Latent variable selection in structural equation models

Yan-Qing Zhang, Guo-Liang Tian, Nian-Sheng Tang

PII: S0047-259X(16)30068-9
DOI: http://dx.doi.org/10.1016/j.jmva.2016.08.004
Reference: YJMVA 4148

To appear in: *Journal of Multivariate Analysis*

Received date: 16 July 2015

1. Propose a PL method to identify the structure of coefficient matrix
2. show the consistency and the oracle property of the proposed estimators
3. propose a minorization--maximization algorithm to facilitate the second M-step

Latent variable selection in structural equation models[☆]

Yan-Qing Zhang[a], Guo-Liang Tian[b], Nian-Sheng Tang[a,*]

[a] *Department of Statistics, Yunnan University, Kunming 650091, China*
[b] *Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China*

Abstract

Structural equation models (SEMs) are often formulated using a prespecified parametric structural equation. In many applications, however, the formulation of the structural equation is unknown, and its misspecification may lead to unreliable statistical inference. This paper develops a general SEM in which latent variables are linearly regressed on themselves, thereby avoiding the need to specify outcome/explanatory latent variables. A penalized likelihood method with a proper penalty function is proposed to simultaneously select latent variables and estimate the coefficient matrix in formulating the structural equation. Under some regularity conditions, we show the consistency and the oracle property of the proposed estimators. We also develop an expectation/conditional maximization (ECM) algorithm involving a minorization–maximization algorithm that facilitates the second M-step. Simulation studies are performed and a real data set is analyzed to illustrate the proposed methods.

Keywords: ECM algorithm, Lasso, SCAD, Structural equation models, Variable selection

2010 MSC: 62H12, 62H25,

[☆] This research was funded by grants from the NSFC (11225103).

[*] Corresponding author
Email address: nstang@ynu.edu.cn (Nian-Sheng Tang)

1. Introduction

Structural Equation Models (SEMs) are widely used, e.g., in biomedical, educational, behavioral, psychological, and social sciences. Many methods have been developed to fit SEMs; see, e.g., [10, 21]. In particular, Song and Lee [19] proposed a Bayesian approach for SEMs with ignorable missing continuous and polytomous data, and Lee and Zhu [15] developed an expectation–maximization (EM; see, e.g., [3]) algorithm together with the Metropolis–Hastings algorithm for maximum likelihood estimation (MLE) of a general nonlinear SEM. More recently, Lee and Tang [14] proposed a Bayesian approach to analyze nonlinear SEMs with variables either categorically ordered or from an exponential distribution family, and Song et al. [22] analyzed longitudinal data in SEMs using a Bayesian approach.

Statistical theory typically assumes a specified formulation of the structural equation or latent variable model. In practice, however, researchers often have no clue as to how the latent variables are related. They then find it is difficult to preassign a specific formulation of the model. In effect, latent variable models are very often exploratory, while the outcome/explanatory latent variables and their relationship need to be identified, model misspecification may lead to unreliable statistical inference.

Therefore, there is a need to develop a general SEM in which the latent variable model is exploratory; see, e.g., [20], p. 200. To this end, some authors tried to select a specific form of latent variable model in an SEM framework by means of model comparison via some criterion such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) or the Deviance Information Criterion (DIC); see, e.g., [13, 18]. Although these methods are popular, the associated computational costs can be very high when many competing SEMs are being compared. In addition, little work has been done on exploring the asymptotic properties of these SEMs.

This paper tackles these issues by considering a general SEM, defined in Eq. (3). The key feature of this model is that latent variables are linearly regressed on themselves with a coefficient matrix and do not specify what are the outcome/explanatory latent variables. Because some of the latent variables are outcome variables while the others are explanatory variables, and considering that some of outcome and explanatory latent variables may be unrelated, the coefficient matrix in the considered structural equation includes many zero components, i.e., it is sparse. Therefore, determining the formulation of the latent variable model is equivalent to identifying the zero components in the coefficient matrix. Motivated by variable selection techniques, we propose here a penalized likelihood method to simultaneously implement latent variable selection and parameter estimation, and then to identify the structure of latent variable model.

Variable selection is vital to complex statistical modeling and has been an important topic in regression analysis. In recent years, many new and efficient variable selection methods have been proposed by various authors; see, e.g., [1, 4, 16, 23, 25, 26, 27]. In particular, Tibshirani [23] introduced the least absolute shrinkage and selection operator (Lasso) method by minimizing the ordinary least squares with the L_1 penalty. Subsequently, Fan and Li [4] proposed the smoothly clipped absolute deviation (SCAD) penalty and proved the oracle property of the regression coefficient estimator. More recently, Bondell et al. [1] proposed simultaneous selection of the fixed and random factors in a linear mixed-effects model using a modified Cholesky decomposition and adaptive Lasso, and they obtained the final estimates by a constrained EM algorithm.

In this paper, we propose a penalized likelihood approach with some proper penalty function for variable selection in SEM, and we establish the consistency and the oracle property of the proposed estimators under some mild regularity conditions. From a computational point of view, we develop an expectation/conditional maximization algorithm (ECM; see, e.g., [17]) that relies on a minorization–maximization algorithm (MM; see, e.g., [9]) for the second M-step.

The rest of this article is organized as follows. Section 2 develops the penalized likelihood approach for variable selection in SEM with a new structural equation and studies the oracle properties of the penalized maximum likelihood estimators. A computational procedure, standard error estimates, and the choice of the tuning parameters are provided in Section 3. In Section 4, simulation studies are performed and in Section 5, a real data set is analyzed to illustrate the proposed methods. Finally, a discussion is presented in Section 6. Some technical details are given in Appendices A and B.

2. Latent variable selection in structural equation model

2.1. Model formulation

Consider the following linear measurement model with n subjects:

$$\boldsymbol{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^\top$ is the observed random vector of subject i, $\boldsymbol{\mu}$ is the intercept vector, $\boldsymbol{\Lambda}$ is a $p \times q$ factor loading matrix, $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iq})^\top$ is the latent random vector for subject i, $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}_\epsilon)$ is independent of $\boldsymbol{\omega}_i$, and $\boldsymbol{\Psi}_\epsilon = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$ is a $p \times p$ diagonal matrix.

In an SEM framework [20], for a confirmatory structural equation (i.e., latent variable model), it is assumed that $\boldsymbol{\omega}_i = (\boldsymbol{\eta}_i^\top, \boldsymbol{\xi}_i^\top)^\top$, where for some integer $q_1 < q$ such that $q_2 = q - q_1 > 0$, $\boldsymbol{\eta}_i$ is a $q_1 \times 1$ vector of outcome latent variables, and $\boldsymbol{\xi}_i$ is a $q_2 \times 1$ vector of explanatory latent variables. It is further assumed that

$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i, \qquad i = 1, \ldots, n, \tag{2}$$

where $\mathbf{B}$ is a $q_1 \times q_1$ matrix of coefficients that allows some outcome latent variables to depend on the other outcome latent variables, $\boldsymbol{\Gamma}$ is a $q_1 \times q_2$ matrix of coefficients measuring the effect of explanatory latent variables on outcome latent variables, and $\boldsymbol{\zeta}_i$ is a $q_1 \times 1$ vector of random residuals.

Generally, outcome and explanatory latent variables in Eq. (2) are preassigned according to some prior information; see, e.g., [2]. In practice, however, researchers may not have access to such prior information. A typical example is the latent curve model [20]. Moreover, misspecification of outcome/explanatory latent variables may lead to unreliable statistical inference. Hence, it is important to develop a reliable and efficient method to identify outcome/explanatory latent variables and the structural relations among latent variables in achieving appropriate estimation. To this end, we consider the following exploratory structural equation:

$$\boldsymbol{\omega}_i = \boldsymbol{\Pi}\boldsymbol{\omega}_i + \boldsymbol{\delta}_i, \qquad i = 1, \ldots, n, \tag{3}$$

where $\boldsymbol{\Pi} = (\pi_{\ell j})$ is a $q \times q$ matrix of coefficients expressing the structural relations among the latent variables in $\boldsymbol{\omega}_i$, and $\boldsymbol{\delta}_i$ is a $q \times 1$ vector of random residuals. It is assumed that $\mathbf{I}_q - \boldsymbol{\Pi}$ is nonsingular and that, for all $i \in \{1, \ldots, n\}$, $\boldsymbol{\delta}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi}_\delta)$ with $\boldsymbol{\Psi}_\delta = \text{diag}(\gamma_1^2, \ldots, \gamma_q^2)$.

In Eq. (3), the latent variable $\omega_{i\ell} = \boldsymbol{\Pi}_\ell \boldsymbol{\omega}_i + \delta_{i\ell}$ is an outcome latent variable if there is at least a nonzero component in $\boldsymbol{\Pi}_\ell$, which is the ℓth row vector of matrix $\boldsymbol{\Pi}$; the latent variable $\omega_{i\ell}$ is an explanatory latent variable if all the components in $\boldsymbol{\Pi}_\ell$ are zero and there is at least a nonzero component in the ℓth column vector of matrix $\boldsymbol{\Pi}$; the latent variable $\omega_{i\ell}$ has no relation with other latent variables if $\pi_{\ell j} = \pi_{j\ell} = 0$ for all $j \in \{1, \ldots, q\}$. Thus, once we properly select nonzero and zero components in $\boldsymbol{\Pi}$, we can obtain outcome/explanatory latent variables. Consequently, the exploratory structural equation (3) becomes confirmatory structural equation (2) by specifying a specific structure of $\boldsymbol{\Pi}$.

Unlike the traditional latent variable models, the model (3) is characterized by the fact that (i) latent variables are linearly regressed on themselves with a coefficient matrix, and (ii) outcome/explanatory latent variables are not specified. This model was considered earlier by Kenneth [12] and Daniel [2] in fitting the dataset from Youth Development Study Analysis in which three latent factors (namely mastery, depression and optimism) were involved and the structure of latent variable model was assumed by specifying a structure for $\boldsymbol{\Pi}$, but these authors did not consider the structure selection of $\boldsymbol{\Pi}$. In what follows, we use a recently developed variable selection technique to perform simultaneous estimation and detection of outcome/explanatory latent variables.

Let $\boldsymbol{\theta} \in \mathbb{R}^k$ be the parameter vector that contains all the unknown parameters in $\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Pi}, \boldsymbol{\Psi}_\epsilon$ and $\boldsymbol{\Psi}_\delta$, where k is the number of unknown parameters in (3). The proposed model is over parameterized if there are no appropriate identification conditions on $\boldsymbol{\theta}$. To address the identification issue, we need to consider some restrictions on $\boldsymbol{\theta}$. It is rather difficult to find a necessary and sufficient condition for identifiability in a SEM. Hence, in many applications, only sufficient conditions for identifiability are provided on a case-by-case basis; see, e.g., [20]. Here, we follow the common practice in SEMs for identifiability by fixing some entries in $\boldsymbol{\Lambda}$ and assuming that the diagonal elements of $\boldsymbol{\Pi}$ are known.

2.2. The method of variable selection

Let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the observed data matrix and $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$ be the matrix of latent factors. It follows from Eqs. (1) and (3) that $\boldsymbol{y}_i$ is Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}_\epsilon$, where $\boldsymbol{\Sigma} = (\mathbf{I}_q - \boldsymbol{\Pi})^{-1}\boldsymbol{\Psi}_\delta(\mathbf{I}_q - \boldsymbol{\Pi}^\top)^{-1}$. Removing the constant term, the observed-data log-likelihood function is

$$L_0(\boldsymbol{\theta}) \equiv \sum_{i=1}^n L_{0i}(\boldsymbol{\theta}) = -\frac{n}{2}\ln|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\mu})^\top \mathbf{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}).$$

To simultaneously select latent variables and estimate the unknown parameters in $\boldsymbol{\Pi}$, we consider the following penalized log-likelihood function

$$\ell(\boldsymbol{\theta}) = L_0(\boldsymbol{\theta}) - \sum_{\ell=1}^q \sum_{j=1, j\neq\ell}^q \phi_{\lambda_n}(|\pi_{\ell j}|), \tag{4}$$

where $\pi_{\ell j}$ is the (ℓ, j)th entry of $\boldsymbol{\Pi}$, ϕ_{λ_n} is a penalty function and λ_n is a tuning parameter that controls the model complexity and can be selected by some data-driven method such as cross validation (CV) or generalized cross validation (GCV). By maximizing $\ell(\boldsymbol{\theta})$ with a suitable choice of penalty function, one can ensure that some $\pi_{\ell j}$ are zero and select latent variables automatically. Thus, the procedure combines variable selection and parameter estimation into one step and reduces the computational burden substantially.

