Rotation to Sparse Loadings using L^p Losses and Related Inference Problems

Xinyi Liu, Gabriel Wallin, Yunxiao Chen, and Irini Moustaki[1]

[1]London School of Economics and Political Science

Problem Setup

Linear Exploratory Factor Model(EFA) A linear EFA model with J indicators and K factors given by

$$\begin{aligned}\boldsymbol{\xi} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}),\\ \mathbf{X}|\boldsymbol{\xi} &\sim \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{\xi}, \boldsymbol{\Omega}),\end{aligned} \tag{1}$$

where $\boldsymbol{\xi}$ is a K-dimensional vector of common factors,$\boldsymbol{\Phi} \in \mathbb{R}^{K\times K}$ has diagonal entries equal to 1 and is symmetric positive definite ($\boldsymbol{\Phi} \succ 0$), $\mathbf{X}$ is a J-dimensional vector of manifest variables, $\boldsymbol{\Lambda} = (\lambda_{jk})_{J\times K}$ is the loading matrix, and $\boldsymbol{\Omega} = (\omega_{ij})_{J\times J}$ denotes the residual covariance matrix. Let $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$, the model in (1) implies the marginal distribution of $\mathbf{X}$

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Omega}. \tag{2}$$

Rotational Indeterminacy Suppose $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, let $\tilde{\boldsymbol{\xi}} = \mathbf{T}'\boldsymbol{\xi}$, $\tilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}(\mathbf{T}')^{-1}$. Then, $\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\xi}} = \boldsymbol{\Lambda}\boldsymbol{\xi}$, and $(\boldsymbol{\Lambda}(\mathbf{T}')^{-1}, \mathbf{T}'\mathbf{T}, \boldsymbol{\Omega})$ and $(\boldsymbol{\Lambda}, \mathbf{I}, \boldsymbol{\Omega})$ result in the same distribution for $\mathbf{X}$. Without further constraints, the problem is unidentifiable.However, it is also an opportunity to find a solution that has simple loading structure among the equivalent rotation class.

Oblique Rotation Method

1. Calculate the MLE with orthogonal latent covariance based on N samples

$$\hat{\boldsymbol{\theta}}_N = (\hat{\boldsymbol{\Lambda}}_N, \mathbf{I}, \hat{\boldsymbol{\Omega}}_N) \in \text{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\boldsymbol{\Phi}=\mathbf{I}}} L(\boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{3}$$

where $L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) = \log \text{det}(2\pi\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \text{tr}(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{S})$, $\boldsymbol{S} = (\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^{\top})/N$.

2. Rotate the estimated loading matrix to find a sparse loading structure.

$$\hat{\mathbf{T}}_N = \text{argmin}_{\mathbf{T}\in\mathcal{M}} Q(\hat{\boldsymbol{\Lambda}}_N\mathbf{T}'^{-1}), \hat{\boldsymbol{\Lambda}}_N = \hat{\boldsymbol{\Lambda}}_N\hat{\mathbf{T}}_N'^{-1} \tag{4}$$

where $\mathcal{M} = \{\mathbf{T} \in \mathbb{R}^{K\times K} : rank(\mathbf{T}) = K,\ (\mathbf{T}'\mathbf{T})_{kk} = 1,\ k = 1, \dots, K\}$
Obtain $\hat{\boldsymbol{\theta}}_R = (\hat{\boldsymbol{\Lambda}}_N, \hat{\mathbf{T}}_N'\hat{\mathbf{T}}_N, \hat{\boldsymbol{\Omega}}_N)$.

