PSYCHOMETRIKA—VOL. 88, NO. 2, 527–553
JUNE 2023
https://doi.org/10.1007/s11336-023-09911-y

ROTATION TO SPARSE LOADINGS USING L^P LOSSES AND RELATED INFERENCE PROBLEMS

XINYI LIU

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

GABRIEL WALLIN

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UMEÅ UNIVERSITY

YUNXIAO CHEN AND IRINI MOUSTAKI

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Researchers have widely used exploratory factor analysis (EFA) to learn the latent structure underlying multivariate data. Rotation and regularised estimation are two classes of methods in EFA that they often use to find interpretable loading matrices. In this paper, we propose a new family of oblique rotations based on component-wise L^p loss functions ($0 < p \leq 1$) that is closely related to an L^p regularised estimator. We develop model selection and post-selection inference procedures based on the proposed rotation method. When the true loading matrix is sparse, the proposed method tends to outperform traditional rotation and regularised estimation methods in terms of statistical accuracy and computational cost. Since the proposed loss functions are nonsmooth, we develop an iteratively reweighted gradient projection algorithm for solving the optimisation problem. We also develop theoretical results that establish the statistical consistency of the estimation, model selection, and post-selection inference. We evaluate the proposed method and compare it with regularised estimation and traditional rotation methods via simulation studies. We further illustrate it using an application to the Big Five personality assessment.

Key words: component loss function, analytic rotation, regularised estimation, model selection, confidence interval.

Researchers have widely used exploratory factor analysis (EFA) to learn the latent structure underlying multivariate data. A major problem in EFA is identifying an interpretable factor structure among infinitely many equivalent solutions that give the same data distribution, where two equivalent solutions differ by a rotation transformation (see Chapters 10–12, Mulaik, 2009). Mathematically, we aim to find a sparse solution for which many entries of the loading matrix are exactly or approximately zero so that each factor can be interpreted based on a small number of manifest variables whose loadings on the factor are not close to zero. This idea dates back to the seminal discussion on simple factor structure in Thurstone (1947).

We can classify methods for obtaining sparse loading structures into two categories – rotation and regularised estimation methods. A rotation method involves two steps. In the first, we obtain an estimate of the loading matrix. Typically, but not necessarily, a maximum likelihood estimator is used in this step (Bartholomew et al., 2011) , under some arbitrary but mathematically convenient constraints that avoid rotational indeterminacy. In the second step, we rotate the estimated loading

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-023-09911-y.

Correspondence should be made to Yunxiao Chen, Department of Statistics, London School of Economics and Political Science, Columbia House, Room 5.16, Houghton Street, London, WC2A 2AE, UK. Email: y.chen186@lse.ac.uk

© 2023 The Author(s)

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

matrix to minimise a certain loss function where a smaller loss function value tends to imply a more interpretable solution. Researchers have proposed different rotation methods that differ by, first, whether the factors are allowed to be correlated, and second, the loss function for measuring sparsity. A rotation method is called an orthogonal rotation when the factors are constrained to be uncorrelated and an oblique rotation otherwise. Different loss functions have been proposed for orthogonal and oblique rotations, including varimax (Kaiser, 1958) , oblimin (Jennrich & Sampson, 1966) , geomin (Yates, 1987) , simplimax (Kiers, 1994) , and component-wise loss (Jennrich, 2004, 2006) , among many others. Among the existing rotation methods, we draw attention to the monotone concave Component Loss Functions (CLFs; Jennrich, 2004, 2006) due to their desired theoretical properties and superior performance in recovering sparse loading matrices. Specifically, Jennrich (2004, 2006) provided some theoretical guarantees to the CLFs when the true loading matrix has a perfect simple structure and further found that the CLFs are often more accurate in recovering sparse loading matrices than other rotation methods under both orthogonal and oblique settings.

In recent years, several regularised estimation methods have been proposed for EFA (e.g., Geminiani et al., 2014; Jin et al., 2017; Trendafilov, 2018; Yamamoto et al., 2021). Slightly different from rotation methods, a regularised estimation method simultaneously estimates the model parameters and produces a sparse solution. It introduces a least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) type sparsity-inducing regularisation term into the loss function for parameter estimation, where the regularisation term imposes sparsity on the estimated loadings. It typically obtains a sequence of candidate models by varying the weight of the regularisation term in the loss function. The final model is chosen from the candidate models, often using an information criterion.

In this paper, we propose a new family of oblique rotations based on component-wise L^p loss functions, for $0 < p \leq 1$. We show the proposed loss functions to be special cases of monotone concave CLFs and that they thus share the same theoretical properties. We note that Jennrich (2004, 2006) considered the L^1 loss function but not the L^p loss functions with $p < 1$. With the proposed rotations, we solve several previously unaddressed problems regarding rotation and regularised estimation methods. First, we establish the statistical consistency of the rotated solution. More specifically, we provide conditions under which the rotated solution converges to the true sparse loading matrix as the sample size goes to infinity. These conditions also provide insights into the choice of p. Seemingly straightforward, this consistency result requires some refined analysis and, to our best knowledge, such results have not been established for other rotation methods. In particular, the theoretical results for the CLFs in Jennrich (2004, 2006) were established concerning the population loading matrix rather than its estimate. Second, we address the difficulty of establishing whether regularised estimation methods outperform rotation methods or vice versa. To gain some insights into this question, we theoretically show that the proposed rotation method can be viewed as the limiting case of a regularised estimator when the weight of the regularisation term converges to zero. In addition, to compare the two methods in terms of model selection, we develop a hard-thresholding procedure that conducts model selection based on a rotated solution. Through computational complexity analysis and simulation studies, we find that the proposed method achieves similar statistical accuracy as regularised estimation given a reasonable sample size and is computationally faster. Third, monotone concave CLFs, including the proposed L^p loss functions, are not smooth everywhere. Consequently, the traditional gradient projection algorithms are no longer applicable. Jennrich (2004, 2006) bypassed the computational issue by replacing a CLF with a smooth approximation and pointed out potential issues with this treatment. We propose an Iteratively Reweighted Gradient Projection (IRGP) algorithm that may better solve this nonsmooth optimisation problem. Finally, uncertainty quantification for the rotated solution affects the interpretation of the factors and, thus, is vital in EFA. However, the delta method, which is used to obtain confidence intervals for rotation methods with a smooth objective

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

function (Jennrich, 1973), is not applicable due to the nonsmoothness of the current loss functions. That is, the delta method requires the loss function to be smooth at the true loading matrix, which is not satisfied for monotone concave CLFs. We tackle this problem by developing a post-selection inference procedure that gives asymptotically valid confidence intervals for loadings in a rotated solution. We evaluate the proposed method and compare it with regularised estimation and traditional rotation methods via simulation studies. We further illustrate it using an application to the Big Five personality assessment.

The rest of the paper is structured as follows. In Sect. 1 we propose L^p criteria for oblique rotation, and draw a connection with regularised estimation. In Sect. 2 we discuss statistical inferences based on the proposed rotation method and establish their asymptotic properties, and in Sect. 3 we develop an iteratively reweighted gradient projection algorithm for solving the optimisation problem associated with the proposed rotation criteria. We evaluate the proposed method via simulation studies in Sect. 4 and an application to the Big Five personality assessment in Sect. 5. We conclude this paper with discussions on the limitations of the proposed method and future directions in Sect. 6. Proof of the theoretical results, additional simulation results, and further details of the real application are given in the supplementary material.

1. L^p Rotation Criteria

1.1. Problem Setup

We consider an exploratory linear factor model with J indicators and K factors given by

$$\mathbf{X}|\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{\xi}, \Omega), \tag{1}$$

where $\mathbf{X}$ is a J-dimensional vector of manifest variables, $\boldsymbol{\Lambda} = (\lambda_{jk})_{J\times K}$ is the loading matrix, $\boldsymbol{\xi}$ is a K-dimensional vector of common factors, and $\boldsymbol{\Omega} = (\omega_{ij})_{J\times J}$ denotes the residual covariance matrix. It is assumed that the common factors are normally distributed with variances fixed to 1, i.e, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$, where $\boldsymbol{\Phi} \in \mathbb{R}^{K\times K}$ has diagonal entries ϕ_{kk}, $k = 1, \ldots, K$, equal to 1 and is symmetric positive definite, denoted by $\boldsymbol{\Phi} \succ 0$. The manifest variables are assumed to be conditionally independent given $\boldsymbol{\xi}$, i.e., the off-diagonal entries of Ω are set to 0. To simplify the notation, we use $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$ to denote all of the unknown parameters. The model in (1) implies the marginal distribution of $\mathbf{X}$

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{2}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Omega}$. Without further constraints, the parameters in (2) are not identifiable due to rotational indeterminacy. That is, two sets of parameters $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\Phi}}, \tilde{\boldsymbol{\Omega}})$ give the same distribution for $\mathbf{X}$ if $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' = \tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Lambda}}'$ and $\boldsymbol{\Omega} = \tilde{\boldsymbol{\Omega}}$. Note that the normality assumptions above are not essential. We adopt them for ease of writing, and the development in the current paper does not rely on these normality assumptions. Throughout this paper, we assume that the number of factors K is known.

An oblique rotation method is a two-step procedure. In the first step, one obtains an estimate of the model parameters, under the constraints that $\boldsymbol{\Phi} = \mathbf{I}$ and other arbitrary but mathematically convenient constraints that fix the rotational indeterminacy. Note that due to the rotational indeterminacy, we can always constrain $\boldsymbol{\Phi} = \mathbf{I}$ and absorb the dependence between the factors into the loading matrix $\boldsymbol{\Lambda}$. We can obtain the estimate using any reasonable estimator for factor analysis, such as the least-square (Jöreskog & Goldberger, 1972), weighted-least-square (Browne,

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

1984) , and maximum likelihood estimators (Jöreskog, 1967) . We denote this estimator by $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{A}}, \mathbf{I}, \hat{\boldsymbol{\Omega}})$. In the second step, we find an oblique rotation matrix $\hat{\mathbf{T}}$, such that the rotated loading matrix $\hat{\boldsymbol{\Lambda}} = \hat{\mathbf{A}}\hat{\mathbf{T}}'^{-1}$ minimises a certain loss function Q that measures the sparsity level of a loading matrix. We will propose the functional form of Q in the sequel. Here, an oblique rotation matrix $\mathbf{T}$ satisfies that $\mathbf{T}$ is invertible and $(\mathbf{T}'\mathbf{T})_{kk} = 1,\ k = 1, \ldots, K$. Consequently, any rotated solution $(\hat{\mathbf{A}}\mathbf{T}'^{-1}, \mathbf{T}'\mathbf{T}, \hat{\boldsymbol{\Omega}})$ is still in the parameter space and gives the same distribution for $\mathbf{X}$. More precisely, we let

$$\mathcal{M} = \{\mathbf{T} \in \mathbb{R}^{K\times K} : rank(\mathbf{T}) = K,\ (\mathbf{T}'\mathbf{T})_{kk} = 1,\ k = 1, \ldots, K\} \tag{3}$$

be the space for oblique rotation matrices, where $rank(\cdot)$ gives the rank of a matrix. Then the oblique rotation problem involves solving the optimisation

$$\hat{\mathbf{T}} \in \underset{\mathbf{T}\in\mathcal{M}}{\arg\min}\, Q(\hat{\mathbf{A}}\mathbf{T}'^{-1}), \tag{4}$$

and the rotated solution is given by $(\hat{\mathbf{A}}\hat{\mathbf{T}}'^{-1}, \hat{\mathbf{T}}'\hat{\mathbf{T}}, \hat{\boldsymbol{\Omega}})$. Equivalently, the rotated loading matrix $\hat{\boldsymbol{\Lambda}}$ satisfies

$$(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Phi}}) \in \underset{\boldsymbol{\Lambda},\boldsymbol{\Phi}}{\arg\min}\, Q(\boldsymbol{\Lambda}),\ \text{such that } \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' = \hat{\mathbf{A}}\hat{\mathbf{A}}',\ \boldsymbol{\Phi} \succ 0,\ \text{and } \phi_{kk} = 1, k = 1, \ldots, K. \tag{5}$$

As explained in Remark 1, the minimiser of (4), or equivalently that of (5), is not unique.

