scaled Lasso. *The Journal of Machine Learning Research*, *14*(1), 3385–3418.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*(1), 267–288.

Turlach, B. A., Venables, W. N., Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, *47*(3), 349–363.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, *17*(1), 98–110.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, *55*(5), 2183–2202.

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica*, *25*(3), 831–851.

Yin, J., Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, *5*(4), 2630.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, *99*, 2261–2286.

Yuan, M., Ekici, A., Lu, Z., Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *69*(3), 329–346.

Zhou, S., van de Geer, S., Bühlmann, P. (2009). Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. arXiv preprint arXiv:0903.2515.