L^p Rotation Criteria

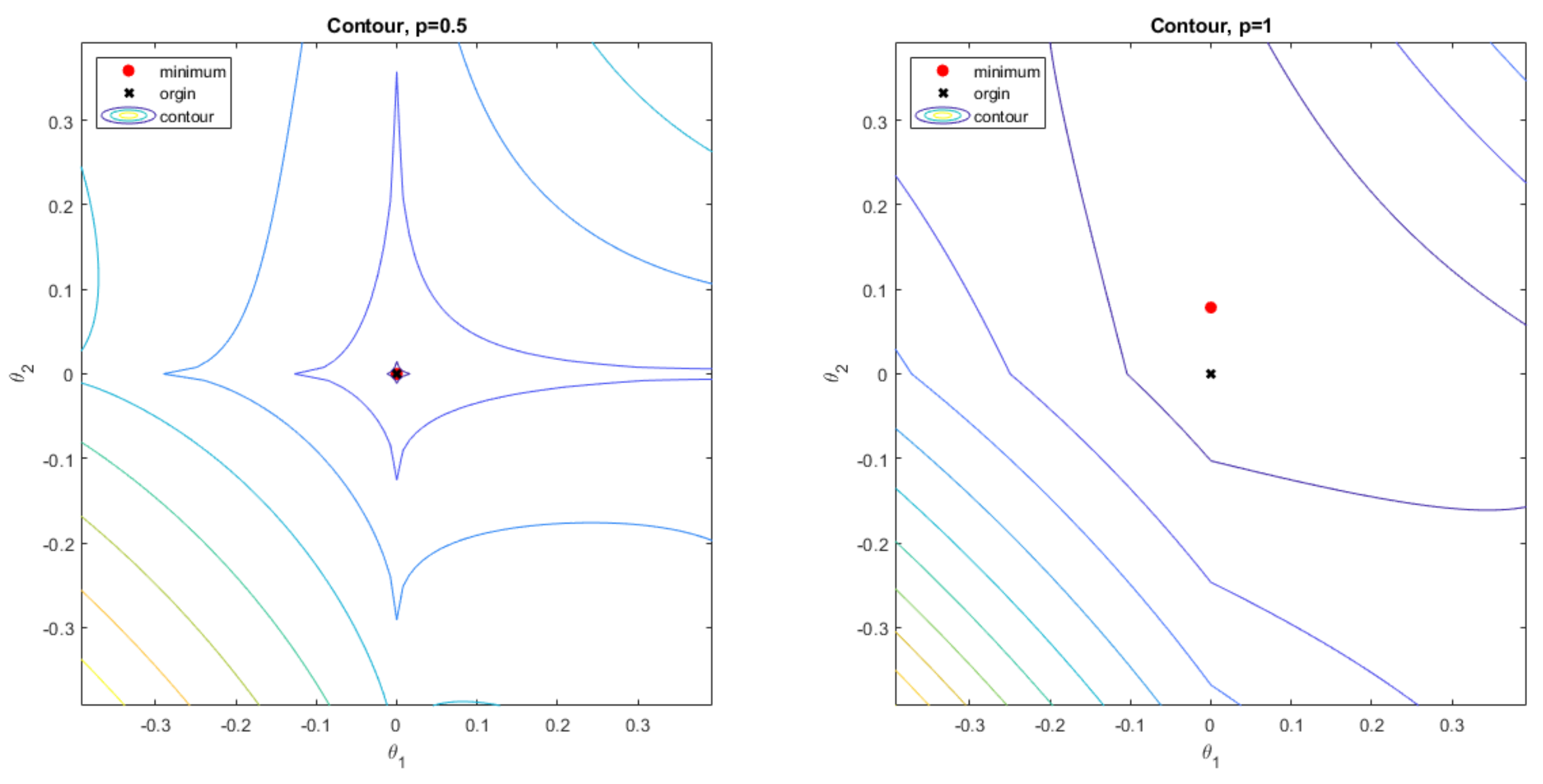
To fill the gap of dealing with true loading matrices with different sparsity levels, we propose a family of L^p component loss functions [1]. More specifically, for each value of $p \in (0, 1]$, the loss function takes the form

$$Q_p(\boldsymbol{\Lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K} |\lambda_{jk}|^p, \tag{5}$$