Remark 1. Let $\mathcal{D}_1$ be the set of all $K \times K$ permutation matrices and $\mathcal{D}_2$ be the set of all $K \times K$ sign flip matrices. For any $\mathbf{D}_1 \in \mathcal{D}_1$, $\mathbf{D}_2 \in \mathcal{D}_2$, and $K \times K$ matrix $\mathbf{T}$, $\mathbf{T}\mathbf{D}_1$ is a matrix whose columns are a permutation of those of $\mathbf{T}$ and, $\mathbf{T}\mathbf{D}_2$ is a matrix whose kth column is either the same as the kth column of $\mathbf{T}$ or the kth column of $\mathbf{T}$ multiplied by -1. Let $\hat{\mathbf{T}}$ be one solution to the optimisation problem (4). It is easy to check that $\hat{\mathbf{T}}\mathbf{D}_1\mathbf{D}_2$ also minimises the objective function (4), for any $\mathbf{D}_1 \in \mathcal{D}_1$ and $\mathbf{D}_2 \in \mathcal{D}_2$. The resulting loading matrix is equivalent to $\hat{\boldsymbol{\Lambda}}$ up to a column permutation and column sign flips.

We conclude the problem setup with two remarks.

Remark 2. The rotation problem not only applies to the linear factor model, but also other settings, such as item factor analysis (Chen et al., 2019; Chen et al., 2021; Reckase, 2009) and machine learning models such as the stochastic blockmodel and latent Dirichlet allocation (see Rohe & Zeng, 2022). These models are all latent variable models involving manifest variables $\mathbf{X}$, latent variables $\boldsymbol{\xi}$, a parameter matrix $\boldsymbol{\Lambda}$, and possible other model parameters. The parameter matrix $\boldsymbol{\Lambda}$ connects $\mathbf{X}$ and $\boldsymbol{\xi}$, playing a similar role to the loading matrix in the linear factor model. We can view these models as extensions of the linear factor model to more general variable types (e.g., binary or categorical) with more flexible assumptions on the distribution of $(\mathbf{X}, \boldsymbol{\xi})$. We can apply the rotation method to learn an interpretable $\boldsymbol{\Lambda}$ in these models.

Remark 3. Although in the current paper we focus on oblique rotations, we note that the proposed criteria can be easily extended to orthogonal rotation, as the latter can be viewed as a special case of the former when $\boldsymbol{\Phi}$ is fixed to be an identify matrix. That is, given a loss function Q, orthogonal rotation solves the problem

$$\min_{\boldsymbol{\Lambda}} Q(\boldsymbol{\Lambda}),\ \text{such that } \boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \hat{\mathbf{A}}\hat{\mathbf{A}}'.$$

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

1.2. Proposed Rotation Criteria

Jennrich (2004, 2006) proposed a family of monotone concave CLFs for the choice of Q in (4), taking the form

$$Q(\mathbf{\Lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K} h(|\lambda_{jk}|), \tag{6}$$

where $\mathbf{\Lambda} = (\lambda_{jk})_{J\times K}$ and h is a concave and monotone increasing function that maps from $[0, \infty)$ to $[0, \infty)$. This family of loss functions is appealing for several reasons. First, a CLF takes a simple form that does not involve products of loadings and their higher-order polynomial terms. Second, the monotone concave CLFs have desirable properties. In particular, Jennrich (2006) proved that a monotone concave CLF is minimised by loadings with a perfect simple structure when such a loading structure exists. Third, simulation studies in Jennrich (2004, 2006) showed that these loss functions tend to outperform traditional rotation methods (e.g., promax, simplimax, quartimin, and geomin) when the true loading matrix is sparse.

Two examples of h are given in Jennrich (2004, 2006), including the linear CLF where $h(|\lambda|) = |\lambda|$ and the basic CLF where $h(|\lambda|) = 1 - \exp(-|\lambda|)$. However, there does not exist a full spectrum of monotone concave CLFs for dealing with true loading matrices with different sparsity levels. To fill this gap, we propose a general family of monotone concave CLFs that we name the L^p CLFs. More specifically, for each value of $p \in (0, 1]$, the loss function takes the form

$$Q_p(\mathbf{\Lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K} |\lambda_{jk}|^p. \tag{7}$$

Proposition 1 below shows that this choice of h yields a monotone concave CLF.

Proposition 1. *The absolute value function $h(x) = |x|^p$, $p \in (0, 1]$ is monotonically increasing and concave on the interval $[0, \infty)$.*

Under very mild regularity conditions, any L^p CLF is uniquely minimised by a loading matrix of perfect simple structure, when such a loading matrix exists. We say the minimiser is unique when all the minimisers of the loss function are equivalent up to column permutation and sign flip transformations (see Remark 1 for these transformations). We summarise this result in Proposition 2 below. This result improves Theorem 1 of Jennrich (2006), as the uniqueness of the perfect simple structure is not established in Jennrich (2006) for the L^1-criterion.

Proposition 2. *Suppose that the true loading matrix $\mathbf{\Lambda}^*$ has perfect simple structure, in the sense that each row has at most one non-zero entry. Further suppose that $\mathbf{\Lambda}^*$ is of full column rank, i.e., $rank(\mathbf{\Lambda}^*) = K$. Then, for any oblique rotation matrix $\mathbf{T} \in \mathcal{M}$,*

$$Q_p(\mathbf{\Lambda}^*\mathbf{T}'^{-1}) \geq Q_p(\mathbf{\Lambda}^*),$$

where the two sides are equal if, and only if, $\mathbf{T}'^{-1} = \mathbf{D}_1\mathbf{D}_2$ for $\mathbf{D}_1 \in \mathcal{D}_1$ and $\mathbf{D}_2 \in \mathcal{D}_2$; see Remark 1 for the definitions of $\mathcal{D}_1$ and $\mathcal{D}_2$.

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Why do we need the loss functions with $p < 1$, given that the choice of $p = 1$ is already available in Jennrich (2004, 2006)? This is because different L^p CLFs may behave differently when the true loading matrix does not have a perfect simple structure but still contains many zero loadings. Such a loading structure is more likely to be recovered by an L^p CLF when $p < 1$ than by the L^1 CLF. In what follows, we elaborate on this point. Let $\mathbf{\Lambda}^*$ be the true sparse loading matrix and $\mathbf{\Phi}^*$ be the corresponding covariance matrix for the factors. For the true loading matrix $\mathbf{\Lambda}^*$ to be recovered by an L^p CLF, a minimum requirement is that

$$Q_p(\mathbf{\Lambda}^*) = \min_{\mathbf{\Lambda}} Q_p(\mathbf{\Lambda}), \text{ such that there exists } \mathbf{\Phi} \succ 0, \phi_{kk} = 1, k = 1, ..., K, \mathbf{\Lambda}^*\mathbf{\Phi}^*\mathbf{\Lambda}^{*'} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}'. \tag{8}$$

In other words, $\mathbf{\Lambda}^*$ needs to be a stationary point of Q_p. In Fig. 1 we give the plots for $|x|^p$ with different choices of p and their derivatives when $x > 0$. We note that when $p < 1$ the derivative of $|x|^p$ converges to infinity as x approaches zero. The smaller the value of p, the faster the convergence speed is. On the other hand, when $p = 1$, the derivative of $|x|$ takes the value one for any $x > 0$. Therefore, when $\mathbf{\Lambda}^*$ is sparse but does not have a perfect simple structure, it is more likely to be a stationary point of Q_p for $p < 1$ than Q_1. We illustrate this point by a numerical example, where

$$(\mathbf{\Lambda}^*)' = \begin{pmatrix} 1.20 & 0 & 0.15 & 0 & 0.25 & 1.05 & 0.18 \\ 0 & 0.27 & 0 & 1.04 & 0.15 & 1.29 & 0.11 \end{pmatrix}$$

and $\mathbf{\Phi}^*$ is set to be an identity matrix. Note that a 2×2 oblique rotation matrix can be reparameterised by

$$\mathbf{T}(\theta_1, \theta_2) = \begin{pmatrix} \cos(\theta_1) & \sin(\theta_2) \\ \sin(\theta_1) & \cos(\theta_2) \end{pmatrix}$$

for $\theta_1, \theta_2 \in [0, 2\pi)$. In Fig. 2 we show the contour plots of $Q_p(\mathbf{\Lambda}^*(\mathbf{T}(\theta_1, \theta_2))^{-1'})$, with $p = 0.5$ and 1, respectively. The point (0, 0), which is indicated by a black cross, corresponds to $\mathbf{\Lambda} = \mathbf{\Lambda}^*$, and the point indicated by a red point corresponds to the $\mathbf{\Lambda}$ matrix such that $Q_p(\mathbf{\Lambda})$ is minimised. As we can see, when $p = 0.5$, the loss function is minimised by $\mathbf{\Lambda}^*$. On the other hand, when $p = 1$, the minimiser of the loss function is not $\mathbf{\Lambda}^*$ and the resulting solution does not contain as many zeros as $\mathbf{\Lambda}^*$.

We emphasise that due to the singularity of the L^p function near zero when $p < 1$, the optimisation for Q_p tends to be more challenging with a smaller value of p. This is also reflected by the contour plots in Fig. 2, where we see $Q_{0.5}$ is very non-convex, even around the minimiser. On the other hand, Q_1 seems locally convex near the minimiser. Therefore, although the L^p-rotation with $p < 1$ may be better at recovering sparse loading matrices, its computation is more challenging than the L^1-rotation. Thus, the choice of p involves a trade-off between statistical accuracy and computational cost. We have noticed that despite the above counter example, the L^1 criterion tends to give similar results as other L^p criteria ($p < 1$) in most simulation and real-data settings that we have encountered. Considering its computational advantage, we recommend users to always start with the L^1 criterion. Some smaller p values (e.g., $p = 0.5$) may be tried in order to validate the L^1-rotation result. More guidance on the choice of p can be found in Sect. 6. We discuss the computation of the proposed rotation criteria in Sect. 3.

Finally, we remark that when the true loading matrix is sparse but does not have a perfect simple structure, rotation criteria with a smooth objective function (e.g., quartimin and geomin)

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

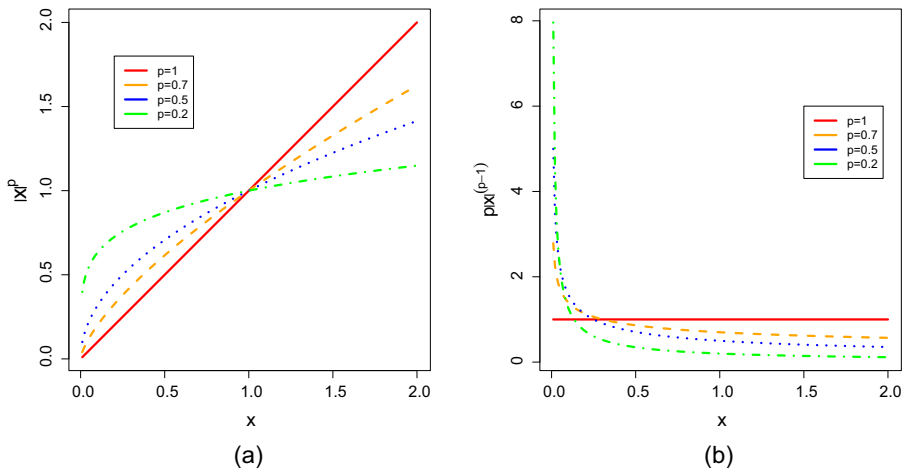


FIGURE 1.
Panel (**a**): Plots of $|x|^p$, for different choices of p. Panel (**b**): Plots of the derivative of $|x|^p$, for different choices of p.

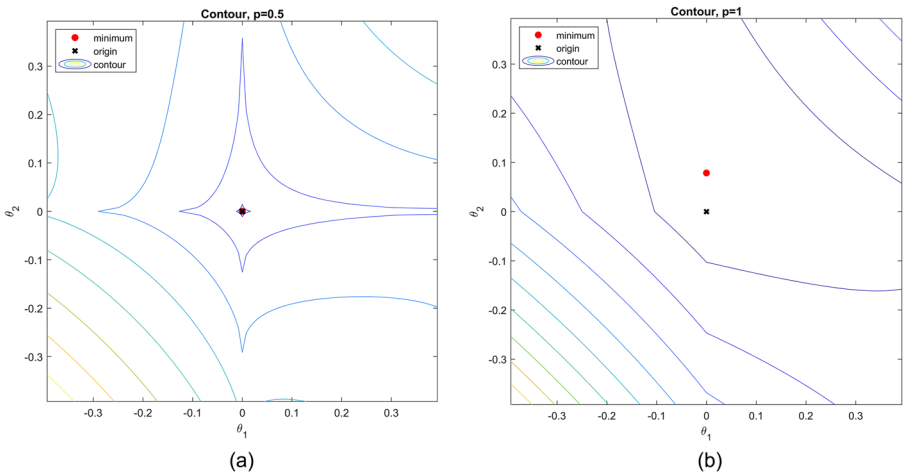


FIGURE 2.
Plots of contours of $|\mathbf{\Lambda}^*\mathbf{T}^{-1'}|^p$, where $\mathbf{T} = [\cos(\theta_1), \sin(\theta_2); \sin(\theta_1), \cos(\theta_2)]$. Panel (**a**): $p = 0.5$. Panel (**b**): $p = 1$.

typically cannot exactly recover the true sparse loading matrix, even when the true loading matrix can be estimated without error. This is due to the fact that a smooth objective function does not discriminate well between zero parameters and close-to-zero parameters. Thus, such rotation criteria do not favour exactly sparse solutions (i.e., with many zero loadings) and only tend to yield approximately sparse solutions (i.e., many small but not exactly zero loadings). Numerical examples illustrating this point are given in Jennrich (2004, 2006) and a new numerical example and associated simulation results are in Appendix H of the supplementary material.