Many penalty functions, e.g., the L_κ penalty for some $\kappa \geq 0$, have been used for penalized likelihood and penalized least squares in various parametric models. For example, L_0 is the entropy penalty, the L_1 penalty results in the Lasso proposed by Tibshirani [23], and ridge regression (see, e.g., [6]) corresponds to L_k with some $\kappa \in (0,1)$. Fan and Li [4] also proposed the SCAD penalty which results in an estimator with oracle properties. Its first derivative is given by

$$\phi'_\lambda(\beta) = \lambda\left\{\mathbf{1}(\beta \le \lambda) + \frac{(a\lambda-\beta)_+}{(a-1)\lambda}\mathbf{1}(\beta > \lambda)\right\} \text{ for some } a > 2 \text{ and } \beta > 0, \quad (5)$$

and $\phi_\lambda(0) = 0$, where $\mathbf{1}(\cdot)$ denotes an indicator function and by definition, $x_+ = x\mathbf{1}(x \ge 0)$. The SCAD penalty has two unknown parameters λ and a. Fan and Li [4] suggested using $a = 3.7$ from the Bayesian point of view. In the following development, we consider the SCAD penalty function. However, our proposed method can accommodate more general penalty functions.

2.3. Asymptotic properties

Let $\boldsymbol{\theta}^o$ be the true value of the parameter vector $\boldsymbol{\theta}$, and denote its components by a superscript, for example, $\boldsymbol{\pi}^o$ represents the true value of parameter vector $\boldsymbol{\pi} = \{\pi_{\ell j} : \ell = 1,\ldots,q, j = 1,\ldots,q, \ell \ne j\}$. It is assumed that $\boldsymbol{\pi}^o$ is sparse, i.e., the majority of its components are exactly zero.

Without loss of generality, we assume that the first q_1 components of $\boldsymbol{\pi}^o$ are nonzero and that the last q_2 components of $\boldsymbol{\pi}^o$ are zero; i.e., we can write $\boldsymbol{\pi}^o = (\boldsymbol{\pi}_1^{o\top}, \boldsymbol{\pi}_2^{o\top})^\top = (\boldsymbol{\pi}_1^{o\top}, \mathbf{0}^\top)^\top$, where $\mathbf{0}$ denotes a q_2-dimensional vector of zeros and $q_1+q_2 = q(q-1)$. The above assumption means that some components of $\boldsymbol{\theta}^o$ are exactly zero.

Similarly, we suppose that the first $s = k - q_2$ components of $\boldsymbol{\theta}^o$ are nonzero and that the last q_2 components of $\boldsymbol{\theta}^o$ are zero; i.e., we can write $\boldsymbol{\theta}^o = (\boldsymbol{\theta}_1^{o\top}, \mathbf{0}^\top)^\top$, which indicates that $\boldsymbol{\theta}_1^o$ contains components of $\boldsymbol{\pi}_1^o$.

Let $L_0(\boldsymbol{\theta}_1)$ and $\ell(\boldsymbol{\theta}_1)$ denote the log-likelihood and the penalized log-likelihood of the first s components of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, respectively; i.e., $L_0(\boldsymbol{\theta}_1) \equiv L_0(\boldsymbol{\theta}_1, \mathbf{0})$ and $\ell(\boldsymbol{\theta}_1) \equiv \ell(\boldsymbol{\theta}_1, \mathbf{0})$. Let $\mathbf{J}(\boldsymbol{\theta}_1) = \mathbf{J}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_2 = \mathbf{0}} = \mathbf{J}(\boldsymbol{\theta}_1, \mathbf{0})$ denote the Fisher information matrix evaluated at $\boldsymbol{\theta}_2 = \mathbf{0}$. Considering a more general nonconcave penalty function, we define

$$a_n = \max\{|\phi'_{\lambda_n}(|\pi^o_{\ell j}|)|/\sqrt{n} : \pi^o_{\ell j} \ne 0 \text{ and } \ell \ne j\}$$

and

$$b_n = \max\{|\phi''_{\lambda_n}(|\pi^o_{\ell j}|)|/n : \ \pi^o_{\ell j} \neq 0 \text{ and } \ell \neq j\},$$

where $\phi'_{\lambda_n}(\alpha)$ and $\phi''_{\lambda_n}(\alpha)$ are the first and second derivatives of $\phi_{\lambda_n}(\alpha)$ with respect to α.

Our asymptotic results are based on the following conditions:

(C1) $\phi_{\lambda_n}(0) = 0$, and $\phi_{\lambda_n}(\alpha)$ is symmetric and nonnegative. It is nondecreasing and twice differentiable for all α in $(0, +\infty)$ with at most a few exceptions.

(C2) As $n \to \infty, b_n = o(1)$.

(C3) $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \ \lambda_n^{-1}\phi'_{\lambda_n}(\theta)/n > 0$.

(C4) The Fisher information matrix $\mathbf{J}(\boldsymbol{\theta}) = \mathrm{E}(-\partial^2 L_0(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top)$ is positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}^o$.

(C5) There exist functions $M_{ij\ell}(\mathbf{Y})$ such that $\left|\partial^3 L_0(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j\partial\theta_\ell\right| \leq M_{ij\ell}(\mathbf{Y})$ for all $\boldsymbol{\theta} \in \Theta$, where Θ is a subset of $\mathbb{R}^k$ containing the true parameter $\boldsymbol{\theta}^o$ and $\mathrm{E}_{\theta^o}\{M_{ij\ell}(\mathbf{Y})\} < +\infty$ for all i, j, ℓ.

Conditions (C1) and (C2) are needed for consistent variable selection. Condition (C3) is used to preserve the sparsity property. Regularity Conditions (C4) and (C5) are required to develop the asymptotic theory.

Theorem 1 (Consistency). *Under Conditions (C4) and (C5), if the penalty function ϕ_{λ_n} satisfies Conditions (C1) and (C2), there exists a local maximizer $\hat{\boldsymbol{\theta}}$ of the penalized log-likelihood function* (4) *such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o\| = O_p\{n^{-1/2}(1+a_n)\}$, where $\|\cdot\|$ represents the Euclidean norm.*

When $a_n = O(1)$, $\hat{\boldsymbol{\theta}}$ has the usual convergence rate $n^{-1/2}$. For example, for the SCAD penalty function (5) with $\lambda_n \to 0$ and $\beta > 0$, we have $\phi'_{\lambda_n}(\beta) = \lambda_n(a\lambda_n - \beta)_+/\{(a-1)\lambda_n\} = 0$, which indicates that $a_n \to 0$ if $\lambda_n \to 0$ as $n \to \infty$. Therefore, there exists a $\sqrt{n}$-consistent penalized estimator for $\boldsymbol{\theta}$. Another important property is sparsity, which enables consistent variable selection.

Theorem 2 (Oracle property). *Under Conditions (C1)–(C5), if* $\lambda_n \to 0$, *and* $\sqrt{n}\,\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to* 1, *the* $\sqrt{n}-$*consistent local maximizers* $\hat{\boldsymbol{\theta}} = \big(\hat{\boldsymbol{\theta}}_1^\top, \hat{\boldsymbol{\theta}}_2^\top\big)^\top$ *in Theorem 1 must satisfy:*

(i) (Sparsity) $\hat{\boldsymbol{\theta}}_2 = \mathbf{0}$.

(ii) (Asymptotic normality)

$$\sqrt{n}\left\{\mathbf{J}(\boldsymbol{\theta}_1^o) + \frac{1}{n}\mathbf{Z}(\boldsymbol{\theta}_1^o)\right\}\left[\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o + \frac{1}{n}\left\{\mathbf{J}(\boldsymbol{\theta}_1^o) + \frac{1}{n}\mathbf{Z}(\boldsymbol{\theta}_1^o)\right\}^{-1}\boldsymbol{d}(\boldsymbol{\theta}_1^o)\right] \rightsquigarrow \mathcal{N}_s\big[\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_1^o)\big],$$

where $\mathbf{J}(\boldsymbol{\theta}_1^o) = \mathbf{J}(\boldsymbol{\theta}_1^o, \mathbf{0})$ *is the Fisher information matrix evaluated at* $\boldsymbol{\theta}_2 = \mathbf{0}$,

$$\boldsymbol{d}(\boldsymbol{\theta}_1) = (d_1(\boldsymbol{\theta}_1), \ldots, d_s(\boldsymbol{\theta}_1))^\top, \quad \mathbf{Z}(\boldsymbol{\theta}_1) = \mathrm{diag}\{Z_1(\boldsymbol{\theta}_1), \ldots, Z_s(\boldsymbol{\theta}_1)\},$$

$$d_j(\boldsymbol{\theta}_1) = \begin{cases} \phi'_{\lambda_n}(|\pi_{\ell t}|)\ \mathrm{sgn}(\pi_{\ell t}), & \text{if } \theta_{1j} = \pi_{\ell t} \in \boldsymbol{\pi}_1, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$Z_j(\boldsymbol{\theta}_1) = \frac{\partial}{\partial \theta_{1j}}\, d_j(\boldsymbol{\theta}_1) = \begin{cases} \phi''_{\lambda_n}(|\pi_{\ell t}|), & \text{if } \theta_{1j} = \pi_{\ell t} \in \boldsymbol{\pi}_1, \\ 0\ , & \text{otherwise.} \end{cases}$$

Note that, for the SCAD penalty function, when $\lambda_n \to 0$ as $n \to \infty$ and Condition (2) holds, we have $a_n \to 0$, $\boldsymbol{d}(\boldsymbol{\theta}_1^o) \to \mathbf{0}$ and $\mathbf{Z}(\boldsymbol{\theta}_1^o) \to \mathbf{0}$. Thus, it follows from Theorem 2 that the SCAD-based penalized likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ has the oracle property. To wit, the zero components in $\boldsymbol{\theta}^0$ are estimated as 0 with probability approaching 1, and the nonzero components in $\boldsymbol{\theta}^0$ are estimated as well as in the case that zero components are known.

3. Computational procedure

3.1. Maximization of the penalized log-likelihood function via an ECM algorithm

Let $\mathbf{X} = \{\mathbf{Y}, \boldsymbol{\Omega}\}$ denote the completed data set. After dropping the normalizing constant, we can write the completed-data log-likelihood function as

$$\begin{aligned} L_c(\boldsymbol{\theta}|\mathbf{X}) \quad = \quad & -\frac{n}{2}\sum_{j=1}^{p}\ln(\tau_j^2) - \frac{n}{2}\sum_{\ell=1}^{q}\ln(\gamma_\ell^2) + n\ln|\mathbf{I} - \boldsymbol{\Pi}| \\ & - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij} - \mu_j - \boldsymbol{\Lambda}_j\boldsymbol{\omega}_i)^2/\tau_j^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{q}(\omega_{i\ell} - \boldsymbol{\Pi}_\ell\boldsymbol{\omega}_i)^2/\gamma_\ell^2, \end{aligned}$$

where $\mathbf{\Lambda}_j$ and $\mathbf{\Pi}_\ell$ denote the jth and ℓth rows of $\mathbf{\Lambda}$ and $\mathbf{\Pi}$, respectively. The E-step of the ECM algorithm is to compute the following Q function:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathrm{E}\big\{L_c(\boldsymbol{\theta}|\mathbf{X})\big|\mathbf{Y},\boldsymbol{\theta}^{(r)}\big\} - \sum_{\ell=1}^{q}\sum_{j=1,j\neq\ell}^{q}\phi_{\lambda_n}(|\pi_{\ell j}|), \tag{6}$$

where the expectation is taken with respect to the conditional distribution of $\mathbf{\Omega}$ given $\mathbf{Y}$ and $\boldsymbol{\theta}^{(r)}$, and $\boldsymbol{\theta}^{(r)}$ is the rth iterated value of $\boldsymbol{\theta}$. Since $\boldsymbol{\omega}_i|\boldsymbol{\theta} \sim \mathcal{N}_q(\mathbf{0},\mathbf{\Sigma})$ and $\boldsymbol{y}_i|(\boldsymbol{\omega}_i,\boldsymbol{\theta}) \sim \mathcal{N}_p(\boldsymbol{\mu}+\mathbf{\Lambda}\boldsymbol{\omega}_i,\mathbf{\Psi}_\epsilon)$, the conditional distribution of $\boldsymbol{\omega}_i$ given $(\boldsymbol{y}_i,\ \boldsymbol{\theta})$ is $\boldsymbol{\omega}_i|\boldsymbol{y}_i,\boldsymbol{\theta} \sim \mathcal{N}_q(\boldsymbol{\alpha}_i,\mathbf{D})$, where $\boldsymbol{\alpha}_i = \mathbf{D}\mathbf{\Lambda}^\top\mathbf{\Psi}_\epsilon^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu})$ and $\mathbf{D} = (\mathbf{\Lambda}^\top\mathbf{\Psi}_\epsilon^{-1}\mathbf{\Lambda}+\mathbf{\Sigma}^{-1})^{-1}$. Thus, we have

$$\mathrm{E}(\boldsymbol{\omega}_i|\boldsymbol{y}_i,\boldsymbol{\theta}) = \boldsymbol{\alpha}_i \tag{7}$$

and

$$\mathrm{E}(\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top|\boldsymbol{y}_i,\boldsymbol{\theta}) = \mathrm{var}(\boldsymbol{\omega}_i|\boldsymbol{y}_i,\boldsymbol{\theta}) + \mathrm{E}(\boldsymbol{\omega}_i|\boldsymbol{y}_i,\boldsymbol{\theta})\mathrm{E}^\top(\boldsymbol{\omega}_i|\boldsymbol{y}_i,\boldsymbol{\theta}) = \mathbf{D}+\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top. \tag{8}$$