Suppose that the true loading matrix with full rank $\boldsymbol{\Lambda}^*$ has perfect simple structure, in the sense that each row has at most one nonzero entry, then any L^p CLF is uniquely minimized by $\boldsymbol{\Lambda}^*$. The minimiser is unique when all the minimisers are equivalent up to column permutation and sign flip transformations. When the true loading matrix is less sparse than perfect simple structure, here is an example that the minimiser of $Q_{0.5}$ contains more zeros than Q_1.Let

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

The following are plots of contours of $|\boldsymbol{\Lambda}^*\mathbf{T}^{-1'}|^p$, where $\mathbf{T} = [\cos(\theta_1), \sin(\theta_2); \sin(\theta_1), \cos(\theta_2)]$. Left: $p = 0.5$. Right: $p = 1$.The point $(0, 0)$, which is indicated by a black cross, corresponds to $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^*$, and the point indicated by a red point corresponds to the $\boldsymbol{\Lambda}$ matrix such that $Q_p(\boldsymbol{\Lambda})$ is minimised.



As we can see, when $p = 0.5$, the loss function is minimised by $\boldsymbol{\Lambda}^*$. On the other hand, when $p = 1$, the minimiser of the loss function does not contain as many zeros as $\boldsymbol{\Lambda}^*$

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 $G_t(\mathbf{T}) = \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}^{(t)} \left((\hat{\mathbf{A}}\mathbf{T}'^{-1})_{jk}\right)^2$, where the weights $w_{jk}^{(t)}$ are given by $w_{jk}^{(t)} = \frac{1}{((\hat{\mathbf{A}}(\mathbf{T}_t')^{-1})_{jk}^2+\epsilon^2)^{\frac{p}{2}}}$.

Step 2: Obtain $\mathbf{T}_{t+1} = \text{Proj}(\mathbf{T}_t - \alpha G_t(\mathbf{T}))$, where the step size α is chosen by line search.

Stop until the convergence criterion is met. Let t_{max} be the final iteration number.
Output: $\mathbf{T}_{t_{max}}$.

Main Results

Link to L^p Penalized Estimation L^p Penalized Estimation based on the loss function $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ is defined as

$$\hat{\boldsymbol{\theta}}_{\gamma,p} \in \text{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \gamma\sum_{j=1}^{J}\sum_{k=1}^{K} |\lambda_{jk}|^p, \tag{6}$$

where $\gamma > 0$ is a tuning parameter.

Proposition 1. *Consider a fixed $p \in (0, 1]$ and a fixed dataset. Suppose the solution path $\hat{\boldsymbol{\theta}}_{\gamma,p}$ converges to $\hat{\boldsymbol{\theta}}_{0,p}$ when $\gamma \to 0+$. Then, $\hat{\boldsymbol{\theta}}_{0,p}$ can also be obtained by oblique rotation method.*

Proposition 1 indicates instead of obtaining a solution path, one can choose a small γ which induces less bias in penalised estimation. Since the solution of penalised estimation converges to rotation solution, one can select methods according to their efficiency.

Consistency, Selection Consistency and Inference The proposed estimator can recover true sparse loading matrix $\boldsymbol{\Lambda}^*$ and the corresponsing true intercorrelation matrix $\boldsymbol{\Phi}^*$ under the following three conditions.

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

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

C3. Define $\mathcal{D}_1$ and $\mathcal{D}_2$ to be the sets of column permutation and sign flip transformations. $(\boldsymbol{\Lambda}^*, \boldsymbol{\Phi}^*) \in \text{argmin}_{\boldsymbol{\Lambda},\boldsymbol{\Phi}} Q_p(\boldsymbol{\Lambda})$ such that $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}^*\boldsymbol{\Phi}^*\boldsymbol{\Lambda}^{*'}$. In addition, for any other $(\boldsymbol{\Lambda}^\dagger, \boldsymbol{\Phi}^\dagger) \in \text{argmin}_{\boldsymbol{\Lambda},\boldsymbol{\Phi}} Q_p(\boldsymbol{\Lambda})$ such that $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}^*\boldsymbol{\Phi}^*\boldsymbol{\Lambda}^{*'}$, 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 $\tilde{\mathbf{D}}^{-1}\mathbf{D}^{-1}\boldsymbol{\Phi}^\dagger(\mathbf{D}^{-1})'(\tilde{\mathbf{D}}^{-1})' = \boldsymbol{\Phi}^*$.

