Towards More Interpretable Factor Analysis: A Focus on Sparsity and Uncertainty

Xinyi Liu

A thesis presented for the degree of

Doctor of Philosophy

Department of Statistics

London School of Economics and Political Science

2025

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree at the London School of Economics and Political Science is solely my own work, except where clearly indicated. In such cases, the extent of any collaboration or joint work with other individuals is explicitly stated.

I declare that this thesis contains 33,641 words.

I confirm that Chapters 2 to 4 were co-authored with Dr. Yunxiao Chen (primary supervisor), Professor Irini Moustaki (secondary supervisor), and Dr. Gabriel Wallin (collaborator). I contributed approximately 80% of the work presented in these chapters.

Chapter 2 is based on the following peer-reviewed publication:

> Liu, X., Wallin, G., Chen, Y., Moustaki, I. (2023). Rotation to Sparse Loadings using L^p Losses and Related Inference Problems. *Psychometrika*, 88(2), 527–553. (Liu et al., 2023)

Xinyi Liu

May 2025

Acknowledgements

As I approach the end of my PhD journey, I find myself reflecting with gratitude on the past four years. I would like to express my heartfelt thanks to my two supervisors, Dr. Yunxiao Chen and Professor Irini Moustaki.

Yunxiao, thank you for consistently addressing my questions with clarity and for helping me navigate challenges with timely and thoughtful guidance. I am deeply grateful to you for introducing me to the field of psychometrics and inspiring my exploration of latent variable models. I have genuinely enjoyed our collaboration on L^p rotation and feel fortunate to have worked on such an intellectually rich topic under your guidance. I deeply admire your approach to research — you work with quiet focus and persistence, cultivating deep expertise with integrity and resilience, undistracted by external noise. I sincerely wish you continued success and many further accomplishments in your academic career.

Irini, thank you for always being encouraging and supportive. You consistently recognized my strengths and uplifted me during difficult times, while offering valuable insights into the practical applications of psychometrics. I greatly admire you as a trailblazing woman in the field, whose achievements have inspired me. Through your subtle influence, you helped me realize that one's professional identity need not be defined by gender, and that it is possible to build a meaningful career with confidence and ambition.

Thank you, Coen, for walking up to me after my talk at IMPS and inviting me to contribute the L^p rotation to the `GPArotation` package. I'm deeply grateful for your meticulous answers to my questions and for discussing the naming and structure of the functions with such care. Without your support, I would never have imagined that my work would be included in one of the most widely used R packages in our field.

Thank you, Gabriel, for your patience, warmth, and enthusiasm. You were always there to celebrate even the smallest milestones in my research. Since your move to Lancaster, I often think back to the time we spent working together.

Thank you, Camilo, for your kindness, humility, and patience — and for not hesitating to ask me questions in return. I deeply appreciate the care and time you've devoted to helping me overcome difficulties.

Thank you, Motto, for igniting my passion for psychology. I truly enjoyed our conversations on neurodiversity, the dark triad, psychometric instruments, and Chinese culture. I believe psychology will remain a lifelong love for me.

I am also thankful to the professors at LSE. Thank you, Tengyao, for welcoming me to your group meetings and for sharing your sharp insights and passion for statistics — your perspective was inspiring. Thank you, Chengchun, for introducing me to reinforcement learning and helping me build meaningful connections at LSE. Thank you, Zoltan, Jouni, Yining, Qiwei and Giulia, for your continued support and care throughout my PhD journey.

I am deeply grateful to my classmates at LSE. You walked with me through the long and winding road of this journey. We grew together, laughed together, and supported one another through difficult yet memorable years. At times I felt vulnerable, but I was always met with kindness, encouragement, and reminders of my strength. Whenever I faced challenges, you were there to lift me up. I will forever treasure these friendships. Thank you to Xuzhi, Tao, Arthur, Xinhui, Shuhan, Jialin, August, Pingfan, Weilin, Yudong, Mingwei, Zetai, Xianghe, Kaifang, Zezhun, Qin, Yirui, Yilin, Ziqi, Ziqing, Di, Liyuan, Jin, Lin, Hongyi, Kai, Kaixing Yutong, Sahoko, Giuseppe, Tim, Alexandros, Shakeel, Pouya, Anica, Tamara, Haziq — and many others whose support I will never forget.

Thank you, Haoling, for showing me through your life experience that we should never limit ourselves — that we must bravely pursue our dreams. Thank you for always providing emotional support. Thank you, Shuning and Tiange, for listening and encouraging me during my most difficult times. Thank you, Jingjie, for reminding me that time brings women wisdom and emotional strength. Thank you, Xuzhi, Xiaowen, and Shuhan, for teaching me to protect my emotions and to express myself bravely.

And thank you to my parents. I know you've worried about me while I've been away, and I feel deeply sorry for not being able to be with you more often. Thank you for always supporting me in doing what I want to do. No matter how late it was, you were always on the other end of the phone, offering your care and presence. For that, I am endlessly grateful.

I would also like to thank my examiners for their time and constructive feedback during the viva, which helped me strengthen this thesis in its final stage.

Abstract

Exploratory Factor Analysis (EFA) is a statistical technique for uncovering latent structures in multivariate data by modeling observed variables as linear combinations of unobserved factors. For interpretability, estimated loading matrices are often rotated to achieve sparsity, but existing rotation methods may lack sufficient accuracy or computational efficiency.

This thesis introduces a new family of rotation criteria for recovering loading matrices with varying sparsity in EFA, based on component-wise L^p loss functions, defined by the objective

$$Q(\mathbf{\Lambda}) = \sum_j \sum_k |\lambda_{jk}|^p.$$

To address the nonsmooth nature of this objective, we develop an iteratively reweighted gradient projection algorithm that achieves high accuracy with significantly reduced computational cost compared to penalized estimation techniques.

We further establish novel identification conditions for the L^p rotation estimator, allowing for a small proportion of non-simple items in the true loading matrix. Empirical results confirm that the L^p rotation criterion consistently outperforms classical rotation methods when the underlying factor structure is sparse.

To support valid inference, we also propose a methodology for computing p-values for factor loadings under the L^p framework. Building on these p-values, we incorporate False Discovery Rate (FDR) control procedures—such as the Benjamini-Yekutieli (BY) and e-value-based Benjamini-Hochberg (eBH) methods—to guide variable selection while controlling the expected proportion of false discoveries. These procedures are demonstrated to remain valid across various experiments.

The proposed L^p rotation framework has been implemented in the R package `GPArotation`, with functions `lpT` and `lpQ` available for orthogonal and oblique solutions, respectively. Overall, this thesis offers a unified approach to estimation, identification, and inference in EFA, contributing both theoretical insights and practical tools for sparse and interpretable factor analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Exploratory Factor Analysis: Model and Formulation

Exploratory Factor Analysis (EFA; Thurstone (1947), Mulaik (2009), Bartholomew et al. (2011)) is a fundamental statistical technique used to uncover latent structures underlying observed multivariate data. It is primarily used as a dimension reduction method, representing observed variables as linear combinations of a smaller number of unobserved latent constructs. In applied work, these latent constructs can represent psychological traits, educational abilities, or consumer preferences, depending on the context. For example, in psychology, EFA is used to identify dimensions such as personality traits or clinical symptoms from item-level survey data (Sellbom and Tellegen, 2019); in education, it helps reveal underlying learning domains based on student assessment results (Harerimana and Mtshali, 2020); and in marketing, it reduces high-dimensional preference data into a few latent factors representing customer tastes (Ghandwani and Hastie, 2024). In the social sciences, EFA is often applied to extract latent attitudes or value systems from survey responses (Williams et al., 2010). EFA is also commonly used in economics and finance, where it helps summarize macroeconomic indicators using a few latent components (Stock and Watson, 1989, 1999). Given N observations of J-dimensional manifest variables $\mathbf{X}_1, \ldots, \mathbf{X}_N$, the EFA model assumes:

$$\mathbf{X}_i = \mathbf{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N, \tag{1.1}$$

where:

- $\mathbf{\Lambda} = (\lambda_{jk})_{J\times K} \in \mathbb{R}^{J\times K}$ is the loading matrix that relates the observed variables to the latent factors;

- $\boldsymbol{\xi}_i \in \mathbb{R}^K$ are i.i.d. latent factors following $\mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$, where $\boldsymbol{\Phi} \in \mathbb{R}^{K \times K}$ is a symmetric positive definite matrix (denoted $\boldsymbol{\Phi} > 0$) with unit diagonal entries $\phi_{kk} = 1$ for $k = 1, \ldots, K$;
- $\boldsymbol{\epsilon}_i \in \mathbb{R}^J$ is the residual noise, assumed to follow $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = (\omega_{ij})_{J \times J}$ is the residual covariance matrix. Conditional independence of the manifest variables given the latent factors is imposed by setting the off-diagonal entries of $\boldsymbol{\Omega}$ to zero.

