PSYCHOMETRIKA—VOL. 83, NO. 3, 628–649
SEPTEMBER 2018
https://doi.org/10.1007/s11336-018-9623-z

APPROXIMATED PENALIZED MAXIMUM LIKELIHOOD FOR EXPLORATORY FACTOR ANALYSIS: AN ORTHOGONAL CASE

SHAOBO JIN

UPPSALA UNIVERSITY

IRINI MOUSTAKI

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

FAN YANG- WALLENTIN

UPPSALA UNIVERSITY

The problem of penalized maximum likelihood (PML) for an exploratory factor analysis (EFA) model is studied in this paper. An EFA model is typically estimated using maximum likelihood and then the estimated loading matrix is rotated to obtain a sparse representation. Penalized maximum likelihood simultaneously fits the EFA model and produces a sparse loading matrix. To overcome some of the computational drawbacks of PML, an approximation to PML is proposed in this paper. It is further applied to an empirical dataset for illustration. A simulation study shows that the approximation naturally produces a sparse loading matrix and more accurately estimates the factor loadings and the covariance matrix, in the sense of having a lower mean squared error than factor rotations, under various conditions.

Key words: factor rotation, LASSO, SCAD, MCP, sparsity, shrinkage.

1. Introduction

Exploratory factor analysis (EFA) is a multi-step method that explains the associations among observed variables in terms of unobserved constructs known as latent variables or factors. Maximum likelihood (ML, e.g., Jöreskog, 1967; Lawley, 1940) is commonly used to estimate the parameters of the EFA model. The classical factor analysis model assumes that factors are orthogonal, but then the estimated factor loading matrix is rotated using an oblique or orthogonal rotation to produce a sparse or simple structure matrix. Some commonly used rotation methods are the varimax rotation (Kaiser, 1958), the quartimax rotation (Neuhaus & Wrigley, 1954), and the quartimin rotation (Carroll, 1953). Browne (2001) provides a review on rotations for EFA. For some recent developments, see Jennrich (2004, 2006).

It is often desirable that the rotated loading matrix is a sparse matrix with a few large loadings and many small loadings. However, the small loadings produced by common rotation methods are generally nonzero which complicates the interpretation of the results. In order to produce a sparse factor structure, researchers need to decide which loadings can be regarded as zeros. In practice, researchers commonly discard loadings smaller than 0.3 or 0.32 (Hair et al., 2010; Tabachnick & Fidell, 2001); and less often use inference tools (hypothesis testing and confidence intervals) (Zhang, 2014) to decide which factor loadings are nonzero. Both methods are useful tools for choosing an interpretable factor solution. However, adopting a hard-thresholding approach in

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-018-9623-z) contains supplementary material, which is available to authorized users.

Correspondence should be made to Shaobo Jin, Department of Statistics, Uppsala University, Uppsala, Sweden. Email: shaobo.jin@statistik.uu.se

which loadings passing a threshold value are considered to be significant requires a subjectively choice of the threshold value and a minor change in the threshold value can possibly lead to a large change in the loading matrix. The same applies to statistical tests and confidence intervals that rely on significance levels and tend to over reject when sample size is large. Such a large change is not embraced by Fan and Li (2001), due to its instability in model prediction.

The main contribution of this paper is to introduce a soft-thresholding approach for orthogonal EFA that continuously shrinks a coefficient toward zero. As a result, if the estimated loading matrix itself consists of zero elements, the nonzero loadings can be regarded as "significant" loadings and no subjective hard-thresholding needs to be applied. A continuous soft-thresholding approach yields a continuous model in the parameter θ. That is, a minor change in θ will not cause a dramatic change in model parameters and model interpretation. As we will discuss later, the continuous thresholding approach is achieved by introducing additional tuning parameters. Consequently, the factor loading estimates are functions of the tuning parameters and can be plotted against the tuning parameters. The plot of the factor loadings against the tuning parameters is referred to as the solution path. The solution path offers the opportunity to study and select among many factor solutions. Creating the solution path is fully computational. In general, methods such as information criteria and cross-validation can be used to choose the optimal tuning parameter. In this way, choosing the optimal tuning parameter is viewed from a model selection perspective and is data-driven. However, the most interpretable factor solution still needs to be picked up by the researcher and cannot be automated.

Similarly to the method discussed above, variable selection aims to detect zero regression coefficients. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) for the linear regression model; it simultaneously performs model estimation and variable selection. The idea of LASSO has been generalized to other types of penalties. See Tibshirani (2011) for a review on the variants of the LASSO. In addition to the linear regression problem, some studies also consider penalized maximum likelihood (PML) estimation in generalized linear models (e.g., Fan & Li, 2001; Zou, 2006) and areas outside of the regression analysis. A field that is closely related with the current work is sparse principal component analysis (PCA). A rotated PCA solution produces typically small but still nonzero loadings, which makes the interpretation of the principal components difficult. Similar to the above-mentioned EFA loadings, an ad hoc but problematic way is to hard-threshold the component loadings. Various alternative methods to the subjective hard-thresholding have been proposed in order to produce a sparse loading matrix. See among others Lu (2012), Shen and Huang (2008), and Witten et al. (2009). The reader is directed to Trendafilov (2014) for a review on sparse PCA. In particular, the LASSO-type penalty has been applied by Zou et al. (2006) and Trendafilov and Adachi (2015) to produce a sparse loading matrix.

Although penalized estimation has been extensively used and developed in linear regression, generalized linear models, and in principal component analysis, studies on penalized EFA are still underdeveloped. Choi et al. (2010) introduced a sparse EFA by incorporating a LASSO penalty in the log-likelihood function. The same problem was also studied in Ning and Georgiou (2011) in which a perturbed approximation is used to handle the LASSO penalty. Hirose and Konishi (2012) proposed a variable selection procedure via weighted group LASSO. To reduce bias introduced by the LASSO, Hirose and Yamamoto (2014, 2015) considered the minimax concave penalty with plus algorithm (Zhang, 2010). Recently, Trendafilov et al. (2017) proposed to penalize the reparametrized loading matrix. An EM-type algorithm is proposed by Choi et al. (2010) to produce exact solutions. However, using the EM algorithm to produce a PML estimator can be computationally intensive, especially when the number of observed variables is large. To improve the computational efficiency, we propose an approximated penalized maximum likelihood (APML), motivated by Zou and Li (2008), which yields approximated solutions. As we will discuss later, APML overcomes some disadvantages of the EM-type algorithm and naturally produces a sparse loading matrix.

The rest of the paper is organized as follows. First, we review PML with the EM algorithm for orthogonal EFA. Second, APML is introduced. After that, some practical issues are discussed, followed by an empirical example. A simulation study is conducted to study the performance of the proposed APML method.

2. Penalized Maximum Likelihood

Let us consider the linear factor analysis model given by

$$\boldsymbol{y} = \boldsymbol{\Lambda} \boldsymbol{f} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is a $p \times 1$ vector of observed variables, $\boldsymbol{\Lambda}$ is a $p \times m$ loading matrix with the (i, j)th element λ_{ij}, $\boldsymbol{f}$ is an $m \times 1$ vector of factors, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$, and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix with diagonal elements ψ_i for $i = 1, 2, \ldots, p$. The common factor $\boldsymbol{f}$ is assumed to be normally distributed with mean $\mathbf{0}$ and correlation matrix $\boldsymbol{\Phi}$. Hence, $\boldsymbol{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. If $\boldsymbol{f}$ follows an orthogonal structure then $\boldsymbol{\Phi}$ is equal to the identity matrix $\boldsymbol{I}$. The ML estimator minimizes the fit function given by

$$\frac{n}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \frac{n}{2} \mathrm{tr}\left[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\right],$$

where $\boldsymbol{\theta}$ contains all parameters in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, n is the sample size, and $\boldsymbol{S}$ is the sample covariance matrix. In the orthogonal EFA, $\boldsymbol{\theta}$ is a $p(m+1) \times 1$ vector. Note that not all elements in $\boldsymbol{\theta}$ are free parameters, due to rotational indeterminacy. As we will explain later, PML removes such indeterminacy. For the purpose of presentation and being consistent with PML, $\boldsymbol{\theta}$ is used for EFA without penalization.

In the current study, the PML estimator minimizes

$$\frac{n}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \frac{n}{2} \mathrm{tr}\left[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\right] + n \sum_{i=1}^{p} \sum_{j=1}^{m} P\left(|\lambda_{ij}|; \boldsymbol{\beta}, \boldsymbol{w}\right), \tag{1}$$

where P is a scalar-valued function, $\boldsymbol{\beta}$ is a scalar/vector of tuning parameters, and $\boldsymbol{w}$ is a vector of possible weights on the factor loadings. Some typical examples of the penalty term with equal weights are

- LASSO (Tibshirani, 1996): $P\left(|\lambda_{ij}|; \boldsymbol{\beta}, \boldsymbol{w}\right) = \beta|\lambda_{ij}|$, where $\boldsymbol{\beta} = \beta > 0$;
- Smoothly clipped absolute deviation (SCAD, Fan & Li, 2001):

$$P\left(|\lambda_{ij}|; \boldsymbol{\beta}, \boldsymbol{w}\right) = \int_0^{|\lambda_{ij}|} I(x \leq \beta_1) + \frac{\max(\beta_1\beta_2 - x, 0)\, I(x > \beta_1)}{\beta_1\beta_2 - \beta_1} \mathrm{d}x,$$

  where $\boldsymbol{\beta} = (\beta_1, \beta_2)$ with $\beta_1 > 0$ and $\beta_2 > 2$;
- Minimax concave penalty with plus algorithm (MCP, Zhang, 2010):

$$P\left(|\lambda_{ij}|; \boldsymbol{\beta}, \boldsymbol{w}\right) = \beta_1 \int_0^{|\lambda_{ij}|} \max\left(1 - \frac{x}{\beta_1\beta_2}, 0\right) \mathrm{d}x,$$

  where $\boldsymbol{\beta} = (\beta_1, \beta_2) > 0$.