The M-step is to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ by updating $\boldsymbol{\theta}^{(r)}$ to obtain $\boldsymbol{\theta}^{(r+1)}$. For $\boldsymbol{\theta}$ except the unknown parameters in $\mathbf{\Pi}$, maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ is equivalent to solving the following system of equations:

$$\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})}{\partial\boldsymbol{\theta}} = \mathrm{E}\left\{\frac{\partial L_c(\boldsymbol{\theta}|\mathbf{X})}{\partial\boldsymbol{\theta}}\middle|\mathbf{Y},\boldsymbol{\theta}^{(r)}\right\} = \mathbf{0}.$$

For $j = 1,\ldots,p$ and $\ell = 1,\ldots,q$, it is easy to verify that

$$\begin{aligned}
\tfrac{\partial}{\partial\mu_j}L_c(\boldsymbol{\theta}|\mathbf{X}) &= \textstyle\sum_{i=1}^{n}(y_{ij}-\mu_j-\mathbf{\Lambda}_j\boldsymbol{\omega}_i)/\tau_j^2,\\
\tfrac{\partial}{\partial\tau_j^2}L_c(\boldsymbol{\theta}|\mathbf{X}) &= -n/(2\tau_j^2)+\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(y_{ij}-\mu_j-\mathbf{\Lambda}_j\boldsymbol{\omega}_i)^2/\tau_j^4,\\
\tfrac{\partial}{\partial\gamma_\ell^2}L_c(\boldsymbol{\theta}|\mathbf{X}) &= -n/(2\gamma_\ell^2)+\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\omega_{i\ell}-\mathbf{\Pi}_\ell\boldsymbol{\omega}_i)^2/\gamma_\ell^4,\\
\tfrac{\partial}{\partial\mathbf{\Lambda}_j}L_c(\boldsymbol{\theta}|\mathbf{X}) &= \textstyle\sum_{i=1}^{n}\boldsymbol{\omega}_i^\top(y_{ij}-\mu_j-\mathbf{\Lambda}_j\boldsymbol{\omega}_i)/\tau_j^2.
\end{aligned} \tag{9}$$

Closed-form solutions are not available for these equations. Based on the ECM idea as described, e.g., in [17], the M-step can be conducted by several computationally simpler conditional maximization steps. Conditionally on other parameters, the solution of each individual equation given in Eq. (9) can be obtained. The solution for the M-step is given as follows, in conjunction with Eq. (7) and Eq. (8):

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\{y_{ij} - \mathbf{\Lambda}_j \mathrm{E}(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{\theta})\}, \quad \hat{\tau}_j^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{(y_{ij} - \mu_j - \mathbf{\Lambda}_j\boldsymbol{\omega}_i)^2|\boldsymbol{y}_i, \boldsymbol{\theta}\},$$

$$\hat{\gamma}_\ell^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{(\omega_{i\ell} - \mathbf{\Pi}_\ell\boldsymbol{\omega}_i)^2|\boldsymbol{y}_i, \boldsymbol{\theta}\}, \quad \ell = 1, \ldots, q,$$

$$\hat{\mathbf{\Lambda}}_j^\top = \left\{\sum_{i=1}^{n}\mathrm{E}(\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top|\boldsymbol{y}_i, \boldsymbol{\theta})\right\}^{-1}\left\{\sum_{i=1}^{n}\mathrm{E}(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{\theta})(y_{ij} - \mu_j)\right\}, \quad j = 1, \ldots, p.$$

Now we consider the estimation of the unknown parameters in $\mathbf{\Pi}$. Since $\phi_{\lambda_n}(|\pi_{\ell j}|)$ satisfying Condition (C1) may be not differentiable at $\pi_{\ell j} = 0$, the Newton–Raphson algorithm cannot be applied directly in the current situation unless it is properly adapted to deal with the single non-smooth point at $\pi_{\ell j} = 0$. Generally, the local linear approximation [28] can be employed to address this issue. However, it is rather difficult to approximate the objective function (6) as a penalized squared loss function with respect to the components of $\mathbf{\Pi}$ because the local approximation of the log-likelihood involves the matrix $(\mathbf{I} - \mathbf{\Pi})^{-2}$. Hence, following the idea of the MM algorithm of Hunter and Li [9], we approximate the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ function specified in Eq. (6) with the following surrogate function

$$\begin{aligned} Q_\zeta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) &= \mathrm{E}\Big\{L_c(\boldsymbol{\theta}|\mathbf{X})|\mathbf{Y}, \boldsymbol{\theta}^{(r)}\Big\} - \sum_{\ell=1}^{q}\sum_{j=1, j\neq\ell}^{q}\phi_{\lambda_n,\zeta}(|\pi_{\ell j}|) \\ &= \mathrm{E}\Big\{L_c(\boldsymbol{\theta}|\mathbf{X})|\mathbf{Y}, \boldsymbol{\theta}^{(r)}\Big\} - \sum_{\ell=1}^{q}\sum_{j=1, j\neq\ell}^{q}\left\{\phi_{\lambda_n}(|\pi_{\ell j}|) - \zeta\int_0^{|\pi_{\ell j}|}\frac{\phi'_{\lambda_n}(t)}{\zeta + t}\,\mathrm{d}t\right\}, \end{aligned}$$

where ζ is a very small perturbation to prevent any component of the estimate from getting stuck at zero. In the neighborhood of the rth approximation value $\mathbf{\Pi}^{(r)}$ of $\mathbf{\Pi}$ obtained by the MM algorithm, we can further approximate $Q_\zeta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ by

$$\begin{aligned} S_{r,\zeta}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) &= \mathrm{E}\{L_c(\boldsymbol{\theta}|\mathbf{X})|\mathbf{Y}, \boldsymbol{\theta}^{(r)}\} \\ &\quad - \sum_{\ell=1}^{q}\sum_{j=1, j\neq\ell}^{q}\left[\phi_{\lambda_n,\zeta}(|\pi_{\ell j}^{(r)}|) + \frac{\phi'_{\lambda_n}(|\pi_{\ell j}^{(r)}|+)}{2(\zeta + |\pi_{\ell j}^{(r)}|)}\{\pi_{\ell j}^2 - (\pi_{\ell j}^{(r)})^2\}\right]. \end{aligned}$$

Therefore, starting from $\pi_{\ell j}^{(r)}$, the one-step update is given by

$$\pi_{\ell j}^{(r+1)} = \arg \max_{\mathbf{\Pi}} \ S_{r,\zeta}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}),$$

which is reduced to a maximization problem on a quadratic function so that the Newton–Raphson algorithm can be applied. Given a tolerance ν, the MM algorithm is deemed to have converge if

$$\forall_{\ell \neq j \in \{1,\ldots,q\}} \quad \left| \frac{\partial Q_\zeta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})}{\partial \pi_{\ell j}} \right| < \frac{\nu}{2}.$$

Finally, the $\pi_{\ell j}$'s satisfying the following condition are set to zero:

$$\left| \frac{\partial Q_\zeta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})}{\partial \pi_{\ell j}} - \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})}{\partial \pi_{\ell j}} \right| = \frac{n\zeta\phi'_{\lambda_n}(|\pi_{\ell j}|+)}{\zeta + |\pi_{\ell j}|} > \frac{\nu}{2}.$$

The ordinary MLEs of $\mathbf{\Pi}$ can be taken as the initial values $\mathbf{\Pi}^{(0)}$ for the MM algorithm. The perturbation ζ should be kept small so that the difference between Q_ζ and Q is negligible. We use the following value suggested by Hunter and Li [9]:

$$\zeta = \frac{\nu}{2n\lambda_n} \min\{|\pi_{\ell j}^{(0)}|\colon \ \pi_{\ell j}^{(0)} \neq 0, \ \ell \neq j, \text{ and } \ell, j = 1, \ldots, q\}.$$

We employ the ECM algorithm, in which $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ is computed and the MM algorithm is used to solve the optimization problem on $\mathbf{\Pi}$. Then we obtain the updated penalized likelihood estimates of the parameters. This process is repeated iteratively until convergence. From the idea of the MM algorithm (see, e.g., [9]), it is easily seen that the above ECM algorithm can be applied to the penalty functions satisfying Condition (C1), such as the hard thresholding, Lasso, SCAD and L_κ penalty with $0 < \kappa \leq 1$.

To identify the model, suitable entries in $\mathbf{\Lambda}$ are fixed at known values. To deal with this situation, in general, we consider the following linear transformations $\mathbf{\Lambda}_j^\top = \mathbf{A}_j \mathbf{\Lambda}_j^{*\top} + \boldsymbol{a}_j$ for $j = 1, \ldots, p$, where $\boldsymbol{a}_j$ is a $q \times 1$ constant column vector, $\mathbf{A}_j$ is a full column rank selection matrix of size $q \times r_j$, and $\mathbf{\Lambda}_j^*$ is a reduced unknown parameter vector of $\mathbf{\Lambda}_j$ of size $1 \times r_j$. We then have

$$\frac{\partial L_c(\boldsymbol{\theta}|\mathbf{X})}{\partial \mathbf{\Lambda}_j^{*\top}} = \mathbf{A}_j^\top \sum_{i=1}^n \frac{\boldsymbol{\omega}_i(y_{ij} - \mu_j - \mathbf{\Lambda}_j \boldsymbol{\omega}_i)}{\tau_j^2},$$

which yields

$$\hat{\mathbf{\Lambda}}_j^{*\top} = \left\{ \mathbf{A}_j^\top \sum_{i=1}^n \mathrm{E}(\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top | \boldsymbol{y}_i, \boldsymbol{\theta}) \mathbf{A}_j \right\}^{-1} \mathbf{A}_j^\top \sum_{i=1}^n \left\{ \mathrm{E}(\boldsymbol{\omega}_i | \boldsymbol{y}_i, \boldsymbol{\theta})(y_{ij} - \mu_j) - \mathrm{E}(\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top | \boldsymbol{y}_i, \boldsymbol{\theta}) \boldsymbol{a}_j \right\}.$$

3.2. Standard error estimates

To construct a confidence region for the nonzero parameter vector $\hat{\boldsymbol{\theta}}_1$, we need to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}_1$, which is a rather difficult task due to the complicated form of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_1$ given in Theorem 2. Following Hunter and Li [9], the sandwich method can be adopted to estimate the standard errors of the components in $\hat{\boldsymbol{\theta}}_1$. Based on the observed-data log-likelihood function, we can estimate the covariance matrix of $\hat{\boldsymbol{\theta}}_1$ by using the corresponding submatrix of the sandwich covariance estimator, viz.