To simplify notation, we let $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$ denote the collection of all unknown model parameters. The model in Equation (1.1) implies the following marginal distribution for the observed variable $\mathbf{X}$:

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{1.2}$$

where

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Omega}.$$

1.1.1 Differences Between EFA and CFA Models

Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA, Jöreskog (1969a)) are both fundamental techniques for modeling the relationships between observed variables and underlying latent factors. However, they differ substantially in their objectives and modeling assumptions.

In EFA, no prior assumptions are made regarding the specific relationships between observed variables and factors. All factor loadings are freely estimated, allowing the data to determine the underlying structure. The primary goal of EFA is to explore potential factor structures by identifying the minimum number of factors needed to explain the observed covariances among variables.

By contrast, Confirmatory Factor Analysis (CFA) is a hypothesis-driven approach in which the researcher specifies in advance which observed variables are expected to load onto which latent factors, often based on theoretical considerations. In CFA models, certain factor loadings are fixed to zero to reflect these pre-specified structures, and the model evaluates how well the hypothesized factor structure fits the observed data. Thus, EFA is typically used in the early stages of research for structure discovery, while CFA is employed for structure confirmation and theory testing. Moreover, CFA imposes more constraints than EFA, providing a stricter framework for evaluating model fit.

CFA has been widely applied in psychology, education, and the social sciences for validating measurement instruments, testing theoretical constructs such as personality traits, and assessing the structure of standardized tests and questionnaires (Brown, 2015). For example, researchers frequently use CFA to confirm the factor structure of scales like the Big Five personality inventory (John and Srivastava, 1999) or academic motivation scales (Vallerand et al., 1992).

1.2. Sparse Loading Estimation in Exploratory Factor Analysis: Rotations and Regularization

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 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 and Sampson, 1966), geomin (Yates, 1987), simplimax (Kiers, 1994), and componentwise 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., Trendafilov, 2014; Yamamoto et al., 2017; Jin et al., 2018; Geminiani 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.

1.3. Summary and Outline of This Dissertation

In this dissertation, we propose a new rotation criterion—the L^p rotation criterion—

$$Q(\mathbf{\Lambda}) = \sum_j \sum_k |\lambda_{jk}|^p, \tag{1.3}$$

for Exploratory Factor Analysis (EFA) models. As demonstrated in Chapter 2, this criterion is more effective at recovering sparse loading matrices than existing rotation methods. Building on this foundation, we further develop identification theory and uncertainty quantification techniques within the EFA framework. The L^p rotation has been implemented in the R package `GPArotation` (Bernaards and Jennrich, 2005), via the functions `lpT` and `lpQ` for orthogonal and oblique solutions, respectively.

Chapter 2 introduces this new family of oblique rotations based on component-wise L^p loss functions. Due to the nonsmooth nature of these loss functions, we develop an iteratively reweighted gradient projection algorithm to solve the resulting optimization problem efficiently. Our results demonstrate that L^p rotation achieves accuracy comparable to penalized estimation with a small tuning parameter, while significantly reducing computational cost.

In Chapter 3, we establish theoretical results that demonstrate the identification conditions for the L^p rotation estimator. These conditions are novel in that they permit a small proportion of non-simple items in the true loading matrix, relaxing the assumptions typically required in the literature. Additionally, we show that the L^p rotation criterion tends to outperform traditional

rotation methods when the underlying loading matrix is sparse.

Chapter 4 builds upon these results by leveraging the L^p rotation framework to identify EFA models under minimal simple item assumptions. We then propose a method for computing p-values in EFA models, offering a more detailed understanding of the relationships between latent variables and observed items. To further evaluate variable selection accuracy, we incorporate False Discovery Rate (FDR) control based on the computed p-values. Additionally, we establish methodologies and conditions under which traditional FDR control procedures—such as the Benjamini-Yekutieli (BY) procedure and the e-value-based Benjamini-Hochberg procedure (eBH)—remain valid.

Overall, this dissertation introduces a novel rotation framework for EFA, providing both theoretical guarantees and practical tools that enhance interpretability and support valid statistical inference.

Chapter 2

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

2.1. Introduction

In this chapter we propose a new family of oblique rotations based on component-wise L^p loss functions, for $0 < p \leqslant 1$. 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 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 Section 2.2 we propose L^p criteria for oblique rotation, and draw a connection with regularised estimation. In Section 2.3 we discuss statistical inferences based on the proposed rotation method and establish their asymptotic properties, and in Section 2.4 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 Section 2.5 and an application to the Big Five personality assessment in Section 2.6. We conclude this paper with discussions on the limitations of the proposed method and future directions in Section 2.7. Proof of the theoretical results, additional simulation results, and further details of the real application are given in the Appendix for Chapter 2.

2.2. L^p Rotation Criteria

2.2.1 Problem Setup

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

$$\mathbf{X}|\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{\xi}, \boldsymbol{\Omega}), \tag{2.1}$$

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 $\boldsymbol{\Omega}$ 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 (2.1) implies the marginal distribution of $\mathbf{X}$

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{2.2}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Omega}$. Without further constraints, the parameters in (2.2) 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 and Goldberger, 1972), weighted-least-square (Browne, 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

$$\mathcal{M} = \{\mathbf{T} \in \mathbb{R}^{K\times K} : rank(\mathbf{T}) = K,\ (\mathbf{T}'\mathbf{T})_{kk} = 1,\ k = 1, \ldots, K\} \tag{2.3}$$

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

$$\hat{\mathbf{T}} \in \arg\min_{\mathbf{T}\in\mathcal{M}} Q(\hat{\mathbf{A}}\mathbf{T}'^{-1}), \tag{2.4}$$

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

$$(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Phi}}) \in \arg\min_{\boldsymbol{\Lambda},\boldsymbol{\Phi}} Q(\boldsymbol{\Lambda}),\ \text{such that } \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' = \hat{\mathbf{A}}\hat{\mathbf{A}}', \boldsymbol{\Phi} \succ 0,\ \text{and } \phi_{kk} = 1, k = 1, \ldots, K. \tag{2.5}$$

As explained in Remark 1, the minimiser of (2.4), or equivalently that of (2.5), 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 (2.4). It is easy to check that $\hat{\mathbf{T}}\mathbf{D}_1\mathbf{D}_2$ also minimises the objective function (2.4), 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 (Reckase, 2009; Chen et al., 2019, 2021) and machine learning models such as the stochastic blockmodel and latent Dirichlet allocation (see Rohe and Zeng, 2022). These models are all latent variable models involving manifest variables $\mathbf{X}$, multi-dimensional continuous 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*

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

2.2.2 Proposed Rotation Criteria

Jennrich (2004, 2006) proposed a family of monotone concave CLFs for the choice of Q in (2.4), taking the form

$$Q(\boldsymbol{\Lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K} h(|\lambda_{jk}|), \tag{2.6}$$

where $\boldsymbol{\Lambda} = (\lambda_{jk})_{J\times K}$ 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

$$Q_p(\boldsymbol{\Lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K} |\lambda_{jk}|^p. \tag{2.7}$$

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, where 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 $\mathbf{T} \in \mathcal{M}$,*

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

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 $\mathcal{D}_2$.