Fan and Li (2001) proposed that a good penalty term should at least satisfy three properties: unbiasedness (the estimator should be nearly unbiased for parameters with large values), sparsity (the estimator automatically shrinks small estimated parameters to zero), and continuity (the estimator continuous shrinks parameters to zero). Although the LASSO penalty does not satisfy all three properties (Fan & Li, 2001), it is still widely used for its simplicity and computational efficiency. In contrast, the SCAD and the MCP penalties possess the above three properties and often improve the performance of the LASSO penalty. The price to pay is an additional tuning parameter.

2.1. Consequences of a Penalty Term

2.1.1. Shrinkage Inclusion of the penalty term shrinks the ML estimators $\lambda_{ij}^{(\mathrm{MLE})}$ toward zero. For example, PML with the LASSO penalty is equivalent to

$$\min_{\boldsymbol{\theta}} \left\{ \frac{n}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \frac{n}{2} \mathrm{tr}\left[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \right] \right\}, \quad \text{s.t.} \frac{\sum_{i=1}^{p} \sum_{j=1}^{m} |\lambda_{ij}|}{\sum_{i=1}^{p} \sum_{j=1}^{m} |\lambda_{ij}^{(\mathrm{MLE})}|} \leq t,$$

for some $t > 0$ (Osborne et al., 2000), the sum of estimated factor loadings should not be too large. For the SCAD and the MCP penalty, the sum of nonlinear functions of the estimated factor loadings should not be too large. Thus, the ordinary ML fit function is minimized subject to a constrain on the factor loadings, which makes the LASSO estimator biased. As shown by Fan and Li (2001) and Zhang (2010), the nonlinear functions in the SCAD and the MCP penalty neutralize the bias introduced by the LASSO penalty for the large parameters. Browne and Du Toit (1992) proposed an automated estimation method to incorporate constraints on the parameters. To be more specific, the constraints proposed in Browne and Du Toit (1992) are of the type $c(\boldsymbol{\theta}) = 0$ or $c(\boldsymbol{\theta}) > 0$, where $c(\boldsymbol{\theta})$ is a continuously differentiable function and the Gauss-Newton method (see, e.g., Björck,1996) is used to minimize the fit function under constraints. A fundamental difference between PML and the method in Browne and Du Toit (1992) is that constraints in PML are often nondifferentiable, whereas constrains in Browne and Du Toit (1992) are continuously differentiable. For example, $|\lambda_{ij}|$ involved in the LASSO, SCAD, and MCP is not differentiable and, consequently, the Gauss-Newton method is not applicable.

2.1.2. Sparsity For all aforementioned penalty terms, the coefficient of a parameter θ may be shrunk to zero. Thus, sparsity and estimation are conducted simultaneously. Furthermore, if a column in the loading matrix consists only of zero loadings, no variance is explained by the corresponding factor and the number of factors is reduced.

2.1.3. Indeterminacy A nondifferentiable penalty term removes rotation indeterminacy. Take the LASSO penalty as an example. The PML estimator with the LASSO penalty minimizes

$$\frac{n}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \frac{n}{2} \mathrm{tr}\left[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \right] + n\beta \sum_{i=1}^{p} \sum_{j=1}^{m} |\lambda_{ij}| . \tag{2}$$

Orthogonal rotation imposes $m(m-1)/2$ equality constrains and keeps $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$ invariant because of the quadratic terms in the cross-product. However, most nonsingular and orthogonal transformations change the value of $\sum_{i=1}^{p} \sum_{j=1}^{m} |\lambda_{ij}|$. Subsequently, PML is rotation-free but still produces a sparse loading matrix. However, a PML estimator is still invariant in terms of permutations of columns and changes of the sign of the entire column.

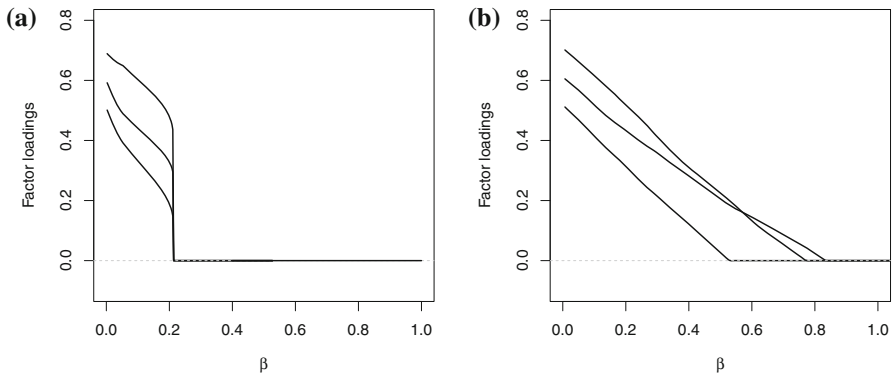

FIGURE 1.
Solution path of some factor loadings of the EM algorithm and the approximated penalized maximum likelihood (APML) with the LASSO penalty. **a** EM algorithm. **b** APML.

2.1.4. Solution Path Fit function (1) is minimized for a fixed $\boldsymbol{\beta}$. Take the fit function (2) as an example. The PML estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, depends on the tuning parameter β. A larger β typically yields a smaller $\hat{\boldsymbol{\theta}}$. Varying β potentially changes the value of $\hat{\theta}_i$, the ith element in $\hat{\boldsymbol{\theta}}$, for all i. $\hat{\theta}_i$ can then be understood as $\hat{\theta}_i(\beta)$, a function of β. Thus, all $\hat{\theta}_i$ can be plotted against β to reflect the change in the values of $\hat{\theta}_i$ as we change β. Such a plot is referred to as a solution path. As we shall illustrate in Sect. 5.1.2, the solution path allows us to identify the position of zero loadings and to extract all possible loading structures suggested by the data. Various penalty terms produce their own solution paths. Examples of a simple linear regression model can be found in Zou (2006).

2.2. EM Algorithm

Choi et al. (2010) proposed an EM-type algorithm to minimize Eq. (2) in the spirit of Rubin and Thayer (1982). A similar EM-type algorithm is implemented in Hirose and Yamamoto (2014, 2015) for the MCP penalty and oblique EFA. See also Garcia et al. (2010) for a more general algorithm that incorporates missing data in regression models. The reader is directed to the above-cited works for a detailed explanation of the EM-type algorithm.

In our experience, the EM-type has several limitations. First, the algorithm is a combination of an EM algorithm and a penalized least squares minimization problem. Hence, the solution path as a function of tuning parameters is inefficiently constructed. In the linear regression model with the LASSO penalty, for example, the solution path can be constructed at the same computational cost of ordinary least squares (Efron et al., 2004). Second, the solution path is not necessarily smooth. For example, Fig. 1a depicts solution paths for some factor loadings with the LASSO penalty using the EM algorithm. It is seen that the solution path is not smooth where several jumps occur for large loadings.

3. Approximate PML Estimator

3.1. Penalized Least Squares

To overcome the above-mentioned limitations, approximated penalized maximum likelihood (APML) can be efficiently deployed using the technique introduced in Zou and Li (2008), which applies to a general penalized likelihood problem. If the ML estimator $\hat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}$, the first two terms of Eq. (1) can be Taylor-expanded as

$$\frac{n}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})|+\frac{n}{2}\mathrm{tr}\left[\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\right]\approx F\left(\hat{\boldsymbol{\theta}}\right)+\frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^{T}\frac{\partial^{2}F\left(\hat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right), \tag{3}$$

where the notation $F(\boldsymbol{\theta})$ is used to denote the left-hand side of Eq. (3). Hence, fit function (1) after replacing its first two terms by their quadratic approximation is approximated by a penalized weighted least squares fit function

$$\frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^{T}\frac{\partial^{2}F\left(\hat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)+n\sum_{i=1}^{p}\sum_{j=1}^{m}P\left(|\lambda_{ij}|;\boldsymbol{\beta},\boldsymbol{w}\right). \tag{4}$$

Many efficient algorithms have been proposed for penalized ordinary least squares. Hence, rewriting the fit function (4) as an ordinary least squares problem can help to efficiently minimize it. Note that the Hessian matrix, evaluated at the ML estimates $\hat{\boldsymbol{\theta}}$, no longer involves unknown parameters. It can be diagonalized as follows:

$$\frac{\partial^{2}F\left(\hat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}=n\boldsymbol{P}^{T}\boldsymbol{D}\boldsymbol{P},$$

where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{D}$ is a diagonal matrix with nonnegative elements. Define $\tilde{\boldsymbol{y}}=\boldsymbol{D}^{1/2}\boldsymbol{P}\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{X}}=\boldsymbol{D}^{1/2}\boldsymbol{P}$. The ordinary least squares function to be minimized is

$$\frac{1}{2}\left(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\theta}\right)^{T}\left(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\theta}\right)+P\left(|\boldsymbol{\Lambda}|;\boldsymbol{\beta},\boldsymbol{w}\right). \tag{5}$$

3.2. Algorithms

The algorithm used to minimize the fit function (5) depends on the form of its penalty term. For commonly used penalty terms, efficient algorithms exist. For the LASSO penalty, the fit function (5) can be efficiently minimized using least angle regression (Efron et al., 2004) or coordinate descent (Friedman et al., 2007, 2010). For the MCP penalty, the fit function (5) can be solved efficiently by the plus algorithm (Zhang, 2010) or by the coordinate descent algorithms (Brehenv & Huang, 2011; Mazumder et al., 2011). For more details on the algorithms we refer the reader to the above-mentioned references.