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\theta}}) = \{\nabla^2 L_0(\hat{\boldsymbol{\theta}}) - \mathbf{B}\}^{-1} \widehat{\mathrm{cov}}\{\mathbf{D}(\hat{\boldsymbol{\theta}})\} \{\nabla^2 L_0(\hat{\boldsymbol{\theta}}) - \mathbf{B}\}^{-1},$$

where

$$\begin{aligned} \widehat{\mathrm{cov}}\{\mathbf{D}(\hat{\boldsymbol{\theta}})\} &= \frac{1}{n} \sum_{i=1}^n \left\{ \nabla L_{0i}(\hat{\boldsymbol{\theta}}) - \mathbf{B}\hat{\boldsymbol{\theta}} \right\} \left\{ \nabla L_{0i}(\hat{\boldsymbol{\theta}}) - \mathbf{B}\hat{\boldsymbol{\theta}} \right\}^\top \\ &\quad - \left\{ \frac{1}{n} \sum_{i=1}^n \nabla L_{0i}(\hat{\boldsymbol{\theta}}) - \mathbf{B}\hat{\boldsymbol{\theta}} \right\} \left\{ \frac{1}{n} \sum_{i=1}^n \nabla L_{0i}(\hat{\boldsymbol{\theta}}) - \mathbf{B}\hat{\boldsymbol{\theta}} \right\}^\top \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \nabla L_{0i}(\hat{\boldsymbol{\theta}}) \right\} \left\{ \nabla L_{0i}(\hat{\boldsymbol{\theta}}) \right\}^\top - \left\{ \frac{1}{n} \sum_{i=1}^n \nabla L_{0i}(\hat{\boldsymbol{\theta}}) \right\} \left\{ \frac{1}{n} \sum_{i=1}^n \nabla L_{0i}(\hat{\boldsymbol{\theta}}) \right\}^\top, \end{aligned}$$

and $\mathbf{B}$ is a diagonal matrix with entry $\phi'_{\lambda_n}(|\pi_{\ell j}|_+)/(\zeta + |\pi_{\ell j}|)$ corresponding to $\pi_{\ell j}$ ($\ell = 1, \ldots, q; j = 1, \ldots, q; \ell \neq j$), and entry 0 corresponding to the other parameters; in the above, ∇L_{0i} and $\nabla^2 L_0$ are the first- and second-order derivatives of L_{0i} and L_0, respectively. Detailed expressions for these derivatives are given in Appendix B. For the estimated standard error of components in $\hat{\boldsymbol{\theta}}_2 = \mathbf{0}$, the sandwich formula gives a zero standard error estimate; see, e.g., [4, 23].

Under the normality assumption, the observed-data likelihood function is a Gaussian distribution function satisfying Conditions (E)–(G) given in Fan and Peng [5]. Therefore, proceeding as in the latter, we can show that the sandwich covariance estimator is a consistent estimator of the covariance matrix for $\hat{\boldsymbol{\theta}}_1$; we omit the details. The sandwich covariance estimator can still provide a consistent estimator of the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_1$ even when the normality assumption is incorrect, although in this situation $\hat{\boldsymbol{\theta}}_1$ may be biased; see, e.g., [7, 11], which provide comprehensive reviews of sandwich estimation.

3.3. Choice of tuning parameters

To ensure that $\hat{\boldsymbol{\theta}}_{\lambda_n}$ has good properties, the tuning parameter λ_n has to be suitably selected. This can be accomplished via minimizing a certain criterion such as the AIC, BIC or GCV in the presence of latent variables. However, it has been shown in Wang et al. [24] that even in the simple linear model, the GCV criterion can lead to a significant overfit. Thus, we use the $\mathrm{IC_Q}$ criterion (an AIC/BIC-type criterion) suggested by Garia et al. [8] to select the optimal tuning parameter λ_n by minimizing

$$\mathrm{IC_Q}(\lambda_n) = -2Q_1(\hat{\boldsymbol{\theta}}_{\lambda_n}|\hat{\boldsymbol{\theta}}_0) + c_n(\hat{\boldsymbol{\theta}}_{\lambda_n}),$$

where $Q_1(\hat{\boldsymbol{\theta}}_{\lambda_n}|\hat{\boldsymbol{\theta}}_0) = \mathrm{E}\{L_c(\hat{\boldsymbol{\theta}}_{\lambda_n}|\mathbf{X})|\mathbf{Y}, \hat{\boldsymbol{\theta}}_0\}$, $\hat{\boldsymbol{\theta}}_0$ is the unpenalized MLE of $\boldsymbol{\theta}$ and $c_n(\boldsymbol{\theta})$ is a function of the data and the fitted model. For instance, if $c_n(\boldsymbol{\theta})$ equals twice the total number of parameters, then we obtain an AIC-type criterion; alternatively, we obtain a BIC-type criterion when $c_n(\boldsymbol{\theta}) = \dim(\boldsymbol{\theta}) \ln n$.

4. Simulation studies

In this section, simulation studies are conducted to investigate the finite-sample performance of the proposed methods.

4.1. Experiment 1

In this experiment, data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are generated from a SEM specified by Eqs. (1) and (2) with six observed variables (i.e., $p = 6$) and three latent variables (i.e., $q = 3$), and with $n = 300$ and 500. For the measurement model (1), we take the intercept vector $\boldsymbol{\mu}$ to be $2\mathbf{1}_6$ and assume the error vector to be Gaussian, i.e., $\boldsymbol{\epsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_6(\mathbf{0}, 0.8\mathbf{I}_6)$. Furthermore, we consider the following structure of the loading matrix $\boldsymbol{\Lambda}$:

$$\boldsymbol{\Lambda}^\top = \begin{pmatrix} 1 & \lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_{42} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{63} \end{pmatrix},$$

where $\lambda_{21} = \lambda_{42} = \lambda_{63} = 0.8$. For the structural equation (2), we set $q_1 = 2$ and $q_2 = 1$, i.e., $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2})^\top$ and $\boldsymbol{\xi}_i = \xi_i$, take $\boldsymbol{\Gamma} = (2, 3)^\top$ and $\mathbf{B} = (b_{jk})$ with $b_{11} = b_{12} = b_{22} = 0$ and $b_{21} = 1$, and assume $\boldsymbol{\zeta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$, $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $\boldsymbol{\zeta}_i$ is independent of ξ_i.

To investigate the performance of the above proposed latent variable selection procedure, we fit the above generated dataset to the SEM defined in Eq. (1) and Eq. (3). To this end, we first note that the above specified model (2) can be written in the form (3) by letting $\boldsymbol{\omega}_i = (\eta_{i1}, \eta_{i2}, \xi_i)^\top$ using

$$\boldsymbol{\Pi} = \begin{pmatrix} 0 & \pi_{12} & \pi_{13} \\ \pi_{21} & 0 & \pi_{23} \\ \pi_{31} & \pi_{32} & 0 \end{pmatrix}$$

and $\boldsymbol{\delta}_i = (\boldsymbol{\zeta}_i^\top, \xi_i)^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}_3(\mathbf{0}, \mathbf{I}_3)$. For identifiability, the 1's and 0's in $\boldsymbol{\Lambda}$ and the diagonal elements in $\boldsymbol{\Pi}$ are treated as known parameters. Thus, the true values of the unknown parameters in the fitted SEM are $\boldsymbol{\mu} = 2\mathbf{1}_6$, $\lambda_{21} = \lambda_{42} = \lambda_{63} = 0.8$, $\pi_{13} = 2.0$, $\pi_{21} = 1.0$, $\pi_{23} = 3.0$ and $\pi_{12} = \pi_{31} = \pi_{32} = 0.0$, $\boldsymbol{\Psi}_\epsilon = \text{diag}(\tau_1^2, \ldots, \tau_p^2) = 0.8\mathbf{I}_6$ and $\boldsymbol{\Psi}_\delta = \text{diag}(\gamma_1^2, \ldots, \gamma_q^2) = \mathbf{I}_3$. We have 24 unknown parameters in $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_6, \lambda_{21}, \lambda_{42}, \lambda_{63}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{23}, \pi_{31}, \pi_{32}, \tau_1^2, \ldots, \tau_6^2, \gamma_1^2, \gamma_2^2, \gamma_3^2)^\top$. For the oracle situation, because $\pi_{12} = \pi_{31} = \pi_{32} = 0.0$ are prespecified, the number of unknown parameters is 21.

Based on the above generated data $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, we use the ECM algorithm introduced in Section 3.1 to compute the Lasso and SCAD penalized MLEs of $\boldsymbol{\theta}$, respectively. For both Lasso and SCAD penalties, the tolerance ν is set to 10^{-5}. For the SCAD penalty, we set $a = 3.7$ as suggested by Fan and Li [4], and the tuning parameter is selected via the BIC-type criterion. For comparison, we compute the unpenalized MLE (denoted UnPenlM), and also consider the oracle estimate of $\boldsymbol{\theta}$ as a benchmark, although it is not feasible in the real data analysis.

To compare the performance of the Lasso and SCAD variable selection methods in the linear SEM with different sample sizes, we first compute the MRRSE, which denotes the median of the ratios of root square errors. The estimates of $\mathbf{\Pi}$ are denoted by $\hat{\mathbf{\Pi}}_{(t)}$ for $t = 0, 1, 2, 3$ corresponding to the unpenalized, Lasso penalized, SCAD penalized MLEs and the oracle estimates, respectively. Based on the root square error

$$\mathrm{RSE}(\hat{\mathbf{\Pi}}_{(t)}) = \sqrt{\sum_{l \neq j} (\hat{\pi}_{(t)lj} - \pi_{\ell j})^2},$$

we compute the ratio $\mathrm{RSE}(\hat{\mathbf{\Pi}}_{(t)})/\mathrm{RSE}(\hat{\mathbf{\Pi}}_{(0)})$ for each simulated data set with $t = 1, 2, 3$. The median of the ratios over the 100 simulated data sets is called MRRSE. Their values are reported in the third column of Table 1. The fourth column of Table 1 labeled "Correct" presents the average number restricted only to the true zero coefficients, and the fifth column labeled "Incorrect" reports the average number of coefficients erroneously set to be 0. Small "Incorrect" values and "Correct" values that are closest to the number of true zero coefficients are preferred.

We also consider two other measures to evaluate the performance of the two variable selection procedures. Let $\mathcal{A}$ be the set containing the subscripts of true nonzero entries in $\mathbf{\Pi}$ and $\widehat{\mathcal{A}}_m$ be the set containing the subscripts of the estimated nonzero entries of $\mathbf{\Pi}$ in the mth simulation. The proportion of cases in which the true model is selected is then defined by

$$\mathrm{TM} = \frac{1}{M} \sum_{m=1}^{M} I(\widehat{\mathcal{A}}_m = \mathcal{A})$$

and the proportion that not all the nonzero entries is selected is given by

$$\mathrm{FM} = \frac{1}{M} \sum_{m=1}^{M} I(\widehat{\mathcal{A}}_m \cap \mathcal{A} \neq \mathcal{A}).$$

It is easy to see that larger TM values and smaller FM values are preferred. Results based on 100 replicates are reported in Table 1.

From Table 1, the MRRSE values for the Lasso and SCAD penalized MLEs are close to 1, indicating that the performance of the proposed SCAD and Lasso estimates is satisfactory, relative to the unpenalized MLEs. In addition, in terms of "Incorrect" and "Correct" values, the SCAD-based latent variable selection procedure outperforms the Lasso-based latent variable selection procedure regardless of the sample size n. The SCAD has larger TM values and smaller FM values than the Lasso, implying that the SCAD method performs uniformly better than the Lasso in terms of the TM and FM criteria. Finally, for the unpenalized MLE method, the corresponding "Correct," "Incorrect," TM and FM values are almost zero, indicating that the unpenalized MLE procedure cannot detect the true outcome/explanatory latent variables.

The frequencies of the identified nonzero components in 100 replicates for the different methods are reported in Table 2. From Table 2, we can observe that the Lasso and SCAD methods select the important variables with high frequency and unimportant variables with low frequency, which implies that both can identify well the structure of Eq. (2) via Eq. (3); furthermore, the SCAD method behaves uniformly better than the Lasso. However, the unpenalized MLE method cannot identify the true structure of the model because the unpenalized MLEs are almost nonzero.

For the unpenalized, Lasso penalized and SCAD penalized MLEs of $\boldsymbol{\theta}$ based on $M = 100$ replicates, we compute the empirical bias (Bias), root mean square (RMS) and empirical standard deviation (SD) between the estimated value and the true value, viz.

$$\begin{aligned}
\mathrm{Bias}(\hat{\theta}_i) &= \left| \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_i(m) - \theta_i \right|, \\
\mathrm{RMS}(\hat{\theta}_i) &= \left[\frac{1}{M} \sum_{m=1}^{M} \{\hat{\theta}_i(m) - \theta_i\}^2 \right]^{1/2}, \\
\mathrm{SD}(\hat{\theta}_i) &= \left[\frac{1}{M-1} \sum_{m=1}^{M} \left\{ \hat{\theta}_i(m) - \frac{\hat{\theta}_i(1) + \cdots + \hat{\theta}_i(M)}{M} \right\}^2 \right]^{1/2},
\end{aligned}$$

and $\hat{\theta}_i(m)$ is the estimate of θ_i in the mth replicate. Figure 1 shows the results for $n = 300$ and 500. From Figure 1, we can observe that (i) the Lasso and SCAD penalized MLEs are quite close to their corresponding true parameter values; (ii) the performance of the latent variable selection procedure affects the accuracy of parameter estimation, i.e., the accuracy of the Lasso and SCAD penalized MLEs is slightly better than that of the unpenalized MLE for true zero components of $\mathbf{\Pi}$ in terms of the RMS and SD values, regardless of sample sizes.