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

$$Q_p(\mathbf{\Lambda}^*) = \min_{\mathbf{\Lambda}} Q_p(\mathbf{\Lambda}), \text{ such that there exists } \mathbf{\Phi} > 0, \phi_{kk} = 1, k = 1, ..., K, \mathbf{\Lambda}^* \mathbf{\Phi}^* \mathbf{\Lambda}^{*'} = \mathbf{\Lambda} \mathbf{\Phi} \mathbf{\Lambda}'. \tag{2.8}$$

In other words, $\mathbf{\Lambda}^*$ needs to be a stationary point of Q_p. In Figure 2.1 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

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

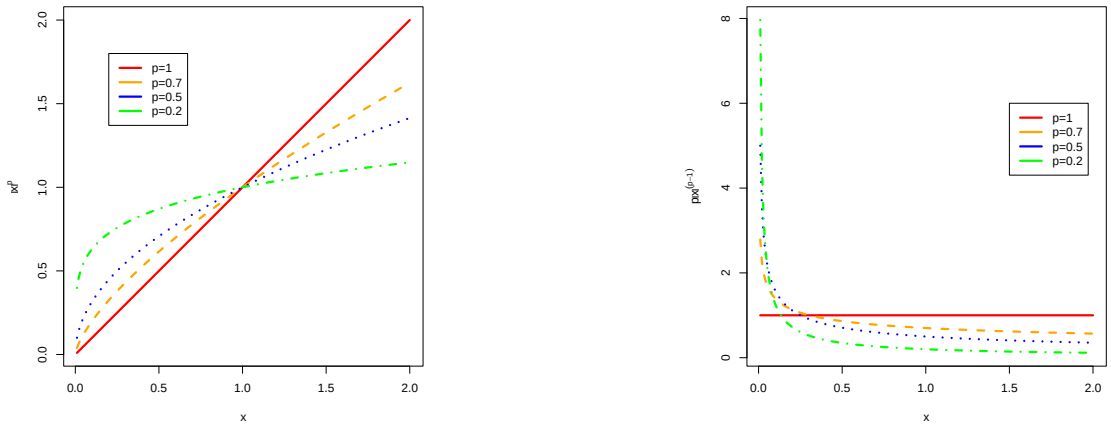
and $\mathbf{\Phi}^*$ is set to be an identity matrix. Note that a 2×2 oblique rotation matrix can be reparameterised by

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

for $\theta_1, \theta_2 \in [0, 2\pi)$. In Figure 2.2 we show the contour plots of $Q_p(\mathbf{\Lambda}^*\mathbf{T}(\theta_1, \theta_2))$, 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 Figure 2.2, 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. Further guidance on the choice of p is provided in Section 2.7 and illustrated through empirical results in Section A2.12. We discuss the computation of the proposed rotation criteria in Section 2.4.

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) 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 A2.8 of the Appendix for Chapter 2.



(a) (b)

Figure 2.1: 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.

2.2.3 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 $\boldsymbol{\Phi}$ is constrained to be an identity matrix. We further 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

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

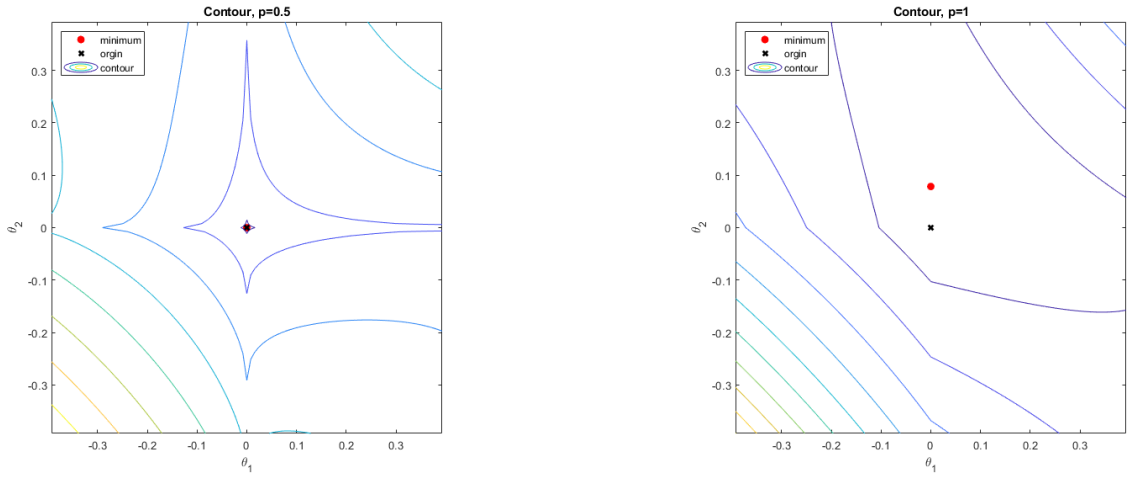
where $\boldsymbol{S} = (\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top)/N$ is the sample covariance matrix.

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

$$\hat{\boldsymbol{\theta}}_{\gamma,p} \in \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \gamma \sum_{j=1}^{J}\sum_{k=1}^{K} |\lambda_{jk}|^p, \tag{2.9}$$

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 (2.9) is also not unique due to column permutation and sign flips similar to the non-uniqueness of optimisation (2.4). We denote the set of minimisers as

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



(a) (b)

Figure 2.2: 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): $p = 1$.

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 (2.9) is similar in spirit to L^p-regularised regression (Mazumder et al., 2011; Lai and Wang, 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*

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

exists, and $\hat{\boldsymbol{\theta}}^{(i)}_{0,p}$ satisfies that $(\hat{\mathbf{\Lambda}}^{(i)}_{0,p}, \hat{\mathbf{\Phi}}^{(i)}_{0,p})$ solves the optimisation problem (2.5) and $\hat{\mathbf{\Omega}}^{(i)}_{0,p} = \hat{\mathbf{\Omega}}$,

where $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{A}}, \mathbf{I}, \hat{\boldsymbol{\Omega}})$ *minimises the loss function* $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$.

We now discuss the implications of this connection. First, if we have a numerical solver for the regularised estimator (2.9), then we can obtain an approximate solution to the L^p-rotation problem (2.5) 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 Section 2.5. We will discuss the computation of these methods in Section 2.4.

2.3. Statistical Inference and Asymptotic Theory

2.3.1 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). Consistency results for the maximum likelihood estimator in continuous factor models are provided in Bartholomew et al. (2011) and Kano (1986). As discussed in Remark 2, the L^p rotation framework is also applicable to other models involving multi-dimensional continuous latent factors. For example, in exploratory multidimensional item response theory (MIRT) models (Haberman, 1977) with ordinal responses,

the estimated loading matrix can benefit from rotation to enhance interpretability, similar to linear factor models. More broadly, the consistency of MLE can be established under the general M-estimation framework described in van der Vaart (2000), which requires standard regularity conditions, including identifiability, continuity of the log-likelihood, compactness (or local compactness) of the parameter space, and a uniform law of large numbers. These results also extend to settings in which the MLE is defined implicitly—such as when the likelihood involves integration over latent variables. For different models, it is important to verify appropriate identification conditions to ensure the uniqueness of the maximizer.

In addition, we require that the EFA model is truly a K-dimensional model, in the sense that condition C2 holds.

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

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. $(\mathbf{\Lambda}^*, \mathbf{\Phi}^*) \in \arg\min_{\mathbf{\Lambda},\mathbf{\Phi}} Q_p(\mathbf{\Lambda})$ such that $\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' = \mathbf{\Lambda}^*\mathbf{\Phi}^*\mathbf{\Lambda}^{*\prime}$. In addition, for any other $(\mathbf{\Lambda}^\dagger, \mathbf{\Phi}^\dagger) \in \arg\min_{\mathbf{\Lambda},\mathbf{\Phi}} Q_p(\mathbf{\Lambda})$ such that $\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' = \mathbf{\Lambda}^*\mathbf{\Phi}^*\mathbf{\Lambda}^{*\prime}$, there exist $\mathbf{D} \in \mathcal{D}_1$ and $\tilde{\mathbf{D}} \in \mathcal{D}_2$, such that $\mathbf{\Lambda}^\dagger\mathbf{D}\tilde{\mathbf{D}} = \mathbf{\Lambda}^*$ and $\tilde{\mathbf{D}}^{-1}\mathbf{D}^{-1}\mathbf{\Phi}^\dagger(\mathbf{D}^{-1})'(\tilde{\mathbf{D}}^{-1})' = \mathbf{\Phi}^*$. 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 $\mathbf{\Lambda}^*$ has a perfect simple structure, i.e., if it has at most one non-zero loading in each row. As we discussed in Section 2.2.2, 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.