For other penalties without exact algorithms, Zou and Li (2008) proposed Taylor-expanding the likelihood function and applying a local linear approximation to the penalty term, in which the penalty term is approximated by

$$P\left(|\lambda_{ij}|;\boldsymbol{\beta},\boldsymbol{w}\right)\approx P\left(|\lambda_{ij,0}|;\boldsymbol{\beta},\boldsymbol{w}\right)+P'\left(|\lambda_{ij,0}|;\boldsymbol{\beta},\boldsymbol{w}\right)\left(|\lambda_{ij}|-|\lambda_{ij,0}|\right),$$

where $P'()$ is the first-order derivative of $P()$. Consequently, the fit function is approximated by a weighted penalized least squares:

$$\frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^{T}\frac{\partial^{2}F\left(\hat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)+n\sum_{i=1}^{p}\sum_{j=1}^{m}P'\left(|\lambda_{ij,0}|;\boldsymbol{\beta},\boldsymbol{w}\right)|\lambda_{ij}|,$$

which can be further refined to

$$\frac{1}{2}\left(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\theta}\right)^{T}\left(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\theta}\right)+n\sum_{i=1}^{p}\sum_{j=1}^{m}P'\left(|\lambda_{ij,0}|;\boldsymbol{\beta},\boldsymbol{w}\right)|\lambda_{ij}|, \tag{6}$$

by decomposing the Hessian matrix. For example, the SCAD penalty term can be approximated by

$$P'\left(|\lambda_{ij,0}|;\boldsymbol{\beta},\boldsymbol{w}\right)|\lambda_{ij}|=\beta_1\left[I\left(|\lambda_{ij,0}|\leq\beta_1\right)+\frac{\max\left(\beta_1\beta_2-x,0\right)I\left(|\lambda_{ij,0}|>\beta_1\right)}{\beta_1\beta_2-\beta_1}\right]|\lambda_{ij}|$$

(Fan & Li, 2001). The fit function (6) is essentially the fit function of an adaptive LASSO (Zou, 2006). The adaptive LASSO can be solved efficiently by the algorithm in Zou (2006) or coordinate descent. For alternatives of Zou and Li (2008), see Fan and Li (2001) and Hunter and Li (2005) for local quadratic approximations.

In the current study, the above-mentioned coordinate descent algorithm is used to minimize the fit functions (5) and (6), where the parameters in $\boldsymbol{\theta}$ are estimated one at a time until the change in the parameter estimates is sufficiently small. Let $\tilde{\boldsymbol{X}}_k$ and $\tilde{\boldsymbol{X}}_{-k}$ be kth column of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{X}}$ excluding the kth column, respectively. After t iterations, $\theta_k^{(t+1)}$ is updated conditional on $\boldsymbol{\theta}_{-k}^{(t)}$, where θ_k is the kth element in $\boldsymbol{\theta}$, $\theta_k^{(t+1)}$ is the value of θ_k after $t+1$ iterations, $\boldsymbol{\theta}_{-k}$ is the vector by excluding θ_k from $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_{-k}^{(t)}=\left(\theta_1^{(t+1)},\cdots,\theta_{k-1}^{(t+1)},\theta_{k+1}^{(t)},\cdots,\theta_q^{(t)}\right)$ if $\boldsymbol{\theta}$ is a vector of size $q\times 1$. Following Mazumder et al. (2011), $\theta_k^{(t+1)}$ can be updated in closed forms. If the LASSO or the SCAD penalty is used, the solution of the fit function (5) satisfies

$$\theta_k^{(t+1)}=\begin{cases}\operatorname{sgn}\left(z^{(t)}\right)\left(\left|z^{(t)}\right|-\frac{\beta_1 w_k}{\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k}\right), & \text{if } \frac{\beta_1 w_k}{\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k}<\left|z^{(t)}\right|,\\ 0, & \text{otherwise.}\end{cases}$$

where $\operatorname{sgn}(\cdot)$ returns the sign of the enclosed value,

$$z^{(t)}=\frac{\left(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}_{-k}\boldsymbol{\theta}_{-k}^{(t)}\right)^{T}\tilde{\boldsymbol{X}}_k}{\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k},$$

and $w_k=1$ if the penalty is the LASSO and

$$w_k=I\left(|\hat{\lambda}_{ij}|\leq\beta_1\right)+\frac{\max\left(\beta_1\beta_2-x,0\right)I\left(|\hat{\lambda}_{ij}|>\beta_1\right)}{\beta_1\beta_2-\beta_1}$$

if the penalty is the SCAD. If the MCP penalty is used, the solution of the fit function (5) is

$$\theta_k^{(t+1)}=\begin{cases}\operatorname{sgn}\left(z^{(t)}\right)\left(\frac{|z^{(t)}|-\frac{\beta_1}{\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k}}{1-1/\beta_2}\right), & \text{if } \frac{\beta_1}{\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k}<\left|z^{(t)}\right|<\frac{\beta_1\beta_2}{\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k},\\ z^{(t)}, & \text{if } \left|z^{(t)}\right|\geq\frac{\beta_1\beta_2}{\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k},\\ 0, & \text{otherwise.}\end{cases}$$

At the $t+1$th iteration, all parameters are cycled through, one at a time, such that $\boldsymbol{\theta}^{(t)}$ is updated to $\boldsymbol{\theta}^{(t+1)}$. As suggested by Mazumder et al. (2011), we start with the largest β_1 and gradually decrease the β_1 value. By doing so, the solution with a larger β_1 is used as the starting value for a smaller β_1.

Compared to the solution path produced by the EM algorithm, the solution path of APML with the LASSO penalty is piecewise linear and no jumps appear in it (Fig. 1b).

4. Practical Considerations

4.1. Tuning Parameters

The tuning parameter $\boldsymbol{\beta}$ is critical in penalization. A larger tuning parameter typically produces a sparser loading matrix and a larger bias, whereas a smaller tuning parameter yields a denser loading matrix and a lower bias. The effect of the tuning parameter is studied in Sect. 5.1. A grid search can be implemented to select the tuning parameter $\boldsymbol{\beta}$. If $\boldsymbol{\beta}$ is a scalar β, as in the LASSO penalty, then a large number of β's, say 100 or 200, is selected. If $\boldsymbol{\beta}$ is a vector $\boldsymbol{\beta} = (\beta_1, \beta_2)$ as in the MCP penalty, then a small number of β_2's is selected and, for every β_2, a large number of β_1s is selected. The optimal tuning parameter can be selected via AIC or BIC. Based on Zou, Hastie, and Tibshirani (2007), Hirose and Yamamoto (2015) introduced the AIC and BIC type criteria

$$\text{AIC} = -2\left\{-\frac{n}{2}\left[\log|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})| + \text{tr}\left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{S}\right)\right]\right\} + 2\,(d+p)\,,$$
$$\text{BIC} = -2\left\{-\frac{n}{2}\left[\log|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})| + \text{tr}\left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{S}\right)\right]\right\} + (\log n)\,(d+p)\,,$$

where d is the number of degrees of freedom. An alternative measure is the mean squared error (MSE) computed using the raw covariance residuals given by

$$\text{MSE} = \sum_i \sum_j \left(\hat{\sigma}_{ij} - s_{ij}\right)^2,$$

where $\hat{\sigma}_{ij}$ is the (i, j)th element of $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ and s_{ij} is the (i, j)th element of the sample covariance matrix. The tuning parameter with the smallest MSE is selected. Finally, following Choi et al. (2010), the chosen tuning parameter minimizes the Kullback–Leibler (K–L) divergence

$$\text{KL} = \frac{1}{2}\log|\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)| + \frac{1}{2}\text{tr}\left[\mathbf{S}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)^{-1}\right] - \frac{1}{2}\log|\mathbf{S}| - \frac{p}{2}\,.$$

The selection methods discussed above rely on a numerical criterion and therefore are analytical methods and fully data-driven. The tuning parameter controls the sparsity of the loading matrix. A larger $\boldsymbol{\beta}$ generally shrinks the factor loadings more and leads to more zero loadings. However, an analytical selection method is not guaranteed to produce an interpretable solution. Furthermore, researchers often want to incorporate their prior knowledge into the analysis. Simply relying on the analytical selection criteria may fail to meet such a need. In this case, a nonanalytical subjective selection can be based on the solution path. An entire solution path contains different loading matrices produced by APML, from a more dense to a more sparse model. Since the proposed method produces all loading structures from the sequence of $\boldsymbol{\beta}$, it is possible to extract all unique loadings structures. The loading structure with the best interpretability can then

be selected subjectively. If the estimates of the factor loadings are also of interest, the LARS-OLS hybrid (Efron et al., 2004) or the relaxed LASSO (Meinshausen, 2007) may be used to reduce the bias created by penalization based on the selected loading structure.