1.3. Connection with Regularised Estimation

The proposed rotation criteria have a close connection with regularised estimators for EFA. In what follows, we establish this connection. Recall that the proposed procedure relies on an initial estimator of the loading matrix for which $\mathbf{\Phi}$ is constrained to be an identity matrix. We further

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

require it to be an M-estimator (Chapter 5, van der Vaart, 2000), obtained by minimising a certain loss function, denoted by $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Note that all the popular EFA estimators are M-estimators. For instance, when the maximum likelihood estimator is used, then the loss function to be minimised is

$$L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) = \log\det(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{S}),$$

where $\boldsymbol{S} = (\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top)/N$ is the sample covariance matrix.

Now we introduce an L^p regularised estimator based on the loss function $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ in the form

$$\hat{\boldsymbol{\theta}}_{\gamma,p} \in \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \gamma \sum_{j=1}^{J}\sum_{k=1}^{K} |\lambda_{jk}|^p, \tag{9}$$

where $\gamma > 0$ is a tuning parameter and the covariance matrix $\boldsymbol{\Phi}$ is estimated rather than constrained to be an identity matrix. We note that the minimiser of (9) is also not unique due to column permutation and sign flips similar to the non-uniqueness of optimisation (4). We denote the set of minimisers as

$$\hat{\mathcal{C}}_{\gamma,p} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \gamma \sum_{j=1}^{J}\sum_{k=1}^{K} |\lambda_{jk}|^p.$$

Note that the regularisation term takes the same form as the L^p CLF. It is used to impose sparsity on the estimate of the loading matrix. When $p = 1$, it becomes a LASSO-regularised estimator that has been considered in, for example, Choi et al. (2010), Hirose and Yamamoto (2014, 2015), Jin et al. (2018), and Geminiani et al. (2021). The regularised estimator (9) is similar in spirit to L^p-regularised regression (Lai & Wang, 2011; Mazumder et al., 2011; Zheng et al., 2017) , where the L^p regularisation with $p < 1$ has been shown to better recover sparse signals under high-dimensional linear regression settings while computationally more challenging (Zheng et al., 2017) .

As summarised in Proposition 3 below, we can view the proposed L^p rotation solution as a limiting case of the L^p-regularised estimator when the tuning parameter γ converges to zero.

Proposition 3. *Consider a fixed $p \in (0, 1]$ and a fixed dataset. Suppose that for any sufficiently small $\gamma > 0$, $\hat{\mathcal{C}}_{\gamma,p}$ only contains $n = 2^K K!$ elements that are equivalent up to column permutation and sign flips of the loading matrix, where $K!$ denotes K factorial that counts the number of all possible permutations and 2^K gives the total number of sign flip transformations. Furthermore, assume that for any sufficiently small $\gamma > 0$, one can label the elements of $\hat{\mathcal{C}}_{\gamma,p}$, denoted by $\hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p}$, $i = 1, ..., n$, such that there exists a sufficiently small constant $\delta > 0$, $\hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p}$ is a continuous and bounded function of γ in $(0, \delta)$, for each i. Then the limit*

$$\hat{\boldsymbol{\theta}}^{(i)}_{0,p} = (\hat{\boldsymbol{\Lambda}}^{(i)}_{0,p}, \hat{\boldsymbol{\Phi}}^{(i)}_{0,p}, \hat{\boldsymbol{\Omega}}^{(i)}_{0,p}) = \lim_{\gamma\to 0} \hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p}$$

exists, and $\hat{\boldsymbol{\theta}}^{(i)}_{0,p}$ satisfies that $(\hat{\boldsymbol{\Lambda}}^{(i)}_{0,p}, \hat{\boldsymbol{\Phi}}^{(i)}_{0,p})$ solves the optimisation problem (5) and $\hat{\boldsymbol{\Omega}}^{(i)}_{0,p} = \hat{\boldsymbol{\Omega}}$, where $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{A}}, \mathbf{I}, \hat{\boldsymbol{\Omega}})$ minimises the loss function $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$.

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

We now discuss the implications of this connection. First, if we have a numerical solver for the regularised estimator (9), then we can obtain an approximate solution to the L^p-rotation problem (5) by using a sufficiently small tuning parameter γ. Second, thanks to this connection, the choice between regularised estimation and rotation becomes the choice of the tuning parameter in regularised estimation. Note that the tuning parameter γ corresponds to a bias-variance trade-off in estimating the model parameters $\boldsymbol{\theta}$. As γ increases, the bias of the regularised estimator also increases and the variance decreases. In applications where the sample size is large relative to the number of model parameters, the optimal choice of the tuning parameter is often close to zero. In that case, it is a good idea to use the rotation method, as the regularised estimator under the optimal tuning parameter may not be substantially more accurate than the rotation solution and searching for the optimal tuning parameter can be computationally costly. We further discuss this point in a simulation study in Sect. 4. We will discuss the computation of these methods in Sect. 3.

2. Statistical Inference and Asymptotic Theory

2.1. Estimation Consistency

We establish the statistical consistency of the proposed estimator based on the L^p rotation. Suppose the true parameter set that we aim to recover is $(\boldsymbol{\Lambda}^*, \boldsymbol{\Phi}^*, \boldsymbol{\Omega}^*)$, where the true loading matrix $\boldsymbol{\Lambda}^*$ is sparse. To emphasise the dependence on the sample size, we attach the sample size N as a subscript to the initial estimator in the first step of the rotation method; that is, $\hat{\boldsymbol{\theta}}_N = (\hat{\mathbf{A}}_N, \mathbf{I}, \hat{\boldsymbol{\Omega}}_N)$. We require the initial estimator to be consistent, in the sense that

C1. $\hat{\mathbf{A}}_N \hat{\mathbf{A}}_N' \xrightarrow{pr} \boldsymbol{\Lambda}^* \boldsymbol{\Phi}^* \boldsymbol{\Lambda}^{*\prime}$ and $\hat{\boldsymbol{\Omega}}_N \xrightarrow{pr} \boldsymbol{\Omega}^*$, where the notation "$\xrightarrow{pr}$" denotes convergence in probability.

This requirement easily holds when the linear factor model is correctly specified and the loss function $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ is reasonable (e.g., the negative log-likelihood). In addition, we require that the EFA model is truly a K-dimensional model, in the sense that condition C2 holds.

C2. $rank(\boldsymbol{\Lambda}^* \boldsymbol{\Phi}^* \boldsymbol{\Lambda}^{*\prime}) = K$.

For the L^p rotation estimator to be consistent, for a specific value of $p \in (0, 1]$, we further require that the true loading matrix uniquely minimises the L^p CLF, in the sense of condition C3 below.

C3. $(\boldsymbol{\Lambda}^*, \boldsymbol{\Phi}^*) \in \arg\min_{\boldsymbol{\Lambda}, \boldsymbol{\Phi}} Q_p(\boldsymbol{\Lambda})$ such that $\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}' = \boldsymbol{\Lambda}^* \boldsymbol{\Phi}^* \boldsymbol{\Lambda}^{*\prime}$. In addition, for any other $(\boldsymbol{\Lambda}^\dagger, \boldsymbol{\Phi}^\dagger) \in \arg\min_{\boldsymbol{\Lambda}, \boldsymbol{\Phi}} Q_p(\boldsymbol{\Lambda})$ such that $\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}' = \boldsymbol{\Lambda}^* \boldsymbol{\Phi}^* \boldsymbol{\Lambda}^{*\prime}$, there exist $\mathbf{D} \in \mathcal{D}_1$ and $\tilde{\mathbf{D}} \in \mathcal{D}_2$, such that $\boldsymbol{\Lambda}^\dagger \mathbf{D} \tilde{\mathbf{D}} = \boldsymbol{\Lambda}^*$ and $\tilde{\mathbf{D}}^{-1} \mathbf{D}^{-1} \boldsymbol{\Phi}^\dagger (\mathbf{D}^{-1})' (\tilde{\mathbf{D}}^{-1})' = \boldsymbol{\Phi}^*$. Recall that $\mathcal{D}_1$ and $\mathcal{D}_2$ are the sets of column permutation and sign flip transformations, respectively, which we gave in Remark 1.

Condition C3 tends to hold when the true loading matrix contains many zeros, as the L^p loss function is a good approximation to the L^0 function that counts the number of non-zero elements. In particular, according to Proposition 2, condition C3 is guaranteed to hold when $\boldsymbol{\Lambda}^*$ has a perfect simple structure, i.e., if it has at most one non-zero loading in each row. As we discussed in Sect. 1.2, this condition is more likely to hold for a smaller value of p, when there are cross-loadings. Conditions C1 through C3 guarantee the estimation consistency of the L^p rotation estimator, up to column permutation and sign flips. We summarise this result in Theorem 1 below.

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Theorem 1. *Suppose that for a given* $p \in (0, 1]$ *conditions C1 through C3 hold. Then there exist* $\mathbf{D}_N \in \mathcal{D}_1$ *and* $\tilde{\mathbf{D}}_N \in \mathcal{D}_2$*, such that* $\hat{\mathbf{\Lambda}}_{N,p}\mathbf{D}_N\tilde{\mathbf{D}}_N \xrightarrow{pr} \mathbf{\Lambda}^*$ *and* $\tilde{\mathbf{D}}_N^{-1}\mathbf{D}_N^{-1}\hat{\mathbf{\Phi}}_{N,p}(\mathbf{D}_N^{-1})'(\tilde{\mathbf{D}}_N^{-1})' \xrightarrow{pr} \mathbf{\Phi}^*$*, where*

$$(\hat{\mathbf{\Lambda}}_{N,p}, \hat{\mathbf{\Phi}}_{N,p}) \in \underset{\mathbf{\Lambda},\mathbf{\Phi}}{\arg\min}\, Q_p(\mathbf{\Lambda}), \text{ such that } \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' = \hat{\mathbf{A}}_N\hat{\mathbf{A}}_N'.$$

2.2. Model Selection

The interpretation of the factors relies on the sign pattern of the loading matrix, so that we can interpret each factor based on the associated manifest variables and their directions (positive or negative associations). Learning this sign pattern is a model selection problem. A regularised estimator may seem advantageous as it yields simultaneous parameter estimation and model selection. We note that, however, we can easily achieve model selection with a rotation method, using a Hard-Thresholding (HT) procedure. Similar HT procedures have been proven to be successful in the model selection for linear regression models (Meinshausen and Yu, 2009).

More precisely, let $\mathbf{\Gamma}^* = \left(\text{sgn}(\lambda^*_{jk})\right)_{J\times K}$ denote the true sign pattern of $\mathbf{\Lambda}^*$, where $\text{sgn}(x)$ returns the sign of a scalar satisfying that

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Given the L^p rotation estimator $\hat{\mathbf{\Lambda}}_{N,p} = \left(\hat{\lambda}^{(N,p)}_{jk}\right)_{J\times K}$, the HT procedure estimates the pattern of $\mathbf{\Gamma}^*$ by $\hat{\mathbf{\Gamma}}_{N,p} = \left(\text{sgn}(\hat{\lambda}^{(N,p)}_{jk}) \times 1_{\{|\hat{\lambda}^{(N,p)}_{jk}|>c\}}\right)_{J\times K}$, where $c > 0$ is a pre-specified threshold. If we choose the threshold c properly, then $\hat{\mathbf{\Gamma}}_{N,p}$ consistently estimates $\mathbf{\Gamma}^*$. We state this result in Theorem 2 below.

C4. The threshold c lies in the interval $(0, c_0)$, where $c_0 = \min\{|\lambda^*_{jk}| : \lambda^*_{jk} \neq 0\}$.