4.2. Experiment 2

In this experiment, we generate the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{500}$ from a SEM defined in Eqs. (1) and (2) with twenty manifest variables (i.e., $p = 20$) and ten latent variables (i.e., $q = 10$). For the measurement model (1), we set $\boldsymbol{\mu} = 2\mathbf{1}_{20}$, assume $\boldsymbol{\epsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{20}(\mathbf{0}, 0.8\mathbf{I}_{20})$, and consider the loading matrix $\mathbf{\Lambda}$:

$$\mathbf{\Lambda}^\top = \begin{pmatrix} 1 & \lambda_{2,1} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \lambda_{4,2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{6,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & \lambda_{20,10} \end{pmatrix},$$

where $\lambda_{2k,k} = 0.8$ for $k = 1, \ldots, 10$. For the structural equation (2), we set $q_1 = 1$ and $q_2 = 9$, take $\eta_i = 0.8\omega_{i2} + \omega_{i3} + \zeta_i$ and $\boldsymbol{\xi}_i = (\omega_{i2}, \ldots, \omega_{i,10})^\top$, which indicates that $\mathbf{B} = b_{11} = 0$, $\mathbf{\Gamma} = (0.8, 1, 0, 0, 0, 0, 0, 0, 0)$, and assume $\zeta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, $\boldsymbol{\xi}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_9(\mathbf{0}, \mathbf{I}_9)$ and ζ_i is independent of $\boldsymbol{\xi}_i$.

We fit the above generated dataset via the SEM defined in Eq. (1) and Eq. (3). In this case, the above specified structural equation (2) can be expressed as the form of structural equation (3) by letting $\boldsymbol{\omega}_i = (\eta_i, \omega_{i2}, \ldots, \omega_{i,10})^\top$, taking

$\boldsymbol{\delta}_i = (\zeta_i, \boldsymbol{\xi}_i^\top)^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{10}(\mathbf{0}, \mathbf{I}_{10})$ and specifying the following structure of $\boldsymbol{\Pi}$:

$$\boldsymbol{\Pi} = \begin{pmatrix} 0 & \pi_{1,2} & \cdots & \pi_{1,9} & \pi_{1,10} \\ \pi_{2,1} & 0 & \cdots & \pi_{2,9} & \pi_{2,10} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \pi_{9,1} & \pi_{9,2} & \cdots & 0 & \pi_{9,10} \\ \pi_{10,1} & \pi_{10,2} & \cdots & \pi_{10,9} & 0 \end{pmatrix},$$

where $\pi_{j,k} = 0$ for $j \neq k$ except for $\pi_{1,2} = 0.8$ and $\pi_{1,3} = 1.0$. For identifiability, the 1's and 0's in $\boldsymbol{\Lambda}$ and the diagonal elements in $\boldsymbol{\Pi}$ are treated as known parameters. Thus, in the fitted SEM, the true values of unknown parameters in $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \lambda_{2,1}, \ldots, \lambda_{20,10}, \pi_{1,2}, \ldots, \pi_{10,9}, \tau_1^2, \ldots, \tau_p^2, \gamma_1^2, \ldots, \gamma_q^2\}$ are $\boldsymbol{\mu} = 2\mathbf{1}_{20}$, $\lambda_{2k,k} = 0.8$ for $k = 1, \ldots, 10$, $\pi_{j,k} = 0$ for $j \neq k$ except for $\pi_{1,2} = 0.8$ and $\pi_{1,3} = 1.0$, $\boldsymbol{\Psi}_\epsilon = \text{diag}(\tau_1^2, \ldots, \tau_p^2) = 0.8\mathbf{I}_{20}$ and $\boldsymbol{\Psi}_\delta = \text{diag}(\gamma_1^2, \ldots, \gamma_q^2) = \mathbf{I}_{10}$. Hence, there are 150 unknown parameters in the fitted SEM. For the oracle situation, the number of unknown parameters is 62.

As in Experiment 1, we compute the oracle estimate, the unpenalized, Lasso penalized, and SCAD penalized MLEs of $\boldsymbol{\theta}$, respectively. Results for 100 replicates are reported in Table 3 and Table 4. From Table 3, we observe that the MRRSE values for the Lasso and SCAD penalized MLEs are less than 1, indicating that the SCAD and Lasso MLEs outperform the unpenalized MLE. The "Correct" value for the SCAD method is more close to the true zero number (i.e., 88), and the "Incorrect" value for the SCAD method is smaller than that for the Lasso. In addition, the SCAD method has a larger TM value and a smaller FM value than the Lasso. Yet the unpenalized MLE cannot identify the structure of the model. From Table 4, we can find that the SCAD method selects the important elements with higher frequency and unimportant elements with lower frequency than the Lasso, and the unpenalized MLEs are almost nonzero, which implies that the Lasso and SCAD methods perform well in terms of latent variable selection whereas the unpenalized MLE method behaves poor. We also plot heatmaps of frequencies that each component in the coefficient matrix $\boldsymbol{\Pi}$ is estimated as zero out of 100 replicates in Figure 2. Examination of Figure 2 shows that the SCAD method can well recover the sparsity structure of the coefficient matrix $\mathbf{\Pi}$.

5. A real example

In the project of WORLD VALUES SURVEY 1981–1984 AND 1990–1993 (World Value Study Group, ICPSR Version), the Inter-university Consortium for Political and Social Research (ICPSR) data were collected in 45 societies around the world on broad topics such as work, religious belief, the meaning and purpose of life, family life, contemporary social issues, etc. Lee and Zhu [15] analyzed a small portion of the data set gathered only from the United Kingdom. To illustrate the above proposed methods, we reanalyze the UK data subset, which can be obtained from the authors upon the approval of the ICPSR funding agencies.

Six variables (i.e., Variables 180, 96, 62, 176, 116 and 117 in the UK data subset) that are related with respondents' job, religious belief, and homelife are taken as manifest variables, denoted by $\boldsymbol{y} = (y_1, \ldots, y_6)^\top$, where

y_1: Overall, how satisfied or dissatisfied are you with your home life? (V180)

y_2: All things considered, how satisfied are you with your life as a whole these days? (V96)

y_3: Thinking about your reasons for doing voluntary work, how important are religious beliefs in your own case? (V62)

y_4: How important is God in your life? (V176)

y_5: Overall, how satisfied or dissatisfied are you with your job? (V116)

y_6: How free are you to make decisions in your job? (V117)

We note that (y_1, y_2) are related to life, (y_3, y_4) are related to religious belief, and (y_5, y_6) are related to job satisfaction. Variable 62 (i.e., y_3) was measured by a 5-point scale, while all others were measured by a 10-point scale. To illustrate the above proposed methodologies, we regard these variables as continuous variables. After deleting cases with missing data, the sample size is 196.

To model the data without prior knowledge of the relationship among the latent factors, we consider the SEM defined in Eq. (1) and Eq. (3) with the following specifications:

$$\mathbf{\Lambda}^{\top} = \begin{pmatrix} 1 & \lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_{42} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{63} \end{pmatrix}, \quad \mathbf{\Pi} = \begin{pmatrix} 0 & \pi_{12} & \pi_{13} \\ \pi_{21} & 0 & \pi_{23} \\ \pi_{31} & \pi_{32} & 0 \end{pmatrix},$$

where the 1's and 0's in $\mathbf{\Lambda}$ and the diagonal entries in $\mathbf{\Pi}$ are treated as fixed known parameters. Thus, we have a total of 24 parameters. For comparison, we compute simultaneously the unpenalized, Lasso penalized and SCAD penalized MLEs. The tuning parameter λ_n in the penalty function is chosen by minimizing the $\mathrm{IC_Q}$ introduced in Section 3.3. In the present case, λ_n is 0.135. The results are reported in Table 5. Latent factors w_1, w_2 and w_3 can be roughly interpreted as "life," "religious belief" and "job satisfaction" factors.

From Table 5, we can see that the structures of $\mathbf{\Pi}$ chosen by the proposed two variable selection approaches are the same, but different from that obtained from the unpenalized method. The SCAD and Lasso methods identify π_{12} and π_{13} as nonzero components. Thus, the "life" factor is detected as an outcome latent variable, while the "religious belief" and "job satisfaction" factors are identified as explanatory latent variables. Moreover, the latter two factors have a positive effect on the "life" factor. By comparing these results with those given in Lee and Zhu [15], one can see that the proposed method is successful in implementing latent variable selection and parameter estimation.

6. Discussion

We have introduced the penalized maximum likelihood approach to identify the structure of latent variable model in the framework of structural equation models. The proposed procedure is shown to be consistent and to have the oracle property given suitable choices of penalty function and tuning parameter. We have regarded the basic latent variables as missing data and applied the ECM algorithm to obtain the penalized maximum likelihood estimates. The second M-step is implemented via the MM algorithm. Moreover, we use the IC_{Q} criterion to choose the tuning parameter and develop the standard error formulas of our estimators. Our simulations and illustration show that the proposed method is effective.

This feature of the proposed method provides two distinct advantages. First, the method does not require any prior knowledge of the nature of the modeled relation but rather relies on a variable selection procedure to identify the outcome and explanatory latent variables and to approximate their relationship. The second advantage is that the proposed procedure provides simultaneously the model and parameter estimates, thereby reducing greatly computational costs.

Although we have only considered selecting outcome/explanatory latent variables, the proposed method can be used in a similar way to simultaneously select outcome/explanatory latent variables and the structure of factor loading matrix.

Acknowledgments

The authors are grateful to the Editor, Christian Genest, FRSC, an Associate Editor, and two referees for their valuable suggestions and comments that greatly improved the manuscript. The research was supported by grants from the National Science Fund for Distinguished Young Scholars of China (11225103).

Appendix A: Proofs of theorems

Proof of Theorem 1. For any given $\varepsilon > 0$, we would like to show that there exists a large positive constant M_ε such that

$$\Pr\left\{\sup_{\|\boldsymbol{u}\|\geq M_\varepsilon} \ell(\boldsymbol{\theta}^o + \alpha_n \boldsymbol{u}) < \ell(\boldsymbol{\theta}^o)\right\} \geq 1-\varepsilon, \tag{A.1}$$

where $\alpha_n = n^{-1/2}(1+a_n)$. The probability inequality in (A.1) implies that with probability at least $1-\varepsilon$ there exists a local maximum in the ball $\{\boldsymbol{\theta}^o + \alpha_n\boldsymbol{u} : \|\boldsymbol{u}\| \leq M_\varepsilon\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o\| = O_p(\alpha_n)$.

