PSYCHOMETRIKA—VOL. 82, NO. 2, 329–354
JUNE 2017
DOI: 10.1007/s11336-017-9566-9

A PENALIZED LIKELIHOOD METHOD FOR STRUCTURAL EQUATION MODELING

PO-HSIEN HUANG

NATIONAL TAIWAN UNIVERSITY
NATIONAL CHENG KUNG UNIVERSITY

HUNG CHEN AND LI-JEN WENG

NATIONAL TAIWAN UNIVERSITY

A penalized likelihood (PL) method for structural equation modeling (SEM) was proposed as a methodology for exploring the underlying relations among both observed and latent variables. Compared to the usual likelihood method, PL includes a penalty term to control the complexity of the hypothesized model. When the penalty level is appropriately chosen, the PL can yield an SEM model that balances the model goodness-of-fit and model complexity. In addition, the PL results in a sparse estimate that enhances the interpretability of the final model. The proposed method is especially useful when limited substantive knowledge is available for model specifications. The PL method can be also understood as a methodology that links the traditional SEM to the exploratory SEM (Asparouhov & Muthén in Struct Equ Model Multidiscipl J 16:397–438, 2009). An expectation-conditional maximization algorithm was developed to maximize the PL criterion. The asymptotic properties of the proposed PL were also derived. The performance of PL was evaluated through a numerical experiment, and two real data illustrations were presented to demonstrate its utility in psychological research.

Key words: structural equation modeling, penalized likelihood, ECM algorithm, oracle property, model selection, factor analysis model.

1. Introduction

Structural equation modeling (SEM) is a statistical method that is aimed toward explaining the moments among observable variables. The application of SEM involves a confirmatory testing of models proposed by researchers based on available theories. However, because of the complexity of human behavior, nearly all SEM models are subject to misspecification (Browne & Cudeck, 1993; MacCallum, 2003). Previous studies have shown that a misfit SEM model may lead to biased parameter estimates (Kaplan, 1988; Yuan, Marshall, & Bentler, 2003), incorrect standard errors (Arminger & Schoenberg, 1989; Yuan & Hayashi, 2006), and hence mistaken conclusions about psychological phenomena. Yet, in practice, an exploratory process seems to be an inevitable task for deriving a relatively plausible SEM model, especially when the development of the substantive theory is still in its infancy.

Exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) is now a major exploratory methodology for SEM. The ESEM replaces the confirmatory factor analysis

The research was supported in part by Grant NSC 99-2410-H-002-083-MY3 from the National Science Council in Taiwan.

Electronic supplementary material The online version of this article (doi:10.1007/s11336-017-9566-9) contains supplementary material, which is available to authorized users.

Correspondence should be made to Li-Jen Weng, Department of Psychology, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan. Email: ljweng@ntu.edu.tw

(CFA; Jöreskog, 1969) measurement model in SEM with the exploratory factor analysis (EFA) model. Hence, the misfit caused by omitting nonzero loadings can be avoided. Several novel applications of the ESEM have been demonstrated (see Marsh, Morin, Parker, & Kaur, 2014 for a review). However, the ESEM cannot be utilized without drawbacks. Because the ESEM must freely estimate the entire loading matrix, prior knowledge of null relations among measurements and factors cannot be imposed, which attenuates the theory-driven nature of SEM methodology. In addition, compared to the usual SEM that assumes a sparse loading matrix (i.e., a loading matrix with many zero elements), the ESEM results in a dense loading matrix that generally includes unnecessary parameters. Previous studies have shown that SEM models with unnecessary model complexity yield less efficient estimators (Bentler & Mooijaart, 1989) and exhibit relatively low generalizability (Browne & Cudeck, 1989; Cudeck & Browne, 1983; Preacher, 2006). Also, a dense loading matrix is less interpretable than a sparse one (Trendafilov & Adachi, 2015).

Considering the potential limitations of ESEM, the current study proposes a penalized likelihood (PL) approach for SEM as a new methodology. Over the past 15 years, PL has become a popular method for statistical learning problems (see Bühlmann & van de Geer, 2011; Hastie, Tibshirani, & Friedman, 2009; Hastie, Tibshirani, & Wainwright, 2015 for reviews). Compared to the usual likelihood, the PL estimation criterion includes a penalty term to control the complexity of the hypothesized model. As the penalty level is chosen appropriately, PL can lead to a final model that balances the tradeoff between the model goodness-of-fit and model complexity (e.g., Hoerl & Kennard, 1970). Hence, the final model is relatively generalizable to other samples according to the rationale of a bias-variance tradeoff. By implementing sparsity-inducing penalties, PL can yield sparse estimates (i.e., estimates with elements that are exactly zero), which not only reduces the model complexity but also enhances the interpretability of estimation results. Furthermore, under the family of generalized linear models (GLM), theoretical results indicate that PL has the capacity to consistently identify all relevant and irrelevant covariates (Fan & Li, 2001). The resulting estimator possesses the so-called oracle property, i.e., it performs as well as if the researcher has known the true sparsity pattern of the parameters in advance. Finally, the PL estimation is proven to be an effective method for handling the problem of $P > N$, where P is the number of variables, and N is the sample size (e.g., Fan & Lv, 2011; Fan & Peng, 2004; Kwon & Kim, 2012).

Under the proposed PL, an SEM model is formulated with a confirmatory part and an exploratory part. The *confirmatory part* includes the functional form of the specified model and the theory-derived free and fixed parameters. The *exploratory part*, wherein a set of penalized parameters are specified to represent the ambiguous relations, is data-driven yet with model complexity controlled by the penalty term. Hence, the proposed PL is not a pure exploratory method. We call it a *semi-confirmatory approach* because the PL for SEM can embrace both existing theories and ambiguous relations that await further exploration. The PL then yields relatively efficient and interpretable sparse parameter estimates (e.g., loadings and path coefficients). A final SEM model with possibly high generalizability can be achieved through the choice of penalty level. As we shall see in Sect. 5, under suitable conditions the PL for SEM could result in an oracle estimator, i.e., an estimator that possesses the oracle property.

Several works have applied the idea of penalization to SEM or related latent variable modeling. Jung (2012) applied the ℓ_2 penalty to SEM to improve the small sample performance of two-stage least squares estimation. Ning and Georgiou (2011) and Choi, Zou, and Oehlert (2011) implemented the ℓ_1 penalty in an orthogonal EFA to obtain a sparse loading matrix. Subsequently, Hirose and Yamamoto (2014, 2015) considered more general concave penalties in an EFA, also for the purpose of obtaining a sparse loadings matrix. Under the generalized linear mixed model (GLMM) framework, several works have proposed PL methods for selecting nonzero fixed effects (Fan & Li, 2012, Groll & Tutz, 2014; Schelldorfer, Bühlmann, & van de Geer, 2011) or both fixed and random effects (Ibrahim, Zhu, Garcia, & Guo, 2011). Although the GLMM is quite flexible, and SEM can be formulated within it (e.g., Ibrahim et al., 2011), these PL methods cannot be used to explore the directional relations among latent variables. PL has also been applied to many other psychometric models, including principal component analysis (Zou, Hastie, & Tibshirani, 2006; Trendafilov & Adachi, 2015), the Rasch model (Tutz & Schauberger, 2015), and the cognitive diagnostic model (Chen, Liu, Xu, & Ying, 2015). However, none of these studies developed a general PL method for SEM. Our work concerns a PL method with several state-of-the-art penalties for a wide class of SEM models. The proposed PL can be used to simultaneously identify the sparsity patterns of many parameter matrices under an SEM framework, including the factor loading matrix, path coefficient matrix, and covariance matrices of latent variables and measurement errors.

This article is organized as follows: Sect. 2 presents the proposed PL for SEM. In Sect. 3, an algorithm for optimizing the PL criterion is described. Section 4 discusses several practical issues when implementing the PL. In Sect. 5, the asymptotic properties of the PL are derived. Section 6 presents two real data illustrations. A numerical experiment is presented in Sect. 7. Finally, merits, limitations, and further directions concerning this study are discussed.

2. A PL Method for SEM

Let Y denote a P-dimensional observable random vector with element $Y_p (p = 1, 2, \dots, P)$. The measurement part of SEM describes the relationships between the manifest variables and the latent variables

$$Y = \nu + \Lambda\eta + \epsilon, \tag{1}$$

where ν is a P-dimensional intercept vector with element ν_p; η is an M-dimensional latent variable vector with element $\eta_m (m = 1, 2, \dots, M)$; Λ is a $P \times M$ factor loading matrix with element λ_{pm}, and ϵ is a P-dimensional measurement error vector with element ϵ_p. The structural part of SEM describes the hypothesized relationships among latent variables

$$\eta = \alpha + \mathrm{B}\eta + \zeta, \tag{2}$$

where α is an M-dimensional intercept vector with element α_m; B is an $M \times M$ path coefficient matrix with element $\beta_{mk} (\beta_{mm} = 0)$, and ζ is an M-dimensional residual vector with element ζ_m. It is assumed that

1. ϵ has a zero mean and positive definite covariance matrix Ψ.
2. ζ has a zero mean and positive definite covariance matrix Φ.
3. ϵ and ζ are uncorrelated.

Under these assumptions, the model-implied mean and covariance of Y are

$$\mu(\theta) = \nu + \Lambda(I_M - \mathrm{B})^{-1}\alpha, \tag{3}$$

$$\Sigma(\theta) = \Lambda(I_M - \mathrm{B})^{-1}\Phi(I_M - \mathrm{B})^{-1^{\mathrm{T}}}\Lambda^{\mathrm{T}} + \Psi, \tag{4}$$

respectively, where $\theta \in \Theta$ is a Q-dimensional parameter vector with element θ_q, $\Theta \subset \Re^Q$ being a parameter space, and I_M is the $M \times M$ identity matrix. The parameter vector θ only contains the unknown parameters in ν, Λ, Ψ, α, B, and Φ.

Consider a random sample $\mathcal{Y}_N = \{Y_n\}_{n=1}^N$, the maximum likelihood (ML) estimation finds an estimate $\hat{\theta}$ that maximizes the log-likelihood function

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \log \varphi_\theta\,(Y_n), \tag{5}$$

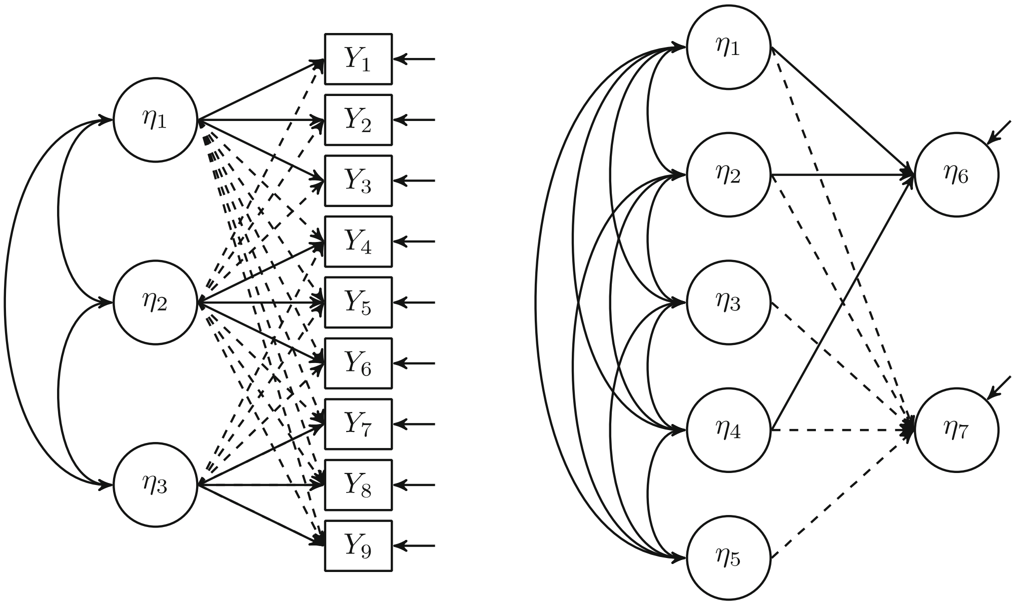

FIGURE 1.
Path diagrams for two possible model specifications under the proposed penalized likelihood method. *Left* A specification for a factor analysis model with three latent factors and nine indicators. *Right* A specification for a full SEM model with five exogenous latent variables and two endogenous latent variables. *Arrows with solid* and *broken lines* represent freely estimated and penalized parameters, respectively.

where $\varphi_\theta(y)$ is the normal density, with mean and covariance as parameterized by θ in Eqs. (3) and (4), i.e., $\log \varphi_\theta(y) = -\frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}(y-\mu(\theta))^{\mathrm{T}}\Sigma(\theta)^{-1}(y-\mu(\theta))$. The log-likelihood function is established under the working assumption that the response variable Y has a multivariate normal distribution with the specified moment structure. In the current study, we consider the PL criterion

$$\mathcal{U}(\theta, \gamma) = \mathcal{L}(\theta) - \mathcal{R}(\theta, \gamma), \tag{6}$$

where $\mathcal{R}(\theta, \gamma) = \sum_{q=1}^{Q} c_q \rho\left(\left|\theta_q\right|, \gamma\right)$ is a penalty term with a penalty function $\rho(t, \gamma)$, a set of penalization indicators $\left\{c_q\right\}_{q=1}^{Q}$, and a regularization parameter γ. A parameter with $c_q = 0$ is freely estimated, and $c_q = 1$ allows θ_q to be explored. The confirmatory part of the hypothesized model includes the functional forms of moment structures, freely estimated parameters, and fixed parameters. The exploratory part is formed by the parameters being penalized. By implementing sparsity-inducing penalties, non-null relationships in the exploratory part can be selected. Because the variance of an exogenous variable should be larger than zero, the proposed method requires c_q to be zero if θ_q is a diagonal element of Ψ or Φ. Each component in $\mathcal{U}(\theta, \gamma)$ plays a different role: $\mathcal{L}(\theta)$ measures the goodness-of-fit; $\mathcal{R}(\theta, \gamma)$ measures the complexity, and γ controls the tradeoff between the previous two components. A PL estimator $\hat{\theta} = \hat{\theta}(\gamma)$ is defined as a local maximizer of $\mathcal{U}(\theta, \gamma)$. Note that different values of γ yield different estimates even under the same data. An optimal value of γ can be determined based on model selection criteria (see Sect. 4).