Theorem 2. *Suppose that for a given* $p \in (0, 1]$ *conditions C1 through C4 hold. Then there exist* $\mathbf{D}_N \in \mathcal{D}_1$ *and* $\tilde{\mathbf{D}}_N \in \mathcal{D}_2$*, such that the probability* $P(\hat{\mathbf{\Gamma}}_{N,p}\mathbf{D}_N\tilde{\mathbf{D}}_N = \mathbf{\Gamma}^*)$ *converges to 1 as the sample size N goes to infinity.*

In practice, the value of c_0 is unknown and thus cannot be used for choosing the threshold c. Instead, we choose c based on the Bayesian Information Criterion (BIC; Schwarz, 1978). We summarise the steps of this procedure in Algorithm 1 below, where we simplify the notation for ease of exposition.

When the candidate values of c are chosen properly (i.e., $\mathcal{C}$ includes values that are below c_0), then Theorem 2 implies that with probability tending to one, the true model will be in the candidate models. Together with the consistency of BIC for parametric models (Shao, 1997; Vrieze, 2012) , the true non-zero pattern can be consistently recovered. We remark that it may not be a good idea to manually select c or use some default thresholds. Unless there is very good substantive knowledge about the latent structure, it is very likely to under- or over-select c, leading to high false-positive and false-negative errors. Even with the proposed procedure, the selection consistency is only guaranteed when the sample size goes to infinity. For a finite sample, the false-positive and false-negative errors likely exist and thus we should look at the selected model with caution. Furthermore, we note that the BIC is not the only information criterion that

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Algorithm 1
Hard-thresholding for model selection based on L^p rotation

Input: A sequence of candidate thresholds $\mathcal{C}$, observed data, and the rotated loading matrix $\hat{\mathbf{\Lambda}} = (\hat{\lambda}_{jk})_{J\times K}$ given by the L^p CLF criterion.

For each value of $c \in \mathcal{C}$, we perform the following two steps:

Step 1: Obtain the corresponding selected loading structure $\hat{\mathbf{\Gamma}}_c = \left(\text{sgn}(\hat{\lambda}_{jk}) \times 1_{\{|\hat{\lambda}_{jk}|>c\}}\right)_{J\times K}$.

Step 2: Fit a Confirmatory Factor Analysis (CFA) model based on $\hat{\mathbf{\Gamma}}_c$ using the maximum likelihood estimator, in which the (i, j)th loading parameter satisfies the sign constraint implied by the corresponding entry of $\hat{\mathbf{\Gamma}}_c$. Calculate the BIC value for this CFA model, denoted by BIC_c.

Obtain $\hat{c} = \arg\min_{c\in\mathcal{C}} \text{BIC}_c$.

Output: The selected sign pattern $\hat{\mathbf{\Gamma}}_{\hat{c}}$.

leads to model selection consistency (Nishii, 1984), but it is probably the most commonly used information criterion with consistency guarantee. Another commonly used information criterion is the Akaike Information Criterion (AIC) which tends to over-select and thus does not guarantee model selection consistency (Shao, 1997).

2.3. *Confidence Intervals*

Often, we are not only interested in the point estimate of the underlying sparse loading matrix, but also in quantifying its uncertainty. We typically achieve uncertainty quantification by constructing confidence intervals for the loadings of the rotated solution. Traditionally, we can do this by establishing the asymptotic normality of the rotated loading matrix using the delta method, which involves calculating the partial derivatives of a rotation algorithm using implicit differentiation (Jennrich, 1973). Unfortunately, this procedure is no longer suitable if the true loading matrix is sparse and the loss function is not differentiable with respect to the zero loadings.

Motivated by a simple but nevertheless well-performing post-selection inference procedure in regression analysis (Zhao et al., 2021), we propose a procedure for constructing confidence intervals for individual loading parameters of the rotated solution. More precisely, this procedure runs a loop over all the manifest variables, $j = 1, ..., J$. Each time, the procedure obtains the confidence intervals for the loading parameters of manifest variable j by fitting a CFA model whose loading structure is determined by the selected sign pattern of the remaining $J-1$ manifest variables. More precisely, the loading parameters of the CFA model satisfy the sign constraints imposed by the selected sign pattern $\hat{\mathbf{\Gamma}}_{\hat{c}}$ from Algorithm 1, for all the items except for j. We impose no constraint on the loading parameters of item j. We obtain confidence intervals for the loading parameters of item j based on the asymptotic normality of the estimator for this CFA model. We summarise this procedure in Algorithm 2 below.

In what follows, we establish the consistency of confidence intervals given by Algorithm 2. To emphasise that the statistics in Algorithm 2 depend on the sample size N, we attach N as a subscript or superscript when describing this consistency result. We require the following conditions:

C5. The selected sign pattern $\hat{\mathbf{\Gamma}}_N$ is consistent. That is, there exist $\mathbf{D}_N \in \mathcal{D}_1$ and $\tilde{\mathbf{D}}_N \in \mathcal{D}_2$, such that the probability $P(\hat{\mathbf{\Gamma}}_{N,p}\mathbf{D}_N\tilde{\mathbf{D}}_N = \mathbf{\Gamma}^*)$ converges to 1 as the sample size N goes to infinity.

Thanks to the consistency of BIC selection, and when we have chosen the candidate thresholds properly, condition C5 holds if $\hat{\mathbf{\Gamma}}_N$ is obtained by Algorithm 1.

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Algorithm 2
Post-selection confidence intervals

Input: The selected sign pattern $\hat{\mathbf{\Gamma}} = (\hat{\gamma}_{jk})_{J\times K}$, observed data, and significance level $\alpha \in (0, 1)$.

For each manifest variable $s = 1, ..., J$, we perform the following two steps:

Step 1: Obtain a CFA model whose loadings λ_{jk} satisfy the constraints that $\text{sgn}(\lambda_{jk}) = \hat{\gamma}_{jk}$ for all $j \neq s$ and for all k.
Step 2: Fit the CFA model and obtain the $(1-\alpha)$-confidence intervals for parameters $\lambda_{s1}, ..., \lambda_{sK}$ using a standard inference procedure for CFA (e.g., based on the maximum likelihood estimator). We denote these confidence intervals by (l_{sk}, u_{sk}). If the CFA model in **Step 1** is not identifiable, we let the confidence intervals be $(-\infty, \infty)$.

Output: Confidence intervals (l_{sk}, u_{sk}), $s = 1, ..., J, k = 1, ..., K$.

C6. For each manifest variable $s = 1, ..., J$, the CFA model whose loading parameters satisfy $\text{sgn}(\lambda_{jk}) = \text{sgn}(\lambda^*_{jk})$ for all $j \neq s$ is identifiable, and using the same procedure in Step 2 of Algorithm 2 leads to consistent confidence intervals for λ_{s1},..., λ_{sK}. That is, let $(l^{*(N)}_{sk}, u^{*(N)}_{sk})$ be the resulting confidence interval for λ_{sk}, then $P(\lambda^*_{sk} \in (l^{*(N)}_{sk}, u^{*(N)}_{sk}))$ converges to $1-\alpha$, as the sample size N goes to infinity.

Note that C6 is a condition imposed on the sign pattern of the true loading matrix. It essentially requires that the factors can be identified by the sign pattern of any $(J-1)$-subset of the manifest variables. Given an identified CFA model, we can easily construct the consistent confidence intervals based on the asymptotic normality of any reasonable estimator of the CFA model, e.g., the maximum likelihood estimator. Under conditions C5 and C6, the following theorem holds.

Theorem 3. *Suppose that conditions C5 and C6 hold for the selected sign pattern $\hat{\mathbf{\Gamma}}_N$ and the true model, where $\mathbf{D}_N \in \mathcal{D}_1$ and $\tilde{\mathbf{D}}_N \in \mathcal{D}_2$ are from condition C5. Suppose we input $\hat{\mathbf{\Gamma}}_N$, observed data from the true model, and significance level α into the true model, and obtain output $(l^{(N)}_{sk}, u^{(N)}_{sk})$, $s = 1, ..., J, k = 1, ..., K$. Then we have $P(\lambda^{*(N)}_{sk} \in (l^{(N)}_{sk}, u^{(N)}_{sk}))$ converges to $1-\alpha$, for all $s = 1, ..., J, k = 1, ..., K$, where $\lambda^{*(N)}_{sk}$ are entries of $\mathbf{\Lambda}^{*(N)} = \mathbf{\Lambda}^*\tilde{\mathbf{D}}_N^{-1}\mathbf{D}_N^{-1}$. Note that $\mathbf{\Lambda}^{*(N)}$ is equivalent to $\mathbf{\Lambda}^*$ up to column permutation and sign flips.*

We remark that under the conditions of Theorem 3, all the CFA models fitted in Step 2 of Algorithm 2 should be identifiable for sufficiently large N. However, in practice, it may happen that some CFA models are not identifiable, either due to the sample size not being large enough or the regularity conditions C5 or C6 not holding. In such cases, we set the corresponding confidence intervals to be $(-\infty, \infty)$ as a conservative choice.

3. Computation

3.1. Proposed IRGP Algorithm

We now discuss the computation for the proposed rotation. Recall that we aim to solve the optimisation problem

$$\hat{\mathbf{T}} \in \underset{\mathbf{T}\in\mathcal{M}}{\arg\min}\, Q_p(\hat{\mathbf{A}}\mathbf{T}'^{-1}),$$

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

where Q_p is the L^p CLF defined in (7). Note that this objective function is not differentiable when $\hat{\mathbf{A}}\mathbf{T}'^{-1}$ has zero elements, as the L^p function is not smooth at zero. Consequently, standard numerical solvers fail, especially when the true solution is approximately sparse. To solve this optimisation problem, we develop an IRGP algorithm that combines the iteratively reweighted least square algorithm (Ba et al., 2013; Daubechies et al., 2010) and the gradient projection algorithm (Jennrich, 2002) .

Similar to Jennrich (2006), the IRGP algorithm also solves a smooth approximation to the objective function $Q_p(\hat{\mathbf{A}}\mathbf{T}'^{-1})$. That is, we introduce a sufficiently small constant $\epsilon > 0$, and approximate the objective function by $Q_{p,\epsilon}(\hat{\mathbf{A}}\mathbf{T}'^{-1})$, where

$$Q_{p,\epsilon}(\mathbf{\Lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K}(\epsilon^2 + \lambda_{jk}^2)^{\frac{p}{2}}.$$

As we discuss in the sequel, the ϵ is introduced to make the computation more robust. The IRGP algorithm alternates between two steps – (1) function approximation step and (2) Projected Gradient Descent (PGD) step. More precisely, let $\mathbf{T}_t$ be the parameter value at the tth iteration. The function approximation step involves approximating the objective function by

$$G_t(\mathbf{T}) = \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}^{(t)}\left((\hat{\mathbf{A}}\mathbf{T}'^{-1})_{jk}\right)^2, \tag{10}$$

where the weights $w_{jk}^{(t)}$ are given by

$$w_{jk}^{(t)} = \frac{1}{((\hat{\mathbf{A}}(\mathbf{T}_t')^{-1})_{jk}^2 + \epsilon^2)^{1-p/2}}.$$

Here $\epsilon > 0$ is a pre-specified parameter that is chosen to be sufficiently small. We provide some remarks about this approximation. First, the small tuning parameter is chosen to stabilise the algorithm when certain $\hat{\mathbf{A}}(\mathbf{T}_t')^{-1})_{jk}$s are close to zero. Without ϵ, the weight $w_{jk}^{(t)}$ can become very large, resulting in an unstable PGD step. Second, supposing that $(\hat{\mathbf{A}}(\mathbf{T}_t')^{-1})_{jk} \neq 0$ for all j and k, then $G_t(\mathbf{T}_t) \approx Q_p(\hat{\mathbf{A}}(\mathbf{T}_t')^{-1})$ when ϵ is sufficiently small, i.e., the function approximation and the objective function value are close to each other at the current parameter value. Lastly, this approximation is similar to the E-step of the Expectation-Maximisation algorithm (Dempster et al., 1977) ; see Ba et al. (2013) for this correspondence.

The PGD step involves updating the parameter value based on the $G_t(\mathbf{T})$ via projected gradient descent. This step is similar to the update in each iteration of the gradient projection algorithm for oblique rotations (Jennrich, 2002). We can perform PGD on $G_t(\mathbf{T})$, as this function approximation is smooth in $\mathbf{T}$. More precisely, we define a projection operator as

$$\text{Proj}(\mathbf{T}) = \mathbf{T}(\text{diag}(\mathbf{T}'\mathbf{T}))^{-\frac{1}{2}}, \tag{11}$$

where $(\text{diag}(\mathbf{T}'\mathbf{T}))^{-\frac{1}{2}}$ is a diagonal matrix whose ith diagonal entry is given by $1/\sqrt{(\mathbf{T}'\mathbf{T})_{ii}}$. This operator projects any invertible matrix into the space of oblique rotation matrices $\mathcal{M}$ as defined in (3). The PGD update is given by

$$\mathbf{T}_{t+1} = \text{Proj}(\mathbf{T}_t - \alpha\nabla G_t(\mathbf{T})), \tag{12}$$

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Algorithm 3
IRGP algorithm for L^p rotation

Input: The initial loading matrix estimate $\hat{\mathbf{A}}$, parameter $\epsilon > 0$, and an initial value $\mathbf{T}_0$.