Further, EFA is often applied to large datasets with large loadings matrices. The routinely used estimation-rotation approach is likely to produce small loadings, regardless whether rotation is done toward a target structure. The most convenient way is perhaps to rely on hypotheses tests or confidence intervals. However, inferential tools are not flawless as the choice of the significance level may yield a dramatic change in the loading matrix. In such a case, APML serves as an alternative either in a data-driven or in a subjective way. In the data-driven manner, the optimal tuning parameter can be selected by the analytical criteria. In the subjective manner, the unique loading structures are much fewer than 200 even though 200 tuning parameters are examined, which makes APML still user-friendly. The most interpretable loading structure can be chosen from the extracted unique loading structures.

4.2. Starting Point of the Solution Path

The ML estimator $\hat{\boldsymbol{\theta}}$ in Eq. (3) is not uniquely defined because of rotation indeterminacy. Hence, different choices of $\hat{\boldsymbol{\theta}}$ yield different $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}}$, and further influence the solution path. A natural starting point would be the ML estimator after some commonly used rotation. Starting from a rotated solution can be understood as the APML continuously shrinking nonzero loadings to zero loadings without setting a cutoff value. Thus, the proposed approach is similar to a three-stage approach. In the first two stages, the factor loadings are estimated and then rotated similarly to the classical EFA and then the zero loadings are determined in the third stage using APML. The APML stage refines the classical ML estimates and determines the zero loading positions without the use of a hard-thresholding. It should be noted that model fit after penalization does not remain the same, which makes the proposed method different from the classical factor analysis approach.

4.3. Correlation Structure

So far we have focused on analyzing the covariance matrix. However, it is straightforward to apply APML to a correlation matrix. Factorizing a correlation matrix is essentially the same as factorizing a covariance with $\boldsymbol{S}$ replaced by the correlation matrix $\boldsymbol{R}$. In the case of a covariance matrix, error variances are also contained in $\boldsymbol{\theta}$ but not penalized. In the case of a correlation matrix, $\boldsymbol{\theta}$ contains only unknown elements in $\boldsymbol{\Lambda}$. The EM algorithms for PML in Choi et al. (2010) and Hirose and Yamamoto (2014, 2015) are applicable to the covariance matrix.

Two possible approaches emerge when factorizing a covariance matrix: (1) factorize the covariance matrix directly and (2) factorize the correlation matrix, select the optimal tuning parameter, and then rescale the results back to the covariance structure using the estimated sample variances. The ML factor analysis enjoys the invariant property, in the sense that factorizing a covariance matrix is essentially equivalent to factorizing a correlation matrix except the scale. However, APML factor analysis no longer keeps the invariant property. If a covariance matrix is directly factorized, $\boldsymbol{\theta}$ contains the factor loadings and the error variances. The covariance matrix is decomposed to $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. $\boldsymbol{\Lambda}$ is penalized but $\boldsymbol{\Psi}$ is freely estimated. Penalized factor loadings and unpenalized error variances lead to a penalized estimator of the variance of every observed variable. That is, the variance estimator $\hat{\boldsymbol{\lambda}}_i^T\hat{\boldsymbol{\lambda}}_i + \hat{\psi}_i$ is not guaranteed to be the same as the sample variance, where $\boldsymbol{\lambda}_i$ is the ith row of the loading matrix. Consequently, the diagonal elements of $\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)$ do not need to be the same as the diagonal elements of $\boldsymbol{S}$ due to penalization. In contrast, if a correlation matrix is factorized, $\boldsymbol{\Sigma}$ is decomposed into $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{I} - \text{diag}\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\right)$, where $\boldsymbol{I}$ is an identity matrix and diag () denotes the diagonal matrix whose diagonal elements are those

of the enclosed matrix. Since the diagonal elements of the estimated correlation matrix need to be 1, the error variances cannot be freely estimated. Rather, they are determined by $1 - \hat{\boldsymbol{\lambda}}_i^T \hat{\boldsymbol{\lambda}}_i$ to maintain the correlation scale. Thus, the variance estimator of the second approach is the same as the variance estimator of the ML factor analysis. Because of the difference in treating the variance components, the above two approaches may produce solutions with different number of zero loadings and parameter estimates.

5. Empirical Example

In this section, APML is applied to a subset of the classic Holzinger and Swineford (1939) dataset. This dataset has been widely used in, for example, Browne (2001), Du Toit et al. (2001), and Jöreskog and Sörbom (1993). Browne (2001) conducted EFA to the full dataset and Du Toit et al. (2001) conducted EFA to a subset. Following Du Toit et al. (2001) and Jöreskog and Sörbom (1993), a subset that consists of nine psychological tests of 145 students from Grant-White School is used here. Tests included in the study are visual perception (VIS PERC), cubes (CUBE), lozenges (LOZENGE), paragraph comprehension (PAR), sentence completion (SEN), word meaning (WORD), speeded addition (ADD), speeded counting of dots (COUNT), and speeded discrimination between straight and curved capitals (S-C CAPS). In the present study, we focus on orthogonal EFA.

For the APML the following penalty terms will be used: (1) LASSO, (2) SCAD, and (3) MCP. The SCAD is approximated using local linear approximation and therefore is an adaptive LASSO (Zou, 2006) with weights depending on the starting points and the tuning parameters. For all the penalty terms, 200 values of β or β_1 are used. For the SCAD penalty, $\beta_2 = 3.7$ is used as recommended in Fan and Li (2001). For the MCP penalty, β_2 is set to 1.1, 2, 5, 10, and 50. For every penalty term, the analytical selection methods used are (1) AIC, (2) BIC, (3) MSE and (4) KL. The number of factors is three in the current study.

5.1. Covariance Structure

5.1.1. Analytical Selection We first fit a model with the varimax rotation. The first panel in Table 1 reports the nonzero loadings from the varimax rotation hard-thresholded with a cutoff value of the standardized loadings of 0.3. Here, hard-thresholding refers to setting all $|\hat{\lambda}_{ij}| < 0.3$ to zero, where $\hat{\lambda}_{ij}$ is the standardized factor loading estimate. The varimax solution without hard-thresholding is used as the starting point for APML. The left panel of Table 2 shows the number of zero loadings identified by analytical selection methods when the covariance matrix is analyzed. As shown in Table 2, the MCP tends to yield a sparser model than the other penalties. However, all penalty terms produce fewer zeros than the varimax solution with hard-thresholding. In the second panel in Table 1, we report the nonzero loadings by the MCP with BIC for illustration. Jöreskog and Sörbom (1993) fitted a confirmatory factor model to this dataset with the nonzero factor loading positions similar to the first panel in Table 1 and all factors allowed to be correlated. The three factors were labeled as "Visual Perception", "Verbal Ability", and "Speed" (Jöreskog and Sörbom, 1993). The varimax solution suggests the same loading structure but with orthogonal factors. The structure suggested by the MCP solution with BIC allows for cross-loadings to account for the correlations among factors. It is easy to notice that many nonzero loadings above the threshold value 0.3 in the MCP solution are shrunk to zero in the hard-thresholded varimax solution.

We also performed the APML using the geomin orthogonal rotation (Bernaards & Jennrich, 2005). The loading matrix after hard-thresholding standardized loadings smaller than 0.3 is reported in the third panel of Table 1. The geomin rotation suggests a different loading structure from the varimax rotation. The first factor appears to be a general factor. If the geomin solution

TABLE 1.
Nonzero factor loadings of the Holzinger and Swineford (1939) dataset under varimax and geomin rotations and APML.