Figure 1 illustrates two possible model specifications under the PL framework. Example 1 is an oblique three factor model with nine observed variables (left side of Fig. 1). Each variable is assumed to be mainly a measure of some latent factor (see the arrows with solid lines in Fig. 1). However, the PL specification does not exclude the possibility of cross loadings. All other loadings are estimated with penalization (see the arrows with broken lines in Fig. 1). The resulting penalty term is $\mathcal{R}(\theta, \gamma) = \sum_{(p,m)\in\mathcal{C}} \rho\left(\left|\lambda_{pm}\right|, \gamma\right)$, where $\mathcal{C}$ is a two-dimensional index set that indicates which loadings are to be penalized.

Example 2 is a full SEM model (right side of Fig. 1). Five exogenous ($\eta_1, \eta_2, \ldots, \eta_5$) and two endogenous ($\eta_5$ and η_6) latent variables are considered. For simplicity, the observed variables are omitted. The measurement part is assumed to be a seven-factor model without cross loadings. For the structural part, prior knowledge states that the exogenous factors are all correlated and that η_6 is influenced by η_1, η_2, and η_4 but not η_3 and η_5. No theory is available to specify the exact relationships among $\eta_1, \eta_2, \ldots, \eta_5$ and η_7. Hence, these effects were set as penalized parameters. The corresponding penalty term is simply $\mathcal{R}(\theta, \gamma) = \sum_{m=1}^{5} \rho\left(\left|\beta_{7m}\right|, \gamma\right)$.

Many penalty functions have been proposed in the literature (see Hastie, Tibshirani, & Wainwright, 2015). ℓ_2 is the most well-known penalty function with the form

$$\rho_{\ell_2}(t, \gamma) = \gamma t^2. \tag{7}$$

In linear regression, ℓ_2 can be used to handle the problem of not full rank design matrix (Hoerl & Kennard, 1970). Another important penalty function is the ℓ_1 penalty,

$$\rho_{\ell_1}(t, \gamma) = \gamma t. \tag{8}$$

A linear regression with ℓ_1, also called the lasso (least absolute shrinkage and selection operator; Tibshirani, 1996), is now a popular method due to its ability to produce sparse estimates. It can be shown that maximizing the PL criterion with ℓ_1 is equivalent to solving the problem of maximizing $\mathcal{L}(\theta)$ subject to $\sum_{q=1}^{Q}\left|\theta_q\right| \leq K$, where K is a positive number depending on γ. Geometrically, $\sum_{q=1}^{Q}\left|\theta_q\right| \leq K$ forms a rhomboid in a Q-dimensional space. The maximizer of $\mathcal{U}(\theta, \gamma)$ is the first place where the contours of $\mathcal{L}(\theta)$ touch the rhomboid. Because the rhomboid has corners on the axes, a sparse estimate arises when the maximizer is located at a corner (Tibshirani, 1996).

Under the ℓ_1 penalty, the tasks of parameter estimation and variable selection can be simultaneously achieved. However, ℓ_1 is only consistent for variable selection under certain restricted conditions (see Zhao & Yu, 2006; Zou, 2006). To overcome this problem, several nonconvex penalties have been proposed. Of these, SCAD (smoothly clipped absolute deviation; Fan & Li, 2001) and MCP (minimax concave penalty; Zhang, 2010) are the two most well known. The functional forms of SCAD and MCP are

$$\rho_{SCAD}(t, \gamma) = \begin{cases} \gamma t & \text{if } t \leq \gamma \\ -\frac{t^2+\gamma^2-2\gamma\delta t}{2(\delta-1)} & \text{if } \gamma < t \leq \gamma\delta \\ \frac{\gamma^2(\delta^2-1)}{2(\delta-1)} & \text{if } \gamma\delta < t \end{cases}, \tag{9}$$

$$\rho_{MCP}(t, \gamma) = \begin{cases} \gamma t - \frac{t^2}{2\delta} & \text{if } t \leq \gamma\delta \\ \frac{1}{2}\gamma^2\delta & \text{if } \gamma\delta < t \end{cases}, \tag{10}$$

respectively. Here, δ is a parameter that controls the convexity of either SCAD or MCP ($\delta > 2$ for SCAD and $\delta > 1$ for MCP). As $\delta \to \infty$ both SCAD and MCP become ℓ_1. On the other hand, a small δ makes them behave similarly to the best subset selection method. In the GLM family, the theoretical results indicate that SCAD and MCP can both yield oracle estimators, even under the case of $P > N$ (Fan & Li, 2001; Kwon & Kim, 2012; Zhang, 2010).

It is worth noting that the PL estimation can be understood as a Bayesian maximum of a posteriori (MAP) method. ℓ_2 and ℓ_1 represent Gaussian and Laplace priors, respectively, and both SCAD and MCP comprise some improper priors (for further discussion, see Meng, 2008; Strawderman, Wells, & Schifano, 2013).

Because a PL with the ℓ_2 penalty cannot select nonzero parameters though sparse estimation, the proposed PL method emphasizes ℓ_1, SCAD, and MCP. In the latter paragraph, "PL-A" is used to indicate a PL method that uses A as a penalty function.

Remark 1. We have mentioned that the exploratory part contains all the ambiguous relations that await further exploration. Although the term "ambiguous relations" is used, in both Examples 1 and 2, the directionality of these relations is clearly specified. When the PL is implemented as an exploratory method, we recommend formulating all the relationships as clearly as possible. PL users should determine whether a relationship between two variables is directional or simply correlated. If the relation is directional, it should be determined which direction is sensible according to the available substantive theories. Blindly using PL is dangerous because it could result in many equivalent models. Under such cases, no theory is available to justify the use of PL, and the resulting model might be nonsense. PL users should avoid such practices.

Remark 2. It is worth noting that for any SEM model, if we replace all the zero loadings with penalized parameters, the PL with a small γ can yield an ESEM-like estimate (see Hirose & Yamamoto, 2015 for such a result in an EFA case). On the other hand, a very large γ results in an exactly traditional SEM result. Therefore, the proposed PL can be seen as a methodology that links the traditional SEM to the ESEM. In theory, through choosing γ, the PL has the capacity to obtain an optimal model on the continuum from the traditional SEM to the ESEM.

3. An ECM Algorithm for Optimizing the PL Criterion

In this section, an expectation-conditional maximization (ECM) algorithm (Meng & Rubin, 1993) for maximizing the PL criterion is proposed. The expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is a popular method by which to obtain estimates in many PL applications with latent variables (e.g., Choi, Zou, & Oehlert, 2011; Garcia, Ibrahim, & Zhu, 2010; Hirose & Yamamoto, 2014; 2015; Ibrahim, Zhu, Garcia, & Guo, 2011). The EM algorithm includes two steps: the E-step calculates the conditional expectation of the complete data likelihood given the observed data, and the M-step maximizes the derived conditional expectation. ECM extends EM by replacing the M-step with several conditional maximizations, called CM-steps. As shown below, each CM-step in the proposed algorithm involves only a maximizer with a closed form expression.

Before introducing the proposed algorithm, some notations must be defined. Let A denote a $P \times M$ matrix with a_{pm} being its (p, m) element. We use $A[p,]$, $A[, m]$, and $A[p, m]$ to denote the pth row, the mth column, and the (p, m) element of matrix A, respectively. Similarly, $A[-p,]$, $A[, -m]$, and $A[-p, -m]$ denote the submatrices of A with its pth row, mth column, and both deleted. If A^{-1} exists, a^{pm} is defined as the (p, m) element of A^{-1}. Also, $b[p]$ is used to denote the pth element in vector b.

Instead of maximizing the PL criterion directly, the proposed ECM alternatively considers the complete data PL criterion

$$\mathcal{U}^C(\theta,\gamma) \propto -\frac{1}{2}\log|\Psi| - \frac{1}{2N}\sum_{n=1}^{N}\epsilon_n^{\mathrm{T}}\Psi^{-1}\epsilon_n - \frac{1}{2}\log|\Phi| - \frac{1}{2N}\sum_{n=1}^{N}\zeta_n^{\mathrm{T}}\Phi^{-1}\zeta_n - \mathcal{R}(\theta,\gamma), \tag{11}$$

where $\epsilon_n = Y_n - \nu - \Lambda\eta_n$ and $\zeta_n = \eta_n - \alpha - \mathrm{B}\eta_n$. Compared to the PL criterion in Eq. (6), which is constructed only based on the marginal normality of Y, the complete data PL treats the latent factor η as observable and is established under the joint normality assumption of Y and η with the relationships defined in Eqs. (1) and (2) (see Appendix A). Note that the normality of Y and η is merely a working assumption. The ECM algorithm can still find a local maximizer when this assumption fails.

The E-step of ECM calculates the conditional expectation of $\mathcal{U}^C(\theta,\gamma)$

$$\mathcal{M}(\theta|\hat{\theta}^{(t)}) = \mathbb{E}[\mathcal{U}^C(\theta,\gamma)|\mathcal{Y}_N,\hat{\theta}^{(t)}], \tag{12}$$

where $\hat{\theta}^{(t)} = \hat{\theta}^{(t)}(\gamma)$ is the estimate at iteration t under the penalty level γ. Define $e_V^{(t+1)} = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} V_n \middle| \mathcal{Y}_N, \hat{\theta}^{(t)}\right]$ and $C_{VW}^{(t+1)} = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} V_n W_n^{\mathrm{T}} \middle| \mathcal{Y}_N, \hat{\theta}^{(t)}\right]$ for any random vectors V_n and W_n. By the functional form of $\mathcal{U}^C(\theta,\gamma)$, it suffices to calculate e_Y, $e_\eta^{(t+1)}$, C_{YY}, $C_{Y\eta}^{(t+1)}$, and $C_{\eta\eta}^{(t+1)}$ to obtain $\mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)$ (see Appendix A).

After obtaining $\mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)$, the qth CM-step increases the value of $\mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)$ through maximizing $\mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)$ under the restriction $g_q(\theta) = g_q\left(\hat{\theta}^{\left(t+\frac{q-1}{Q}\right)}\right)$, where $g_q(\theta) = \left(\theta_1,\ldots,\theta_{q-1},\theta_{q+1},\ldots,\theta_Q\right)$ and $\hat{\theta}^{\left(t+\frac{q-1}{Q}\right)}$ is a maximizer of $\mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)$ at the $(q-1)$th CM-step. Hence, each CM-step increases the value of $\mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)$ in each coordinate. The main idea here is similar to that of the coordinate descent method, which is now the major method for maximizing PL in regressions (e.g., Breheny, & Huang, 2011; Friedman, Hastie, Höfling, & Tibshirani, 2007; Mazumder & Friedman, 2011).

SEM considers two types of parameters: regression-type coefficients and variance and covariance parameters. Regression-type coefficients include the parameters in Λ, B, ν, and α, and the variance and covariance parameters include the parameters in Ψ and Φ. For the regression-type coefficients, the estimates under $\gamma = 0$ can be updated through

$$\begin{aligned}
\hat{\lambda}_{pj}^{(t+1)} &= \frac{1}{C_{\eta\eta}^{(t+1)}[j,j]}\Bigg[\left(C_{Y\eta}^{(t+1)}[p,j] - \hat{\nu}_p^{(t*)}e_\eta^{(t)}[j] - \hat{\Lambda}^{(t*)}[p,-j]C_{\eta\eta}^{(t+1)}[-j,j]\right)\\
&\quad + \sum_{p'\neq p}\frac{\hat{\psi}^{pp'(t*)}}{\hat{\psi}^{pp(t*)}}\left(C_{Y\eta}^{(t+1)}[p',j] - \hat{\nu}_{p'}^{(t*)}e_\eta^{(t+1)}[j] - \hat{\Lambda}^{(t*)}[p',]C_{\eta\eta}^{(t+1)}[,j]\right)\Bigg],
\end{aligned} \tag{13}$$

$$\begin{aligned}
\hat{\beta}_{mj}^{(t+1)} &= \frac{1}{C_{\eta\eta}^{(t+1)}[j,j]}\Bigg[\left(C_{\eta\eta}^{(t+1)}[m,j] - \hat{\alpha}_m^{(t*)}e_\eta^{(t+1)}[j] - \hat{\mathrm{B}}^{(t*)}[m,-j]C_{\eta\eta}^{(t+1)}[-j,j]\right)\\
&\quad + \sum_{m'\neq m}\frac{\hat{\phi}^{mm'(t*)}}{\hat{\phi}^{mm(t*)}}\left(C_{\eta\eta}^{(t+1)}[m',j] - \hat{\alpha}_{m'(t*)}e_\eta^{(t+1)}[j] - \hat{\mathrm{B}}^{(t*)}[m',]C_{\eta\eta}^{(t+1)}[,j]\right)\Bigg],
\end{aligned}$$

TABLE 1.