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*

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

2.3.2 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

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 $\hat{\mathbf{\Gamma}}_{\hat{c}}$.

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 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).

2.3.3 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.

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 λ_{s1}, ..., λ_{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}), $s = 1, ..., J, k = 1, ..., K$.

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.

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* $(l_{sk}^{(N)}, u_{sk}^{(N)})$*,* $s = 1, ..., J, k = 1, ..., K$*. Then we have* $P(\lambda_{sk}^{*(N)} \in (l_{sk}^{(N)}, u_{sk}^{(N)}))$ *converges to* $1-\alpha$*, for all* $s = 1, ..., J, k = 1, ..., K$*, where* $\lambda_{sk}^{*(N)}$ *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.*

Another remark regarding Algorithm 2 is that it is primarily designed to provide confidence intervals for all factor loadings, including those that are zero. For other parameters—such as the factor covariance matrix and the unique variances of items—we do not recommend using this algorithm for inference. This is because, for each row, we fit a separate CFA model with slightly different identification conditions, making it necessary to aggregate multiple confidence intervals in order to obtain valid inference for these additional parameters.

We refer interested readers to Algorithm 5 and Algorithm 6 in Chapter 4, which provide a more efficient alternative for computing confidence intervals for all parameters in the EFA model. These methods rely on milder assumptions, are computationally less intensive, and yield a single valid confidence interval for each parameter, thus avoiding the need for post hoc aggregation.

2.4. Computation

2.4.1 Proposed IRGP Algorithm

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

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

where Q_p is the L^p CLF defined in (2.7). 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 T_t be the parameter value at the tth iteration.

The function approximation step involves approximating the objective function by

$$G_t(\mathbf{T}) = \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}^{(t)}\left((\hat{\mathbf{A}}\mathbf{T}'^{-1})_{jk}\right)^2, \tag{2.10}$$

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

$$\text{Proj}(\mathbf{T}) = \mathbf{T}(\text{diag}(\mathbf{T}'\mathbf{T}))^{-\frac{1}{2}}, \tag{2.11}$$

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 (2.3). The PGD update is given by

$$\mathbf{T}_{t+1} = \text{Proj}(\mathbf{T}_t - \alpha \boldsymbol{\nabla} G_t(\mathbf{T})), \tag{2.12}$$

where $\alpha > 0$ is a step size chosen by line search and $\boldsymbol{\nabla} G_t(\mathbf{T})$ 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 summarise the IRGP algorithm below.

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 (2.10).

Step 2: Obtain $\mathbf{T}_{t+1}$ using equation (2.12), 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: $\mathbf{T}_{t_{max}}$.

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 Section 2.2.2, 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., $p = 1$).

2.4.2 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 and

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:

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

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

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

We define a proximal operator for the loading matrix as

$$\text{Prox}_{\alpha,\gamma}(\tilde{\boldsymbol{\Lambda}}_t) = \arg\min_{\boldsymbol{\Lambda}} \quad \frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{K} (\lambda_{jk} - \tilde{\lambda}_{jk}^{(t)})^2 + \alpha\gamma \sum_{j=1}^{J} \sum_{k=1}^{K} |\lambda_{jk}|, \tag{2.13}$$

where $\alpha > 0$ will be a step size and $\tilde{\boldsymbol{\Lambda}}_t = (\tilde{\lambda}_{jk}^{(t)})_{J \times K}$ will be the value of $\boldsymbol{\Lambda}$ from the previous step in the proximal gradient algorithm. Note that (2.13) has a closed-form solution given by soft-thresholding (Parikh and 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 $\boldsymbol{\Lambda}$ 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 Appendix for Chapter 2.

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 $\text{Proj}(\cdot)$ is defined in (2.11), 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: $(\mathbf{\Lambda}_{t_{max}}, \mathbf{T}_{t_{max}}, \mathbf{v}_{t_{max}})$.

2.5. Simulation Study

2.5.1 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, quartmin, simplimax, geomin, and promax methods. These methods have been considered in the simulation studies in Jennrich (2006). They are implemented using the `GPArotation` package (Bernaards and 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 Appendix for Chapter 2. 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 1,600. 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 bth 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:

$$\overline{\text{TPR}}_c = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{j,k} \mathbb{1}_{\{\hat{\lambda}_{jk}^{(b,c)} \neq 0, \lambda_{jk}^* \neq 0\}}}{\sum_{j,k} \mathbb{1}_{\{\lambda_{jk}^* \neq 0\}}}, \tag{2.14}$$

   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:

$$\overline{\text{TNR}}_c = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{j,k} \mathbb{1}_{\{\hat{\lambda}_{jk}^{(b,c)} = 0, \lambda_{jk}^* = 0\}}}{\sum_{j,k} \mathbb{1}_{\{\lambda_{jk}^* = 0\}}}. \tag{2.15}$$

   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

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

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

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

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

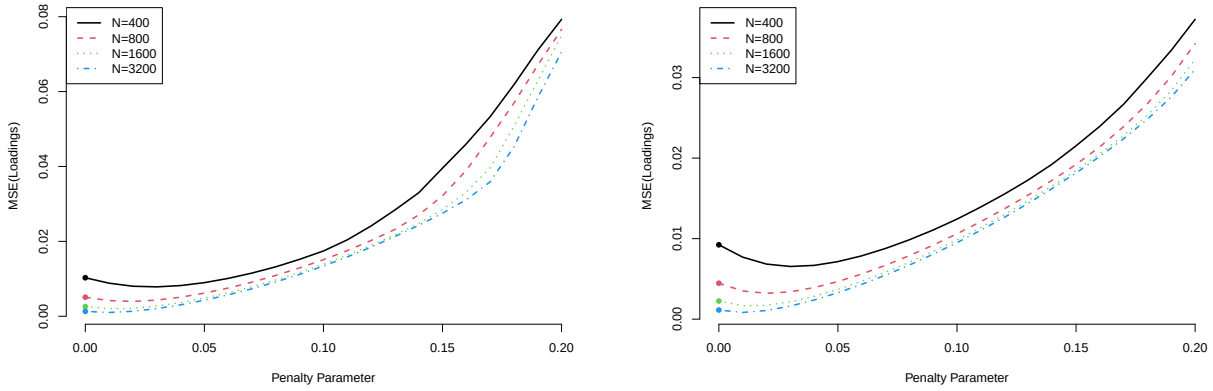
$$\mathrm{ECIC}_{jk} = \frac{\sum_{b=1}^{B}\mathbb{1}_{\{\lambda_{jk}^{*(N)}\in(l_{jk}^{(N)},u_{jk}^{(N)})\}}}{B}.$$

Results on point estimation. In Table 2.1, we present the MSE of the estimated loading matrix, for both simulation settings and $N \in \{400, 800, 1,600\}$. 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 2.2, 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 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.

	15×3			30×5		
	$N = 400$	$N = 800$	$N = 1,600$	$N = 400$	$N = 800$	$N = 1,600$
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

Table 2.1: MSE obtained by using different rotation criteria under various settings, Study I.



(a) (b)

Figure 2.3: The MSE (for loadings) as a function of the tuning parameter γ in the LASSO-regularised estimator. Panel (a): 15×3 settings. Panel (b): 30×5 settings. The dots at $\gamma = 0$ correspond to the L^1 rotation solutions.

Results on confidence intervals. In Figure 2.4, we show boxplots of the ECIC for the L^p rotations, for $p = 0.5$ and $p = 1$ and $N \in \{400, 800, 1,600\}$. 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

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

	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
				$N = 1,600$				
$L^{0.5}$ 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

Table 2.2: The AUC, TR, TPR, and TNR for the L^p-based rotation estimator and the regularised estimator, Study I.