Indicator	Varimax			MCP with BIC and Varimax			Geomin			MCP with BIC and Geomin		
	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3
VIS PERC	**0.72**	0.30	0.22	0.74	0.41		**0.80**	0.04	0.13	0.83		
CUBE	**0.53**	0.17	0.08	0.49	0.26		**0.56**	−0.02	0.01	0.52		
LOZENGE	**0.64**	0.28	0.10	0.58	0.39		**0.70**	0.05	0.02	0.68		
PAR	0.17	**0.96**	0.08		0.98		**0.49**	**0.84**	0.01	0.52	0.81	
SEN	0.12	**0.94**	0.21		0.94	0.15	**0.45**	**0.85**	0.13	0.49	0.81	0.07
WORD	0.18	**0.91**	0.07		0.94		**0.48**	**0.80**	0.01	0.50	0.77	
ADD	−0.09	0.19	**0.78**		0.12	0.81	0.06	0.24	**0.77**		0.25	0.81
COUNT	0.24	0.03	**0.83**	0.41		0.76	**0.33**	−0.02	**0.80**	0.32		0.74
S-C CAPS	**0.45**	0.25	**0.56**	0.52	0.29	0.45	**0.56**	0.10	**0.49**	0.59		0.46

For the varimax and geomin rotation solutions, the factor loadings whose standardized loadings are larger than 0.3 or lower than -0.3 are bold faced. For the APML solution with the MCP penalty, the blank spaces correspond to exact zero factor loadings. For MCP with BIC and Varimax, the chosen tuning parameters are $(\beta_1, \beta_2) = (0.014, 50)$ and the BIC value is 1117.817. For MCP with BIC and Geomin, the chosen tuning parameters are $(\beta_1, \beta_2) = (0.007, 1.1)$ and the BIC value is 1114.491.

TABLE 2.
Number of zero factor loadings, APML on the covariance/correlation matrix under combinations of penalty terms and analytical selection methods for the Holzinger and Swineford (1939) dataset.

Penalty	Rotation	Covariance				Correlation			
		AIC	BIC	MSE	KL	AIC	BIC	MSE	KL
LASSO	Varimax	4	7	3	3	7	9	7	7
	Geomin	5	7	3	3	6	10	6	6
SCAD	Varimax	4	7	3	3	5	8	3	3
	Geomin	6	7	3	3	7	10	3	3
MCP	Varimax	10	10	3	3	10	10	8	8
	Geomin	11	11	3	3	11	11	7	8
–	Varimax				17				
–	Geomin				13				

without hard-thresholding is used as the starting point of APML, it may produce loading structures of different sparsity (Table 2). The loading matrix produced by the MCP penalty with BIC is shown in the fourth panel of Table 1. The results shown in the second and the fourth panels of Table 1 produced a general factor, but the solution in the second panel found in addition to the general factor, a factor closely related to the observed indicators of "Visual Perception" whereas the solution in the fourth panel found a factor closely related to the indicators of "Verbal Ability".

5.1.2. Solution Path Based Selection One advantage of PML and APML is the solution path that contains all loading structures suggested by the data. Thus, the covariance matrix is analyzed 200 times corresponding to the 200 values of β. Consequently, 200 estimates arise for every factor loading. As mentioned previously, the solution path is referred to the plot of the 200 estimates for each factor loading against the 200 β values. Figure 2a, c gives the solution path of the LASSO for the nine factor loadings plotted together when the varimax and the geomin rotations are used as starting points, respectively. As we can see from these figures, the estimated factor loadings tend to zero and more factor loadings are estimated as zeros as β increases. The loading structure

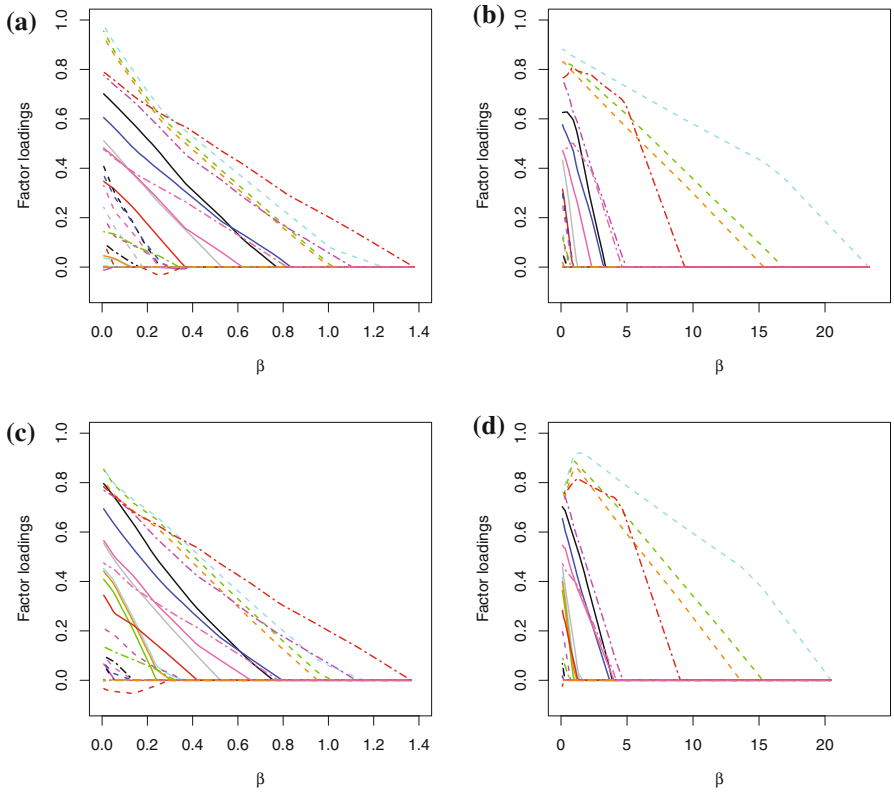

FIGURE 2.
The LASSO solution path of factor loadings of the Holzinger and Swineford (1939) dataset. The solid, dashed, and dot-dashed lines are the indicators of f_1, f_2, and f_3, respectively. The black, gray, blue, cyan, green, orange, magenta, red, and pinks lines correspond to visual perception (VIS PERC), cubes (CUBE), lozenges (LOZENGE), paragraph comprehension (PAR), sentence completion (SEN), word meaning (WORD), speeded addition (ADD), speeded counting of dots (COUNT), and speeded discrimination straight and curved capitals (S-C CAPS), respectively. **a** Varimax with covariance structure. **b** Varimax with correlation structure. **c** Geomin with covariance structure. **d** Geomin with correlation structure (Color figure online).

suggested by the LASSO at any value of the tuning parameter β can be extracted from Fig. 2a, c. For example, from Fig. 2a and at the value of $\beta = 0.5$, there are ten nonzero factor loadings. If we decrease β, the path from f_1 to "ADD" (magenta and solid) is added to the model first. If we increase β instead, the path from f_1 to "CUBE" (gray and solid) will be removed. The solution path when a geomin rotation is used is given in Fig. 2c and it is similar to that from a varimax rotation.

By scanning through all the 200 β values, 200 loading matrices suggested by the LASSO can be extracted. However, the unique patterns of the loading matrices in this example is only 26, if the rotation is the Varimax, which is far less than 200. For example, all β values between 0.4 and 0.5 produce the same loading structure, although the values of the nonzero loadings are different. Further, all unique patterns can be automatically extracted from the entire solution path. The loading structure with the best interpretability can then be chosen.

5.2. Correlation Structure

5.2.1. Analytical Selection The right panel of Table 2 presents the number of zero loadings identified by analytical selection methods if the correlation matrix is factorized and penalized

instead. Similar to the case where a covariance matrix is factorized, the MCP tends to yield a sparser model than the other two penalties. It is also seen that factorizing a correlation matrix often produces a different loading structure than when factorizing a covariance matrix.

5.2.2. Solution Path Based Selection The solution paths of the LASSO are given in Fig. 2b, d when the varimax and the geomin rotations are used as starting points, respectively. It is seen that the solution paths are different from the ones based on the covariance matrix, especially the sequence of zero loadings.

6. Simulation Study

In this section, a simulation study is conducted to study the performance of APML and shed some light on the choice of penalty terms and selection criteria. The APML solution is compared with the PML solution and the commonly used rotation methods such as the varimax, the geomin, and their pairwise rotation solutions. For the rotation methods, standardized loadings less than 0.3 in absolute value are considered zero. For factor rotations see Jennrich (2007) and the references therein. The main purpose of the simulation is to illustrate that the proposed method produces reasonable factor solutions, rather than showing that it uniformly dominates the traditional approach.

6.1. Simulation Design

Two EFA models are considered in the simulation, one without cross-loadings and one with a cross-loading.

6.1.1. Model 1 An EFA model with nine indicators and three factors is considered, where

$$\boldsymbol{\Lambda}_1 = \begin{pmatrix} 0.8 & 0.8 & 0.8 & & & & & & \\ & & & 0.8 & 0.8 & 0.8 & & & \\ & & & & & & 0.8 & 0.8 & 0.8 \end{pmatrix}^T.$$

The matrix $\boldsymbol{\Psi}_1$ is set such that $\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1^T + \boldsymbol{\Psi}_1$ is a correlation matrix. In the case of factorizing a covariance structure, the correlation matrix is scaled to a covariance matrix with diagonal elements (3.50, 3.51, 4.90, 3.98, 3.81, 3.87, 4.66, 3.29, 3.39). A factor rotation is expected to recover the perfect simple structure.