Shrinkage formulae for $\hat{\lambda}_{pj}^{(t+1)}(\gamma)$, $\hat{\beta}_{mj}^{(t+1)}(\gamma)$, $\hat{\nu}_{p}^{(t+1)}(\gamma)$, $\hat{\alpha}_{m}^{(t+1)}(\gamma)$, $\hat{\psi}_{pj}^{(t+1)}(\gamma)$, and $\hat{\phi}_{mj}^{(t+1)}(\gamma)$ with selected penalty functions.

Penalty function	Shrinkage formula
ℓ_1	$S\left(\hat{\theta}_q^{(t+1)}, w_q^{(t*)}\right)$
SCAD	$\begin{cases} S\left(\hat{\theta}_q^{(t+1)}, w_q^{(t*)}\gamma\right) & \text{if } \left\lvert\hat{\theta}_q^{(t+1)}\right\rvert \leq \gamma\left(1+w_q^{(t*)}\right) \\ \frac{S\left(\hat{\theta}_q^{(t+1)}, w_q^{(t*)}\gamma\delta/(\delta-1)\right)}{1-w_q^{(t*)}/(\delta-1)} & \text{if } \gamma\left(1+w_q^{(t*)}\right) < \left\lvert\hat{\theta}_q^{(t+1)}\right\rvert \leq \gamma\delta \\ \hat{\theta}_q^{(t+1)} & \text{if } \gamma\delta < \left\lvert\hat{\theta}_q^{(t+1)}\right\rvert \end{cases}$
MCP	$\begin{cases} \frac{S\left(\hat{\theta}_q^{(t+1)}, w_q^{(t*)}\gamma\right)}{1-w_q^{(t*)}\delta} & \text{if } \left\lvert\hat{\theta}_q^{(t+1)}\right\rvert \leq \gamma\delta \\ \hat{\theta}_q^{(t+1)} & \text{if } \gamma\delta < \left\lvert\hat{\theta}_q^{(t+1)}\right\rvert \end{cases}$

$\hat{\theta}_q^{(t+1)} = \hat{\theta}_q^{(t+1)}(0)$; $w_q^{(t*)} = \frac{c_q}{\hat{\psi}^{pp(t*)} C_{\eta\eta}^{(t)}[j,j]}$ if $\theta_q = \lambda_{pj}$; $w_q^{(t*)} = \frac{c_q}{\hat{\phi}^{mm(t*)} C_{\eta\eta}^{(t)}[j,j]}$ if $\theta_q = \beta_{mj}$; $w_q^{(t*)} = \frac{c_q}{\hat{\psi}^{pp(t*)}}$ if $\theta_q = \nu_p$; $w_q^{(t*)} = \frac{c_q}{\hat{\phi}^{mm(t*)}}$ if $\theta_q = \alpha_m$; $w_q^{(t*)} = \frac{c_q}{\hat{\tilde{\psi}}_p^{2(t*)} v_{l(j)l(j)}^{(t*)}(p)}$ if $\theta_q = \psi_{pj}$; $w_q^{(t*)} = \frac{c_q}{\hat{\tilde{\phi}}_m^{2(t*)} v_{l(j)l(j)}^{(t*)}(m)}$ if $\theta_q = \phi_{mj}$; $S(\theta, \gamma) = \operatorname{sign}(\theta) \max\{|\theta| - \gamma, 0\}$.

(14)

$$\hat{\nu}_p^{(t+1)} = e_Y[p] - \hat{\Lambda}^{(t*)}[p,]e_\eta^{(t+1)} + \sum_{p^{'} \neq p} \frac{\hat{\psi}^{pp'(t*)}}{\hat{\psi}^{pp(t*)}} \left(e_Y[p^{'}] - \hat{\nu}_{p^{'}}^{(t*)} - \hat{\Lambda}^{(t*)}[p',]e_\eta^{(t+1)}\right),$$

(15)

$$\begin{aligned} \hat{\alpha}_m^{(t+1)} &= e_\eta^{(t+1)}[m] - \hat{\mathrm{B}}^{(t*)}[m,]e_\eta^{(t+1)} \\ &\quad + \sum_{m' \neq m} \frac{\hat{\phi}^{mm'(t*)}}{\hat{\phi}^{mm(t*)}} \left(e_\eta^{(t+1)}[m'] - \hat{\alpha}_{m^{'}}^{(t*)} - \hat{\mathrm{B}}^{(t*)}[m',]e_\eta^{(t+1)}\right), \end{aligned} \tag{16}$$

where $\hat{\theta}^{(t*)} = \hat{\theta}^{\left(t+\frac{q-1}{Q}\right)}$ stands for the newly updated estimate (see Appendix B for the derivation). When $\gamma > 0$, the CM-steps for updating $\hat{\lambda}_{pj}^{(t+1)}(\gamma)$, $\hat{\beta}_{mj}^{(t+1)}(\gamma)$, $\hat{\nu}_p^{(t+1)}(\gamma)$, and $\hat{\alpha}_m^{(t+1)}(\gamma)$ involve shrinkage steps, and the shrinkage functions depend on the choice of penalties. These shrinkage formulae are given in Table 1. Any fixed coefficient is kept at the pre-specified value in each iteration. For the variance and covariance parameters, the CM-steps are derived based on the iterative conditional fitting method (ICF; Chaudhuri, Drton, & Richardson, 2007). Compared to Newton-type methods, the ICF can ensure the semi-positive definiteness of $\hat{\Psi}$ and $\hat{\Phi}$. We extend the original ICF method by allowing the covariance parameters to be penalized (see Appendix C for the details).

In summary, the proposed ECM algorithm can be briefly described as

1. Initialize $e_Y = \frac{1}{N}\sum_{n=1}^N Y_n$, $C_{YY} = \frac{1}{N}\sum_{n=1}^N Y_n Y_n^{\mathrm{T}}$, and $\hat{\theta}^{(t)} \in \Theta$ such that $\Sigma\left(\hat{\theta}^{(t)}\right)$ is positive definite with $t = 0$.
2. Compute $e_\eta^{(t+1)}$, $C_{Y\eta}^{(t+1)}$, and $C_{\eta\eta}^{(t+1)}$ according to Appendix A.
3. For $q = 1, 2, \ldots, Q$

(a) If θ_q is a regression-type coefficient, compute $\hat{\theta}_q^{(t+1)}$ by Eqs. (13)–(16) and the shrinkage formulae in Table 1.

(b) Otherwise, compute $\hat{\theta}_q^{(t+1)}$ by the regularized ICF method described in Appendix C and the shrinkage formulae in Table 1.

4. Repeat 2 and 3 until $\sum_{q=1}^{Q}\left(\hat{\theta}_q^{(t-1)} - \hat{\theta}_q^{(t)}\right)^2 / \sum_{q=1}^{Q}\left(\hat{\theta}_q^{(t)}\right)^2 < \varepsilon$ for some small $\varepsilon > 0$.

4. Practical Considerations in Implementing PL

In this section, we discuss several practical considerations when implementing PL. The first issue is related to the scale setting of the latent variables. After adding a penalty, the invariance property of ML no longer holds (e.g., PL analysis results based on covariance and correlation are not equivalent). Moreover, the scale of variables affects the shrinkage level directly (see Table 1). In SEM, the scale of latent variables is often set by fixing the values of certain factor loadings or factor variances. We suggest fixing the scaling loadings deliberately such that all the latent variables have variances of around one. Approximate unit variance can be achieved using a two-step method by first running a measurement model restricting $\phi_m^2 = 1$ for every m and then conducting PL with scaling loadings fixed at the corresponding estimates obtained in step one.

The second consideration concerns model identifiability. If the specified model $\tau(\theta) = \left(\mu(\theta)^{\mathrm{T}}, \mathrm{vech}\left(\Sigma(\theta)\right)^{\mathrm{T}}\right)^{\mathrm{T}}$ is locally identified around θ, PL can yield a locally unique estimate in a neighborhood of θ, where $\mathrm{vech}\left(\cdot\right)$ is an operator that stacks nonduplicated elements of a symmetric matrix. However, even when $\tau(\theta)$ is not identified, it is still possible to obtain a locally unique PL estimate because the penalty term introduces additional constraints. Motivated by the results of McDonald (1982) and Shapiro and Browne (1983), one possible way for checking the local identifiability of $\tau(\theta)$ under the restriction of the penalty term is to examine whether $\frac{\partial\tau(\hat{\theta})}{\partial\theta^{\mathrm{T}}_{\hat{\mathcal{A}}(\gamma)}}$ is of full column rank, where $\theta_{\hat{\mathcal{A}}(\gamma)}$ is the subvector of θ formed by $\left\{\theta_q\right\}_{q\in\hat{\mathcal{A}}(\gamma)}$, and $\hat{\mathcal{A}}(\gamma) = \left\{q \,|\, \hat{\theta}_q \neq 0\right\}$ is the support of $\hat{\theta}$.

The third practical consideration involves the selection of an optimal regularization parameter γ. Let $\mathcal{D}(\hat{\theta})$ denote the sample discrepancy evaluated at $\hat{\theta}$, i.e., $\mathcal{D}(\hat{\theta}) = \mathcal{D}_{ML}\left(\tau(\hat{\theta}), t\right) = -\log\left|\Sigma(\hat{\theta})^{-1}S\right| + \mathrm{tr}\left(\Sigma(\hat{\theta})^{-1}S\right) - P + \left(\bar{Y} - \mu(\hat{\theta})\right)^{\mathrm{T}}\Sigma(\theta)^{-1}\left(\bar{Y} - \mu(\hat{\theta})\right)$, where $t = \left(\mathrm{vech}(S)^{\mathrm{T}}, \bar{Y}^{\mathrm{T}}\right)^{\mathrm{T}}$ with $S = \frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \bar{Y}\right)\left(Y_n - \bar{Y}\right)^{\mathrm{T}}$ and $\bar{Y} = \frac{1}{N}\sum_{n=1}^{N} Y_n$. The Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978) are two commonly used selectors,[1] defined as

$$AIC(\gamma) = \mathcal{D}(\hat{\theta}) + \frac{2}{N}e(\gamma), \tag{17}$$

$$BIC(\gamma) = \mathcal{D}(\hat{\theta}) + \frac{\log N}{N}e(\gamma), \tag{18}$$

where $e(\gamma)$ is the number of nonzero estimated parameters in $\hat{\theta} = \hat{\theta}(\gamma)$. Given a candidate set $\Gamma = \{\gamma_l\}_{l=0}^{L}$ with $0 = \gamma_0 < \gamma_1 < \ldots < \gamma_L$, an optimal γ can be determined through selecting

[1] In the statistical learning literature, cross validation (CV; Stone, 1974) is another commonly used method for choosing regularization parameters. Since CV generally cannot yield a consistent selection result for the true model, we do not consider it here. However, interested users could still use CV for selecting γ, especially when the primary goal of analysis is to find a model for predicting sample moments.

a $\hat{\gamma}$ that minimizes AIC or BIC on Γ. A method to construct Γ is to choose L and γ_L first and then set $\gamma_l = \frac{l}{L}\gamma_L$. We may simply choose a γ_L such that all penalized parameters are estimated as zero. According to the updating formula in Table 1, $\gamma_L = 1$ is large enough if both manifest and latent variables are standardized. Using the continuity of PL criterion, both $\{AIC(\gamma)\}_{\gamma\in\Gamma}$ and $\{BIC(\gamma)\}_{\gamma\in\Gamma}$ can be computed efficiently if $\hat{\theta}(\gamma_l)$ is used as a warm start for calculating $\hat{\theta}(\gamma_{l-1})$ (see Friedman et al., 2007). When there are multiple minima on Γ, the largest minimizer is selected according to the principle of Occam's razor. If SCAD or MCP is used, an optimal pair of (γ, δ) can be determined with a similar strategy.

Model evaluation is the forth issue of practical concern. We suggest using goodness-of-fit indices for such purposes. Let $df(\hat{\gamma}) = P(P+3)\big/2 - e(\hat{\gamma})$ be the degrees of freedom of the final model, respectively, for the SEM framework presented in Sect. 2. When appropriate, these quantities, $\mu\left(\hat{\theta}(\hat{\gamma})\right)$, $\Sigma\left(\hat{\theta}(\hat{\gamma})\right)$, and $df(\hat{\gamma})$, can be substituted into an existing formula to calculate fit indices for assessing model-data fit. Yet, the validity of this tentative approach calls for a thorough investigation.