Let $\boldsymbol{\theta} = \boldsymbol{\theta}^o + \alpha_n\boldsymbol{u}$ and $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_1^o + \alpha_n\boldsymbol{u}_I$, where $\boldsymbol{u}_I$ is a subvector of $\boldsymbol{u}$ with corresponding $\boldsymbol{\pi}_1^o$. By the definition of $\ell(\boldsymbol{\theta})$, we have

$$\begin{aligned}
\Delta_n(\boldsymbol{u}) &= \ell(\boldsymbol{\theta}^o + \alpha_n\boldsymbol{u}) - \ell(\boldsymbol{\theta}^o) \\
&= \{L_0(\boldsymbol{\theta}^o + \alpha_n\boldsymbol{u}) - L_0(\boldsymbol{\theta}^o)\} - \sum_{\ell=1}^{q}\sum_{j=1, j\neq\ell}^{q}\{\phi_{\lambda_n}(|\pi_{\ell j}|) - \phi_{\lambda_n}(|\pi_{\ell j}^o|)\} \\
&\leq \{L_0(\boldsymbol{\theta}^o + \alpha_n\boldsymbol{u}) - L_0(\boldsymbol{\theta}^o)\} - \sum_{\ell=1}^{q_1}\{\phi_{\lambda_n}(|\boldsymbol{\pi}_{1(\ell)}^o + \alpha_n\boldsymbol{u}_{I(\ell)}|) - \phi_{\lambda_n}(|\boldsymbol{\pi}_{1(\ell)}^o|)\},
\end{aligned}$$

where $\boldsymbol{\pi}_{1(\ell)}^o$ and $\boldsymbol{u}_{I(\ell)}$ are the ℓth component in $\boldsymbol{\pi}_1^o$ and $\boldsymbol{u}_I$, respectively. By the Taylor's expansion, we have

$$\begin{aligned}
\Delta_n(\boldsymbol{u}) \leq\ & \alpha_n L_0'(\boldsymbol{\theta}^o)^\top\boldsymbol{u} + \tfrac{1}{2}\alpha_n^2\boldsymbol{u}^\top L_0''(\boldsymbol{\theta}^o)\boldsymbol{u}\{1+o_p(1)\} \\
& - \sum_{\ell=1}^{q_1}\left[\alpha_n\phi_{\lambda_n}'(|\boldsymbol{\pi}_{1(\ell)}^o|)\boldsymbol{u}_{I(\ell)} + \tfrac{\alpha_n^2}{2}\phi_{\lambda_n}''(|\boldsymbol{\pi}_{1(\ell)}^o|)\boldsymbol{u}_{I(\ell)}^2\{1+o_p(1)\}\right],
\end{aligned} \tag{A.2}$$

where $L_0'(\boldsymbol{\theta}^o)$, $L_0''(\boldsymbol{\theta}^o)$, $\phi_{\lambda_n}'(|\boldsymbol{\pi}_{1(\ell)}^o|)$ and $\phi_{\lambda_n}''(|\boldsymbol{\pi}_{1(\ell)}^o|)$ denote the first- and second-order partial derivatives of $L_0(\boldsymbol{\theta})$ and $\phi_{\lambda_n}(|\boldsymbol{\pi}_{1(\ell)}|)$ evaluated at $\boldsymbol{\theta}^o$ and $\boldsymbol{\pi}^o$, respectively. For each component, $L_0'(\boldsymbol{\theta}^o)$ satisfies

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial L_0(\boldsymbol{\theta})}{\partial\boldsymbol{\mu}}\right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^o} &= \mathrm{E}\left\{\sum_{i=1}^{n}\mathbf{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right\}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^o} = \mathbf{0}, \\
\mathrm{E}\left(\frac{\partial L_0(\boldsymbol{\theta})}{\partial\theta_i}\right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^o} &= \mathrm{E}\left[\frac{n}{2}\mathrm{tr}\left\{\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}(\mathbf{V}^{-1}\mathbf{S}_n - \mathbf{I})\right\}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^o} = 0,
\end{aligned}$$

where tr denotes the trace for matrix,

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top,$$

and θ_i is a component of $\boldsymbol{\theta}$ except the components of $\boldsymbol{\mu}$. Hence, we have $L_0'(\boldsymbol{\theta}^o) = O_p(\sqrt{n})$ and $n^{-1}L_0''(\boldsymbol{\theta}^o) = -\mathbf{J}(\boldsymbol{\theta}^o) + o_p(1)$, where $\mathbf{J}(\boldsymbol{\theta}^o)$ is the Fisher information evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^o$. Similarly with (A.2), we have

$$\begin{aligned} L_0(\boldsymbol{\theta}^o + \alpha_n \boldsymbol{u}) - L_0(\boldsymbol{\theta}^o) &= \alpha_n L_0'(\boldsymbol{\theta}^o)^\top \boldsymbol{u} + \tfrac{1}{2}\alpha_n^2 \boldsymbol{u}^\top L_0''(\boldsymbol{\theta}^o)\boldsymbol{u}\{1 + o_p(1)\} \\ &= \sqrt{n}\,\alpha_n O_p(1)\boldsymbol{u} - \tfrac{1}{2} n\alpha_n^2 \boldsymbol{u}^\top \mathbf{J}(\boldsymbol{\theta}^o)\boldsymbol{u}\{1 + o_p(1)\}. \end{aligned} \tag{A.3}$$

Thus, the first term in (A.3) is of order $O_p(n^{1/2}\alpha_n) = O_p(n\alpha_n^2)$. By choosing a sufficiently large positive constant M_ε, the second term dominates the first term uniformly in $\|\boldsymbol{u}\| = M_\varepsilon$. Since $b_n = o_p(1)$, we have

$$\begin{aligned} &\left|\sum_{\ell=1}^{q_1}\left[\alpha_n \phi_{\lambda_n}'(|\boldsymbol{\pi}_{1(\ell)}^o|)\boldsymbol{u}_{I(\ell)} + \frac{\alpha_n^2}{2}\phi_{\lambda_n}''(|\boldsymbol{\pi}_{1(\ell)}^o|)\boldsymbol{u}_{I(\ell)}^2\{1 + o_p(1)\}\right]\right| \\ \leq\;& \alpha_n \sum_{\ell=1}^{q_1}\left|\phi_{\lambda_n}'(|\boldsymbol{\pi}_{1(\ell)}^o|)\boldsymbol{u}_{I(\ell)}\right| + \frac{\alpha_n^2}{2}\sum_{\ell=1}^{q_1}\left|\phi_{\lambda_n}''(|\boldsymbol{\pi}_{1(\ell)}^o|)\boldsymbol{u}_{I(\ell)}^2\{1 + o_p(1)\}\right| \\ \leq\;& \alpha_n \max\left|\phi_{\lambda_n}'(|\boldsymbol{\pi}_{1(\ell)}^o|)\right|\sqrt{q_1}\|\boldsymbol{u}_I\| + \frac{\alpha_n^2}{2}\max\left|\phi_{\lambda_n}''(|\boldsymbol{\pi}_{1(\ell)}^o|)\right|\|\boldsymbol{u}_I\|^2\{1 + o_p(1)\} \\ \leq\;& \sqrt{q_1}\alpha_n\sqrt{n}\,a_n\|\boldsymbol{u}\| + \frac{\alpha_n^2}{2} n b_n \|\boldsymbol{u}\|^2\{1 + o_p(1)\} \\ =\;& \sqrt{q_1}(1 + a_n)a_n\|\boldsymbol{u}\| + \frac{1}{2}(1 + a_n)^2\|\boldsymbol{u}\|^2\{1 + o_p(1)\}. \end{aligned}$$

This is also dominated by the second term of (A.3). Hence, by choosing a sufficiently large positive constant M_ε, (A.1) holds. This completes the proof. □

Proof of Theorem 2. (i) It suffices to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\theta}_1$ satisfying $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^o = O_p(n^{-1/2})$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \ldots, k$,

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} < 0 & \text{for } 0 < \theta_j < \varepsilon_n, \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} > 0 & \text{for } -\varepsilon_n < \theta_j < 0. \end{cases}$$

That is,

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \pi_{\ell j}} < 0 & \text{for } 0 < \pi_{\ell j} < \varepsilon_n, \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \pi_{\ell j}} > 0 & \text{for } -\varepsilon_n < \pi_{\ell j} < 0, \end{cases} \tag{A.4}$$

395 where $\pi_{\ell j} \in \boldsymbol{\pi}_2$. A Taylor's expansion about $\partial L_0(\boldsymbol{\theta})/\partial \pi_{\ell j}$ yields

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \pi_{\ell j}} &= \frac{\partial L_0(\boldsymbol{\theta})}{\partial \pi_{\ell j}} - \phi'_{\lambda_n}(|\pi_{\ell j}|)\mathrm{sgn}(\pi_{\ell j}) \\ &= \frac{\partial L_0(\boldsymbol{\theta}^o)}{\partial \pi_{\ell j}} + \sum_{m=1}^{k} \frac{\partial^2 L_0(\boldsymbol{\theta}^o)}{\partial \pi_{\ell j}\partial \theta_m}(\theta_m - \theta_m^o) \\ &\quad + \frac{1}{2}\sum_{m=1}^{k}\sum_{t=1}^{k} \frac{\partial^3 L_0(\boldsymbol{\theta}^*)}{\partial \pi_{\ell j}\partial \theta_m \partial \theta_t}(\theta_m - \theta_m^o)(\theta_t - \theta_t^o) - \phi'_{\lambda_n}(|\pi_{\ell j}|)\mathrm{sgn}(\pi_{\ell j}), \end{aligned}$$

where $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^o\| \le \|\boldsymbol{\theta} - \boldsymbol{\theta}^o\| = O_p(n^{-1/2})$. From the proof of Theorem 1, it is easy to see that

$$\frac{\partial L_0(\boldsymbol{\theta}^o)}{\partial \pi_{\ell j}} = O_p(\sqrt{n}) \qquad \text{and} \qquad \frac{\partial^2 L_0(\boldsymbol{\theta}^o)}{n\partial \pi_{\ell j}\partial \theta_m} = \mathrm{E}\left\{\frac{\partial^2 L_0(\boldsymbol{\theta}^o)}{\partial \pi_{\ell j}\partial \theta_m}\right\} + o_p(1).$$

Then, we have

$$\partial \ell(\boldsymbol{\theta})/\partial \pi_{\ell j} = n\lambda_n \left\{-n^{-1}\lambda_n^{-1}\phi'_{\lambda_n}(|\pi_{\ell j}|)\mathrm{sgn}(\pi_{\ell j}) + O_p(n^{-1/2}/\lambda_n)\right\}.$$

By Condition (C3) and the fact that $\sqrt{n}\,\lambda_n \to \infty$, the sign of the derivative is completely determined by that of $\pi_{\ell j}$. Hence (A.4) follows. Part (i) is thus proved.

(ii) From Theorem 1, there exists a local maximizer $\hat{\boldsymbol{\theta}}_1$ of $\ell(\boldsymbol{\theta}_1)$ such that $\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^o\| = O_p(n^{-1/2})$, and $\hat{\boldsymbol{\theta}}_1$ satisfies the likelihood equations

$$\left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\right|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1^\top,\mathbf{0})} = \left.\frac{\partial L_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\right|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1^\top,\mathbf{0})} - \boldsymbol{d}(\hat{\boldsymbol{\theta}}_1) = \mathbf{0},$$

where $\boldsymbol{d}(\boldsymbol{\theta}_1) = (d_1(\boldsymbol{\theta}_1), \ldots, d_s(\boldsymbol{\theta}_1))^\top$ and

$$d_j(\boldsymbol{\theta}_1) = \begin{cases} \phi'_{\lambda_n}(|\pi_{\ell t}|)\mathrm{sgn}(\pi_{\ell t})\ , & \text{if } \theta_{1j} = \pi_{\ell t} \in \boldsymbol{\pi}_1, \\ 0\ , & \text{otherwise.} \end{cases}$$

Using Taylor's expansion, we have

$$\begin{aligned}\frac{\partial L_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\bigg|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1^\top,\mathbf{0})} - \boldsymbol{d}(\hat{\boldsymbol{\theta}}_1) &= \frac{\partial L_0(\boldsymbol{\theta}_1^o)}{\partial\boldsymbol{\theta}_1} + \left\{\frac{\partial^2 L_0(\boldsymbol{\theta}_1^o)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^\top}+o_p(1)\right\}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o)\\ &\quad - \boldsymbol{d}(\boldsymbol{\theta}_1^o) - \{\mathbf{Z}(\boldsymbol{\theta}_1^o)+o_p(1)\}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o)\\ &= \mathbf{0}\,,\end{aligned}$$

where $\mathbf{Z}(\boldsymbol{\theta}_1) = \mathrm{diag}\big\{Z_1(\boldsymbol{\theta}_1),\ldots,Z_s(\boldsymbol{\theta}_1)\big\}$ and

$$Z_j(\boldsymbol{\theta}_1) = \frac{\partial d_j(\boldsymbol{\theta}_1)}{\partial\theta_{1j}} = \begin{cases}\phi''_{\lambda_n}(|\pi_{\ell t}|)\ , & \text{if } \theta_{1j}=\pi_{\ell t}\in\boldsymbol{\pi}_1,\\ 0\ , & \text{otherwise.}\end{cases}$$

Hence, we get

$$\begin{aligned}&\sqrt{n}\{\mathbf{J}(\boldsymbol{\theta}_1^o)+n^{-1}\mathbf{Z}(\boldsymbol{\theta}_1^o)\}\Big[\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o+n^{-1}\{\mathbf{J}(\boldsymbol{\theta}_1^o)+n^{-1}\mathbf{Z}(\boldsymbol{\theta}_1^o)\}^{-1}\boldsymbol{d}(\boldsymbol{\theta}_1^o)\Big]\\ &\qquad\qquad\qquad\qquad = n^{-1/2}\partial L_0(\boldsymbol{\theta}_1^o)/\partial\boldsymbol{\theta}_1 + o_p(n^{-1/2}).\end{aligned}$$