For iterations $t = 0, 1, 2, ...$, we iterate between the following two steps:

Step 1: Construct $G_t(\mathbf{T})$ using equation (10).
Step 2: Obtain $\mathbf{T}_{t+1}$ using equation (12), where the step size α is chosen by line search.

Stop when the convergence criterion is met. Let t_{max} be the final iteration number.

Output: $\mathbf{T}_{t_{max}}$.

where $\alpha > 0$ is a step size chosen by line search and $\nabla G_t(\mathbf{T})$ is a $K \times K$ matrix whose (i, j)th entry is the partial derivative of $G_t(\mathbf{T})$ with respect to the (i, j)th entry of $\mathbf{T}$. We summarize the IRGP algorithm in Algorithm 3.

Under reasonable regularity conditions (Ba et al., 2013), every limit point of $\{\mathbf{T}_t\}_{t=1}^{\infty}$ will be a stationary point of the approximated objective function $Q_{p,\epsilon}(\hat{\mathbf{A}}\mathbf{T}'^{-1})$. In addition, the algorithm has local linear convergence when $p = 1$ and super-linear convergence when $0 < p < 1$.

We remark on the choice of initial value $\mathbf{T}_0$ when $0 < p < 1$. As discussed previously in Sect. 1.2, when $0 < p < 1$, the objective function $Q_p(\hat{\mathbf{A}}\mathbf{T}'^{-1})$ is highly non-convex and thus may contain many stationary points. To avoid the algorithm getting stuck at a local optimum, the choice of $\mathbf{T}_0$ is important. When solving the optimisation for a smaller value of p, we recommend using the solution from a larger value of p as the starting point (e.g., $p = 1$).

3.2. Comparison with Regularised Estimation

To compare the computation of the proposed rotation method and that of regularised estimation, we also describe a proximal gradient algorithm for the L^1 regularised estimator. The proximal algorithm is a state-of-the-art algorithm for solving nonsmooth optimisation problems (Parikh & Boyd, 2014) . We can view it as a generalisation of projected gradient descent. As we will discuss below, each iteration of the algorithm can be computed easily. In principle, we can also apply the proximal algorithm to the L^p regularised estimator, for $0 < p < 1$. However, it is computationally much more costly than the case when $p = 1$, and thus, will not be discussed here.

The L^1 regularised estimator, also referred to as the LASSO estimator, solves the following optimisation problem:

$$\min_{\boldsymbol{\theta}} \; L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \gamma \sum_{j=1}^{J} \sum_{k=1}^{K} |\lambda_{jk}|.$$

To apply the proximal gradient algorithm, we reparameterise the covariance matrix $\boldsymbol{\Phi}$ by $\mathbf{T}'\mathbf{T}$, where we let $\mathbf{T}$ be an upper triangular matrix to ensure its identifiability. We also reparameterise the diagonal entries of the diagonal matrix $\boldsymbol{\Omega}$ by $\mathbf{v} = (v_1, ..., v_J)$, where $v_i = \log(\omega_{ii})$. With slight abuse of notation, we can write the optimisation problem as

$$\min_{\boldsymbol{\Lambda},\mathbf{T},\mathbf{v}} \; L(\boldsymbol{\Sigma}(\boldsymbol{\Lambda}, \mathbf{T}, \mathbf{v})) + \gamma \sum_{j=1}^{J} \sum_{k=1}^{K} |\lambda_{jk}|.$$

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Algorithm 4
Proximal gradient algorithm for L^1 regularised estimation.

Input: The initial values $\mathbf{\Lambda}_0$, $\mathbf{T}_0$, and $\mathbf{v}_0$.

For iterations $t = 0, 1, 2, ...$, we iterate between the following two steps:

Step 1: Calculate the gradients of $L(\mathbf{\Sigma}(\mathbf{\Lambda}, \mathbf{T}, \mathbf{v}))$ with respect to $\mathbf{\Lambda}$, $\mathbf{T}$, and $\mathbf{v}$, respectively, at $(\mathbf{\Lambda}_t, \mathbf{T}_t, \mathbf{v}_t)$. Denote these gradients by $\nabla L_{t,\mathbf{\Lambda}}$ $\nabla L_{t,\mathbf{T}}$, and $\nabla L_{t,\mathbf{v}}$.
Step 2: Update the parameters by

$$\mathbf{\Lambda}_{t+1} = \text{Prox}_{\alpha,\gamma}(\mathbf{\Lambda}_t - \alpha \nabla L_{t,\mathbf{\Lambda}}),$$
$$\mathbf{T}_{t+1} = \text{Proj}(\mathbf{T}_t - \alpha \nabla L_{t,\mathbf{T}}),$$

and

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \alpha \nabla L_{t,\mathbf{v}}.$$

Recall that the operator Proj$(\cdot)$ is defined in (11), and α is a step size chosen by line search.

Stop when the convergence criterion is met. Let t_{max} be the final iteration number.

Output: $(\mathbf{\Lambda}_{t_{max}}, \mathbf{T}_{t_{max}}, \mathbf{v}_{t_{max}})$.

We define a proximal operator for the loading matrix as

$$\text{Prox}_{\alpha,\gamma}(\tilde{\mathbf{\Lambda}}_t) = \arg\min_{\mathbf{\Lambda}} \ \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}(\lambda_{jk} - \tilde{\lambda}_{jk}^{(t)})^2 + \alpha\gamma \sum_{j=1}^{J}\sum_{k=1}^{K}|\lambda_{jk}|, \tag{13}$$

where $\alpha > 0$ is the step size and $\tilde{\mathbf{\Lambda}}_t = (\tilde{\lambda}_{jk}^{(t)})_{J\times K}$ is the value of $\mathbf{\Lambda}$ from the previous step in the proximal gradient algorithm. Note that (13) has a closed-form solution given by soft-thresholding (Parikh & Boyd, 2014) that we can easily compute. We summarise the proximal gradient algorithm in Algorithm 4 below.

Under suitable conditions, this proximal gradient algorithm converges to stationary points of the objective function and has a local linear convergence rate (Karimi et al., 2016) . We notice that when $p = 1$, Algorithms 3 and 4 have similar convergence properties. However, their per-iteration computational complexities are different. In particular, Algorithm 4 involves parameters $\mathbf{\Lambda}$ and $\mathbf{v}$, which substantially increases its computational complexity. In fact, the per-iteration complexity for Algorithm 3 is $O(K^3 + K^2J)$, while that for Algorithm 4 is $O(J^3 + J^2K + K^2J + K^3)$. The difference can be substantial when J is much larger than K. We give the derivation of these computational complexities in the supplementary material.

4. Simulation Study

4.1. Study I

In this study, we evaluate the performance of $L^{0.5}$ and L^1 rotations and compare them with some traditional rotation methods and L^1-regularised estimation. We consider several traditional oblique rotation methods, including the oblimin, quartimin, simplimax, geomin, and promax

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

methods. These methods have been considered in the simulation studies in Jennrich (2006). They are implemented using the `GPArotation` package (Bernaards & Jennrich, 2005) in R.

Settings. We consider two simulation settings, one with $J = 15$ manifest variables and $K = 3$ factors, and the other with $J = 30$ and $K = 5$. The first setting has nine manifest variables each loading on a single factor (three variables for each factor), and six manifest variables each loading on two factors. The second setting has 15 manifest variables each loading on a single factor (three variables for each factor), 10 manifest variables each loading on two factors, and 5 manifest variables each loading on three factors. We give the true model parameters in the supplementary material. By numerical evaluations, the true loading matrices satisfy condition C3 for both $L^{0.5}$ and L^1 criteria. Under each setting, we consider three sample sizes, including $N = 400$, 800, and 1600. For each setting and each sample size, we run $B = 500$ independent replications.

Evaluation criteria. We evaluate the proposed method from three aspects. First, we compare all estimators in terms of accuracy of point estimation. Second, we compare the proposed method and the L^1-regularised estimator in terms of their model selection accuracy. Finally, we examine the coverage rate of the proposed method for constructing confidence intervals.

When evaluating the performance of different estimators, we take into account the indeterminacy due to column permutations and sign flips. Let $\tilde{\mathbf{\Lambda}}^{(b)}$ be the loading matrix estimate given by a rotation or regularised estimation method in the b-th replication. We then find

$$\hat{\mathbf{\Lambda}}^{(b)} = \arg\min_{\mathbf{\Lambda}} \{\|\mathbf{\Lambda} - \mathbf{\Lambda}^*\|^2 : \mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}^{(b)}\mathbf{D}\tilde{\mathbf{D}}, \mathbf{D} \in \mathcal{D}_1, \tilde{\mathbf{D}} \in \mathcal{D}_2\},$$

which is the one closest to the true loading matrix $\mathbf{\Lambda}^*$ among all the loading matrices that are equivalent to $\tilde{\mathbf{\Lambda}}^{(b)}$. Our evaluation criteria are constructed based on $\hat{\mathbf{\Lambda}}^{(b)}$:

1. The accuracy of point estimation is estimated by the mean squared error (MSE):

$$\text{MSE} = \frac{||\hat{\mathbf{\Lambda}}^{(b)} - \mathbf{\Lambda}^*||_F^2}{JK},$$

   where $\hat{\mathbf{\Lambda}}^{(b)}$ is obtained by a certain rotation or regularisation method in the b-th replication.
2. The model selection accuracy is assessed using the area under the curve (AUC) from the corresponding receiver operating characteristic (ROC) curve. For each threshold c, we compute the average true positive rate ($\overline{\text{TPR}}_c$), which is the proportion of successfully identified non-zero elements in the true loading matrix:

$$\overline{\text{TPR}}_c = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{j,k} \mathbb{1}_{\{\hat{\lambda}_{jk}^{(b,c)} \neq 0, \lambda_{jk}^* \neq 0\}}}{\sum_{j,k} \mathbb{1}_{\{\lambda_{jk}^* \neq 0\}}}, \tag{14}$$

   where $\{\hat{\lambda}_{jk}^{(b,c)}\}_{J\times K} = \hat{\mathbf{\Lambda}}^{(b,c)}$ is the estimated loading matrix in the b-th replication from a CFA model based on $\hat{\mathbf{\Gamma}}_c$ using the maximum likelihood estimator. Similarly, we calculate the average true negative rate ($\overline{\text{TNR}}_c$), which is the success rate of identifying zero elements:

$$\overline{\text{TNR}}_c = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{j,k} \mathbb{1}_{\{\hat{\lambda}_{jk}^{(b,c)} = 0, \lambda_{jk}^* = 0\}}}{\sum_{j,k} \mathbb{1}_{\{\lambda_{jk}^* = 0\}}}. \tag{15}$$

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

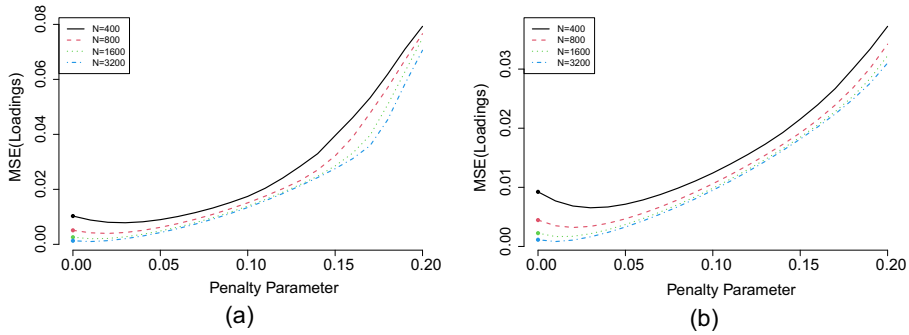


FIGURE 3.
The MSE (for loadings) as a function of the tuning parameter γ in the LASSO-regularised estimator. **a** 15×3 settings. **b** 30×5 settings. The dots at $\gamma = 0$ correspond to the L^1 rotation solutions.