The final consideration concerns statistical inference. Inference in PL is generally difficult because of the difficulty of post-model-selection inference (e.g., Leeb & Pötscher, 2006). The theorems derived in the next section show that a heuristic estimator for the covariance matrix of $\hat{\theta}_{\hat{\mathcal{A}}(\gamma)}$ is $\hat{\mathcal{V}}\left(\hat{\theta}_{\hat{\mathcal{A}}(\gamma)}\right) = \frac{1}{N}\hat{\mathcal{F}}^{-1}_{\hat{\mathcal{A}}(\gamma)}\hat{\mathcal{H}}_{\hat{\mathcal{A}}(\gamma)}\hat{\mathcal{F}}^{-1}_{\hat{\mathcal{A}}(\gamma)}$, where $\hat{\mathcal{H}}_{\hat{\mathcal{A}}(\gamma)} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \log \varphi_{\hat{\theta}}(y_n)}{\partial \theta_{\hat{\mathcal{A}}(\gamma)}}\frac{\partial \log \varphi_{\hat{\theta}}(y_n)}{\partial \theta^{\mathrm{T}}_{\hat{\mathcal{A}}(\gamma)}}$, and $\hat{\mathcal{F}}_{\hat{\mathcal{A}}(\gamma)} = -\frac{\partial^2 \mathcal{L}(\hat{\theta})}{\partial \theta_{\hat{\mathcal{A}}(\gamma)}\partial \theta^{\mathrm{T}}_{\hat{\mathcal{A}}(\gamma)}}$. The square root of the diagonal of $\hat{\mathcal{V}}\left(\hat{\theta}_{\hat{\mathcal{A}}(\gamma)}\right)$ provides a standard error estimate. However, the method does not give standard errors for parameters shrunk to zero. The empirical performance of the standard error formula thus requires further evaluation.

5. Asymptotic Properties of the PL Method

In this section, asymptotic properties of the PL method for SEM are derived. Three theorems are presented. Theorems 1 and 2 concern the oracle property of the PL estimator, and Theorem 3 describes the asymptotic behavior of AIC and BIC with regard to choosing the regularization parameter. The big picture here is that under suitable conditions, PL with SCAD/MCP and BIC selector could result in an oracle estimator asymptotically.

Before presenting the derived theorems, some notations are introduced. For a vector $x \in \Re^P$, $\|x\|_0$ is used to denote the ℓ_0 norm of x, i.e., $\|x\|_0 = \sum_{p=1}^{P} 1\left\{x_p \neq 0\right\}$, where $1\{\cdot\}$ is an indicator function. For a square matrix $A \in \Re^{P\times P}$, $\omega_{\min}(A)$ is used to denote the smallest eigenvalue of A. Given an index set $\mathcal{J} \subset \{1, 2, \ldots, P\}$, $x_{\mathcal{J}}$ is defined as the $|\mathcal{J}|$-dimensional vector formed by $\left\{x_p\right\}_{p\in\mathcal{J}}$, and $A_{\mathcal{J}}$ is the $|\mathcal{J}| \times |\mathcal{J}|$ matrix formed by $\left\{a_{pp'}\right\}_{p,p'\in\mathcal{J}}$, where $|\mathcal{J}|$ is the number of elements in $\mathcal{J}$.

To derive the asymptotic properties of the PL method, the following regularity conditions are assumed:

Condition A. $\mathcal{Y}_N = \{Y_n\}_{n=1}^{N}$ is a random sample from some distribution F that satisfies (1) $\mathbb{E}(Y) = \mu^*$; (2) $\mathbb{V}\mathrm{ar}(Y) = \Sigma^* \succ 0$; i.e., Σ^* is positive definite; (3) there exists an $\varepsilon > 0$ such that $\mathbb{E}\left(\left|Y_p\right|^{4+\varepsilon}\right) < \infty$ for all p.

Condition B. For each $\theta \in \Theta$ and any combination of q, q', and $q''(q, q', q'' = 1, 2, \ldots, Q)$, $\frac{\partial^3 \tau(\theta)}{\partial\theta_q \partial\theta_{q'} \partial\theta_{q''}}$ exists.

Condition C. There exists a quasi-true parameter $\theta^* \in \Theta$ such that (1) $\theta^* \in \operatorname{argmax}_{\theta\in\Theta}\mathbb{E}\left(\mathcal{L}(\theta)\right)$; (2) $\|\theta^*\|_0 < \|\theta\|_0$ for any $\theta \in \operatorname{argmax}_{\theta\in\Theta}\mathbb{E}\left(\mathcal{L}(\theta)\right)$, but $\theta \neq \theta^*$; (3) θ^* is the unique maximizer of $\mathbb{E}(\mathcal{L}(\theta))$ on $\Theta_{\mathcal{A}^*}$, where $\mathcal{A}^* = \left\{q \,|\, \theta_q^* \neq 0\right\}$ is the support of θ^*; $\Theta_{\mathcal{A}^*} = \Theta \cap \left(\prod_{q=1}^{Q} \mathfrak{X}_q\right)$ is the restricted parameter space with $\mathfrak{X}_q = \mathfrak{R}$ if $q \in \mathcal{A}^*$, and $\mathfrak{X}_q = \{0\}$ otherwise; (4) there exists a neighborhood of θ^* on $\Theta_{\mathcal{A}^*}$, denoted by $\Omega_{\mathcal{A}^*}(\theta^*)$ and a constant $\kappa_1 > 0$ such that $\omega_{\min}\left(\mathcal{F}_{\mathcal{A}^*}(\theta)\right) > \kappa_1$ for all $\theta \in \Omega_{\mathcal{A}^*}(\theta^*)$, where $\mathcal{F}_{\mathcal{A}^*}(\theta) = \mathbb{E}\left(-\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_{\mathcal{A}^*} \partial \theta_{\mathcal{A}^*}^{\mathrm{T}}}\right)$.

Condition D. For each combination of q, q', and q'', there exists an F-integrable random function $K_{qq'q''}(y)$ such that $\left|\frac{\partial^3 \log \varphi_\theta(y)}{\partial \theta_q \partial \theta_{q'} \partial \theta_{q''}}\right| < K_{qq'q''}(y)$ for all y and θ in the neighborhood of θ^*.

Condition E. The penalty term $\mathcal{R}(\theta, \gamma) = \sum_{q=1}^{Q} c_q \rho\left(\left|\theta_q\right|, \gamma\right)$ satisfies (1) $c_q = 1$ if $\theta_q^* = 0$; (2) $\rho(t, \gamma)$ is increasing and concave in $t > 0$; (3) $\frac{\partial \rho(t,\gamma)}{\partial t}$ is continuous in both t and γ; (4) $\frac{\partial \rho(0+,\gamma)}{\partial t} = \gamma$; (5) $\frac{\partial \rho(t,\gamma)}{\partial t} = 0$ if $t > \delta\gamma$.

Condition F. θ^* is the unique maximizer of $\mathbb{E}(\mathcal{L}(\theta))$ on Θ, and there exists a neighborhood of θ^* on Θ, denoted by $\Omega(\theta^*)$, and a constant $\kappa_2 > 0$ such that $\omega_{\min}(\mathcal{F}(\theta)) \geq \kappa_2$ for all $\theta \in \Omega(\theta^*)$, where $\mathcal{F}(\theta) = \mathbb{E}\left(-\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}\right)$.

Conditions A, B, and D are standard (e.g., Browne, 1984). Condition C requires the existence and the uniqueness of a quasi-true parameter θ^* on a restricted parameter space $\Theta_{\mathcal{A}^*}$ even when $\tau(\theta)$ is not identifiable on the entire parameter space Θ. However, the positive definiteness of $\mathcal{F}_{\mathcal{A}^*}(\theta)$ on $\Omega_{\mathcal{A}^*}(\theta^*)$ implies that $\tau(\theta)$ should be at least locally identified on $\Theta_{\mathcal{A}^*}$. Condition E requires that the penalization indicators must be one for all estimated true-zero parameters. Condition E also restricts the shape of the penalty function. Both SCAD and MCP satisfy these properties, but ℓ_1 does not. Finally, Condition F is a restricted version of Condition C and is required to establish a global theoretical result.

Theorem 1. (Local oracle property) *If Conditions A–E are true, γ satisfies $\gamma \to 0$, and $\sqrt{N}\gamma \to \infty$ as $N \to \infty$, then there exists a strictly local maximizer of $\mathcal{U}(\theta, \gamma)$, denoted by $\hat{\theta} = \hat{\theta}(\gamma)$, such that*

(a) *$lim_{N\to\infty} \mathbb{P}\left(\hat{\mathcal{A}}(\gamma) = \mathcal{A}^*\right) = 1$;*

(b) *$\sqrt{N}\left(\hat{\theta}_{\mathcal{A}^*} - \theta_{\mathcal{A}^*}^*\right) \longrightarrow_{\mathcal{D}} \mathcal{N}\left(0, \mathcal{F}_{\mathcal{A}^*}^{*\ -1} \mathcal{H}_{\mathcal{A}^*}^* \mathcal{F}_{\mathcal{A}^*}^{*\ -1}\right)$, where $\mathcal{F}_{\mathcal{A}^*}^* = \mathbb{E}\left(-\frac{\partial^2 \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*} \partial \theta_{\mathcal{A}^*}^{\mathrm{T}}}\right)$ and $\mathcal{H}_{\mathcal{A}^*}^* = \mathbb{E}\left(\frac{1}{N} \sum_{n=1}^{N} \frac{\partial \log \varphi_{\theta^*}(Y_n)}{\partial \theta_{\mathcal{A}^*}} \frac{\partial \log \varphi_{\theta^*}(Y_n)}{\partial \theta_{\mathcal{A}^*}^{\mathrm{T}}}\right)$.*

All the proofs of derived theorems can be found in the online supplemental materials. Part (a) of Theorem 1 states that the PL method has the potential to correctly identify zero and nonzero model parameters. Hence, it is possible to obtain an estimation result that performs as well as if the sparsity pattern is known in advance, as part (b) shows. Because Theorem 1 does not require Condition F to hold, it could still be true when the specified model is not identifiable on the entire parameter space.

Theorem 1 only ensures that asymptotically, there exists a local maximizer $\hat{\theta}$ such that the oracle property is satisfied. In practice, choosing such a local maximizer is difficult. Hence, it is hoped that a global result can be derived so that the obtained PL estimator is assuredly an oracle estimator. To derive such a global result, Condition F, which ensures the uniqueness of the quasi-true parameter on the whole parameter space, is required.

Theorem 2. (Global oracle property) *Under Conditions A–F, γ satisfies $\gamma \to 0$ and $\sqrt{N}\gamma \to \infty$ as $N \to \infty$. Asymptotically, there exists a unique global maximizer of $\mathcal{U}(\theta, \gamma)$, denoted by $\hat{\theta}$, such that*

(a) $lim_{N\to\infty}\mathbb{P}\left(\hat{\mathcal{A}}(\gamma)=\mathcal{A}^*\right)=1$;
(b) $\sqrt{N}\left(\hat{\theta}_{\mathcal{A}^*}-\theta^*_{\mathcal{A}^*}\right)\longrightarrow_{\mathcal{D}}\mathcal{N}\left(0,{\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\mathcal{H}^*_{\mathcal{A}^*}{\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\right)$.

If Y is normally distributed and $\tau(\theta)$ is correctly specified, the information equality holds (i.e., ${\mathcal{F}^*_{\mathcal{A}^*}}^{-1}=\mathcal{H}^*_{\mathcal{A}^*}$) and Theorem 2 reduces to Corollary 1 below. The main implication of Corollary 1 is that the PL estimator can achieve the Cramér–Rao lower bound, even when the true sparsity pattern is unknown beforehand. Furthermore, Corollary 1 implies that the test statistic $N\cdot\mathcal{D}(\hat{\theta})$ is asymptotically distributed as a Chi-square random variable. Therefore, it is easy to construct an asymptotic $1-\alpha$ level test for examining the null hypothesis $\tau=\tau(\theta)$ versus the alternative $\tau\neq\tau(\theta)$. The asymptotic χ^2 distribution of the test statistic can be also used to justify the application of the goodness-of-fit indices described in Sect. 4.

Corollary 1. *Under Conditions A–F and γ satisfies $\gamma\to 0$ and $\sqrt{N}\gamma\to\infty$ as $N\to\infty$. If the density of Y is actually $\varphi_\theta(y)$, then asymptotically, there exists a unique global maximizer of $\mathcal{U}(\theta,\gamma)$, denoted by $\hat{\theta}$, such that*

(a) $lim_{N\to\infty}\mathbb{P}\left(\hat{\mathcal{A}}(\gamma)=\mathcal{A}^*\right)=1$;
(b) $\sqrt{N}\left(\hat{\theta}_{\mathcal{A}^*}-\theta^*_{\mathcal{A}^*}\right)\longrightarrow_{\mathcal{D}}\mathcal{N}\left(0,{\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\right)$;
(c) $N\cdot\mathcal{D}(\hat{\theta})\longrightarrow_{\mathcal{D}}\chi^2_{df^*}$, *where* $df^*=P(P+3)/2-|\mathcal{A}^*|$.