6.1.2. Model 2 Model 2 also has nine indicators and three factors where

$$\boldsymbol{\Lambda}_2 = \begin{pmatrix} 0.8 & 0.8 & 0.8 & & & & & & 0.3 \\ & & & 0.8 & 0.8 & 0.8 & & & \\ & & & & & & 0.8 & 0.8 & 0.8 \end{pmatrix}^T.$$

The matrix $\boldsymbol{\Psi}_2$ is set such that $\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_2$ is a correlation matrix. In the case of factorizing a covariance structure the correlation matrix is scaled to a covariance matrix with the same diagonal elements as Model 1. The rotation solutions hard-thresholded at 0.3 are expected to mistakenly shrink the cross-loading from 0.3 to zero.

6.2. Simulation Specifics

Normally distributed data were simulated from a three-factor model with nine observed variables. Sample sizes $n = 100$ and 200, which are commonly encountered in empirical studies, are used. For each simulation condition, we generated 1000 datasets. APML used the unpenalized ML solution, with varimax or geomin rotation employed as the starting point, which avoided inadmissible solutions.

The methods compared here can produce four types of outcomes: (1) truly zero loadings are correctly identified as zero; (2) truly zero loadings are falsely identified as nonzero; (3) truly nonzero loadings are correctly identified as nonzero; and (4) truly nonzero loadings are falsely identified as zero. A good method is expected to produce a good deal of first and third outcomes and yield none or few of the fourth outcome.

The statistics to be reported are: the empirical percentage recovery of the correct loading structure,

$$\frac{1}{R}\sum_{i=1}^{R} I\,(\text{the loading structure is correctly recovered at } i\text{th replication}),$$

the empirical percentage of correctly identifying a truly zero loading as zero

$$\frac{1}{R}\sum_{i=1}^{R}\frac{\text{number of correctly identified zeros at } i\text{th replication}}{\text{number of zero elements in the true loading matrix}},$$

and the empirical percentage of falsely identifying a truly nonzero loading as zero

$$\frac{1}{R}\sum_{i=1}^{R}\frac{\text{number of falsely identified zeros at } i\text{th replication}}{\text{number of nonzero elements in the true loading matrix}},$$

where R is the number of admissible replications.

Since a sparse loading matrix may be biased the average mean squared error (AMSE) is used to compare APML with PML under different penalty terms and analytical selection criteria. The AMSE is defined as

$$\text{AMSE} = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{q}\sum_{j=1}^{q}\left(\hat{\theta}_j^{(i)} - \theta_j\right)^2,$$

where q is the number of parameters and $\hat{\theta}_j^{(i)}$ is the estimate of θ_j at the ith replication. The AMSE of factor loadings and the covariance (or correlation) matrix $\boldsymbol{\Sigma}$ are computed. For the comparison with the standard factor rotations, solutions with and without hard-thresholding are considered. The cutoff value of 0.3 is applied for the rotated solutions with hard-thresholding.

6.3. Penalty Terms Included in the Study

Three penalty terms are considered in the simulation study (LASSO, SCAD, and MCP). For all penalty terms, the varimax and geomin solution were used as starting points. The SCAD is approximated using local linear approximation. For all penalty terms, 200 values of β_1 (or β if

there is only one tuning parameter) are used. For the SCAD, β_2 is set to 3.7, as recommended by Fan and Li (2001). For the MCP, five values of β_2 are used: 1.1, 2, 5, 10, and 50. Note that the same settings are applied in the above empirical example. The β_1 values are chosen such that the models suggested by PML and APML may have less than three factors and less than nine indicators (some of the items might have zero loadings on all factors).

For each penalty term, different analytical selection criteria are applied. If a covariance matrix is directly factorized, AIC, BIC, MSE, and KL are used to select the tuning parameters. As mentioned previously, an alternative is, for example, to construct the solution path for the correlation matrix, use AIC to select the tuning parameter, extract the estimates from the solution path, and rescale the extracted estimates to the covariance scale. The above approach is referred to as AICR in the present study, which stands for AIC followed by rescaling. Likewise, we can introduce the acronyms BICR, MSER, and KLR. Thus, eight selection criteria will be used for a covariance matrix, namely AIC, BIC, MSE, KL, AICR, BICR, MSER, and KLR. If a correlation matrix is factorized, four analytical criteria are used (AIC, BIC, MSE, and KL). In the case of factorizing a covariance matrix with rescaling, the optimal tuning parameter is chosen from a correlation structure. Hence, factorizing a correlation matrix produces the same loading structure as factorizing a covariance matrix but a different AMSE.

6.4. Simulation Results

Due to space limitation, only selected results are discussed here. More results can be found in the supplementary materials. In particular, BIC and BICR outperform the other selection criteria in the sense of greater sparsity and lower AMSE. Therefore, attention is mostly paid to BIC and BICR, unless otherwise stated. The syntax is based on the R package Rcpp (Eddelbuettel, 2013; Eddelbuettel & François, 2011) and will be made public in the future.

6.4.1. Sparsity of Factorizing a Covariance Matrix The percentages of recovering the correct loading structure are given in Table 3. APML starting from varimax and geomin produces a similar pattern, so only the results with varimax rotation as the starting point are included in Table 3. The two loading structures studied show some similarities. Columns 4 and 6 of Table 3 show the results when the covariance matrix is directly factorized. If the penalty term is the LASSO or the MCP then PML and APML produce a similar recovery percentage. However, the LASSO rarely recovers the correct loading structure. If the penalty term is the SCAD, PML achieves a higher percentage recovery than APML. Columns 5–7 give the results when the correlation matrix is factorized and rescaled to the covariance scale. For APML, the percentage recovery is higher for all penalty terms but for PML this is not always true. The two loading structures also show some discrepancies. The standard rotations with hard-thresholding at 0.3 work satisfactorily and outperform all the APML solutions when $\mathbf{\Lambda} = \mathbf{\Lambda}_1$ but this is not the case for $\mathbf{\Lambda} = \mathbf{\Lambda}_2$. For example, under the geomin rotation and for $n = 200$, the percentage recovery is 40.20%, whereas for APML with the MCP penalty and BICR the corresponding number is 79.50% (Table 3).

Table 4 shows that all the penalty terms have a high percentage of identifying a truly zero factor loading as zero. The corresponding percentages for the standard rotation methods with hard-thresholding at 0.3 are close to 100%. However, the falsely zero recovery rate is found to be higher for the standard rotation methods than for the PML and APML when $\mathbf{\Lambda} = \mathbf{\Lambda}_2$ (Table 5). This is mostly due to the small cross-loading in $\mathbf{\Lambda}_2$. When $\mathbf{\Lambda} = \mathbf{\Lambda}_1$, all methods produce a zero falsely zero recovery rate. Therefore, PML and APML tend to overfit the model by including more nonzero factor loadings than the true data generation process. However, results in Hirose and Yamamoto (2015) do not suggest that over-fitting tendency.

As mentioned previously, the solution path which is a function of the tuning parameter contains all the possible path diagrams. Although the analytical selection fails to distinguish the

TABLE 3.
Percentage of recovering the correct loading structure, covariance matrix is factorized. APML uses the varimax rotation as a starting point.

n	Penalty	Method	$\mathbf{\Lambda} = \mathbf{\Lambda}_1$		$\mathbf{\Lambda} = \mathbf{\Lambda}_2$	
			BIC	BICR	BIC	BICR
100	LASSO	PML	2.40	0.00	1.40	0.00
		APML	0.70	46.70	0.30	36.40
	SCAD	PML	77.50	32.40	60.30	24.40
		APML	0.50	69.00	0.20	52.60
	MCP	PML	56.90	62.60	55.50	50.80
		APML	50.20	80.60	49.50	56.00
	–	Varimax	99.90		25.00	
	–	Geomin	99.50		41.60	
200	LASSO	PML	3.60	0.00	3.30	0.00
		APML	1.50	58.10	0.70	47.30
	SCAD	PML	86.90	78.30	81.10	54.00
		APML	1.40	87.10	1.00	75.10
	MCP	PML	70.90	82.50	72.50	73.20
		APML	66.30	91.70	68.10	79.50
	–	Varimax	100.00		17.40	
	–	Geomin	100.00		40.20	

APML with different starting points produce similar results. The rotations do not rely on the selection criterion. The pairwise rotation with varimax or geomin as the analytical criterion produces similar results to the varimax or geomin rotation, respectively.

TABLE 4.
Percentage of correctly setting a truly zero factor loading as zero in the estimated loading matrix, covariance matrix is factorized. APML using varimax rotation as the starting point.

n	Penalty	Method	$\mathbf{\Lambda} = \mathbf{\Lambda}_1$		$\mathbf{\Lambda} = \mathbf{\Lambda}_2$	
			BIC	BICR	BIC	BICR
100	LASSO	PML	71.06	29.39	68.42	27.70
		APML	63.35	94.73	60.51	92.88
	SCAD	PML	98.18	93.01	97.00	90.92
		APML	63.07	97.44	61.55	96.14
	MCP	PML	96.81	96.64	96.56	94.82
		APML	95.90	98.20	95.75	96.29
	–	Varimax	99.99		99.99	
	–	Geomin	99.97		99.94	
200	LASSO	PML	74.52	30.28	72.47	29.07
		APML	68.88	96.29	65.79	94.52
	SCAD	PML	99.07	98.59	98.54	95.79
		APML	68.56	99.13	66.94	98.12
	MCP	PML	98.11	98.68	98.07	97.27
		APML	97.72	99.41	97.72	98.36
	–	Varimax	100.00		100.00	
	–	Geomin	100.00		100.00	

APML with different starting points produce similar results. The rotations do not rely on the selection criterion. The pairwise rotation with varimax or geomin as the analytical criterion produces similar results to the varimax or geomin rotation, respectively.