Since $\mathrm{E}\{\partial L_0(\boldsymbol{\theta}_1)/\partial\boldsymbol{\theta}_1\}=\mathbf{0}$ as in the proof of Theorem 1, it follows from the Multivariate Central Theorem that

$$\frac{1}{\sqrt{n}}\frac{\partial L_0(\boldsymbol{\theta}_1^o)}{\partial\boldsymbol{\theta}_1}\rightsquigarrow\mathcal{N}_s[(\mathbf{0},\mathbf{J}(\boldsymbol{\theta}_1^o)].$$

Therefore,

$$\sqrt{n}\left\{\mathbf{J}(\boldsymbol{\theta}_1^o)+\frac{1}{n}\mathbf{Z}(\boldsymbol{\theta}_1^o)\right\}\left[\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^o+\frac{1}{n}\left\{\mathbf{J}(\boldsymbol{\theta}_1^o)+\frac{1}{n}\mathbf{Z}(\boldsymbol{\theta}_1^o)\right\}^{-1}\boldsymbol{d}(\boldsymbol{\theta}_1^o)\right]\rightsquigarrow\mathcal{N}_s[\mathbf{0},\mathbf{J}(\boldsymbol{\theta}_1^o)].$$

This completes the proof of Part (ii). □

Appendix B: Partial derivatives

The second partial derivatives of the observed-data log-likelihood $L_0(\boldsymbol{\theta})$ with respect to parameters can be obtained by using some basic matrix derivatives. The explicit expressions of the second partial derivatives are listed as follows:

$$\begin{aligned}\frac{\partial^2 L_0(\boldsymbol{\theta})}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^\top} &= -n\mathbf{V}^{-1},\\ \frac{\partial^2 L_0(\boldsymbol{\theta})}{\partial\boldsymbol{\mu}\partial\theta_i} &= -\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{V}^{-1}\sum_{r=1}^{n}(\boldsymbol{y}_r-\boldsymbol{\mu}),\\ \frac{\partial^2 L_0(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} &= \frac{n}{2}\mathrm{tr}\left\{\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}(\mathbf{I}-2\mathbf{V}^{-1}\mathbf{S}_n)+\mathbf{V}^{-1}\frac{\partial^2\mathbf{V}}{\partial\theta_i\partial\theta_j}(\mathbf{V}^{-1}\mathbf{S}_n-\mathbf{I})\right\},\end{aligned}$$

where

$$\mathbf{S}_n = \frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{y}_r - \boldsymbol{\mu})(\boldsymbol{y}_r - \boldsymbol{\mu})^\top,$$

θ_i and θ_j are the components of $\boldsymbol{\theta}$ except the components of $\boldsymbol{\mu}$, and for different θ_i and θ_j, $\partial\mathbf{V}/\partial\theta_i$ and $\partial^2\mathbf{V}/\partial\theta_i\partial\theta_j$ have different expressions as follows:

$$\begin{aligned}
\frac{\partial\mathbf{V}}{\partial\tau_i^2} &= \mathbf{E}_{ii}, \quad \frac{\partial\mathbf{V}}{\partial\gamma_\ell^2} = \boldsymbol{\Lambda}\frac{\partial\boldsymbol{\Sigma}}{\partial\gamma_\ell^2}\boldsymbol{\Lambda}^\top, \quad \frac{\partial\boldsymbol{\Sigma}}{\partial\gamma_\ell^2} = (\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell\ell}(\mathbf{I}-\boldsymbol{\Pi}^\top)^{-1}, \\
\frac{\partial\mathbf{V}}{\partial\boldsymbol{\Lambda}_{i\ell}} &= (\boldsymbol{\Lambda}\boldsymbol{\Sigma}\mathbf{E}_{i\ell}^\top)^\top + \boldsymbol{\Lambda}\boldsymbol{\Sigma}\mathbf{E}_{i\ell}^\top, \quad \frac{\partial\mathbf{V}}{\partial\pi_{\ell k}} = \boldsymbol{\Lambda}\frac{\partial\boldsymbol{\Sigma}}{\partial\pi_{\ell k}}\boldsymbol{\Lambda}^\top, \\
\frac{\partial\boldsymbol{\Sigma}}{\partial\pi_{\ell k}} &= \{(\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell k}\boldsymbol{\Sigma}\}^\top + (\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell k}\boldsymbol{\Sigma}, \\
\frac{\partial^2\mathbf{V}}{\partial\tau_i^2\partial\tau_j^2} &= \frac{\partial^2\mathbf{V}}{\partial\gamma_\ell^2\partial\gamma_k^2} = \frac{\partial^2\mathbf{V}}{\partial\tau_i^2\partial\gamma_\ell^2} = \frac{\partial^2\mathbf{V}}{\partial\tau_i^2\partial\boldsymbol{\Lambda}_{j\ell}} = \frac{\partial^2\mathbf{V}}{\partial\tau_i^2\partial\pi_{\ell k}} = 0, \\
\frac{\partial^2\mathbf{V}}{\partial\boldsymbol{\Lambda}_{i\ell}\partial\boldsymbol{\Lambda}_{jk}} &= (\mathbf{E}_{i\ell}\boldsymbol{\Sigma}\mathbf{E}_{jk}^\top)^\top + \mathbf{E}_{i\ell}\boldsymbol{\Sigma}\mathbf{E}_{jk}^\top, \quad \frac{\partial^2\mathbf{V}}{\partial\boldsymbol{\Lambda}_{i\ell}\partial\gamma_k^2} = (\boldsymbol{\Lambda}\frac{\partial\boldsymbol{\Sigma}}{\partial\gamma_k^2}\mathbf{E}_{i\ell}^\top)^\top + \boldsymbol{\Lambda}\frac{\partial\boldsymbol{\Sigma}}{\partial\gamma_k^2}\mathbf{E}_{i\ell}^\top, \\
\frac{\partial^2\mathbf{V}}{\partial\boldsymbol{\Lambda}_{i\ell}\partial\pi_{km}} &= (\boldsymbol{\Lambda}\frac{\partial\boldsymbol{\Sigma}}{\partial\pi_{km}}\mathbf{E}_{i\ell}^\top)^\top + \boldsymbol{\Lambda}\frac{\partial\boldsymbol{\Sigma}}{\partial\pi_{km}}\mathbf{E}_{i\ell}^\top, \\
\frac{\partial^2\mathbf{V}}{\partial\gamma_\ell^2\partial\pi_{km}} &= \boldsymbol{\Lambda}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\gamma_\ell^2\partial\pi_{km}}\boldsymbol{\Lambda}^\top, \quad \frac{\partial^2\mathbf{V}}{\partial\pi_{\ell t}\partial\pi_{km}} = \boldsymbol{\Lambda}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\pi_{\ell t}\partial\pi_{km}}\boldsymbol{\Lambda}^\top, \\
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\gamma_\ell^2\partial\pi_{km}} &= \{(\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell\ell}(\mathbf{I}-\boldsymbol{\Pi}^\top)^{-1}\mathbf{E}_{km}^\top(\mathbf{I}-\boldsymbol{\Pi}^\top)^{-1}\}^\top \\
&\quad +(\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell\ell}(\mathbf{I}-\boldsymbol{\Pi}^\top)^{-1}\mathbf{E}_{km}^\top(\mathbf{I}-\boldsymbol{\Pi}^\top)^{-1}, \\
\frac{\partial^2\boldsymbol{\Sigma}}{\partial\pi_{\ell t}\partial\pi_{km}} &= \Big\{(\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{km}(\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell t}\boldsymbol{\Sigma} + (\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell t}\frac{\partial\boldsymbol{\Sigma}}{\partial\pi_{km}}\Big\}^\top \\
&\quad +(\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{km}(\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell t}\boldsymbol{\Sigma} + (\mathbf{I}-\boldsymbol{\Pi})^{-1}\mathbf{E}_{\ell t}\frac{\partial\boldsymbol{\Sigma}}{\partial\pi_{km}}.
\end{aligned}$$

for $i, j = 1, \ldots, p$, $\ell, k, m, t = 1, \ldots, q$, in which $\mathbf{E}_{ij}$ denotes the matrix with 1 in the (i, j)th cell and zeros elsewhere.

References

[1] H.D. Bondell, A. Krishna, S.K. Ghosh, Joint variable selection for fixed and random effects in linear mixed-effects models, Biometrics 66 (2010) 1069–1077.

[2] J.B. Daniel, A semiparametric approach to modeling nonlinear relations among latent variables, Structural Equation Modeling: A Multidisciplinary Journal 12 (2005) 513–535.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc. Ser. B 39 (1977) 1–38.

[4] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.

[5] J. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, Ann. Statist. 32 (2004) 928–961.

[6] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–135.

[7] D.A. Freedman, On the so-called “Huber sandwich estimator” and “robust standard errors”, Amer. Statist. 60 (2006) 299–302.

[8] R.I. Garcia, J.G. Ibrahim, H. Zhu, Variable selection for regression models with missing data, Statistica Sinica 20 (2010) 149–165.

[9] D.R. Hunter, R. Li, Variable selection using MM algorithms, Ann. Statist. 33 (2005) 1617–1642.

[10] K.G. Jöreskog, A general method for estimating a linear structural equation system, In: A.S. Goldberger and O.D. Duncan (Ed.), Structural Equation Models in the Social Sciences (pp. 85–112), Seminar Press, New York, 1973.

[11] G. Kauermann, R.J. Carroll, A note on the efficiency of sandwich covariance matrix estimation, J. Amer. Statist. Assoc. 96 (2001) 1387–1396.

[12] A.B. Kenneth, Structural equation models that are nonlinear in latent variables: A least-squares estimator, Sociol. Meth. 25 (1995) 223–251.

[13] S.Y. Lee, X.Y. Song, Model comparison of nonlinear structural equation models with fixed covariates, Psychometrika 68 (2003) 27–47.

[14] S.Y. Lee, N.S. Tang, Bayesian analysis of structural equation models with mixed exponential family and ordered categorical data, British J. Math. Statist. Psych. 59 (2006) 151–172.

[15] S.Y. Lee, H. Zhu, Maximum likelihood estimation of nonlinear structural equation models, Psychometrika 67 (2002) 189–210.

[16] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, Journal of Machine Learning Research 11 (2010) 2287–2322.

[17] X. Meng, D.B. Rubin, Maximum likelihood estimation via the ECM algorithm: A general framework, Biometrika 80 (1993) 267–278.

[18] A.E. Raftery, Bayesian model selection in structural equation models, In: K.A. Bollen and J.S. Long (Ed.), Testing Structural Equation Models (pp. 163–180), Beverly Hills, CA: Sage, 1993.

[19] X.Y. Song, S.Y. Lee, Analysis of structural equation model with ignorable missing continuous and polytomous data, Psychometrika 67 (2002) 261–288.

[20] X.Y. Song, S.Y. Lee, Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences, Wiley, Chichester, 2012.

[21] X.Y. Song, Z. Lu, J. Cai, E.H. Ip, A Bayesian modeling approach for generalized semiparametric structural equatuion models, Psychometrika 78 (2013) 624–647.

[22] X.Y. Song, Z. Lu, Y.I. Hser, S.Y. Lee, A Bayesian approach for analyzing longitudinal structural equation models, Structural Equation Modeling: A Multidisciplinary Journal 18 (2011) 183–194.

[23] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Statist. Soc. Ser. B 58 (1996) 267–288.

[24] H. Wang, R. Li, C.L. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika 94 (2007) 553–568.

[25] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. Roy. Statist. Soc. Ser. B 68 (2006) 49–67.

[26] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.

[27] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Statist. Soc. Ser. B 67 (2005) 301–320.

[28] H. Zou, R. Li, One-step sparse estimates in nonconcave penalized likelihood models, Ann. Statist. 36 (2008) 1509–1533.