$0.36s$ 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 Section 2.5.2 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 Figure 2.3. The two panels in Figure 2.3 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

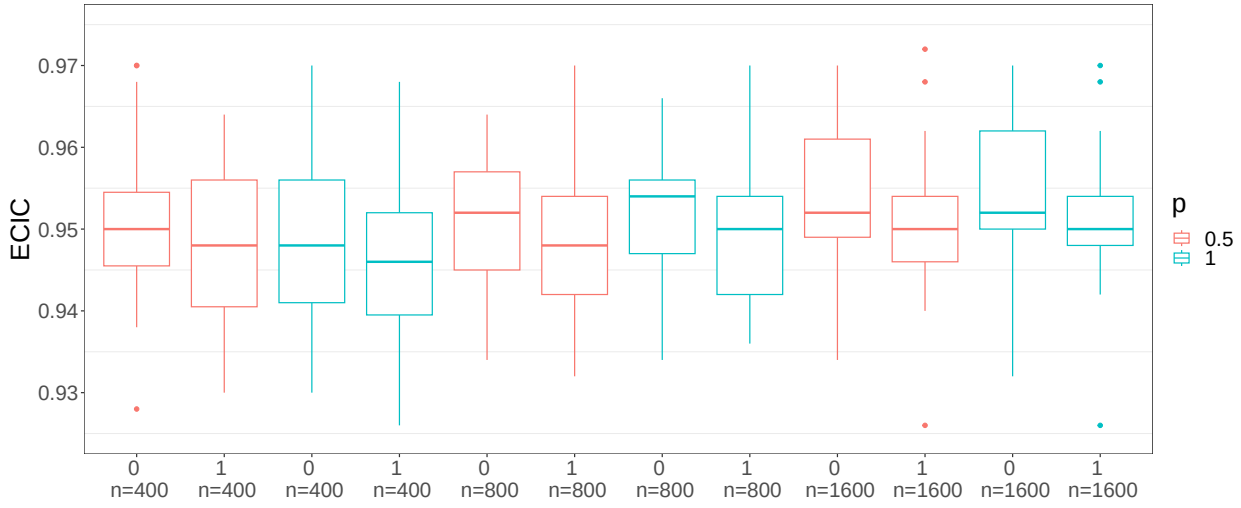
$\gamma = 0$ correspond to the L^1 rotation solutions, as the L^1-rotation estimator is the limit of the LASSO estimator when γ converges to zero (see Proposition 3). As γ increases, the estimation bias increases, and the variance decreases, which results in a U-shaped curve for the MSE – a well-known phenomenon in statistical learning theory (see Chapter 2, Hastie et al., 2009). However, the U-shaped curves in Figure 2.3 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 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.

2.5.2 Study II

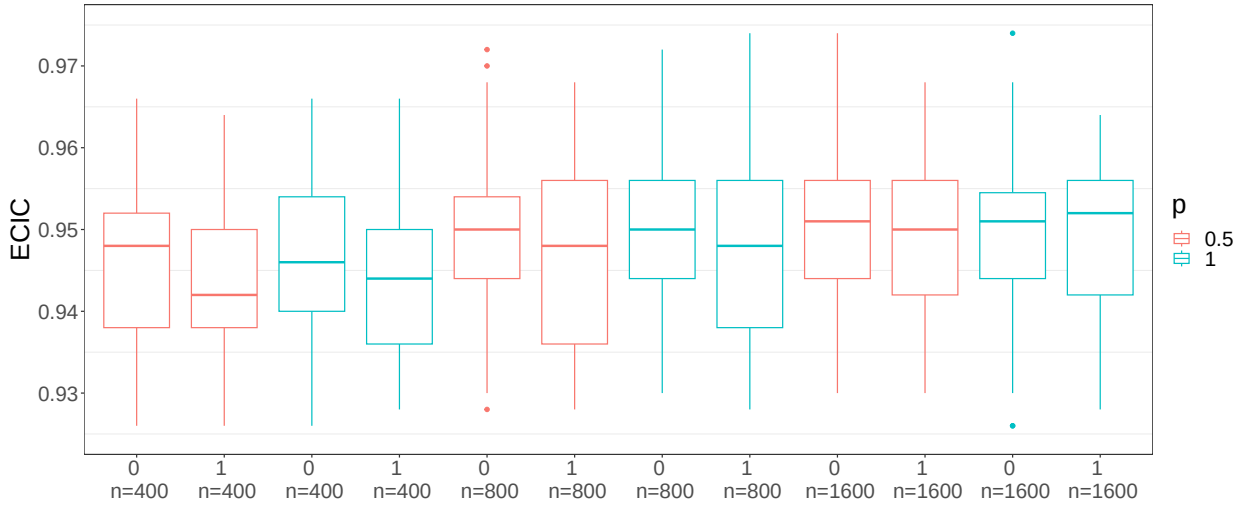
In this study we compare the $L^{0.5}$ and L^1 rotations, under a setting where condition C3 holds for the $L^{0.5}$ rotation but not the L^1 rotation. We chose the setting to somewhat exaggerate the differences, in order to show the consequence of misspecifying p.

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 Appendix for Chapter 2. 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 $N = 3,000$. 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 2.3. 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 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.



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



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

Figure 2.4: Boxplots of ECIC_{jk}. The label 0 means that $\lambda^*_{jk} = 0$ and the label 1 means that $\lambda^*_{jk} \neq 0$.

2.6. 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 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.

	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

Table 2.3: The MSE, AUC, TR, TPR, and TNR for the L^p-based rotation estimator, Study II.

Responses to the items are on a five-level Likert scale, which we treat as continuous variables in the current analysis. To assess the validity of this approximation, we compared the polychoric correlation matrix with the Pearson correlation matrix for the data used in the experiment. The resulting mean squared error (MSE) between the two matrices was 0.000914, indicating that treating the ordinal responses as continuous introduces only a negligible difference.

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 2.4 through 2.7. In Table 2.4 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 and Hughes, 2014; Gow et al., 2005).

In Tables 2.5 through 2.7 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 Appendix for Chapter 2. In Tables 2.5-2.7, 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

[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 "(−)".

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.

	$p = 0.5$					$p = 1$				
	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

Table 2.4: Estimated correlation matrices based on $L^{0.5}$ and L^1 rotations, Big Five personality test

2.7. Concluding Remarks

In this paper we propose a new family of oblique rotations based on component-wise L^p loss functions ($0 < p \leqslant 1$) 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

[4]The R code for the proposed method is available from `https://github.com/yunxiaochen/Lp_rot1129` and the R package `GPArotation` (Bernaards and Jennrich, 2005)

	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)

Table 2.5: 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
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)

Table 2.6: Part II: Point estimates and confidence intervals constructed by $L^{0.5}$, Big Five personality test.

	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)

Table 2.7: Part III: Point estimates and confidence intervals constructed by $L^{0.5}$, Big Five personality test.

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. As discussed earlier and demonstrated by the sensitivity analysis results in Section A2.12, the choice of p in L^p rotation involves a trade-off between statistical accuracy and computational efficiency. Theoretically, a 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 et al., 2019, 2020; Zhang et al., 2020; Chen and Li, 2022). 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.

Appendix for Chapter 2

[List of]**Symbols**

$\mathbf{\Lambda}^*$: $J \times K$ sparse true loading matrix that satisfies Assumption C3.

$\mathbf{\Phi}^*$: $K \times K$ true covariance matrix of the common factors.

$\mathbf{A}^*$**:** $J \times K$ matrix such that $\mathbf{A}^*\mathbf{A}^{*\prime} = \mathbf{\Lambda}^*\mathbf{\Phi}^*\mathbf{\Lambda}^{*\prime}$.

$\hat{\mathbf{A}}$: Initial estimator of the loading matrix.

$\mathbf{T}$: $K \times K$ rotation matrix.

$\mathcal{M}$: The space of oblique rotation matrices, such that

$$\mathcal{M} = \{\mathbf{T} \in \mathbb{R}^{K\times K} : \mathbf{T}'\mathbf{T} \succ \mathbf{0},\ (\mathbf{T}'\mathbf{T})_{ii} = 1,\ i = 1, \ldots, K\}.$$

Q_p : The family of monotone concave CLFs of the form

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

$\hat{\mathbf{T}}$: The solution to the optimisation problem

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

g : A bivariate function for a fixed p such that $g : \mathbb{R}^{J\times K} \times \mathcal{M} \to \mathbb{R}$, which maps $g(\mathbf{A}, \mathbf{T}) \to Q_p(\mathbf{A}\mathbf{T}'^{-\mathbf{1}})$.

$\mathbf{D}$: $K \times K$ matrix such that the columns of $\mathbf{TD}$ are a permutation of those of $\mathbf{T}$.