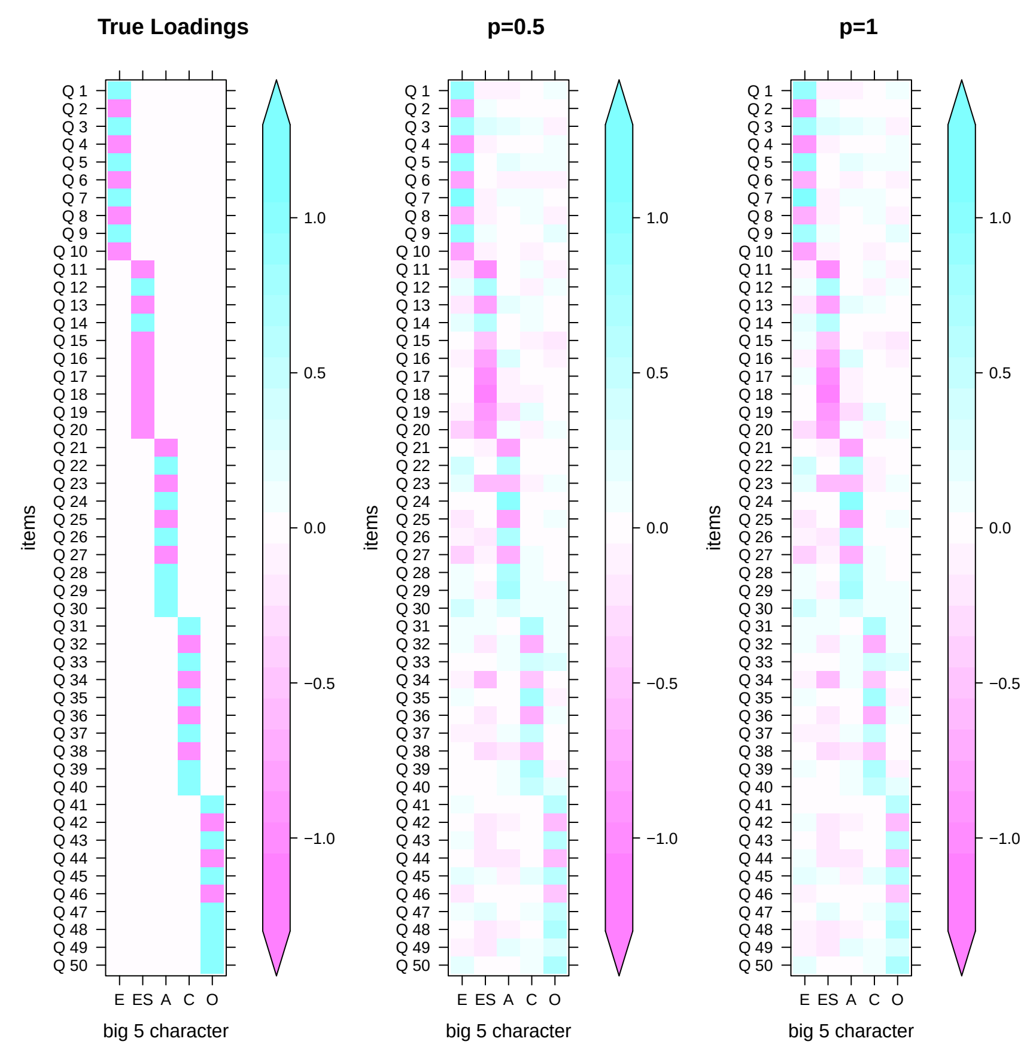
Theorem 1. *Suppose that for a given $p \in (0, 1]$ conditions C1 through C3 hold. Then there exist $\mathbf{D}_N \in \mathcal{D}_1$ and $\tilde{\mathbf{D}}_N \in \mathcal{D}_2$, such that $\hat{\boldsymbol{\Lambda}}_{N,p}\mathbf{D}_N\tilde{\mathbf{D}}_N \overset{pr}{\to} \boldsymbol{\Lambda}^*$ and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