Next, the asymptotic properties of AIC and BIC are derived under the proposed PL. Given a model $\tau(\theta)$, for any index set $\mathcal{A}\subset\{1,2,\ldots,Q\}$, the minimum discrepancy function (MDF) value of $\tau(\theta)$ on $\Theta_{\mathcal{A}}$ is defined as $\mathcal{D}^*_{\mathcal{A}}=\min_{\theta\in\Theta_{\mathcal{A}}}\mathcal{D}_{ML}\left(\tau(\theta),\tau^*\right)$, where $\tau^*=\left(\text{vech}\,(\Sigma^*)^{\mathrm{T}},{\mu^*}^{\mathrm{T}}\right)^{\mathrm{T}}$. Hence, by examining the values of $\mathcal{D}^*_{\mathcal{A}}$ and $\mathcal{D}^*_{\mathcal{A}'}$, the correctness of $\tau(\theta)$ restricted on $\Theta_{\mathcal{A}}$ and $\Theta_{\mathcal{A}'}$ can be compared. According to the definition of $\mathcal{A}^*$, $\mathcal{D}^*_{\mathcal{A}^*}\leq\mathcal{D}^*_{\mathcal{A}}$ for any $\mathcal{A}\subset\{1,2,\ldots,Q\}$. If some $\mathcal{A}$ satisfies $\mathcal{D}^*_{\mathcal{A}^*}=\mathcal{D}^*_{\mathcal{A}}$, Condition C indicates that $\mathcal{A}^*$ must be more parsimonious than $\mathcal{A}$, i.e., $|\mathcal{A}^*|<|\mathcal{A}|$. Given a random sample $\mathcal{Y}_N$, the set of regularization parameters is partitioned into three subsets

$$\Gamma^*=\left\{\gamma\,|\,\mathcal{D}^*_{\hat{\mathcal{A}}(\gamma)}=\mathcal{D}^*_{\mathcal{A}^*},|\hat{\mathcal{A}}(\gamma)|=\left|\mathcal{A}^*\right|\right\},\tag{19}$$

$$\Gamma^+=\left\{\gamma\,|\,\mathcal{D}^*_{\hat{\mathcal{A}}(\gamma)}=\mathcal{D}^*_{\mathcal{A}^*},|\hat{\mathcal{A}}(\gamma)|>\left|\mathcal{A}^*\right|\right\},\tag{20}$$

$$\Gamma^-=\left\{\gamma\,|\,\mathcal{D}^*_{\hat{\mathcal{A}}(\gamma)}>\mathcal{D}^*_{\mathcal{A}^*}\right\}.\tag{21}$$

The subset Γ^* contains all the values of γ where the optimal model $\mathcal{A}^*$ is attained. On the other hand, Γ^+ and Γ^- are formed by γ such that the corresponding models are overfitted and underfitted, respectively. Note that $\hat{\mathcal{A}}(\gamma)$ with $\gamma\in\Gamma^+$ may not be really "overfitting" in the usual sense. An overfitting model is generally used to refer to a model that explains the phenomenon perfectly but contains unnecessary parameters. However, "overfitting" here is merely used to emphasize that $\hat{\mathcal{A}}(\gamma)$ contains unnecessary parameters because it is possible that $\mathcal{D}^*_{\hat{\mathcal{A}}(\gamma)}>0$.

Theorem 3. *Let $\hat{\gamma}^{AIC}$ and $\hat{\gamma}^{BIC}$ denote the selection results based on AIC and BIC, respectively. Under Conditions A–F, we have*

(a) $lim_{N\to\infty}\mathbb{P}\left(\hat{\gamma}^{AIC}\in\Gamma^-\right)=0$ *and* $lim_{N\to\infty}\mathbb{P}\left(\hat{\gamma}^{AIC}\in\Gamma^+\right)>0$;
(b) $lim_{N\to\infty}\mathbb{P}\left(\hat{\gamma}^{BIC}\in\Gamma^*\right)=1$.

Theorems 3 shows that asymptotically, both AIC and BIC select a model that attains the smallest MDF value $\mathcal{D}^*_{\mathcal{A}^*}$. However, only BIC yields a consistent selection result with respect to $\mathcal{A}^*$. AIC may suffer from the problem of overfitting. Of course, if Γ^+ is empty, AIC can also select the quasi-true model with the probability one. The derived results are consistent with the typical behavior of AIC and BIC in parametric regression models (e.g., Zhang, Li, & Tsai, 2012; Shao, 1997).

6. Real Data Illustrations

6.1. Real Data Illustration 1: Nine Psychological Tests

The nine psychological tests adopted from Holzinger and Swineford (1939) are often used to demonstrate factor analysis methods (e.g., Jöreskog, 1969; Rosseel, 2012). This data set contains the responses of 301 seventh and eighth grade students on nine mental tests. These tests are thought of as measuring three correlated abilities: visualization (tests 1, 2, 3), verbal intelligence (tests 4, 5, 6), and speed (tests 7, 8, 9). Under the CFA framework, all of the tests are assumed to be pure measures, and the associated measurement errors are uncorrelated. The left part of Table 2 presents the corresponding ML estimation result. The overall pattern of parameter estimates is reasonable; however, the goodness-of-fit indices show that the model does not fit the data well ($\chi^2 = 85.30$; $df = 24$; $p < .001$; RMSEA $= 0.092$; CFI $= 0.930$; NNFI $= 0.896$).

One possible cause of this misfit is that some tests are not pure measures. Nevertheless, it is difficult to specify the correct sparsity pattern of a loading matrix a priori. In such circumstances, PL is a plausible method by which to explore the underlying loading pattern. When using PL, the model is specified according to Example 1 in Sect. 2. The SCAD is utilized. Based on the value of $BIC(\gamma)$, $\left(\hat{\gamma}, \hat{\delta}\right) = (0.1, 2.5)$ is selected from $\Gamma \times \Delta = \{.01, .02, \ldots, .20\} \times \{2.5, 3.5, 4.5\}$. The central part of Table 2 shows the final PL estimation result. Three penalized loadings are identified as nonzero ($\hat{\lambda}_{71}$, $\hat{\lambda}_{91}$, and $\hat{\lambda}_{12}$). By using the standard error formula described in Sect. 4, all of the nonzero estimates are significantly different from zero at the 0.05 level. The smallest singular value of $\frac{\partial \tau\left(\hat{\theta}(\hat{\gamma})\right)}{\partial \theta^{\mathrm{T}}_{\hat{\mathcal{A}}(\hat{\gamma})}}$ is 0.15, indicating that the final model is locally identified. Goodness-of-fit indices show that the final model fits the data relatively well compared to the CFA result ($\chi^2 = 35.53$; $df(\hat{\gamma}) = 21$; $p = .006$; RMSEA $= 0.048$; CFI $= 0.984$; NNFI $= 0.972$).

An oblique EFA model is also used to fit the data. The right part of Table 2 shows the ML estimation result with a promax rotation (power $= 4$). The estimated loading matrix is quite similar to the PL solution. Goodness-of-fit measures show that the EFA model also performs better than the CFA model ($\chi^2 = 22.41$; $df = 12$; $p = .033$; RMSEA $= 0.055$; CFI $= 0.987$; NNFI $= 0.963$). However, the EFA model contains many nearly zero factor loadings, which means that the model is unnecessarily complex. As a result, the EFA model performs slightly worse than the PL on RMSEA and NNFI. It shows that a complex model does not always outperform a simple one if the considered fit index takes into account the issue of model complexity.

6.2. Real Data Illustration 2: Five Facets of Mindfulness and the Negative/Positive Affect

Chang, Lin, and Huang (2010) collected responses from 231 undergraduate students on the Five Facets Mindfulness Questionnaire (FFMQ; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006) and Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988). The FFMQ is designed to measure five facets of mindfulness: acting with awareness (AA), non-judgment (NJ), describing (DE), non-reactivity (NR), and observing (OB). PANAS measures the levels of positive affect (PA) and negative affect (NA) of an individual, respectively. In this illustration, the question "how do the five facets of mindfulness predict an individual's PA and NA?"

TABLE 2.
Parameter estimates of nine psychological tests under confirmatory, semi-confirmatory (PL), and exploratory factor analysis.

Loading and uniqueness	Confirmatory				Semi-confirmatory				Exploratory			
	Visual	Verbal	Speed	Uniqueness	Visual	Verbal	Speed	Uniqueness	Visual	Verbal	Speed	Uniqueness
Visual perception	0.90			0.55	*0.90*	0.25		0.68	0.74	0.15	0.01	0.70
Cubes	0.50			1.14	0.70			1.07	0.64	−0.02	−0.17	1.02
Lozenges	0.66			0.85	0.92			0.74	0.82	−0.16	0.00	0.70
Paragraph comprehension		0.99		0.37		0.99		0.37	0.03	0.97	0.01	0.38
Sentence completion		1.10		0.45		*1.10*		0.45	−0.08	1.15	0.01	0.41
Word meaning		0.92		0.36		0.91		0.36	0.10	0.87	−0.02	0.37
Speeded addition			0.62	0.80	−0.35		0.85	0.58	−0.19	0.06	0.80	0.59
Speeded counting of dots			0.73	0.49			*0.73*	0.50	0.10	−0.05	0.71	0.48
Speeded discrimination			0.67	0.57	0.41		0.50	0.54	0.38	−0.01	0.46	0.55
Covariance of factors	Visual	Verbal	Speed		Visual	Verbal	Speed		Visual	Verbal	Speed	
Visual	*1.00*	–	–		0.63	–	–		*1.00*	–	–	
Verbal	0.46	*1.00*	–		0.24	1.01	–		0.41	*1.00*	–	
Speed	0.47	0.23	*1.00*		0.31	0.27	0.99		0.34	0.23	*1.00*	

Parameter fixed for scale setting are italic. A table cell that is blank means that the corresponding estimate is zero.

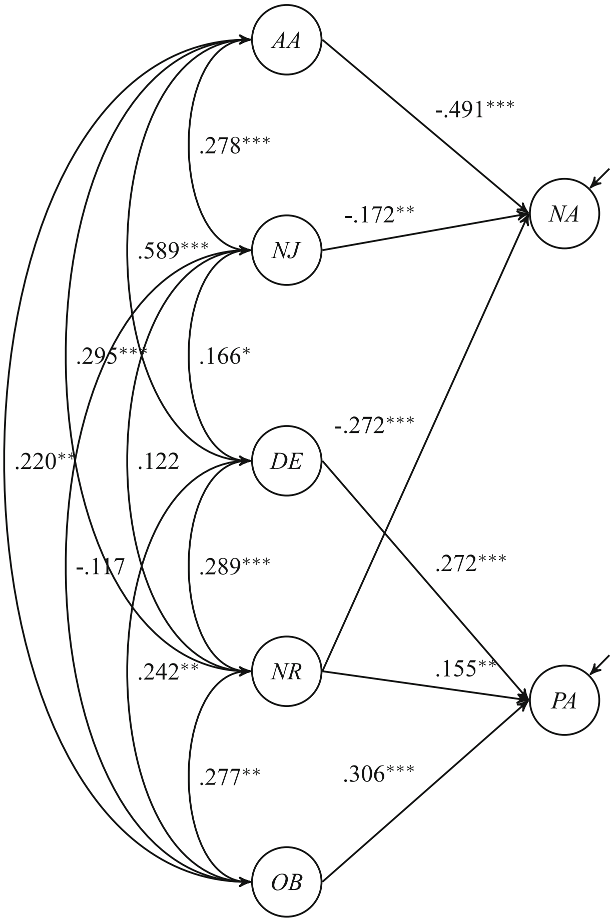

FIGURE 2.
Final path coefficient estimates for the structural model of real data illustration 2. *AA* acting with awareness, *NJ* non-judgment, *DE* describing, *NR* non-reactivity, *OB* observing, *NA* negative affect, *PA* positive affect. $^{*}p < .05$;$^{**}p < .01$;$^{***}p < .001$.

is investigated. Because the main purpose of the data analysis is to understand the relationships among latent variables, 21 parcels from the original items are formed. The measurement model is specified according to the relationships between the parcels and the latent constructs. Each factor is measured using three indicators. The loading of the first indicator for each factor is fixed. The fixed values are .941, .890, .906, .803, .757, .857, and .832, respectively. For the structural model, based on the empirical results of Baer et al. (2006), NA is expected to be influenced by AA, NJ, and NR, forming the confirmatory part of the structural model. However, the relationships between the five facets of mindfulness and the PA are unclear. Hence, these effects are set as penalized parameters. The model specification of the structural part is the same as that of Example 2 in Sect. 2.