The AUC is consequently calculated by plotting $\overline{\text{TPR}}_c$ against $1 - \overline{\text{TNR}}_c$ by varying the threshold c. We also use the overall selection accuracy, i.e., the true selection rate (TR), to evaluate the model selection procedure described in Algorithm 1. The TR is calculated as

$$\text{TR} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{j,k} \mathbb{1}_{\{\hat{\lambda}_{jk}^{(b,\hat{c})} \neq 0, \lambda_{jk}^* \neq 0\}} + \sum_{j,k} \mathbb{1}_{\{\hat{\lambda}_{jk}^{(b,\hat{c})} = 0, \lambda_{jk}^* = 0\}}}{JK},$$

where $\hat{c}$ is the BIC selected threshold value from Algorithm 1. Correspondingly, we calculate the TPR and TNR of the selected model as

$$\text{TPR} = \overline{\text{TPR}}_{\hat{c}} \text{ and } \text{TNR} = \overline{\text{TNR}}_{\hat{c}}.$$

3. The entry-wise 95% confidence interval coverage rate (ECIC) is calculated to evaluate the performance of our post-selection confidence interval procedure in Algorithm 2. For each entry of the loading matrix, the empirical probability of the true loading falling within the estimated confidence interval is calculated as

$$\text{ECIC}_{jk} = \frac{\sum_{b=1}^{B} \mathbb{1}_{\{\lambda_{jk}^{*(N)} \in (l_{jk}^{(N)}, u_{jk}^{(N)})\}}}{B}.$$

Results on point estimation. In Table 1, we present the MSE of the estimated loading matrix, for both simulation settings and $N \in \{400, 800, 1600\}$. In the first five rows we show the results based on traditional oblique rotation criteria, followed by the results of the proposed L^p loss function for two choices of p, and finally those of the LASSO estimator for five choices of γ. For both settings and all sample sizes, geomin performed the best among the traditional rotation methods. The geomin results were very similar to those of L^p rotation and the LASSO estimator with sufficiently small tuning parameter γ. For the LASSO estimator, the MSE increased as γ increased. For L^p rotation, we observed only very small differences between $p = 0.5$ and $p = 1$. In addition, their MSEs were close to those of the LASSO estimator with $\gamma = 0.01$ and $\gamma = 0.05$.

Results on model selection. In Table 2, we present the AUC, TR, TPR, and TNR for the L^p rotations and the LASSO estimator with different tuning parameters. For both scenarios and

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

TABLE 1.
MSE obtained by using different rotation criteria under various settings, Study I.

	15×3			30×5		
	$N = 400$	$N = 800$	$N = 1600$	$N = 400$	$N = 800$	$N = 1600$
Oblimin	0.012	0.007	0.004	0.012	0.008	0.006
GeominQ	0.010	0.005	0.002	0.010	0.005	0.002
Promax	0.013	0.007	0.005	0.014	0.009	0.007
$L^{0.5}$ rotation	0.011	0.005	0.003	0.009	0.005	0.002
L^1 rotation	0.010	0.005	0.003	0.010	0.004	0.002
LASSO, $\gamma = 0.01$	0.009	0.004	0.002	0.008	0.003	0.002
LASSO, $\gamma = 0.05$	0.009	0.006	0.005	0.007	0.005	0.004
LASSO, $\gamma = 0.1$	0.017	0.015	0.014	0.012	0.011	0.010
LASSO, $\gamma = 0.2$	0.079	0.076	0.074	0.038	0.034	0.032
LASSO, $\gamma = 0.5$	0.244	0.244	0.244	0.144	0.149	0.150

all sample sizes, the AUC and TR were very similar for the rotation estimator with $p = 0.5$ and $p = 1$. The AUC of the LASSO estimator with a small tuning parameter is similar to that of the L^1 rotation method. We noted that the model selection performance was poor for the LASSO estimator when γ became large. This is due to the presence of many false negative selections (i.e., non-zero loading parameters selected as zeros), as a result of over-regularisation.

Results on confidence intervals. In Fig. 4, we show boxplots of the ECIC for the L^p rotations, for $p = 0.5$ and $p = 1$ and $N \in \{400, 800, 1600\}$. For both $p = 0.5$ and $p = 1$, the ECIC_{jk}s are close to the 95% nominal level, supporting the consistency of the proposed procedure for constructing confidence intervals.

Some remarks. The computation for the proposed L^p rotation is fast. On a single core of a data science workstation,[1] the mean time for solving the L^1 rotation criterion is within $0.29s$ for the 15×3 settings and within $0.54s$ for the 30×5 settings. Using the L^1 solution as the starting point, the mean time for solving the $L^{0.5}$ criterion is within $0.13s$ for 15×3 settings and within $0.36s$ for the 30×5 settings. Under the current simulation settings, condition C3 is satisfied by both the $L^{0.5}$ and L^1 criteria, in which cases the two criteria tend to perform similarly. As we will show in Sect. 4.2 below, the performance of the two criteria can be substantially different when C3 holds for one criterion but not the other. In addition, we see that the LASSO estimator with a small tuning parameter performed similarly to the L^1 rotation method. We expected this, since the L^1 rotation solution can be viewed as the limiting case of the LASSO estimator when the tuning parameter goes to zero. The LASSO estimator performed poorly for large tuning parameters, due to the bias brought by the regularisation. This bias-variance trade-off is visualised in Fig. 3. The two panels in Fig. 3 correspond to the 15×3 and 30×5 loading matrix settings, respectively. For each panel, the x-axis shows the tuning parameter γ, and the y-axis shows the MSE (for the loading matrix) of the corresponding LASSO estimator. The dots at $\gamma = 0$ correspond to the L^1 rotation solutions, as the L^1-rotation estimator is the limit of the LASSO estimator when γ converges to zero (see Proposition 3). As γ increases, the estimation bias increases, and the variance decreases, which results in a U-shaped curve for the MSE – a well-known phenomenon in statistical learning theory (see Chapter 2, Hastie et al., 2009). However, the U-shaped curves in Fig. 3 are very asymmetric – the MSE only decreases slightly before increasing. This means that the estimators with small γ values including the rotation solution have similar estimation accuracy to the optimal choice of the tuning parameter (i.e., the value of γ at which the MSE

[1] CPU configuration: Intel Xeon 6246R 3.4GHz 2933MHz.

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

TABLE 2.
The AUC, TR, TPR, and TNR for the L^p-based rotation estimator and the regularised estimator, Study I.

	15×3				30×5			
	AUC	TR	TPR	TNR	AUC	TR	TPR	TNR
				$N = 400$				
$L^{0.5}$ rotation	0.996	0.979	0.979	0.978	0.988	0.964	0.937	0.977
L^1 rotation	0.997	0.979	0.979	0.979	0.988	0.964	0.937	0.977
LASSO, $\gamma = 0.01$	0.997	0.981	0.979	0.983	0.989	0.967	0.941	0.979
LASSO, $\gamma = 0.05$	0.997	0.984	0.979	0.988	0.992	0.971	0.944	0.985
LASSO, $\gamma = 0.1$	0.992	0.983	0.973	0.991	0.987	0.970	0.937	0.986
LASSO, $\gamma = 0.2$	0.869	0.848	0.784	0.905	0.903	0.917	0.800	0.975
LASSO, $\gamma = 0.5$	0.500	0.534	0.001	1.000	0.520	0.683	0.075	0.987
				$N = 800$				
$L^{0.5}$ rotation	1.000	0.993	0.992	0.993	0.998	0.986	0.980	0.990
L^1 rotation	1.000	0.993	0.993	0.992	0.998	0.987	0.981	0.990
LASSO, $\gamma = 0.01$	1.000	0.992	0.992	0.993	0.998	0.989	0.983	0.992
LASSO, $\gamma = 0.05$	1.000	0.993	0.992	0.993	0.999	0.990	0.985	0.993
LASSO, $\gamma = 0.1$	0.996	0.992	0.988	0.995	0.995	0.989	0.979	0.994
LASSO, $\gamma = 0.2$	0.880	0.862	0.816	0.902	0.919	0.932	0.824	0.987
LASSO, $\gamma = 0.5$	0.500	0.533	0.000	1.000	0.506	0.672	0.024	0.996
				$N = 1600$				
$L^{0.5}$ rotation	1.000	0.997	0.999	0.995	1.000	0.996	0.996	0.996
L^1 rotation	1.000	0.997	0.999	0.995	1.000	0.996	0.996	0.996
LASSO, $\gamma = 0.01$	1.000	0.997	1.000	0.995	1.000	0.996	0.997	0.996
LASSO, $\gamma = 0.05$	1.000	0.997	0.999	0.996	1.000	0.997	0.998	0.996
LASSO, $\gamma = 0.1$	0.998	0.997	0.995	0.999	0.998	0.996	0.993	0.997
LASSO, $\gamma = 0.2$	0.886	0.870	0.831	0.904	0.929	0.941	0.838	0.993
LASSO, $\gamma = 0.5$	0.500	0.533	0.000	1.000	0.501	0.668	0.003	0.999

curve achieves the minimum value). In that case, it may not be worth searching for the optimal tuning parameter, as constructing a LASSO solution path is typically computationally intensive. Instead, using the rotation method or a LASSO estimator with a sufficiently small tuning parameter is computationally more affordable and yields a sufficiently accurate solution.

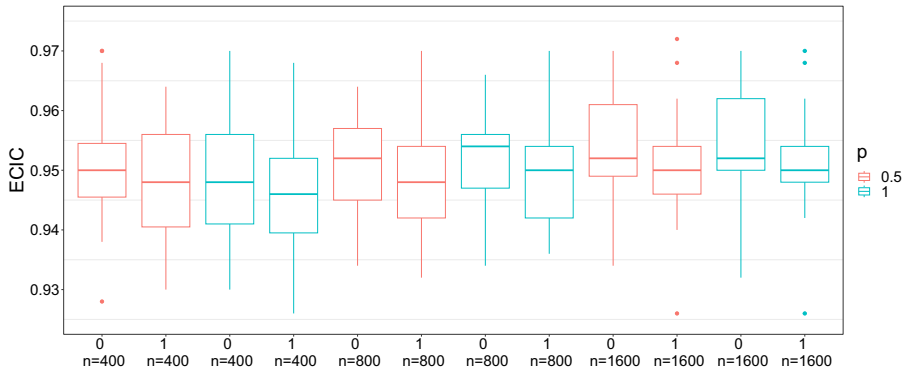
4.2. Study II

In this study we compare the $L^{0.5}$ and L^1 rotations, under a setting where condition C3 holds for the $L^{0.5}$ rotation but not the L^1 rotation. We chose the setting to somewhat exaggerate the differences, in order to show the consequence of misspecifying p.

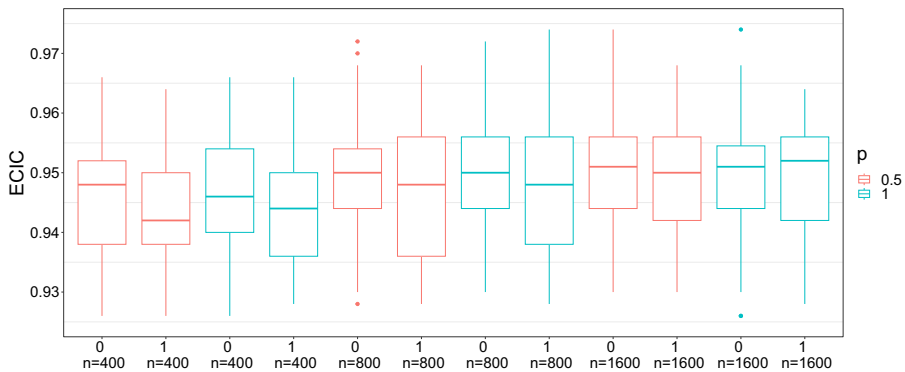
Setting and evaluation criteria. The true loading matrix is of dimension $J = 18$ and $K = 3$. Each item is set to load on two factors, so that no item has a perfect simple structure. Given the loading structure, the model is identifiable as a confirmatory factor analysis model. We present the true model parameters in the supplementary material. By grid search, we checked that C3 holds for the $L^{0.5}$ criterion but not the L^1 criterion. We chose the sample size to be $N = 3000$. Similar to Study I, we compare the two rotation criteria using the MSE, AUC, TR, TPR, and TNR by running $B = 500$ independent replications.

Results. We present the results in Table 3. The $L^{0.5}$ criterion performed better in terms of both point estimation and model selection, as its MSE was lower and the AUC, TR, TPR, and TNR were higher. In particular, we noted that the $L^{0.5}$ rotation achieved a much higher TNR than

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press



(a) Boxplots of ECIC_{jk} for the 15×3 setting



(b) Boxplots of ECIC_{jk} for the 30×5 setting

FIGURE 4.
Boxplots of ECIC_{jk}. The label 0 means that $\lambda^*_{jk} = 0$ and the label 1 means that $\lambda^*_{jk} \neq 0$.