$\tilde{\mathbf{D}}$: $K \times K$ matrix such that the k:th column of $\mathbf{T}\tilde{\mathbf{D}}$ is either the same as the k:th column of $\mathbf{T}$ or the k:th column of $\mathbf{T}$ multiplied by -1.

$\mathcal{D}_1$: The set of all $K \times K$ permutation matrices.

$\mathcal{D}_2$: The set of all $K \times K$ sign flip matrices.

$\mathcal{T}^*$: The solution to $\arg\min_{\mathbf{T}\in\mathcal{M}} g(\mathbf{A}^*, \mathbf{T})$. If $\mathbf{T}^* = \mathbf{\Phi}^{*1/2}$, then $\mathbf{T}^*$ is the minimiser of $g(\mathbf{A}^*, \mathbf{T})$, and by Conditions C2 and C3, $\mathcal{T}^* = \{\mathbf{T}^*\mathbf{D}\tilde{\mathbf{D}} : \mathbf{D} \in \mathcal{D}_1, \tilde{\mathbf{D}} \in \mathcal{D}_2\}$.

B_ϵ : $B_\epsilon(\mathbf{T}_0) = \{\mathbf{T} \in \mathcal{M} : ||\mathbf{T}_0 - \mathbf{T}||_2 < \epsilon\}$ denotes the ϵ ball around $\mathbf{T}_0$, and $B_\epsilon(\mathcal{T}^*) = \bigcup_{\mathbf{T} \in \mathcal{T}_*} B_\epsilon(\mathbf{T})$ is the union of the ϵ balls around the elements in $\mathcal{T}^*$.

A2.1. Proof of Proposition 1

Proof. On the interval $(0, \infty)$, $h'(x) = px^{p-1} \geqslant 0$ and $h''(x) = p(p-1)x^{p-2} \leqslant 0$ for $p \in (0, 1]$. The function h is hence monotonically increasing and concave on $[0, \infty)$. □

A2.2. Proof of Proposition 2

Proof. The inequality in Proposition 2 is already implied by Theorem 1 in Jennrich (2006) combined with Proposition 1. Here, the focus is thus mainly on the equality condition. It is easy to check that if $\mathbf{T}'^{-1} = \mathbf{D}\tilde{\mathbf{D}}$, then $\mathbf{\Lambda}^*\mathbf{T}'^{-1}$ possesses perfect simple structure and $Q_p(\mathbf{\Lambda}^*\mathbf{T}'^{-1}) = Q_p(\mathbf{\Lambda}^*)$. On the other hand, suppose that $\boldsymbol{A} = \mathbf{\Lambda}^*\mathbf{T}'^{-1}$ for some $\mathbf{T} \in \mathcal{M}$ and $Q_p(\boldsymbol{A}) = \min_{\mathbf{T} \in \mathcal{M}} Q_p(\mathbf{\Lambda}^*\mathbf{T}'^{-1}) = Q_p(\mathbf{\Lambda}^*)$. Due to $\mathbf{\Lambda}^* = \boldsymbol{A}\mathbf{T}'$ and since $(\mathbf{T}'\mathbf{T})_{kk} = 1$, $k = 1, \ldots, K$, implies that $||t_k||_2 = 1$ for all columns in $\mathbf{T}$, each row in Λ^* can be expressed as

$$\lambda_j^* = a_{j1}t_1' + a_{j2}t_2' + ... + a_{jK}t_K',$$

$j = 1, \ldots, J$. By evaluating the left and right hand side in terms of their ℓ_2 norm, and by applying the triangle inequality, we get that

$$||\lambda_i^*||_2 \leqslant \sum_k |a_{ik}|||t_k'||_2 = \sum_k |a_{ik}|. \tag{2.16}$$

Now, let λ_{is}^* be the only non-zero entry in λ_i^*. By raising it to the p-th power $(0 < p \leqslant 1)$ and applying Lemma 2 in Jennrich (2006),

$$|\lambda_{is}^*|^p \leqslant (\sum_k |a_{ik}|)^p \leqslant \sum_k |a_{ik}|^p. \tag{2.17}$$

Therefore, to achieve $Q_p(\boldsymbol{A}) = Q_p(\mathbf{\Lambda}^*)$, (2.17) needs to hold as an equality for all i, which further implies that (2.16) holds as an equality for all i as well. However, since $t_1, t_2, t_3, ..., t_K$ are linearly independent, (2.16) holds as an equality if and only if exactly one of $a_{i1}, a_{i2}, ...a_{ik}$ is nonzero for a certain i. Suppose $a_{ij} \neq 0$. Since $\lambda_i^* = a_{ij}t_j'$ and t_j has unit length,

$$t_j = (0, 0, \ldots \frac{\lambda_{is}^*}{a_{ij}}, \ldots, 0)' \in \{+e_s, -e_s\},$$

where e_s is a column vector of length K with 1 on its s:th entry. Since $\text{rank}(\mathbf{T}) = K$, the only possible form of $\mathbf{T}$ is a permutation of $[\pm e_1, \pm e_2, ..., \pm e_K]$. Therefore, $\mathbf{T}$ can be written as $\mathbf{D}\tilde{\mathbf{D}}$ for some $\mathbf{D} \in \mathcal{D}_1$ and $\tilde{\mathbf{D}} \in \mathcal{D}_2$. As $\mathbf{T}'^{-1} = \mathbf{D}'^{-1}\tilde{\mathbf{D}}'^{-1}$, we can easily verify that $\mathbf{D}'^{-1} \in \mathcal{D}_1$ and $\tilde{\mathbf{D}}'^{-1} \in \mathcal{D}_2$ and arrive at the result. □

A2.3. Proof of Proposition 3

Proof. For any $\gamma > 0$, $\hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p}$ and $\hat{\boldsymbol{\theta}}$ achieve the minimum of $L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \gamma Q_p(\boldsymbol{\Lambda})$ and $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ respectively. It follows that

$$L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})) + \gamma Q_p(\hat{\boldsymbol{\Lambda}}) \geqslant L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p})) + \gamma Q_p(\hat{\boldsymbol{\Lambda}}^{(i)}_{\gamma,p}) \geqslant L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p})) \geqslant L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})). \tag{2.18}$$

Therefore, when $\gamma \to 0+$, we have that $L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})) + \gamma Q_p(\hat{\boldsymbol{\Lambda}}) \to L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))$. By the Squeeze theorem (page 104, Sohrab, 2003), $L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p})) \to L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))$ when $\gamma \to 0+$. Since $L(\boldsymbol{\Sigma}(\cdot))$ is a continuous function,

$$L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(i)}_{0,p})) = \lim_{\gamma \to 0+} L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(i)}_{\gamma,p})) = L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})) = \min_{\boldsymbol{\theta}} L(\boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

If $\hat{\boldsymbol{\theta}}^{(i)}_{0,p}$ does not solve the optimisation problem in (2.5) in the main article, there exists a

$$\boldsymbol{\theta}' = (\boldsymbol{\Lambda}', \boldsymbol{\Phi}', \boldsymbol{\Omega}') \text{ s.t. } Q_p(\boldsymbol{\Lambda}') < Q_p(\hat{\boldsymbol{\Lambda}}^{(i)}_{0,p}), \text{ and } L(\boldsymbol{\Sigma}(\boldsymbol{\theta}')) = \min_{\boldsymbol{\theta}} L(\boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

Since Q_p is a continuous function, there exists a γ_0, s.t. $Q_p(\boldsymbol{\Lambda}') < Q_p(\hat{\boldsymbol{\Lambda}}^{(i)}_{\gamma_0,p})$ and $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}')) \leqslant L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(i)}_{\gamma_0,p}))$, where the latter is because $\boldsymbol{\theta}'$ minimises $L(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Therefore,

$$L(\boldsymbol{\Sigma}(\boldsymbol{\theta}')) + \gamma_0 Q_p(\boldsymbol{\Lambda}') < L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(i)}_{\gamma_0,p})) + \gamma_0 Q_p(\hat{\boldsymbol{\Lambda}}^{(i)}_{\gamma_0,p}),$$

which contradicts that $\hat{\boldsymbol{\theta}}^{(i)}_{\gamma_0,p}$ achieves the minimum of $L(\boldsymbol{\Sigma}(\boldsymbol{\theta})) + \gamma_0 Q_p(\boldsymbol{\Lambda})$. □

A2.4. Proof of Theorem 1

We fix p throughout the proof and suppress it as a subscript for all estimators and some of the functions for ease of notation. We will add subscript N when we are considering an estimator applied to a sample of size N. Let $\mathcal{D}(\mathbf{A})$ be the set of all column permutations and sign flips of the matrix $\mathbf{A}$, i.e, $\mathcal{D}(\mathbf{A}) = \{\mathbf{AD}\tilde{\mathbf{D}} : \mathbf{D} \in \mathcal{D}_1, \tilde{\mathbf{D}} \in \mathcal{D}_2\}$. Let $||\cdot||_{max}$ denote the maximum entry in the matrix, $||\mathbf{A}||_{max} = \max_{i,j} |A_{ij}|$. Let $||\cdot||_2$ denote the matrix norm induced by the vector 2-norm,

$||\mathbf{A}||_2 = \max_{||x||_2=1} ||\mathbf{A}x||_2 = \sqrt{\max \text{eig}(\mathbf{A}'\mathbf{A})} = d_1(\mathbf{A})$, where we use $d_k(\mathbf{A})$ to represent the k:th largest singular value of $\mathbf{A}$.