$$(\hat{\boldsymbol{\Lambda}}_{N,p}, \hat{\boldsymbol{\Phi}}_{N,p}) \in \textit{argmin}_{\boldsymbol{\Lambda},\boldsymbol{\Phi}} Q_p(\boldsymbol{\Lambda}),\ \textit{such that } \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' = \hat{\boldsymbol{\Lambda}}_N\hat{\boldsymbol{\Lambda}}_N'.$$

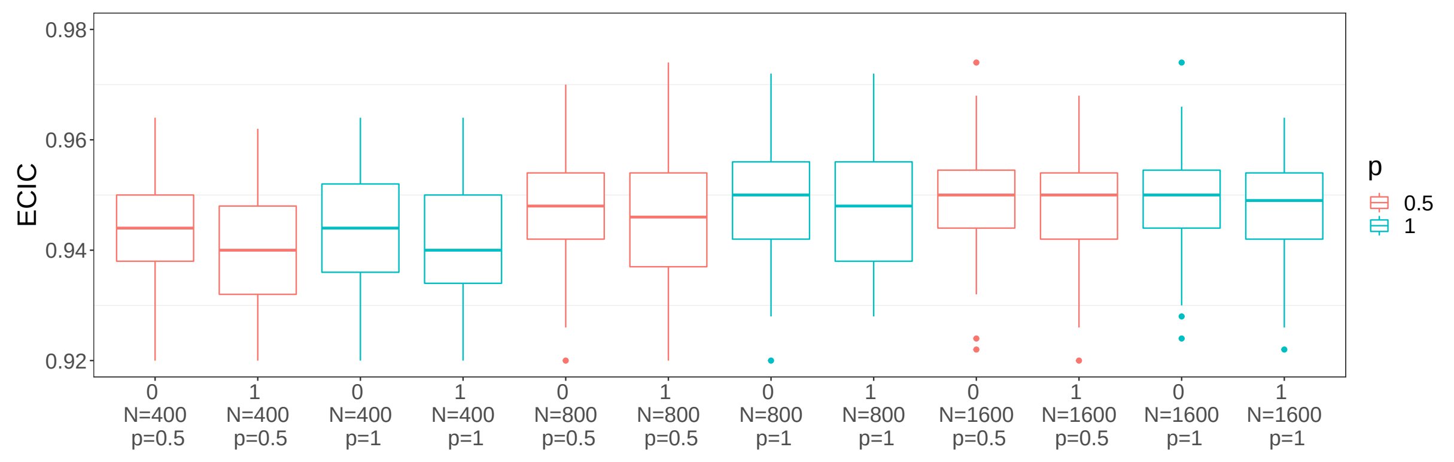
Based on Theorem 1, we can achieve model selection by a Hard-Thresholding procedure, provided the threshold parameter c is smaller than $\min\{|\lambda_{jk}^*| : \lambda_{jk}^* \neq 0\}$.In practice, we choose c based on the Bayesian Information Criterion.
Given the selection consistency, we can construct confidence interval for loadings using a standard inference procedure for CFA which is asymptotically valid.

Experiments

Estimation Consistency, Real Data Example We illustrate the consistency of our method by Big-Five Personality Test[1]. 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. The true loading matrix summarizes the answer key of the online survey.



Entry-wise Confidence Interval Coverage(ECIC) Rate, Simulation We illustrate the validity of the proposed post-selection inference methods by 500 simulations. The true loading matrix is of size 30×5, sparse and with very few cross-loadings. Following is the boxplots of ECIC$_{jk}$. The label 0 means that $\lambda_{jk}^* = 0$ and the label 1 means that $\lambda_{jk}^* \neq 0$.



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.

[1] Robert I Jennrich. “Rotation to simple loadings using component loss functions: The oblique case”. In: *Psychometrika* 71.1 (2006), pp. 173–191.