Figure 2 shows the path coefficient estimates in the structural model of Illustration 2. MCP is used as the penalty function, and $\left(\hat{\gamma}, \hat{\delta}\right) = (.1, 1.5)$ is chosen from $\Gamma \times \Delta = \{.01, .02, \ldots, .20\} \times \{1.5, 2.5, 3.5\}$ according to the value of $BIC(\gamma)$. The final model is locally identified because the smallest singular value of $\frac{\partial\tau\left(\hat{\theta}(\hat{\gamma})\right)}{\partial\theta^{\mathrm{T}}_{\hat{\mathcal{A}}(\hat{\gamma})}}$ is 0.34. Goodness-of-fit indices indicate that the final model fitted the data reasonably ($\chi^2 = 289.09$; $df(\hat{\gamma}) = 173$; RMSEA$(\hat{\gamma}) = 0.054$; CFI$(\hat{\gamma}) = 0.961$; NNFI$(\hat{\gamma}) = 0.953$). For the confirmatory part, AA, NJ, and NR were shown to have negative effects on an individual's negative affect, as expected. For the exploratory part, only DE, NR, and OB were shown to have positive effects on individual PA.

7. Numerical Experiment

In this section, a numerical experiment was conducted to evaluate the performance of the PL method. Two different types of models were considered: an oblique factor analysis model and a complete SEM model. The simulations were conducted in an exploratory manner to assess the effectiveness of PL in recovering the underlying structure. Two questions were to be addressed in the simulation. First, how well does the PL estimator perform compared to the non-oracle ML (NML) and the oracle ML (OML) estimators? Second, how often does the PL method select the true model? The NML treats all the penalized parameters in PL as free parameters, and the OML is the usual ML estimation under the true model. Because the NML tends to yield a relatively unstable estimator, comparing the PL estimator with the NML estimator shows how PL improves the quality of estimation by controlling model complexity. On the other hand, the OML yields an oracle estimator that can be taken as a golden standard. Based on the theory of PL in Sect. 5, we expected that (a) the PL estimators will generally outperform the NML estimator, (b) the estimators from PL with SCAD/MCP and BIC, the so-called *oracle PL*, will perform similarly to the OML under large samples, and (c) the oracle PL will select the true model with high probability asymptotically.

For each model, normally distributed data under the true covariance structure were generated with sample sizes being 100, 200, 400, 800, or 1600 and 500 converged replications that were included under each condition. All the simulations were conducted in R (R Core Team, 2016) with the lsl package used for PL analysis (Huang, 2015).

Model 1 is an oblique factor analysis model with 20 indicators and 4 common factors. The true relationship among the indicators and factors is given in Fig. 3. The measurement error covariance matrix was set as a diagonal matrix such that each indicator had unit variance. Each data set was analyzed through PL with ℓ_1/SCAD/MCP, NML, and OML. The three methods estimated the loading matrix Λ in different ways. For the PL method, loadings in the "independent cluster" part, including $\left\{\lambda_{pm} \middle| p = 5(m-1)+k, m = 1, \ldots, 4, k = 1, \ldots, 5\right\}$, were freely estimated, but loadings in the "non-independent cluster" part were set as penalized parameters. An optimal pair $\left(\hat{\gamma}, \hat{\delta}\right)$ was chosen from $\Gamma \times \Delta$ through the value of AIC and BIC, where $\Gamma = \{.01, .02, \ldots, .30\}$ and $\Delta = \{2.2, 2.6, 3.0\}$ for SCAD, and $\Delta = \{1.2, 1.6, 2.0\}$ for MCP. For the NML method, an initial loading matrix was freely estimated and was then obliquely rotated to a partially specified target (Browne, 1972) by assuming the "non-independent cluster" part should be as small as possible. Finally, the OML only estimated the true nonzero parameters.

Model 2 is an SEM model with 27 manifest and 9 latent variables. The true measurement model was specified as

$$\Lambda = I_9 \otimes \begin{bmatrix} 0.80 \\ 0.75 \\ 0.70 \end{bmatrix}, \quad \Psi = I_9 \otimes \begin{bmatrix} .36 & 0 & 0 \\ 0 & .4375 & 0 \\ 0 & 0 & .51 \end{bmatrix}. \tag{22}$$

The nonzero parameters in the path coefficient matrix B were $\beta_{21} = \beta_{32} = \beta_{54} = \beta_{65} = \beta_{87} = \beta_{98} = 0.35$ and $\beta_{41} = \beta_{74} = \beta_{52} = \beta_{85} = \beta_{63} = \beta_{96} = 0.45$. The residual covariance matrix Φ was chosen such that all the latent variables had unit variances. When a data set is analyzed through a PL method, all coefficients in the lower triangular of B are estimated with ℓ_1,

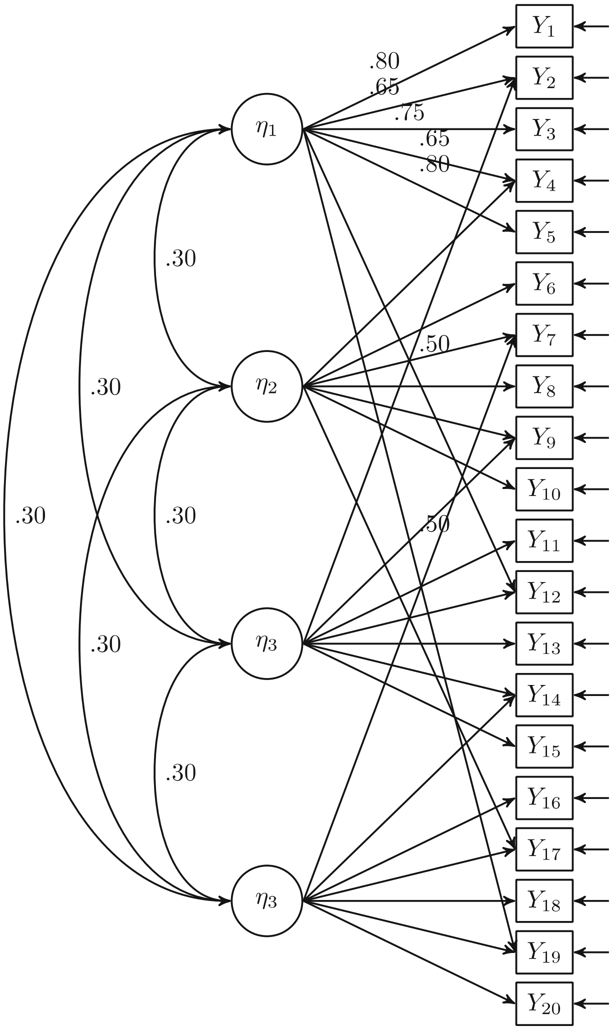

FIGURE 3.
Path diagram for the true model of model 1 in the numerical experiment: an oblique factor analysis model. For simplicity, only values of factor correlations and loadings associated with the first factor are presented. Values of other factor loadings are set in a similar way.

SCAD, or MCP. An optimal pair $\left(\hat{\gamma}, \hat{\delta}\right)$ was chosen from $\Gamma \times \Delta$ using AIC and BIC, where $\Gamma = \{.01, .02, \ldots, .30\}$ and $\Delta = \{2.2, 2.6, 3.0\}$ for SCAD, and $\Delta = \{1.2, 1.6, 2.0\}$ for MCP. NML freely estimates all the lower diagonal paths, and OML estimates the true nonzero parameters only.

The performance of the proposed PL was evaluated with five criteria. First, the overall estimation quality was assessed by the estimated mean squared error (MSE), defined as

$$\widehat{\text{MSE}}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}^{(r)}(\hat{\gamma}) - \theta^*\right)^{\text{T}} \left(\hat{\theta}^{(r)}(\hat{\gamma}) - \theta^*\right), \tag{23}$$

where $\hat{\theta}^{(r)}(\hat{\gamma}) = \left(\hat{\theta}_1^{(r)}(\hat{\gamma}), \ldots, \hat{\theta}_Q^{(r)}(\hat{\gamma})\right)^{\text{T}}$ denoted the final estimate in the rth replication; θ^* was the true parameter, and $R = 500$ was the number of replications. Second, the degree of bias of an estimator was evaluated by the estimated squared bias (SB)

TABLE 3.
Performances of PL, NML, and OML in Model 1: a factor analysis model.

	PL-ℓ_1		PL-SCAD		PL-MCP		NML	OML
	AIC	BIC	AIC	BIC	AIC	BIC		
PCTM								
$N = 100$	0.000	0.008	0.008	0.016	0.034	0.074	–	–
$N = 200$	0.000	0.022	0.122	0.188	0.122	0.544	–	–
$N = 400$	0.002	0.034	0.436	0.726	0.164	0.896	–	–
$N = 800$	0.000	0.082	0.542	0.968	0.228	0.952	–	–
$N = 1600$	0.000	0.044	0.520	0.986	0.308	0.968	–	–
TDR								
$N = 100$	1.000	1.000	1.000	1.000	1.000	1.000	–	–
$N = 200$	1.000	1.000	1.000	1.000	1.000	1.000	–	–
$N = 400$	1.000	1.000	1.000	1.000	1.000	1.000	–	–
$N = 800$	1.000	1.000	1.000	1.000	1.000	1.000	–	–
$N = 1600$	1.000	1.000	1.000	1.000	1.000	1.000	–	–
FDR								
$N = 100$	0.274	0.116	0.160	0.091	0.163	0.054	–	–
$N = 200$	0.272	0.084	0.084	0.034	0.141	0.014	–	–
$N = 400$	0.270	0.073	0.057	0.007	0.124	0.003	–	–
$N = 800$	0.280	0.062	0.049	0.001	0.105	0.001	–	–
$N = 1600$	0.289	0.066	0.043	0.000	0.075	0.001	–	–
MSE								
$N = 100$	0.805	0.965	0.580	0.627	0.651	0.508	0.696	0.471
$N = 200$	0.400	0.524	0.249	0.262	0.294	0.228	0.440	0.223
$N = 400$	0.209	0.282	0.118	0.116	0.140	0.112	0.330	0.110
$N = 800$	0.110	0.163	0.060	0.058	0.069	0.058	0.273	0.057
$N = 1600$	0.058	0.094	0.030	0.029	0.034	0.029	0.249	0.029
SB								
$N = 100$	0.277	0.477	0.012	0.049	0.006	0.002	0.169	0.002
$N = 200$	0.149	0.300	0.003	0.008	0.003	0.001	0.194	0.001
$N = 400$	0.083	0.174	0.001	0.001	0.001	0.000	0.209	0.000
$N = 800$	0.043	0.106	0.000	0.000	0.001	0.000	0.214	0.000
$N = 1600$	0.024	0.065	0.000	0.000	0.001	0.000	0.219	0.000

PL penalized likelihood, *NML* non-oracle maximum likelihood, *OML* oracle maximum likelihood, *AIC* Akaike information criterion, *BIC* Bayesian information criterion, *PCTM* proportion choosing the true model, *TDR* true discovery rate, *FDR* false discovery rate, *MSE* mean squared errors, *SB* squared bias, *N* sample size.

$$\widehat{SB}(\hat{\theta}) = \left(\bar{\hat{\theta}} - \theta^*\right)^{\mathrm{T}} \left(\bar{\hat{\theta}} - \theta^*\right), \tag{24}$$

where $\bar{\hat{\theta}} = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}^{(r)}(\hat{\gamma})$ represented the empirical mean of $\hat{\theta}(\hat{\gamma})$. The third criterion, the proportion choosing the true model (PCTM), assessed selection consistency via the proportion of the true model that was chosen over the replications. The chance of correctly identifying the true nonzero parameters was the fourth criterion as evaluated by the estimated true discovery rate (TDR),

$$\widehat{\mathrm{TDR}}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R} \frac{\sum_{q \in \mathrm{B}} 1\left\{\hat{\theta}_q^{(r)}(\hat{\gamma}) \neq 0\right\}}{|\mathcal{B}|}, \tag{25}$$

TABLE 4.
Performances of PL, NML, and OML in Model 2: An SEM Model.