TABLE 3.
The MSE, AUC, TR, TPR, and TNR for the L^p-based rotation estimator, Study II.

	MSE	AUC	TR	TPR	TNR
$L^{0.5}$ rotation	0.003	0.984	0.954	0.943	0.974
L^1 rotation	0.025	0.953	0.865	0.936	0.725

the L^1 rotation, meaning that the L^1 rotation tended to make many false positive selections (i.e., zero loading parameters selected as non-zeros), as a consequence of violating condition C3.

5. An Application to the Big Five Personality Test

We illustrate the proposed method through an application to the Big Five personality test. We consider the Big Five Factor Markers from the International Personality Item Pool (Goldberg, 1992) , which contains 50 items designed to measure five personality factors, namely Extraversion (E), Emotional Stability (ES), Agreeableness (A), Conscientiousness (C), and Intellect/Imagination (I). Each item is a statement describing a personality pattern like "*I am the life*

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

TABLE 4.
Estimated correlation matrices based on $L^{0.5}$ and L^1 rotations, Big Five personality test.

	$p = 0.5$					$p = 1$				
	E	ES	A	C	I	E	ES	A	C	I
E	1					1				
ES	0.154	1				0.184	1			
A	0.193	−0.017	1			0.197	−0.001	1		
C	0.016	0.010	0.023	1		0.022	0.148	0.038	1	
I	0.050	0.018	−0.005	−0.046	1	0.161	0.040	0.023	−0.019	1

of the party" and "*I get stressed out easily*", designed to primarily measure one personality factor. We can divide the 50 items into five equal-sized groups, with each group mainly measuring one personality factor. Responses to the items are on a five-level Likert scale, which we treat as continuous variables in the current analysis.

Although the Big Five personality test was designed to have a perfect simple structure, cross-loadings are often found in empirical data (e.g., Gow et al., 2005). To better understand the loading structure of this widely used personality test, we applied the proposed $L^{0.5}$ and L^1 rotations to a dataset[2] on this test. To avoid possible complexities brought by measurement non-invariance, we selected the subset of male respondents from the United Kingdom, which has a sample size $N = 609$. In the analysis, the number of factors is set to be $K = 5$.

After applying the proposed rotations, we further adjusted the estimates by column permutation and sign flip transformations, so that the resulting factors correspond to the E, ES, A, C, and I factors, respectively. We give our results in Tables 4 through 7. In Table 4 we show the estimated covariance matrices from the two rotations. The estimated correlation matrices from the two criteria are similar to each other. In particular, all the signs of the correlations are consistent, except for the correlation between A and I, in which case both correlations are close to zero. In addition, for each pair of factors, the correlations obtained by the two criteria are close. The sign pattern of the correlations between the Big Five factors is largely consistent with those found in the literature (e.g., Booth & Hughes, 2014; Gow et al., 2005).

In Tables 5 through 7 we show the estimated loading parameters and the corresponding 95% confidence intervals obtained from the $L^{0.5}$ rotation. We indicate by asterisks the loadings that are significantly different from zero according to the 95% confidence intervals. The results of the L^1 rotation are similar and thus we give them in the supplementary material. In Tables 5, 6 and 7, the items are labelled based on the personality factor that they are designed to measure, and their scoring keys.[3] The estimated loading matrix is largely consistent with the International Personality Item Pool (IPIP) scoring key, where all the items have relatively strong loadings on the factors that they are designed to measure, and the signs of the loadings are consistent with the scoring keys. The confidence intervals shed additional light on the uncertainty of each loading. Specifically, we notice that many loadings are statistically insignificantly different from zero, suggesting that the true loading structure is sparse. There are also items with fairly strong cross-loadings.

[2]The dataset is downloaded from: http://personality-testing.info/_rawdata/.

[3]Positively scored items are indicated by "(+)" and negatively scored items are indicated by "(−)".

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

TABLE 5.
Part I: Point estimates and confidence intervals constructed by $L^{0.5}$, Big Five personality test. The loadings that are significantly different from zero according to the 95% confidence intervals are indicated by asterisks.

	E	ES	A	C	I
E1(+)	0.887*	−0.069	−0.068*	0.004	0.066
	(0.795, 0.984)	(−0.158, 0.005)	(−0.182,−0.016)	(−0.113, 0.059)	(−0.030, 0.140)
E2(−)	−0.851*	0.131*	0.003	0.047	0.001
	(−0.969,−0.765)	(0.049, 0.228)	(−0.057, 0.126)	(−0.021, 0.168)	(−0.071, 0.118)
E3(+)	0.780*	0.276*	0.204*	0.142*	−0.107*
	(0.703, 0.879)	(0.190, 0.344)	(0.121, 0.277)	(0.060, 0.220)	(−0.200,−0.042)
E4(−)	−0.914*	−0.058	−0.022	0.002	0.105*
	(−1.022,−0.844)	(−0.139, 0.012)	(−0.075, 0.077)	(−0.066, 0.094)	(0.050, 0.205)
E5(+)	0.898*	−0.024	0.155*	0.100	0.064
	(0.814, 0.991)	(−0.116, 0.034)	(0.060, 0.212)	(−0.001, 0.155)	(−0.016, 0.140)
E6(−)	−0.754*	−0.001	−0.088	−0.061	−0.123*
	(−0.854,−0.662)	(−0.066, 0.106)	(−0.163, 0.010)	(−0.152, 0.027)	(−0.200,−0.023)
E7(+)	1.119*	−0.078*	0.083*	0.092*	−0.042
	(1.025, 1.228)	(−0.187,−0.019)	(0.002, 0.174)	(0.005, 0.184)	(−0.175, 0.002)
E8(−)	−0.724*	−0.086	0.028	0.115*	−0.051
	(−0.829,−0.634)	(−0.173, 0.002)	(−0.036, 0.142)	(0.026, 0.208)	(−0.129, 0.056)
E9(+)	0.862*	0.051	0.000	−0.010	0.226*
	(0.751, 0.958)	(−0.048, 0.136)	(−0.110, 0.075)	(−0.127, 0.067)	(0.110, 0.301)
E10(−)	−0.828*	−0.117*	−0.049	−0.126*	0.020
	(−0.935,−0.745)	(−0.189,−0.021)	(−0.124, 0.046)	(−0.212,−0.036)	(−0.043, 0.132)
ES1(−)	−0.132*	−0.971*	0.006	0.001	−0.101
	(−0.215,−0.028)	(−1.065,−0.869)	(−0.117, 0.082)	(−0.133, 0.054)	(−0.175, 0.003)
ES2(+)	0.147*	0.671*	0.001	−0.029	0.082
	(0.039, 0.220)	(0.587, 0.768)	(−0.066, 0.112)	(−0.113, 0.064)	(−0.008, 0.170)
ES3(−)	−0.186*	−0.780*	0.231*	0.063	0.046
	(−0.277,−0.095)	(−0.880,−0.696)	(0.128, 0.306)	(−0.041, 0.138)	(−0.008, 0.170)
ES4(+)	0.225*	0.565*	0.002	0.110*	0.006
	(0.116, 0.314)	(0.468, 0.664)	(−0.071, 0.122)	(0.024, 0.224)	(−0.105, 0.090)
ES5(−)	0.013	−0.473*	−0.042	−0.152*	−0.226*
	(−0.075, 0.137)	(−0.566,−0.356)	(−0.163, 0.046)	(−0.272,−0.059)	(−0.337,−0.125)
ES6(−)	−0.130*	−0.806*	0.257*	−0.088*	−0.130*
	(−0.205,−0.023)	(−0.903,−0.716)	(0.147, 0.328)	(−0.223,−0.039)	(−0.209,−0.033)
ES7(−)	0.022	−0.962*	−0.112*	−0.124*	0.004
	(−0.051, 0.119)	(−1.051,−0.867)	(−0.224,−0.050)	(−0.244,−0.064)	(−0.073, 0.089)
ES8(−)	0.000	−1.131*	−0.135*	−0.169*	0.000
	(−0.085, 0.100)	(−1.227,−1.029)	(−0.258,−0.075)	(−0.294,−0.103)	(−0.078, 0.095)
ES9(−)	−0.033	−0.862*	−0.293*	0.097	−0.016
	(−0.134, 0.048)	(−0.949,−0.764)	(−0.394,−0.211)	(−0.002, 0.183)	(−0.095, 0.082)
ES10(−)	−0.344*	−0.837*	0.069	−0.172*	0.104*
	(−0.439,−0.256)	(−0.930,−0.742)	(−0.026, 0.157)	(−0.284,−0.101)	(0.032, 0.206)

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

TABLE 6.
Part II: Point estimates and confidence intervals constructed by $L^{0.5}$, Big Five personality test. The loadings that are significantly different from zero according to the 95% confidence intervals are indicated by asterisks.

	E	ES	A	C	I
A1(−)	0.003	−0.127*	−0.778*	0.010	0.045
	(−0.114, 0.087)	(−0.201,−0.011)	(−0.875,−0.669)	(−0.095, 0.103)	(−0.060, 0.136)
A2(+)	0.439*	−0.007	0.557*	−0.038	0.035
	(0.361, 0.526)	(−0.097, 0.054)	(0.464, 0.626)	(−0.132, 0.024)	(−0.042, 0.113)
A3(−)	0.193*	−0.575*	−0.566*	−0.130*	0.134*
	(0.080, 0.286)	(−0.663,−0.456)	(−0.691,−0.479)	(−0.265,−0.054)	(0.026, 0.230)
A4(+)	0.013	0.001	0.979*	−0.002	−0.002
	(−0.035, 0.144)	(−0.102, 0.038)	(0.895, 1.050)	(−0.081, 0.047)	(−0.051, 0.083)
A5(−)	−0.155*	−0.039	−0.815*	−0.012	0.090*
	(−0.250,−0.091)	(−0.102, 0.049)	(−0.894,−0.724)	(−0.085, 0.069)	(0.017, 0.170)
A6(+)	−0.059	−0.182*	0.717*	0.002	0.014
	(−0.159, 0.020)	(−0.272,−0.105)	(0.629, 0.811)	(−0.101, 0.070)	(−0.063, 0.110)
A7(−)	−0.367*	−0.089*	−0.733*	0.043	0.036
	(−0.456,−0.300)	(−0.159,−0.015)	(−0.800,−0.639)	(−0.023, 0.125)	(−0.032, 0.115)
A8(+)	0.111*	−0.038	0.692*	0.085*	0.025
	(0.039, 0.185)	(−0.128, 0.010)	(0.617, 0.771)	(0.010, 0.152)	(−0.035, 0.107)
A9(+)	0.123*	−0.110*	0.751*	0.066	0.110*
	(0.040, 0.199)	(−0.204,−0.054)	(0.668, 0.836)	(−0.018, 0.137)	(0.041, 0.195)
A10(+)	0.439*	0.071	0.321*	0.133*	0.045
	(0.354, 0.517)	(−0.010, 0.143)	(0.245, 0.402)	(0.043, 0.201)	(−0.037, 0.121)
C1(+)	0.105	0.111	−0.037	0.695*	0.129*
	(−0.001, 0.179)	(−0.005, 0.177)	(−0.099, 0.088)	(0.597, 0.785)	(0.055, 0.238)
C2(−)	0.080	−0.201*	0.107	−0.670*	0.142*
	(−0.014, 0.194)	(−0.284,−0.073)	(−0.028, 0.177)	(−0.800,−0.585)	(0.013, 0.217)
C3(+)	0.023	0.007	0.114*	0.407*	0.280*
	(−0.080, 0.082)	(−0.094, 0.065)	(0.050, 0.210)	(0.315, 0.482)	(0.213, 0.378)
C4(−)	−0.123*	−0.613*	0.048	−0.544*	−0.039
	(−0.202,−0.036)	(−0.671,−0.495)	(−0.057, 0.114)	(−0.656,−0.483)	(−0.149, 0.018)
C5(+)	0.074	0.057	0.000	0.782*	−0.052
	(−0.005, 0.188)	(−0.051, 0.158)	(−0.041, 0.163)	(0.687, 0.882)	(−0.133, 0.061)
C6(−)	0.021	−0.195*	0.045	−0.718*	0.087
	(−0.085, 0.130)	(−0.276,−0.058)	(−0.090, 0.128)	(−0.848,−0.625)	(−0.028, 0.188)
C7(+)	−0.129*	−0.128*	0.110*	0.520*	0.042
	(−0.225,−0.047)	(−0.236,−0.059)	(0.041, 0.220)	(0.427, 0.608)	(−0.015, 0.166)
C8(−)	−0.000	−0.284*	−0.242*	−0.549*	−0.000
	(−0.086, 0.097)	(−0.349,−0.166)	(−0.361,−0.179)	(−0.654,−0.466)	(−0.119, 0.063)
C9(+)	0.031	−0.003	0.123*	0.722*	−0.076
	(−0.061, 0.129)	(−0.140, 0.060)	(0.059, 0.248)	(0.623, 0.816)	(−0.157, 0.034)
C10(+)	−0.001	−0.006	0.127*	0.528*	0.235*
	(−0.110, 0.057)	(−0.120, 0.046)	(0.070, 0.236)	(0.433, 0.605)	(0.170, 0.338)

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

TABLE 7.
Part III: Point estimates and confidence intervals constructed by $L^{0.5}$, Big Five personality test. The loadings that are significantly different from zero according to the 95% confidence intervals are indicated by asterisks.