TABLE 5.
Percentage of falsely setting a nonzero factor loading as zero, covariance matrix is factorized and $\mathbf{\Lambda} = \mathbf{\Lambda}_2$. APML uses the marimax rotation as the starting point.

Penalty	Method	$n = 100$		$n = 200$	
		BIC	BICR	BIC	BICR
LASSO	PML	0.04	0.03	0.00	0.00
	APML	0.04	0.42	0.00	0.00
SCAD	PML	0.71	0.19	0.02	0.00
	APML	0.02	0.70	0.00	0.02
MCP	PML	0.60	0.49	0.05	0.04
	APML	0.41	0.74	0.04	0.01
–	Varimax	7.50		8.26	
–	Geomin	5.83		5.98	

APML with different starting points produces similar results. The rotations do not rely on the selection criterion. The pairwise rotation with varimax or geomin as the analytical criterion produces similar results to the varimax or geomin rotation, respectively.

TABLE 6.
Percentage of containing the correct loading structure in the solution path. APML uses the varimax rotation as starting point.

n	Method	$\mathbf{\Lambda} = \mathbf{\Lambda}_1$			$\mathbf{\Lambda} = \mathbf{\Lambda}_2$		
		LASSO	SCAD	MCP	LASSO	SCAD	MCP
Factorizing a covariance matrix							
100	PML	90.60	100.00	100.00	86.30	93.30	96.40
	APML	100.00	98.70	100.00	93.30	89.80	95.80
200	PML	99.90	100.00	100.00	99.50	99.50	99.90
	APML	100.00	100.00	100.00	99.60	99.10	99.90
Factorizing a correlation matrix							
100	PML	95.40	100.00	99.20	85.60	92.30	93.60
	APML	100.00	100.00	100.00	91.40	90.20	92.40
200	PML	100.00	100.00	100.00	99.10	99.60	99.60
	APML	100.00	100.00	100.00	99.40	99.50	99.50

APML with different starting points produces similar results.

correct loading structure, the correct loading structure may still be contained in the solution path. Table 6 shows that both PML and APML have high rates of producing the correct loading structure at some step in the solution path for both $\mathbf{\Lambda} = \mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. Rescaling generally produces a similar result to factorizing a covariance matrix in this regard. Thus, the selection of the tuning parameter is crucial in finding the correct loading structure.

6.4.2. Sparsity of Factorizing a Correlation Matrix If a correlation matrix is factorized, the percentage recovery of the correct loading structure using BIC is the same when using BICR and factorizing a covariance matrix. Hence, Tables 3, 4, 5, and 6 already reveal the sparsity of factorizing a correlation matrix. Thus, conclusions are not restated here.

6.4.3. Estimation Accuracy of Factorizing a Covariance Matrix Similar conclusions can be drawn from $n = 100$ and $n = 200$. Thus, only the AMSE for the factor loadings and the covariance

TABLE 7.
Average mean squared error (AMSE) of factor loading and covariance matrix, covariance matrix is factorized, sample size is 100.

Penalty	Method	$\mathbf{\Lambda} = \mathbf{\Lambda}_1$				$\mathbf{\Lambda} = \mathbf{\Lambda}_2$			
		BIC		BICR		BIC		BICR	
		Varimax	Geomin	Varimax	Geomin	Varimax	Geomin	Varimax	Geomin
AMSE of factor loading multiplied by 100									
LASSO	PML	3.43		2.44		3.34		2.49	
	APML	3.23	3.23	1.68	1.69	3.14	3.17	1.86	1.83
SCAD	PML	1.19		1.24		1.40		1.45	
	APML	3.21	3.22	1.49	1.50	3.12	3.14	1.68	1.67
MCP	PML	1.40		1.31		1.54		1.54	
	APML	1.49	1.49	1.24	1.25	1.58	1.57	1.42	1.43
–	Rot.	2.43	2.46	2.43	2.46	2.61	2.50	2.61	2.50
–	Rot.-Hard	1.09	1.10	1.09	1.10	1.93	1.77	1.93	1.77
AMSE of covariance matrix multiplied by 100									
LASSO	PML	23.34		15.94		23.12		16.15	
	APML	22.40	22.34	10.11	10.14	21.95	22.15	11.26	11.15
SCAD	PML	8.51		8.82		9.74		10.02	
	APML	21.62	21.73	9.45	9.47	21.38	21.50	10.68	10.63
MCP	PML	9.55		9.09		10.38		10.32	
	APML	9.93	9.92	8.58	8.59	10.46	10.45	9.83	9.87
–	Rot.	17.45	17.45	17.45	17.45	17.52	17.52	17.52	17.52
–	Rot.-Hard	8.02	8.07	8.02	8.07	12.43	11.68	12.43	11.68

Rot. = Rotation without hard-thresholding. Rot.-Hard = Rotation with hard-thresholding. PML does not depend on rotations, whereas the starting points of APML depend on rotations. Varimax and geomin rotations are applied to Columns 3, 5, 7, 9, and Columns 4, 6, 8, 10, respectively, if needed.

matrix when $n = 100$ are shown in Table 7. The AMSE of factor loading and AMSE of covariance matrix share similar patterns. Rows 7, 8, 15, and 16 show that different rotations tend to produce a similar AMSE when $\mathbf{\Lambda} = \mathbf{\Lambda}_1$, but different AMSE values when $\mathbf{\Lambda} = \mathbf{\Lambda}_2$. Nevertheless, APML starting from different rotations often yields a similar AMSE. Columns 3, 4, 7, and 8 show that, when the covariance matrix is directly factorized, PML and APML produce a similar AMSE if the penalty term is the LASSO or MCP, but PML produces a lower AMSE than APML if the penalty term is the SCAD. When rescaling is employed, APML tends to produce a lower AMSE than PML for LASSO but PML and APML often produce a similar AMSE for SCAD and MCP (Columns 5, 6, 9, and 10 of Table 7). Compared to the rotated solutions, both PML and APML produce a higher AMSE than the rotations with hard-thresholding when $\mathbf{\Lambda} = \mathbf{\Lambda}_1$, but a proper choice of the penalty term and the analytical selection method produces a lower AMSE than the rotations without hard-thresholding. When $\mathbf{\Lambda} = \mathbf{\Lambda}_2$, PML with the SCAD penalty, PML with the MCP penalty, and APML with BICR tend to produce a lower AMSE than the rotations, regardless of hard-thresholding.

For both loading structures, rescaling does not have a strong effect for PML with the SCAD penalty, PML with the MCP penalty, and APML with the MCP penalty. In contrast, rescaling tends to produce a lower AMSE than nonrescaling for LASSO. Considering all combinations of the penalty term, numerical algorithm, and analytical selection criterion, PML with SCAD and BIC or BICR, PML with MCP and BIC or BICR, and APML with MCP and BIC or BICR often produce a similar AMSE that is lower than the AMSE value produced by the other combinations (including those selection criteria that are not presented here).

TABLE 8.
Average mean squared error (AMSE) of factor loading and correlation matrix, correlation matrix is factorized, sample size is 100.

Penalty	Method	$\mathbf{\Lambda} = \mathbf{\Lambda}_1$		$\mathbf{\Lambda} = \mathbf{\Lambda}_2$	
		Varimax	Geomin	Varimax	Geomin
AMSE of factor loading multiplied by 100					
LASSO	PML	4.55		4.68	
	APML	2.39	2.40	2.92	2.82
SCAD	PML	1.36		2.05	
	APML	1.91	1.92	2.49	2.45
MCP	PML	1.59		2.30	
	APML	1.40	1.41	1.96	1.98
–	Rot.	4.54	4.63	4.99	4.71
–	Rot.-Hard	1.02	1.04	3.52	3.05
AMSE of correlation matrix multiplied by 100					
LASSO	PML	7.00		7.04	
	APML	2.47	2.49	3.31	3.23
SCAD	PML	1.49		2.54	
	APML	1.83	1.85	2.77	2.73
MCP	PML	1.74		2.69	
	APML	1.31	1.32	2.28	2.31
–	Rot.	8.10	8.10	8.00	8.00
–	Rot.-Hard	0.94	0.98	4.61	3.94

Rot. = Rotation without hard-thresholding. Rot.-Hard = Rotation with hard-thresholding. PML does not depend on rotations, whereas the starting points of APML depends on rotations. Varimax and geomin rotations are applied to Columns 3, 5, and Columns 4, 6, respectively, if needed.