	PL-ℓ_1		PL-SCAD		PL-MCP		NML	OML
	AIC	BIC	AIC	BIC	AIC	BIC		
PCTM								
$N = 100$	0.000	0.000	0.044	0.114	0.038	0.180	–	–
$N = 200$	0.000	0.000	0.264	0.574	0.072	0.538	–	–
$N = 400$	0.000	0.002	0.346	0.918	0.060	0.772	–	–
$N = 800$	0.000	0.000	0.378	0.950	0.092	0.816	–	–
$N = 1600$	0.000	0.000	0.358	0.960	0.100	0.870	–	–
TDR								
$N = 100$	0.997	0.995	0.964	0.953	0.962	0.939	–	–
$N = 200$	1.000	1.000	0.995	0.992	0.994	0.988	–	–
$N = 400$	1.000	1.000	1.000	1.000	1.000	0.999	–	–
$N = 800$	1.000	1.000	1.000	1.000	1.000	1.000	–	–
$N = 1600$	1.000	1.000	1.000	1.000	1.000	1.000	–	–
FDR								
$N = 100$	0.511	0.342	0.236	0.101	0.197	0.073	–	–
$N = 200$	0.504	0.329	0.163	0.031	0.164	0.033	–	–
$N = 400$	0.490	0.308	0.142	0.006	0.159	0.013	–	–
$N = 800$	0.492	0.311	0.126	0.003	0.139	0.009	–	–
$N = 1600$	0.498	0.305	0.134	0.002	0.147	0.006	–	–
MSE								
$N = 100$	1.099	1.104	1.305	1.161	1.296	1.151	2.053	0.725
$N = 200$	0.549	0.561	0.531	0.441	0.565	0.462	0.883	0.363
$N = 400$	0.267	0.279	0.244	0.184	0.268	0.195	0.418	0.177
$N = 800$	0.135	0.145	0.118	0.091	0.128	0.094	0.199	0.088
$N = 1600$	0.068	0.074	0.058	0.045	0.064	0.046	0.099	0.044
SB								
$N = 100$	0.114	0.243	0.017	0.025	0.022	0.028	0.029	0.002
$N = 200$	0.067	0.141	0.004	0.003	0.005	0.004	0.007	0.001
$N = 400$	0.033	0.075	0.001	0.000	0.002	0.000	0.002	0.000
$N = 800$	0.020	0.043	0.001	0.000	0.001	0.000	0.002	0.000
$N = 1600$	0.010	0.023	0.000	0.000	0.001	0.000	0.001	0.000

PL penalized likelihood, *NML* non-oracle maximum likelihood, *OML* oracle maximum likelihood, *AIC* Akaike information criterion, *BIC* Bayesian information criterion, *PCTM* proportion choosing the true model, *TDR* true discovery rate, *FDR* false discovery rate, *MSE* mean squared errors, *SB* squared bias, *N* sample size.

where $\mathcal{B} = \left\{ q \,|\, \theta_q^* \neq 0, c_q = 1 \right\}$, and $|\mathcal{B}|$ was the size of B. Lastly, the degree to which the true-zero parameters were incorrectly identified as nonzero was examined by the estimated false discovery rate (FDR)

$$\widehat{\text{FDR}}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \frac{\sum_{q \in \mathcal{B}^c} 1 \left\{ \hat{\theta}_q^{(r)}(\hat{\gamma}) \neq 0 \right\}}{|\mathcal{B}^c|}, \tag{26}$$

where $\mathcal{B}^c = \left\{ q \,|\, \theta_q^* = 0, c_q = 1 \right\}$.

Tables 3 and 4 show the simulation results for models 1 and 2, respectively. The relative performance of the different PL implementations, NML, and OML, were similar across the two models. The OML estimator performed the best, with the lowest MSE and SB, and the NML estimator performed the worst, with a relatively large MSE. The PL estimators uniformly outperformed the NML estimator, supporting our expectations (a). Within the class of PL estimators, the estimators under the oracle PL performed better than those under ℓ_1 or AIC. Especially, under relatively large sample sizes ($N \geq 800$), the oracle PL estimators performed similarly to those of the OML estimator, confirming our expectations (b). The PCTM showed that the oracle PL was more likely to select the true model than the PL using ℓ_1 or AIC. The high PCTM of the oracle PL under large sample sizes supported our expectations (c). If the oracle PL did not select the true model, the TDR of nearly one showed that it tended to choose relatively complex models. However, the selected models were not very complex because most FDRs were still low. For example, FDRs were all less than 0.05 if the sample size was at least moderate ($N \geq 200$). Therefore, the model selected by the oracle PL was usually close to the true model.

Remark 3. Note that the performance of NML can be improved by hard-thresholding small parameter estimates or using model modification indices. However, in contrast to the hard-threshold or the modification method, PL can result in a continuous estimator by continuous thresholding. Theoretically, a continuous estimation is expected to be relatively stable than a noncontinuous one (see Breiman, 1996; Fan & Li, 2001). We generally recommend using continuous methods whenever they are available.

8. Conclusion and Future Direction

A penalized likelihood method for structural equation modeling is proposed in the present study. The major advantage of the PL method lies in its capacity to incorporate the confirmatory and the exploratory elements in a model, allowing for a simultaneous consideration of existing theories and ambiguous inter-variable relations that await exploration. The proposed method can be applied to a wide class of SEM models to efficiently reach an optimal model. The derived theorems show that the PL can asymptotically yield an oracle estimator. The numerical experiment further confirms such theoretical results. The better performance of PL over NML across all the sample sizes also demonstrates the benefit of controlling model complexity in an exploratory analysis. This preliminary evidence suggests that when uncertainty in relationships among variables exists, the PL method provides an option for conducting SEM.

The present study develops an ECM algorithm for optimizing the PL criteria. The ECM extends the algorithms of Choi, Zou, and Oehlert (2011) and Hirose and Yamamoto (2014, 2015) for EFA in the sense that the current algorithm allows structural parameters and covariances to be penalized. Because SEM can be formulated under GLMM or a regression with incomplete data, the ECM can be also seen as a variant of the algorithms developed by Garcia, Ibrahim, Zhu (2010) and Ibrahim, Zhu, Garcia, and Guo (2011) by considering SEM-specific parameters to be penalized. Nevertheless, the algorithms of Ibrahim and his colleagues could be more general in the sense that they consider any type of response belonging to an exponential family. In the future, it would be worth to generalizing our algorithm to SEM with more flexible response variables.

The novel PL method for SEM presented in this study is not without limitations. Future research in several directions can further our understanding of PL and provide improvements over the current development. First, the standard errors and goodness-of-fit indices under PL are proposed without numerical evaluation. Although both of them were used in our real data illustrations, further studies are required to examine their empirical performances. Second, our numerical experiment only considers relatively standard conditions. Future simulations can be extended to evaluate the PL under more realistic settings, such as non-normal data (e.g., Micceri, 1989), candidate sets without containing the true model (e.g., Cudeck & Henly, 1991), the presence of small parameters (e.g., Vrieze, 2012), and the case of $P > N$. Third, the PL method was not compared with model modification methods (e.g., Chou & Bentler, 1990; MacCallum, 1986). Future studies can be designed to compare these methods in order to understand their relative advantages and disadvantages. Finally, the present PL considers only linear SEM. In reality, the relationships among observed variables and latent variables might be nonlinear (e.g., Lee & Zhu, 2002). Developing PL methods for more complex SEM situations could enhance the applicability of the framework proposed herein.

Appendix A: The E-Step of the ECM Algorithm

Under the normal assumption of ϵ and ζ and the model defined in Eqs. (1) and (2), the joint distribution of Y and η is

$$\begin{pmatrix} Y \\ \eta \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu \\ \kappa \end{pmatrix}, \begin{pmatrix} \Sigma & \Xi\Lambda^{\mathrm{T}} \\ \Lambda\Xi & \Xi \end{pmatrix}\right],$$

where $\Sigma = \Lambda\Xi\Lambda^{\mathrm{T}} + \Psi$; $\Xi = (I_M - \mathrm{B})^{-1}\Phi(I_M - \mathrm{B})^{-1^{\mathrm{T}}}$; $\mu = \nu + \Lambda\kappa$, and $\kappa = (I_M - \mathrm{B})^{-1}\alpha$. Therefore, the complete data log-likelihood in Eq. (11) can be obtained using the distribution of Y given η and the distribution of η, or equivalently, the distribution of ζ.

To derive the expression for the E-step, the conditional moments of the latent factor given the data are necessary. The joint normality of Y and η implies that the conditional mean and variance of η given Y are

$$\begin{aligned} \mathbb{E}\,(\eta|\,Y) &= \kappa + \Xi\Lambda^{\mathrm{T}}\Sigma^{-1}(Y - \mu), \\ \mathbb{V}\mathrm{ar}\,(\eta|\,Y) &= \Xi - \Xi\Lambda^{\mathrm{T}}\Sigma^{-1}\Lambda\Xi. \end{aligned}$$

Hence, we have

$$\begin{aligned} e_{\eta}^{(t+1)} &= J^{(t)} + K^{(t)}e_Y, \\ C_{Y\eta}^{(t+1)} &= e_Y J^{(t)^{\mathrm{T}}} + C_{YY}K^{(t)^{\mathrm{T}}}, \\ C_{\eta\eta}^{(t+1)} &= \hat{\Xi}^{(t)} - \hat{\Xi}^{(t)}\hat{\Lambda}^{(t)^{\mathrm{T}}}\hat{\Sigma}^{(t)^{-1}}\hat{\Lambda}^{(t)}\hat{\Xi}^{(t)} + J^{(t)}J^{(t)^{\mathrm{T}}} + J^{(t)}e_Y^{\mathrm{T}}K^{(t)^{\mathrm{T}}} \\ &\quad + K^{(t)}e_Y J^{(t)^{\mathrm{T}}} + K^{(t)}C_{YY}K^{(t)^{\mathrm{T}}}, \end{aligned}$$

where $J^{(t)} = \hat{\kappa}^{(t)} - \hat{\Xi}^{(t)}\hat{\Lambda}^{(t)^{\mathrm{T}}}\hat{\Sigma}^{(t)^{-1}}\hat{\mu}^{(t)}$ and $K^{(t)} = \hat{\Xi}^{(t)}\hat{\Lambda}^{(t)^{\mathrm{T}}}\hat{\Sigma}^{(t)^{-1}}$. The derivation of $C_{\eta\eta}^{(t+1)}$ is based on the identity $Var\,(\eta|\,Y) = E\left(\eta\eta^{\mathrm{T}}\middle|\,Y\right) - E\,(\eta|\,Y)\,E\left(\eta^{\mathrm{T}}\middle|\,Y\right)$.

Appendix B: The CM-Steps for Regression-Type Coefficients

Consider the updating scheme for λ_{pj}. If $\left|\lambda_{pj}\right| > 0$, the first partial derivative of $\mathcal{M}\left(\theta|\,\hat{\theta}^{(t)}\right)$ with respect to λ_{pj} is

$$\frac{\partial\mathcal{M}\left(\theta|\,\hat{\theta}^{(t)}\right)}{\partial\lambda_{pj}} = \psi^{pp}\left(C_{Y\eta}^{(t+1)}[p,j] - \nu_p e_{\eta}^{(t+1)}[j] - \lambda_{pj}C_{\eta\eta}^{(t+1)}[j,j] - \Lambda[p,-j]C_{\eta\eta}^{(t+1)}[-j,j]\right)$$

$$+\sum_{p^{'}\neq p}\psi^{pp^{'}}\left(C_{Y\eta}^{(t+1)}[p^{'},j]-\nu_{p'}e_{\eta}^{(t+1)}[j]-\Lambda[p^{'},]C_{\eta\eta}^{(t+1)}[,j]\right)$$
$$-\,c_q\frac{\partial\rho\left(\left|\lambda_{pj}\right|,\gamma\right)}{\partial t}\operatorname{sign}\left(\lambda_{pj}\right),$$

where c_q is the penalization indicator for λ_{pj}, and $\operatorname{sign}(t)=1\{t>0\}-1\{t<0\}$. With other parameters fixed at their newly updated values, the maximizer $\hat{\lambda}_{pj}^{(t+1)}(\gamma)>0$ should satisfy

$$\hat{\lambda}_{pj}^{(t+1)}(\gamma)+w_q^{(t*)}\frac{\partial\rho(|\hat{\lambda}_{pj}^{(t+1)}|,\gamma)}{\partial t}\operatorname{sign}\left(\hat{\lambda}_{pj}^{(t+1)}(\gamma)\right)=\hat{\lambda}_{pj}^{(t+1)},$$

where $w_q^{(t*)}=\frac{c_q}{\hat{\psi}^{pp(t*)}C_{\eta\eta}^{(t)}[j,j]}$ and $\hat{\lambda}_{pj}^{(t+1)}=\hat{\lambda}_{pj}^{(t+1)}(0)$ with the expression

$$\hat{\lambda}_{pj}^{(t+1)}=\frac{1}{C_{\eta\eta}^{(t)}[j,j]}\left[\left(C_{Y\eta}^{(t+1)}[p,j]-\hat{\nu}_p^{(t*)}e_{\eta}^{(t+1)}[j]-\hat{\Lambda}^{(t*)}[p,-j]C_{\eta\eta}^{(t+1)}[-j,j]\right)\right.$$
$$\left.+\sum_{p'\neq p}\frac{\hat{\psi}^{pp'(t*)}}{\hat{\psi}^{pp(t*)}}\left(C_{Y\eta}^{(t+1)}[p',j]-\hat{\nu}_{p'}^{(t*)}e_{\eta}^{(t+1)}[j]-\hat{\Lambda}^{(t*)}[p',]C_{\eta\eta}^{(t+1)}[,j]\right)\right].$$

The ℓ_1 penalty yields an updating scheme satisfying

$$\hat{\lambda}_{pj}^{(t+1)}(\gamma)=\hat{\lambda}_{pj}^{(t+1)}-w_q^{(t*)}\gamma\operatorname{sign}\left(\hat{\lambda}_{pj}^{(t+1)}(\gamma)\right).$$

Hence, the solution is the soft-thresholding rule

$$\hat{\lambda}_{pj}^{(t+1)}(\gamma)=S\left(\hat{\lambda}_{pj}^{(t+1)},w_q^{(t*)}\gamma\right),$$

where $S(\theta,\gamma)=\operatorname{sign}(\theta)\max\{|\theta|-\gamma,0\}$ (see Donoho & Johnstone, 1994). For the SCAD and the MCP, the updating schemes are relatively complex. The SCAD gives the updating rule

$$\hat{\lambda}_{pj}^{(t+1)}(\gamma)=\begin{cases}S\left(\hat{\lambda}_{pj}^{(t+1)},w_q^{(t*)}\gamma\right) & \text{if } \hat{\lambda}_{pj}^{(t+1)}\leq\gamma\left(1+w_q^{(t*)}\right),\\ \frac{S\left(\hat{\lambda}_{pj}^{(t+1)},w_q^{(t*)}\gamma\delta\big/(\delta-1)\right)}{1-w_q^{(t*)}\big/(\delta-1)} & \text{if } \gamma\left(1+w_q^{(t*)}\right)<\hat{\lambda}_{pj}^{(t+1)}\leq\gamma\delta,\\ \hat{\lambda}_{pj}^{(t+1)} & \text{if } \gamma\delta<\hat{\lambda}_{pj}^{(t+1)}.\end{cases}$$

The solution form can be obtained by setting the first derivative to be zero. The domain for each solution form can be solved based on the continuity of the derivative. Similarly, the MCP yields

$$\hat{\lambda}_{pj}^{(t+1)}(\gamma)=\begin{cases}\frac{S\left(\hat{\lambda}_{pj}^{(t+1)},w_q^{(t*)}\gamma\right)}{1-w_q^{(t*)}\big/\delta} & \text{if } \hat{\lambda}_{pj}^{(t+1)}\leq\gamma\delta,\\ \hat{\lambda}_{pj}^{(t+1)} & \text{if } \gamma\delta<\hat{\lambda}_{pj}^{(t+1)}.\end{cases}$$

The CM-step for β_{mj}, ν_p, or α_m is almost the same as this for λ_{pj} and can be found in Table 1.