Proof. By Lemma 4, we can find a δ for any $\epsilon_{dist} > 0$, so that as long as $||\mathbf{A} - \mathbf{A}^*||_2 \leqslant \delta$, $\hat{\mathbf{T}} \in B_{\epsilon_{dist}}(\mathcal{T}^*)$. By Lemma 5, there exists a sequence of orthogonal matrices $\{\mathbf{O}_N\}$, such that

$$\hat{\mathbf{A}}_N \mathbf{O}_N \xrightarrow{pr} \mathbf{A}^* \tag{2.19}$$

Therefore, for any $\epsilon_{prob} > 0$, there exists an N_0 so that when $N > N_0$, $\mathbb{P}(||\hat{\mathbf{A}}_N \mathbf{O}_N - \mathbf{A}^*||_2 \leqslant \delta) \geqslant 1 - \epsilon_{prob}$. Consequently, $\mathbb{P}(\hat{\mathbf{T}}_N \in B_{\epsilon_{dist}}(\mathcal{T}^*)) \geqslant 1 - \epsilon_{prob}$, where

$$\hat{\mathbf{T}}_N = \text{argmin}_{\mathbf{T} \in \mathcal{M}} g(\hat{\mathbf{A}}_N \mathbf{O}_N, \mathbf{T}).$$

Thus, by Condition C3, there exists $\mathbf{D}_N \in \mathcal{D}_1$ and $\tilde{\mathbf{D}}_N \in \mathcal{D}_2$ so that $\hat{\mathbf{T}}_N \mathbf{D}_N \tilde{\mathbf{D}}_N \xrightarrow{pr} \mathbf{\Phi}^{*1/2}$. By the continuous mapping theorem, $\hat{\mathbf{T}}_N'^{-1} \mathbf{D}_N'^{-1} \tilde{\mathbf{D}}_N'^{-1} \xrightarrow{pr} \mathbf{\Phi}^{*'-1/2}$. Combined with (2.19) and Slutsky's theorem, we have that

$$\hat{\mathbf{A}}_N (\mathbf{O}_N \hat{\mathbf{T}}_N)'^{-1} \mathbf{D}_N'^{-1} \tilde{\mathbf{D}}_N'^{-1} \xrightarrow{pr} \mathbf{\Lambda}^*,$$

since $\mathbf{O}_N \hat{\mathbf{T}}_N = \text{argmin}_{\mathbf{T} \in \mathcal{M}} g(\hat{\mathbf{A}}_N, \mathbf{T})$ and $\hat{\mathbf{\Lambda}}_N = \hat{\mathbf{A}}_N (\mathbf{O}_N \hat{\mathbf{T}}_N)'^{-1}$. Lastly,

$$\hat{\mathbf{\Phi}}_N = \hat{\mathbf{T}}_N' \hat{\mathbf{T}}_N \xrightarrow{pr} \tilde{\mathbf{D}}_N'^{-1} \mathbf{D}_N'^{-1} \mathbf{\Phi}^* \mathbf{D}_N^{-1} \tilde{\mathbf{D}}_N^{-1}$$

where $\mathbf{D}_N'^{-1} \in \mathcal{D}_1, \tilde{\mathbf{D}}_N'^{-1} \in \mathcal{D}_2$, which concludes the proof. □

A2.4.1 Proof of Lemmata 1 to 5

To prove Lemma 4, we will use the property of a continuous function on a compact set. Firstly, let $\mathcal{M}' = \{\mathbf{T} \in \mathbb{R}^{K \times K} : (\mathbf{T}'\mathbf{T})_{kk} = 1, k = 1, \ldots, K\}$. Note that the space of oblique rotation matrices $\mathcal{M}$ can be written as

$$\mathcal{M} = \mathcal{M}' \cap \{\mathbf{T} \in \mathbb{R}^{K \times K} : \text{rank}(\mathbf{T}) = K\}.$$

It follows that $\mathcal{M}$ is not a compact set since $\mathbf{T}$ is invertible, as $d_K(\mathbf{T}) > 0$. In Corollary 1, we therefore first show that if the initial matrix $\hat{\mathbf{A}}$ is in a neighborhood of $\mathbf{A}^*$, i.e, in

$$\bar{\mathcal{B}} = \left\{ \mathbf{A} : ||\mathbf{A} - \mathbf{A}^*||_2 \leqslant \frac{d_K(\mathbf{A}^*)}{2} \right\},$$

then $\hat{\mathbf{T}}$ lies in a compact subset $\overline{\mathcal{M}}$ of $\mathcal{M}$, where

$$\overline{\mathcal{M}} = \mathcal{M}' \cap \left\{ \mathbf{T} \in \mathbb{R}^{K \times K} : d_K(\mathbf{T}) \geqslant \min\left(\frac{d_K(\mathbf{A}^*)}{4\sqrt{JK}\, g_{\max}^{1/p}}, 1\right) \right\}.$$

The maximum $g_{max} = \max_{\mathbf{A} \in \bar{\mathcal{B}}} g(\mathbf{A}, \mathbf{I})$ is attainable since g is continuous and $\bar{\mathcal{B}}$ is compact. Note that $\overline{\mathcal{M}}$ is not empty since $\mathbf{I} \in \overline{\mathcal{M}}$, and $d_K(\mathbf{I}) = 1$.

To prove Corollary 1, we need to prove that if $\mathbf{T}$ is nearly invertible, i.e, its smallest singular value is very small, then it can not be the minimizer of $g(\mathbf{A}, \mathbf{T})$ if $\mathbf{A} \in \bar{\mathcal{B}}$. To make this argument, we will use the matrix inequality in Lemma 1, and Weyl's bound in Lemma 2.

Lemma 1. $||\mathbf{A}\mathbf{T}'^{-1}||_{max} \geqslant \frac{d_K(\mathbf{A})}{\sqrt{JK}} ||\mathbf{T}^{-1}||_2$

Proof. By the norm equivalence of a matrix (chapter 10.4.4, page 62, Petersen and Pedersen, 2012),

$$||\mathbf{A}\mathbf{T}'^{-1}||_{max} \geqslant \frac{1}{\sqrt{JK}} ||\mathbf{A}\mathbf{T}'^{-1}||_2 \tag{2.20}$$