6.4.4. Estimation Accuracy of Factorizing a Correlation Matrix Similar conclusions can be drawn from $n = 100$ and $n = 200$. Thus, Table 8 only shows the AMSE of factor loadings and the AMSE of the correlation matrix when $n = 100$ and when the correlation matrix is factorized. Similar conclusions can be drawn regarding the AMSE of factor loading and AMSE of correlation matrix. If the penalty term is the LASSO, APML generally produces a lower AMSE than PML. For the SCAD, PML tends to produces a slightly lower AMSE than APML, whereas, for the MCP, PML tends to produces a slightly higher AMSE than APML.

For the SCAD and MCP, PML and APML always yield a lower AMSE than the rotations without hard-thresholding across all loading structures. Compared to the rotations with hard-thresholding, PML and APML generally produce a lower AMSE if $\mathbf{\Lambda} = \mathbf{\Lambda}_2$, except PML with LASSO. If $\mathbf{\Lambda} = \mathbf{\Lambda}_1$, PML and APML always produce a higher AMSE than the rotations with hard-thresholding. Across all combinations of the penalty term, numerical algorithm, and selection criterion, the SCAD with BIC and MCP with BIC are generally preferred to the other combinations (including those selection criteria that are not presented here), no matter whether PML or APML is used.

6.4.5. Elapsed Time The computational efficiency of PML and APML is compared. As an illustration, the median elapsed time of PML for one replication with 200 tuning parameter is 0.34 s when the penalty term is the LASSO, $\mathbf{\Lambda} = \mathbf{\Lambda}_1$, $n = 100$, and a covariance matrix is directly factorized. In contrast, the median elapsed time of APML for one replication is 0.03 s under the same condition. See Figure 8 in the supplementary material for the elapsed time of other settings.

7. Discussion

In this paper, an approximation method for PML is introduced to conduct EFA. Penalization naturally produces a sparse loading matrix by choosing an appropriate penalty term. The shrinkage process is continuous and therefore no subjective decision has to be made on the cutoff values. Hence, small loadings in the estimation-rotation procedure are well avoided. As an approximation to PML, APML inherits the sparsity of PML and constructs a solution path more efficiently. Our simulation results also suggest that APML may produce a higher percentage of recovery of the correct loading structure than PML with proper penalty terms and selection criteria.

In the simulation study APML with analytical selection criteria shows a lower percentage recovery of the true loading structure and a lower percentage of false recovery of a nonzero loading as zero but a higher percentage recovery of the truly zero loadings. Although analytical selection criteria do not always recover the loading structure, the correct loading matrix is frequently contained in the solution path of the APML. The entire solution path of APML contains the trajectory of all loadings as a function of the tuning parameter. All the loading structures identified by PML or APML can be seen from the solution path. A subjective choice of the optimal tuning parameter is made by looking at the number of nonzero loadings and the interpretability of the retrieved loading matrix. Model selection based on the solution path has been demonstrated using a subset of the Holzinger and Swineford (1939) dataset. Similarly to varying the value of the tuning parameter, it is also possible to vary the cutoff value of 0.3 to produce a solution path of loadings. Consequently, sparsity of the loading matrix varies for different choices of the cutoff value and the model with a good interpretability may be preferred. Alternatively, inference tools (hypothesis testing and confidence intervals) can be used to select statistically significant loadings. However, truncating loadings that are less than a prefixed value and statistical testing are both hard-threshold methods leading to a discontinuous loading selection procedure.

In addition, our simulation results show that both the PML and APML estimators frequently produce a lower AMSE than the rotation solutions without hard-thresholding for factor loadings and the covariance/correlation matrix. APML often produces a similar AMSE to PML and even outperforms PML sometimes. In the case of factorizing a covariance matrix, rescaling helps to improve the estimation accuracy of the LASSO. The correlation scale restricts all factor loadings on the same scale, whereas the factor loadings are less bounded in the covariance scale and the larger factor loadings are suspected to dominate the small ones. Thus, it is a good choice to work with a correlation matrix even though a covariance matrix is available. If the covariance scale is preferred, the APML estimator from the correlation analysis can be rescaled back to the estimator corresponding to the covariance matrix. Considering all the combinations of the penalty terms and analytical selection criteria, the MCP with BIC for a correlation matrix is often a reasonable choice for loading structure recovery and accurate estimation. If a covariance matrix is directly factorized, the SCAD with BIC is a reasonable choice. The above discussion reflects the fact that penalizing a covariance matrix is radically different from penalizing a correlation matrix. Thus, it is important to choose the scale before the analysis. Further, APML with the varimax rotation is likely to produce a different solution to APML with the geomin rotation. This reflects the different natures of factor rotation methods. Different rotations are performed to meet different criteria. The nondifferentiable penalty terms make the APML solutions not equivalent. Hence, APML should be interpreted as the continuous thresholding conditional on the rotation method. The researchers therefore need to carefully choose the rotation method as if they were conducting the traditional EFA.

EFA involves two major decisions mainly selecting the number of factors and the rotation method. The current study focuses on the latter. The number of factors is not fixed but the maximum number of factors is fixed to be the true number in the current study. This reflects that PML and APML can be used to select the number of factors. Since PML and APML produce zero loadings,

they are able to shrink the entire column of loadings to zero and the number of nonzero columns is the number of factors retained. Hence, we can start from the maximum number of factors (as long as the model is identified) and then fit the model using PML or APML. The retained sparse loading matrix indicates both the number of factors and the important links suggested by the penalty term. Selecting the number of factors using PML or APML is a noneigenvalue-based approach and can be viewed from a model selection perspective. It will be investigated in a future project.

Equation (4) indicates that the APML estimator relies on the properties of the ML estimator. The effect of choosing a different starting point can also be seen from the empirical example. The solution path starts from the ML estimator and shrinks some of the "unimportant" loadings to zero. Hence, further studies are needed to provide guidelines for selecting a suitable starting point. The present study focuses on the orthogonal EFA. It is natural to extend APML to an oblique EFA and that will be investigated in the future. Moreover, since APML naturally factorizes a correlation matrix, it can be also applied to ordinal data using a polychoric correlation matrix. It deserves a future study to investigate the performance of APML with ordinal data.

Acknowledgments

Shaobo Jin and Fan Yang-Wallentin are partly supported by the Vetenskapsrådet (Swedish Research Council) under contract 2017-01175. We would like to thank the reviewers for providing valuable comments. We also would like to thank Måns Thulin for giving valuable comments on the early version of the manuscript.

References

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, *65*, 676–696.

Björck, A. (1996). *Numerical methods for least squares problems*. Philadelphia, PA: SIAM.

Brehenv, 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.

Browne, M. V. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.

Browne, M. W., & Du Toit, S. H. C. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, *27*, 269–300.

Carroll, J. B. (1953). An analytic rotation for approximating simple structure in factor analysis. *Psychometrika*, *18*, 23–38.

Choi, J., Zou, H., & Oehlert, G. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface*, *3*, 429–436.

Du Toit, M., Du Toit, S., & Hawkins, D. M. (2001). *Interactive LISREL: User's guide*. Linconwood, IL: Scientific Software International.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York: Springer.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*, 1–18.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–840.

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

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *321*, 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

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

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hirose, K., & Konishi, S. (2012). Variable selection via the weighted group lasso for factor analysis models. *Canadian Journal of Statistics*, *40*, 345–361.

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

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via non-concave penalized likelihood in 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, IL: University of Chicago Press.

Hunter, D., & Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, *33*, 1617–1642.
Jennrich, R. I. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika*, *69*, 257–273.
Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, *71*, 173–191.
Jennrich, R. I. (2007). Rotation algorithms: From beginning to end. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 45–63). Amsterdam, The Netherlands: Elsevier.
Johnstone, I. M., & Lu, A. Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, *104*, 682–693.
Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*, 443–482.
Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Linconwood, IL: Scientific Software International.
Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–240.
Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, *60*, 64–82.
Mazumder, R., Friedman, J. H., & Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, *106*, 1125–1138.
Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, *52*, 374–393.
Neuhaus, J. O., & Wrigley, C. (1954). The quartimax method: An analytical approach to orthogonal simple structure. *British Journal of Mathematical and Statistical Psychology*, *7*, 81–91.
Ning, L., & Georgiou, T. T. (2011, December). Sparse factor analysis via likelihood and l_1-regularization. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on decision and control and european control conference* (pp. 5188–5192).
Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, *9*, 319–337.
Rubin, D., & Thayer, D. (1982). EM algorithms for ML factor analysis. *Psychometrika*, *47*, 69–76.
Shen, H., & Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, *99*, 1015–1034.
Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA: Allyn and Bacon.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*, 267–288.
Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 273–282.
Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics*, *29*, 431–454.
Trendafilov, N. T., & Adachi, K. (2015). Sparse versus simple structure loadings. *Psychometrika*, *80*, 776–790.
Trendafilov, N. T., Fontanella, S., & Adachi, K. (2017). Sparse exploratory factor analysis. *Psychometrika*, *82*, 778–794.
Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, *10*, 515–534.
Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.
Zhang, G. (2014). Estimating standard errors in exploratory factor analysis. *Multivariate Behavioral Research*, *49*, 339–353.
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.
Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, *35*, 1849–2311.
Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, *36*, 1509–1533.

Manuscript Received: 27 JUN 2016
Final Version Received: 13 APR 2018
Published Online Date: 6 JUN 2018