Appendix C: The CM-Steps for Variance and Covariance Parameters

Consider the CM-steps for parameters in Ψ. Let $\mathrm{sp}(p) = \left\{k|\, \psi_{pk} \neq 0, k \in \{1, 2, \ldots, P\} \backslash p\right\}$ denote the set of spouses of p and $\epsilon_{-p} = \left(\epsilon_1, \ldots, \epsilon_{p-1}, \epsilon_{p+1}, \ldots, \epsilon_P\right)^{\mathrm{T}}$. Given p and $j \in \mathrm{sp}(p)$, the factorization $\mathbb{P}(\epsilon) = \mathbb{P}\left(\epsilon_p \middle| \epsilon_{-p}\right) \mathbb{P}\left(\epsilon_{-p}\right)$ shows that the conditional maximization of ψ_{pj} only involves the conditional distribution of ϵ_p, a normal distribution with mean $\sum_{k \in \mathrm{sp}(p)} \psi_{pk} \tilde{\epsilon}_{l(k)}(p)$ and variance $\tilde{\psi}_p^2$, where $\tilde{\epsilon}_{l(k)}(p)$ is the $l(k)$th element of $\Psi\left[-p, -p\right]^{-1} \epsilon_{-p}, l(k)$ is the column index of ψ_{pk} in $\Psi[p, -p]$, and $\tilde{\psi}_p^2 = \psi_p^2 - \Psi[p, -p] \Psi\left[-p, -p\right]^{-1} \Psi\left[-p, p\right]$. By taking a derivative, $\frac{\partial \mathcal{M}(\theta | \hat{\theta}^{(t)})}{\partial \psi_{pj}} = \frac{1}{\tilde{\psi}_p^2} \left(u_{l(j)}^{(t*)}(p) - \psi_{pj} v_{l(j)l(j)}^{(t*)}(p) - \sum_{k \in \mathrm{sp}(p) \backslash j} \psi_{pk} v_{l(j)l(k)}^{(t*)}(p)\right) - c_q \frac{\partial \rho(|\psi_{pj}|, \gamma)}{\partial t} \mathrm{sign}\left(\psi_{pj}\right)$, where $u_{l(j)}^{(t*)}(p)$ is the $l(j)$th element of $\hat{\Psi}^{(t*)}\left[-p, -p\right]^{-1} C_{\epsilon\epsilon}^{(t+1)}\left[-p, p\right]$, $v_{jk}^{(t*)}(p)$ is the $(l(j), l(k))$ element of $\hat{\Psi}^{(t*)}\left[-p, -p\right]^{-1} C_{\epsilon\epsilon}^{(t+1)}\left[-p, -p\right] \hat{\Psi}^{(t*)}\left[-p, -p\right]^{-1}$, and

$$\begin{aligned} C_{\epsilon\epsilon}^{(t+1)} = {} & C_{YY} - e_Y \hat{\nu}^{(t*)^{\mathrm{T}}} - C_{Y\eta}^{(t+1)} \hat{\Lambda}^{(t*)^{\mathrm{T}}} - \hat{\nu}^{(t*)} e_Y^{\mathrm{T}} + \hat{\nu}^{(t*)} \hat{\nu}^{(t*)^{\mathrm{T}}} + \hat{\nu}^{(t*)} e_\eta^{(t+1)^{\mathrm{T}}} \hat{\Lambda}^{(t*)^{\mathrm{T}}} \\ & - \hat{\Lambda}^{(t*)} C_{Y\eta}^{(t+1)^{\mathrm{T}}} + \hat{\Lambda}^{(t*)} e_\eta^{(t+1)} \hat{\nu}^{(t*)^{\mathrm{T}}} + \hat{\Lambda}^{(t*)} C_{\eta\eta}^{(t+1)} \hat{\Lambda}^{(t*)^{\mathrm{T}}}. \end{aligned}$$

Hence, under $\gamma = 0$, ψ_{pj} can be updated by

$$\hat{\psi}_{pj}^{(t+1)} = \left(u_{l(j)}^{(t*)}(p) - \sum_{k \in \mathrm{sp}(p) \backslash j} \hat{\psi}_{pk}^{(t*)} v_{l(j)l(k)}^{(t*)}(p)\right) / v_{l(j)l(j)}^{(t*)}(p).$$

If $\gamma > 0$, $\hat{\psi}_{pj}^{(t+1)}(\gamma)$ can be obtained through the shrinkage formula in Table 1 with $w_q^{(t*)} = \frac{c_q}{\hat{\tilde{\psi}}_p^{2(t*)} v_{l(j)l(j)}^{(t*)}(p)}$. Using a similar technique, we obtain

$$\hat{\tilde{\psi}}_p^{2^{(t+1)}} = \left(C_{\epsilon\epsilon}^{(t+1)}\left[p, p\right] - 2 \sum_{k \in \mathrm{sp}(p)} \hat{\psi}_{pk}^{(t*)} u_{l(k)}^{(t*)}(p) + \sum_{k \in \mathrm{sp}(p)} \sum_{j \in \mathrm{sp}(p)} \hat{\psi}_{pk}^{(t*)} \hat{\psi}_{pj}^{(t*)} v_{l(k)l(j)}^{(t*)}(p)\right),$$

and then

$$\hat{\psi}_p^{2^{(t+1)}} = \hat{\tilde{\psi}}_p^{2^{(t+1)}} + \hat{\Psi}^{(t*)}[p, -p] \hat{\Psi}^{(t*)}\left[-p, -p\right]^{-1} \hat{\Psi}^{(t*)}\left[-p, p\right].$$

The CM-steps for parameters in Φ can be derived similarly by replacing ϵ, Ψ, and $C_{\epsilon\epsilon}^{(t+1)}$ with ζ, Φ, and $C_{\zeta\zeta}^{(t+1)}$ respectively, where

$$\begin{aligned} C_{\zeta\zeta}^{(t+1)} &= C_{\eta\eta}^{(t+1)} - e_{\eta}^{(t+1)}\hat{\alpha}^{(t*)^{\mathrm{T}}} - C_{\eta\eta}^{(t+1)}\hat{\mathrm{B}}^{(t*)^{\mathrm{T}}} - \hat{\alpha}^{(t*)}e_{\eta}^{(t)^{\mathrm{T}}} + \hat{\alpha}^{(t*)}\hat{\alpha}^{(t*)^{\mathrm{T}}} \\ &\quad + \hat{\alpha}^{(t*)}e_{\eta}^{(t+1)^{\mathrm{T}}}\hat{\mathrm{B}}^{(t*)^{\mathrm{T}}} - \hat{\mathrm{B}}^{(t*)}C_{\eta\eta}^{(t+1)} + \hat{\mathrm{B}}^{(t*)}e_{\eta}^{(t+1)}\hat{\alpha}^{(t*)^{\mathrm{T}}} + \hat{\mathrm{B}}^{(t*)}C_{\eta\eta}^{(t+1)}\hat{\mathrm{B}}^{(t*)^{\mathrm{T}}}. \end{aligned}$$

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397–438.

Arminger, G., & Schoenberg, R. (1989). Pseudo maximum likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika*, *54*, 409–426.

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, *13*, 27–45.

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin*, *106*, 315–317.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, *5*, 232–253.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*, 2350–2383.

Browne, M. W. (1972). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, *25*, 207–212.

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, *24*, 445–455.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high dimensional data: Methods, theory and applications*. Heidelberg, Berlin: Springer.

Chang, J. H., Lin, Y. C., & Huang, C. L. (2010). Exploring the mechanism of mindfulness: From attention to self-integration. In *11th annual meeting of the society for personality and social psychology*, Las Vegas, USA.

Chaudhuri, S., Drton, M., & Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, *94*, 199–216.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*, 850–866.

Choi, J., Zou, H., & Oehlert, G. (2011). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface*, *3*, 429–436.

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structural modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research*, *25*, 115–136.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*, 147–167.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance-structures analysis and the "problem" of sample size—A clarification. *Psychological Bulletin*, *109*, 512–519.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Donoho, D. L., & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*, 425–455.

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

Fan, J.-Q., & Lv, J.-C. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE - Information Theory*, *57*, 5467–5484.

Fan, Y.-Y., & Li, R.-Z. (2012). Variable selection in linear mixed effects models. *Annals of Statistics*, *40*, 2043–2068.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*, 928–961.

Friedman, J., Hastie, H., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*, 302–332.

Garcia, R. I., Ibrahim, J. G., & Zhu, H.-T. (2010). Variable selection for regression models with missing covariate data. *Statistica Sinica*, *20*, 149–165.

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by ℓ_1-penalized estimation. *Statistics and Computing*, *24*, 137–154.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York, NY: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. London: CRC Press.

Hirose, K., & Yamamoto, M. (2014). Estimation of oblique structure via penalized likelihood factor analysis. *Computational Statistics and Data Analysis*, *79*, 120–132.

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in a factor analysis model. *Statistics and Computing*, *25*, 863–875.

Holzinger, K., & Swineford, F. (1939). *A study in factor analysis: The stability of a bifactor solution.* Supplementary Educational Monograph, no. 48. Chicago: University of Chicago Press.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression-biased estimation for nonorthogonal problems. *Technometrics*, *42*, 80–86.

Huang, P.-H. (2015). lsl: Latent Structure Learning. *R package version 0.5.0*.

Ibrahim, J. G., Zhu, H.-T., Garcia, R. I., & Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, *67*, 495–503.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Jung, S. (2012). Structural equation modeling with small sample sizes using two-stage ridge least-squares estimation. *Behavior Research Methods*, *45*, 75–81.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*, 69–86.

Kwon, S., & Kim, Y. (2012). Large sample properties of the SCAD-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*, *22*, 629–653.

Lee, S. Y., & Zhu, H. T. (2002). Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika*, *67*, 189–210.

Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, *34*, 2554–2591.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107–120.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113–139.

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110.

Mazumder, R., Friedman, J., & Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, *106*, 1125–1138.

McDonald, R. P. (1982). A note on the investigation of local and global identifiability. *Psychometrika*, *47*, 101–103.

Meng, X.-L. (2008). Discussion: one-step sparse estimates in nonconcave penalized likelihood models: Who cares if it is a white cat or a black cat? *Annals of Statistics*, *36*, 1542–1552.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*, 267–278.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Ning, L., & Georgiou, T. T. (2011). Sparse factor analysis via likelihood and ℓ_1 regularization. In *50th IEEE conference on decision and control and european control conference* (pp. 5188–5192).

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*, 227–259.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *19*, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*, 221–242.

Shapiro, A., & Browne, M. W. (1983). On the investigation of local identifiability—A counterexample. *Psychometrika*, *48*, 303–304.

Schelldorfer, J., Bühlmann, P., & van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *The Scandinavian Journal of Statistics*, *38*, 197–214.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of Royal Statistical Society, Series B*, *36*, 111–147.

Strawderman, R. L., Wells, M. T., & Schifano, E. D. (2013). Hierarchical Bayes, maximum a posteriori estimators, and minimax concave penalized likelihood estimation. *Electronic Journal of Statistics*, *7*, 973–990.

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

Trendafilov, N. T., & Adachi, K. (2015). Sparse versus simple structure loadings. *Psychometrika*, *80*, 776–790.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*, 21–43.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228–243.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070.

Yuan, K.-H., Marshall, L. L., & Bentler, E. M. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology*, *33*, 241–265.

Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *The British Journal of Mathematical and Statistical Psychology*, *59*, 397–417.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.

Zhang, Y.-Y., Li, R.-Z., & Tsai, C.-L. (2012). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*, 312–323.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*, 265–286.

Manuscript Received: 28 NOV 2013
Final Version Received: 6 DEC 2016
Published Online Date: 17 APR 2017