Table 1. Simulation results of linear SEM for unpenalized (UnpenlM), Lasso and SCAD and oracle approaches with $n = 300$ and 500 in experiment 1

n	Method	MRRSE	Correct	Incorrect	TM	FM
300	UnpenlM	1.000	0	0	0	0
	Lasso	1.032	2.69	0.14	0.76	0.13
	SCAD	0.979	2.83	0.16	0.87	0.11
	Oracle	0.924	3	0	–	–
500	UnpenlM	1.000	0.01	0	0	0
	Lasso	1.043	2.72	0.07	0.76	0.07
	SCAD	1.008	2.87	0.02	0.94	0.02
	Oracle	0.926	3	0	–	–

Note: $\text{TM} = \frac{1}{M}\sum_{m=1}^{M} I(\widehat{\mathcal{A}}_m = \mathcal{A})$, $\text{FM} = \frac{1}{M}\sum_{m=1}^{M} I(\widehat{\mathcal{A}}_m \cap \mathcal{A} \neq \mathcal{A})$.

Table 2. Frequencies of the identified nonzero components for unpenalized, Lasso and SCAD approaches with $n = 300$ and 500 in experiment 1

n	Method	π_{12}	π_{13}^*	π_{21}^*	π_{23}^*	π_{31}	π_{32}
300	UnpenlM	1.00	1.00	1.00	1.00	1.00	1.00
	Lasso	0.19	0.88	0.99	0.99	0.07	0.05
	SCAD	0.09	0.91	0.93	1.00	0.04	0.04
500	UnpenlM	0.99	1.00	1.00	1.00	1.00	1.00
	Lasso	0.22	0.95	0.99	0.99	0.03	0.03
	SCAD	0.04	0.98	1.00	1.00	0.04	0.05

Note: * The true values of π_{13}, π_{21} and π_{23} are set to be nonzero.

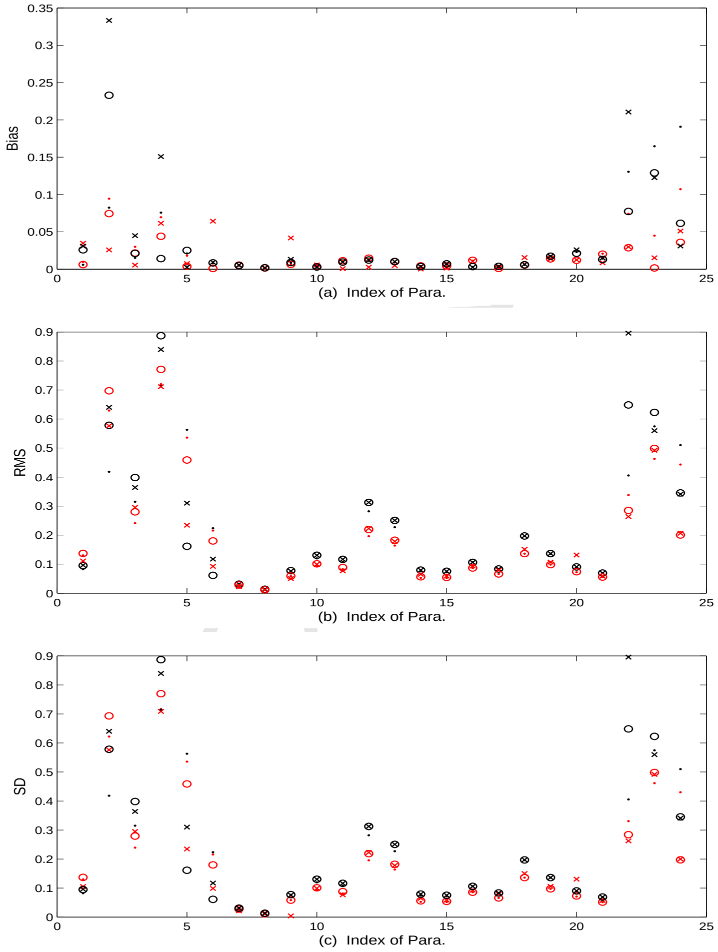

Figure 1. Plots of Bias, RMS and SD values for unpenalized ('·'), Lasso penalized ('x') and SCAD penalized ('o') MLEs with $n = 300$ (black symbols) and 500 (red symbols) in experiment 1. The index of 24 parameters is π_{12}, π_{13}, π_{21}, π_{23}, π_{31}, π_{32}, λ_{21}, λ_{42}, λ_{63}, μ_1, …, μ_6, τ_1^2, …, τ_6^2, γ_1^2, γ_2^2 and γ_3^2.

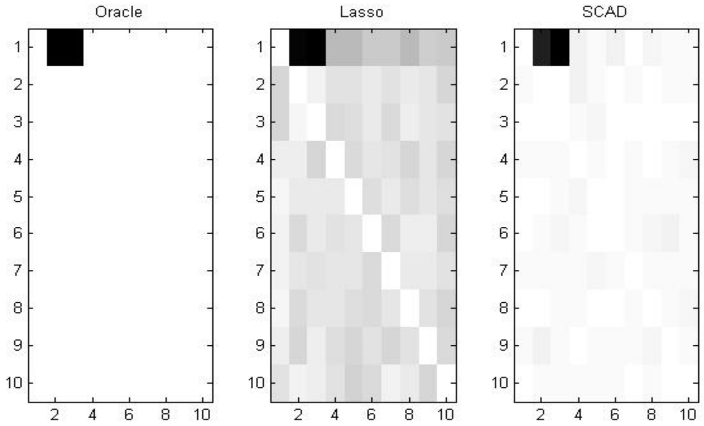

Figure 2. Heatmaps of zeros identified in the coefficient matrix out of 100 replications in experiment 2 (White color is 100/100 zeros identified, black is 0/100)

Table 3. Simulation results of linear SEM for for unpenalized (UnpenlM), Lasso and SCAD and oracle approaches with $n = 500$ in experiment 2

Method	MRRSE	Correct	Incorrect	TM	FM
UnpenlM	1.000	0.01	0	0	0
Lasso	0.152	80.57	0.62	0.27	0.32
SCAD	0.096	86.19	0.15	0.70	0.14
Oracle	0.066	88	0	–	–

485

Note: TM $= \frac{1}{M}\sum_{m=1}^{M} I(\widehat{\mathcal{A}}_m = \mathcal{A})$, FM $= \frac{1}{M}\sum_{m=1}^{M} I(\widehat{\mathcal{A}}_m \cap \mathcal{A} \neq \mathcal{A})$.

Table 4. Frequencies of the identified nonzero components for unpenalized, Lasso and SCAD approaches with $n = 500$ in experiment 2

Method		$\pi_{\cdot,1}$	$\pi_{\cdot,2}$	$\pi_{\cdot,3}$	$\pi_{\cdot,4}$	$\pi_{\cdot,5}$	$\pi_{\cdot,6}$	$\pi_{\cdot,7}$	$\pi_{\cdot,8}$	$\pi_{\cdot,9}$	$\pi_{\cdot,10}$
Lasso	$\pi_{1,\cdot}$	–	0.68^*	0.70^*	0.19	0.19	0.15	0.15	0.19	0.14	0.15
	$\pi_{2,\cdot}$	0.12	–	0.04	0.08	0.08	0.06	0.08	0.06	0.07	0.11
	$\pi_{3,\cdot}$	0.11	0.03	–	0.10	0.09	0.06	0.10	0.05	0.07	0.08
	$\pi_{4,\cdot}$	0.05	0.05	0.11	–	0.10	0.07	0.08	0.11	0.07	0.12
	$\pi_{5,\cdot}$	0.03	0.06	0.06	0.06	–	0.09	0.06	0.09	0.07	0.09
	$\pi_{6,\cdot}$	0.04	0.10	0.06	0.08	0.07	–	0.10	0.05	0.05	0.12
	$\pi_{7,\cdot}$	0.04	0.07	0.08	0.07	0.07	0.10	–	0.06	0.06	0.08
	$\pi_{8,\cdot}$	0.03	0.10	0.07	0.07	0.09	0.10	0.07	–	0.08	0.12
	$\pi_{9,\cdot}$	0.05	0.11	0.05	0.09	0.11	0.08	0.12	0.08	–	0.10
	$\pi_{10,\cdot}$	0.08	0.04	0.05	0.08	0.13	0.10	0.04	0.06	0.11	–
SCAD	$\pi_{1,\cdot}$	–	0.86^*	0.99^*	0.05	0.02	0.06	0.00	0.04	0.02	0.02
	$\pi_{2,\cdot}$	0.02	–	0.01	0.05	0.02	0.01	0.02	0.01	0.03	0.02
	$\pi_{3,\cdot}$	0.00	0.00	–	0.02	0.04	0.01	0.01	0.00	0.00	0.00
	$\pi_{4,\cdot}$	0.02	0.02	0.02	–	0.02	0.01	0.02	0.00	0.02	0.04
	$\pi_{5,\cdot}$	0.00	0.01	0.03	0.04	–	0.01	0.02	0.02	0.02	0.02
	$\pi_{6,\cdot}$	0.01	0.02	0.04	0.02	0.01	–	0.02	0.04	0.05	0.02
	$\pi_{7,\cdot}$	0.02	0.02	0.02	0.02	0.04	0.03	–	0.02	0.02	0.02
	$\pi_{8,\cdot}$	0.00	0.01	0.02	0.03	0.01	0.01	0.02	–	0.02	0.04
	$\pi_{9,\cdot}$	0.02	0.05	0.02	0.01	0.02	0.02	0.03	0.04	–	0.02
	$\pi_{10,\cdot}$	0.01	0.03	0.03	0.02	0.02	0.02	0.01	0.02	0.00	–
UnpenlM	$\pi_{1,\cdot}$	–	1.00^*	1.00^*	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\pi_{2,\cdot}$	1.00	–	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\pi_{3,\cdot}$	1.00	1.00	–	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\pi_{4,\cdot}$	1.00	1.00	1.00	–	1.00	1.00	1.00	1.00	1.00	1.00
	$\pi_{5,\cdot}$	0.99	1.00	1.00	1.00	–	1.00	1.00	1.00	1.00	1.00
	$\pi_{6,\cdot}$	1.00	1.00	1.00	1.00	1.00	–	1.00	1.00	1.00	1.00
	$\pi_{7,\cdot}$	1.00	1.00	1.00	1.00	1.00	1.00	–	1.00	1.00	1.00
	$\pi_{8,\cdot}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	1.00	1.00
	$\pi_{9,\cdot}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	1.00
	$\pi_{10,\cdot}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–

Note: * The true values of π_{12} and π_{13} are set to be nonzero.

Table 5. The unpenalized (UnpenlM), Lasso and SCAD penalized MLEs (ESTs) and their standard errors (SEs) in the ICPSR data

	UnpenlM		Lasso		SCAD	
Para.	EST	SE	EST	SE	EST	SE
π_{12}	0.1581	0.0532	0.2785	0.0904	0.2797	0.0904
π_{13}	0.2964	0.2103	0.4331	0.1158	0.4360	0.1180
π_{21}	0.1700	0.0801	0	0	0	0
π_{23}	-0.0894	0.1227	0	0	0	0
π_{31}	0.3236	0.3236	0	0	0	0
π_{32}	-0.0908	0.0674	0	0	0	0
λ_{21}	0.8707	0.1515	0.8779	0.1534	0.8768	0.1515
λ_{42}	1.9890	0.4797	2.0312	0.4490	2.0343	0.4477
λ_{63}	0.7393	0.1763	0.7308	0.1725	0.7310	0.1722
μ_1	8.4591	0.1180	8.4592	0.1180	8.4592	0.1180
μ_2	7.8571	0.1212	7.8571	0.1212	7.8571	0.1212
μ_3	2.3724	0.1120	2.3724	0.1120	2.3724	0.1120
μ_4	5.5560	0.2254	5.5561	0.2254	5.5561	0.2254
μ_5	7.5663	0.1592	7.5663	0.1592	7.5663	0.1592
μ_6	7.3877	0.1740	7.3878	0.1740	7.3878	0.1740
τ_1^2	0.6719	0.3680	0.6874	0.3676	0.6860	0.3642
τ_2^2	1.3181	0.2402	1.3055	0.2404	1.3079	0.2385
τ_3^2	0.6038	0.4513	0.6413	0.4107	0.6442	0.4084
τ_4^2	2.6137	1.6679	2.4633	1.5412	2.4516	1.5357
τ_5^2	1.5688	0.8095	1.5228	0.7895	1.5293	0.7915
τ_6^2	4.0728	0.6280	4.0914	0.6256	4.0925	0.6243
γ_1^2	1.3697	0.4570	1.2714	0.4436	1.2660	0.4407
γ_2^2	1.7143	0.4810	1.8174	0.4445	1.8147	0.4423
γ_3^2	2.6295	1.2459	3.4455	0.9408	3.4409	0.9475