	E	ES	A	C	I
I1(+)	0.085	0.001	−0.046	0.009	0.621*
	(−0.020, 0.145)	(−0.102, 0.067)	(−0.148, 0.014)	(−0.088, 0.078)	(0.537, 0.713)
I2(−)	−0.000	−0.222*	−0.087*	−0.020	−0.581*
	(−0.052, 0.120)	(−0.289,−0.119)	(−0.183,−0.014)	(−0.103, 0.071)	(−0.675,−0.498)
I3(+)	0.076	−0.152*	0.024	−0.000	0.587*
	(−0.028, 0.137)	(−0.244,−0.087)	(−0.061, 0.095)	(−0.103, 0.063)	(0.503, 0.670)
I4(−)	0.023	−0.204*	−0.154*	0.008	−0.572*
	(−0.027, 0.145)	(−0.269,−0.101)	(−0.228,−0.062)	(−0.074, 0.098)	(−0.663,−0.487)
I5(+)	0.240*	0.068*	−0.058	0.189*	0.575*
	(0.131, 0.272)	(0.003, 0.139)	(−0.130, 0.002)	(0.095, 0.240)	(0.501, 0.648)
I6(−)	−0.217*	−0.001	−0.047	0.020	−0.505*
	(−0.275,−0.104)	(−0.065, 0.102)	(−0.120, 0.046)	(−0.053, 0.119)	(−0.597,−0.421)
I7(+)	0.076	0.168*	−0.035	0.117*	0.449*
	(−0.018, 0.123)	(0.089, 0.226)	(−0.100, 0.036)	(0.032, 0.176)	(0.376, 0.520)
I8(+)	−0.014	−0.163*	−0.108*	−0.003	0.656*
	(−0.163, 0.023)	(−0.262,−0.082)	(−0.198,−0.020)	(−0.130, 0.058)	(0.572, 0.769)
I9(+)	−0.056	−0.213*	0.239*	0.097*	0.260*
	(−0.153, 0.008)	(−0.305,−0.148)	(0.157, 0.317)	(0.006, 0.167)	(0.188, 0.350)
I10(+)	0.246*	−0.004	−0.000	0.107*	0.680*
	(0.130, 0.276)	(−0.108, 0.039)	(−0.072, 0.067)	(0.009, 0.159)	(0.606, 0.761)

6. Concluding Remarks

In this paper we propose a new family of oblique rotations based on component-wise L^p loss functions ($0 < p \leq 1$) and establish the relationship between the proposed rotation estimator and the L^p regularised estimator for EFA. We develop point estimation, model selection, and post-selection inference procedures and establish their asymptotic theories. We also develop an iteratively reweighted gradient projection algorithm for the computation.[4] We demonstrate the power of the proposed method via simulation studies and an application to Big Five personality assessment.

We note that the proposed procedures do not rely on the normality assumption of the EFA model, though we make such an assumption in the problem setup for ease of exposition. Specifically, in the rotation, we only need to obtain a consistent initial estimator for EFA in the sense of condition C1, which we can obtain with any reasonable loss function for factor analysis. In the model selection, only the BIC uses the likelihood function based on the normal model. Note that the likelihood function is a valid loss function under the linear factor model, even if the normality assumption does not hold (Chapter 7, Bollen, 1989). Therefore, the BIC still yields consistent model selection under the misspecification of the normality assumption (Machado, 1993) . Finally, the confidence intervals are based on the asymptotic distributions of CFA models. If we use a robust method (i.e., a sandwich estimator) for computing the asymptotic variance, then the resulting confidence intervals are valid when the normality assumption does not hold.

As each value of $p \in (0, 1]$ leads to a sensible rotation criterion, which L^p criterion should we use? We do not recommend trying too many values of p. From the previous discussion, we see that there is a statistical and computational trade-off underlying the choice of p. Theoretically, a

[4]The R code for the proposed method is available from https://github.com/yunxiaochen/Lp_rot1129

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

smaller value of p is more likely to recover a sparse loading matrix, but the associated optimisation problem is computationally more challenging. The L^1 criterion is the easiest to compute. Although we gave an example earlier in which the L^1 criterion fails to recover the sparest loading structure, the L^1 criterion can accurately recover the true loading structure under most simulation settings. For several real-world datasets we have encountered, different p values also give very similar results. We thus believe that the L^1 criterion is robust and recommend users to always start with the L^1 criterion. To check the result of the L^1 criterion, users may try some smaller p values (e.g., $p = 0.5$) and compare their results with the L^1 result in terms of model fitting and substantive interpretations. If they give similar results, then the best fitting solution should be reported. If the result from a smaller p value substantially differs from the L^1 result, then the value of p should be further decreased until the result stabilises. Computationally, when solving the optimisation with a smaller value of p, we recommend using the solution from the previous larger value of p as the starting point, so that the algorithm is less likely to get stuck at a local optimum.

Our complexity analysis and simulation results suggest that obtaining a solution path for the L^1-regularised estimator has little added value over the L^1 rotation when the sample size is reasonably large. That is, obtaining the solution path of the regularised estimator is computationally more intensive, while the best tuning parameter is often very close to zero and thus the corresponding solution is very similar to the rotation solution. Therefore, when the sample size is reasonably large, we do not recommend running a solution path for the L^1 regularised estimator to learn the loading structure in EFA. Instead, users can obtain a point estimate by either applying the L^1 rotation or running the L^1 regularised estimator with a single small tuning parameter. Model selection can be done by applying hard-thresholding to this point estimate. Furthermore, although an L^p regularised estimator is mathematically well-defined with $p < 1$, algorithms remain to be developed for its computation. On the other hand, L^p rotation can be computed by the proposed IRGP algorithm for all $p \in (0, 1]$. However, when the sample size is small and the number of items is large, the regularised estimators may outperform their rotation counterparts. In that case, an optimally tuned regularised estimator may be substantially more accurate than those with very small tuning parameters or the rotation-based estimator, and thus, better learn the sparse loading structure.

The current work has several limitations that require future investigation. First, the way the confidence intervals are constructed may be improved. That is, accurate model selection (condition C5) and identifiability conditions on the true model (condition C6) are required for the confidence intervals to have good coverage rate, while the uncertainty in the model selection step is not taken into account in the proposed procedure. Consequently, although the proposed confidence intervals are shown to be asymptotically valid, they may not perform well when the sample size is small. This issue may be addressed by future researchers developing bootstrap procedures for constructing confidence intervals, as such procedures may still be valid even when the objective function is nonsmooth (Sen et al., 2010).

The current theoretical results only consider a low-dimensional setting where the numbers of manifest variables and factors are fixed and the sample size goes to infinity. As factor analysis is commonly used by those analysing high-dimensional multivariate data, it is of interest to generalise the current results to a high-dimensional regime where the numbers of manifest variables, factors, and observations all grow to infinity (Chen & Li, 2022; Chen et al., 2019, 2020; Zhang et al., 2020). In particular, it will be of interest to see how the rotation methods work with the joint maximum likelihood estimator for high-dimensional factor models (Chen et al., 2019, 2020).

Finally, as is an issue with any simulation study, we can only examine a small number of simulation settings, and thus, may not be able to provide a complete picture of the proposed methods. Future researchers need to investigate more simulation settings by varying the numbers of manifest variables, factors, and observations, the sign pattern of the true loading matrix, and the generation mechanism of the true model parameters.

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Acknowledgments

Gabriel Wallin is supported by the Vetenskapsrådet (Grant No. 2020-06484)

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ba, D., Babadi, B., Purdon, P. L., & Brown, E. N. (2013). Convergence and stability of iteratively reweighted least squares algorithms. *IEEE Transactions on Signal Processing, 62*(1), 183–195.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Wiley.

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65*(5), 676–696.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Booth, T., & Hughes, D. J. (2014). Exploratory structural equation modeling of personality data. *Assessment, 21*(3), 260–271.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*(1), 62–83.

Chen, Y., & Li, X. (2022). Determining the number of factors in high-dimensional generalized latent factor models. *Biometrika, 109*(3), 769–782.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2021). Item response theory-a statistical framework for educational and psychological measurement. arXiv preprint arXiv:2108.08604.

Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika, 84*(1), 124–146.

Chen, Y., Li, X., & Zhang, S. (2020). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association, 115*(532), 1756–1770.

Choi, J., Oehlert, G., & Zou, H. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and its Interface, 3*(4), 429–436.

Daubechies, I., DeVore, R., Fornasier, M., & Güntürk, C. S. (2010). Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 63*(1), 1–38.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Geminiani, E., Marra, G., & Moustaki, I. (2021). Single-and multiple-group penalized factor analysis: A trust-region algorithm approach with integrated automatic multiple tuning parameter selection. *Psychometrika, 86*(1), 65–95.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*(1), 26.

Gow, A. J., Whiteman, M. C., Pattie, A., & Deary, I. J. (2005). Goldberg's 'IPIP' big-five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences, 39*(2), 317–329.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

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

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing, 25*(5), 863–875.

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika, 38*(4), 593–604.

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika, 67*(1), 7–19.

Jennrich, R. I. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika, 69*(2), 257–273.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika, 71*(1), 173–191.

Jennrich, R. I., & Sampson, P. (1966). Rotation for simple loadings. *Psychometrika, 31*(3), 313–323.

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press

Jin, S., Moustaki, I., & Yang-Wallentin, F. (2018). Approximated penalized maximum likelihood for exploratory factor analysis: An orthogonal case. *Psychometrika, 83*(3), 628–649.
Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*(4), 443–482.
Jöreskog, K. G., & Goldberger, A. S. (1972). Factor analysis by generalized least squares. *Psychometrika, 37*(3), 243–260.
Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187–200.
Karimi, H., Nutini, J., & Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 795–811).
Kiers, H. A. (1994). Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika, 59*(4), 567–579.
Lai, M.-J., & Wang, J. (2011). An unconstrained lq minimization with $0<q=1$ for sparse solution of underdetermined linear systems. *SIAM Journal on Optimization, 21*(1), 82–101.
Machado, J. A. (1993). Robust model selection and M-estimation. *Econometric Theory, 9*(3), 478–493.
Mazumder, R., Friedman, J. H., & Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association, 106*(495), 1125–1138.
Meinshausen, N., & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics, 37*(1), 246–270.
Mulaik, S. A. (2009). *Foundations of factor analysis*. CRC Press.
Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 758–D765.
Parikh, N., & Boyd, S. (2014). Proximal Algorithms. *Foundations and trends in optimization, 1*(3), 127–239.
Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory*. Springer.
Rohe, K., & Zeng, M. (2022). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society - Series B* (to appear).
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.
Sen, B., Banerjee, M., & Woodroofe, M. (2010). Inconsistency of bootstrap: The grenander estimator. *The Annals of Statistics, 38*(4), 1953–1977.
Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 221–242.
Thurstone, L. L. (1947). *Multiple-Factor Analysis. A Development and Expansion of the Vectors of Mind*. University of Chicago Press.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.
Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics, 29*(3), 431–454.
van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.
Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) And the Bayesian Information Criterion (BIC). *Psychological Methods, 17*(2), 228.
Yamamoto, M., Hirose, K., & Nagata, H. (2017). Graphical tool of sparse factor analysis. *Behaviormetrika, 44*(1), 229–250.
Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Suny Press.
Zhang, H., Chen, Y., & Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika, 85*(2), 358–372.
Zhao, S., Witten, D., & Shojaie, A. (2021). In defense of the indefensible: A very Naive approach to high-dimensional inference. *Statistical Science, 36*(4), 562–577.
Zheng, L., Maleki, A., Weng, H., Wang, X., & Long, T. (2017). Does l_p-minimization outperform l_1-minimization? *IEEE Transactions on Information Theory, 63*(11), 6896–6935.

Manuscript Received: 18 JUN 2022
Published Online Date: 31 MAR 2023

https://doi.org/10.1007/s11336-023-09911-y Published online by Cambridge University Press