Denote the thin singular value decomposition of $\mathbf{A}$ as $\mathbf{U}\mathbf{D}\mathbf{V}'$, where $\mathbf{U}$ is a $J \times K$ matrix with orthogonal columns, $\mathbf{D}$ is a $K \times K$ diagonal matrix whose diagonal entries $\mathbf{D}_{kk} = d_i(\mathbf{A})$, where $d_k(\mathbf{A})$ is the k:th largest singular value of $\mathbf{A}$, and $\mathbf{V}$ is a $K \times K$ orthogonal matrix. When $d_K(\mathbf{A}) = 0$, the statement is trivial, and when $d_K(\mathbf{A}) > 0$, $\mathbf{D}$ is invertible. Therefore

$$\begin{aligned}
||\mathbf{A}\mathbf{T}'^{-1}||_2 &= \sup_{||\mathbf{x}||_2=1} |\mathbf{x}'\mathbf{T}^{-1}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{T}'^{-1}\mathbf{x}|^{1/2} \\
&= ||\mathbf{D}\mathbf{V}'\mathbf{T}'^{-1}||_2 \\
&= \sup_{||\mathbf{x}||_2=1} ||\mathbf{x}'\mathbf{D}||_2 || \frac{\mathbf{x}'\mathbf{D}}{||\mathbf{x}'\mathbf{D}||_2} \mathbf{V}'\mathbf{T}'^{-1}||_2 \\
&\geqslant \inf_{||\mathbf{x}||_2=1} ||\mathbf{x}'\mathbf{D}||_2 \cdot \sup_{||\mathbf{x}||_2=1} || \frac{\mathbf{x}'\mathbf{D}}{||\mathbf{x}'\mathbf{D}||_2} \mathbf{V}'\mathbf{T}'^{-1}||_2 \\
&= d_K \sup_{||\mathbf{y}||_2=1} ||\mathbf{y}'\mathbf{V}'\mathbf{T}'^{-1}||_2 \\
&= d_K ||\mathbf{V}'\mathbf{T}'^{-1}||_2 \\
&= d_K ||\mathbf{T}^{-1}||_2
\end{aligned}$$

Plug $||\mathbf{A}\mathbf{T}'^{-1}||_2 = d_K ||\mathbf{T}^{-1}||_2$ into (2.20) and we get the result. □

Lemma 2 (Weyl's bound, (Weyl, 1912))**.** *For a $J \times K$ matrix $\mathbf{A}$, suppose $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, where $\mathbf{E}$*

represents a perturbation matrix, then we have

$$\max_{1\leqslant k\leqslant \min\{J,K\}} |d_k(\mathbf{A}) - d_k(\hat{\mathbf{A}})| \leqslant ||\mathbf{A} - \hat{\mathbf{A}}||_2.$$

We refer interested readers to Theorem 7 in O'Rourke et al. (2018)

Corollary 1. *Under condition C2,* $\mathbf{A}^*$ *is full rank, so* $d_K(\mathbf{A}^*) > 0$. *Then, when* $\mathbf{A} \in \bar{\mathcal{B}}$,

$$\arg\min_{\mathbf{T}\in\mathcal{M}} g(\mathbf{A}, \mathbf{T}) \subseteq \overline{\mathcal{M}}$$

Proof. If $\mathbf{T} \in \mathcal{M}\backslash\overline{\mathcal{M}}$, then $||\mathbf{T}^{-1}||_2 = \frac{1}{d_K(\mathbf{T})} > \frac{4\sqrt{JK}g_{\max}^{1/p}}{d_K(\mathbf{A}^*)}$. Thus, by Lemma 1,

$$g(\mathbf{A}, \mathbf{T}) \geqslant (||\mathbf{A}\mathbf{T}'^{-\mathbf{1}}||_{max})^p \geqslant (\frac{d_K(\mathbf{A})}{\sqrt{JK}}||\mathbf{T}'^{-\mathbf{1}}||_2)^p > (\frac{4d_K(\mathbf{A})g_{\max}^{1/p}}{d_K(\mathbf{A}^*)})^p.$$

When $\mathbf{A} \in \bar{\mathcal{B}}$, by Lemma 2, $|d_K(\mathbf{A}) - d_K(\mathbf{A}^*)| \leqslant ||\mathbf{A} - \mathbf{A}^*||_2 \leqslant \frac{d_K(\mathbf{A}^*)}{2}$, so $d_K(\mathbf{A}) \geqslant \frac{d_K(\mathbf{A}^*)}{2}$. Thus,

$$g(\mathbf{A}, \mathbf{T}) > 2^p g_{max} \geqslant g(\mathbf{A}, \mathbf{I}) \geqslant \min_{\mathbf{T}\in\mathcal{M}} g(\mathbf{A}, \mathbf{T})$$

which contradicts that $\mathbf{T}$ is a minimizer. □

Next, we will prove that if $\mathbf{T}$ is not in a neighborhood of $\mathcal{T}^*$, then there will be a gap between $g(\mathbf{A}^*, \mathbf{T})$ and the minimum.

Lemma 3. *Define* $\epsilon_0 = sup\{\epsilon > 0 : \overline{\mathcal{M}} \cap B_\epsilon(\mathcal{T}^*)^{\mathrm{C}} \neq \varnothing\}$, *which is achievable since* $\overline{\mathcal{M}}$ *is compact. Then, for all positive* $\epsilon < \epsilon_0$, *there exists a* $\delta > 0$, *such that if* $\mathbf{T} \in \overline{\mathcal{M}} \cap B_\epsilon(\mathcal{T}^*)^{\mathrm{C}}$,

$$g(\mathbf{A}^*, \mathbf{T}) - c^*_{min} \geqslant \delta \tag{2.21}$$

where $c^*_{min} \coloneqq \min_{T\in\mathcal{M}} g(\mathbf{A}^*, \mathbf{T})$.

Proof. If the statement does not hold, there exists an $\epsilon' < \epsilon_0$ for all $\delta_m = \frac{1}{m}, m \in \mathbb{N}$, such that $\mathbf{T}_m \in \overline{\mathcal{M}} \cap B_{\epsilon'}(\mathcal{T}^*)^{\mathrm{C}}$. However, $0 < g(\mathbf{A}^*, \mathbf{T}_m) - c^*_{min} < \frac{1}{m}$. Since $\overline{\mathcal{M}} \cap B_{\epsilon'}(\mathcal{T}^*)^{\mathrm{C}}$ is a closed subset of a compact set, it is compact. Therefore, by the Bolzano–Weierstrass theorem (Fitzpatrick, 2009, p.52), there exists a sub-sequence $\{\mathbf{T}_{m_k}\} \subseteq \{\mathbf{T}_m\}$ and a point $\mathbf{T}_0 \in \overline{\mathcal{M}} \cap B_{\epsilon'}(\mathcal{T}^*)^{\mathrm{C}}$, which satisfies $\mathbf{T}_{m_k} \to \mathbf{T}_0$ when $k \to \infty$. However, since $g(\mathbf{A}^*, \mathbf{T})$ is a continuous function of $\mathbf{T}$ when $\mathbf{A}^*$ is fixed, $g(\mathbf{A}^*, \mathbf{T}_0) = \lim_{k\to\infty} g(\mathbf{A}^*, \mathbf{T}_{m_k}) = c^*_{min}$, so $\mathbf{T}_0 \in \mathcal{T}^* \subseteq B_{\epsilon'}(\mathcal{T}^*)$, which makes a contradiction.

□

We can now prove that $\mathbf{T}$ must be close to $\mathcal{T}^*$ if $\mathbf{A}$ is close enough to $\mathbf{A}^*$. We present this result in Lemma 4.

Lemma 4. *Under condition C2, for any $\epsilon < \epsilon_0$, there exists a $\delta > 0$, s.t. if $||\mathbf{A} - \mathbf{A}^*||_2 \leqslant \delta$, then $\mathbf{T} \in B_\epsilon(\mathcal{T}^*)$.*

Proof. For any $\epsilon < \epsilon_0$, let δ_1 be the lower bound of $g(\mathbf{A}^*, \mathbf{T}) - c^*_{min}$ for $\mathbf{T} \in \overline{\mathcal{M}} \cap B_\epsilon(\mathcal{T}^*)^{\mathrm{C}}$ in Lemma 3. Because $\mathbf{\Omega} = \bar{\mathcal{B}} \times \overline{\mathcal{M}}$ is a compact set in the domain of $(\mathbf{A}, \mathbf{T})$ and g is continuous on $\mathbf{\Omega}$, g is uniformly continuous on $\mathbf{\Omega}$. Therefore, for $\frac{\delta_1}{3}$, there exists a $\delta_2 > 0$ s.t. whenever $||\mathbf{A} - \mathbf{A}^*||_2 \leqslant \delta_2$, $|g(\mathbf{A}, \mathbf{T}) - g(\mathbf{A}^*, \mathbf{T})| < \frac{\delta_1}{3}$, for all $\mathbf{T} \in \overline{\mathcal{M}}$. Let $\delta = \min(\frac{d_K(\mathbf{A}^*)}{2}, \delta_2)$. When $||\mathbf{A} - \mathbf{A}^*||_2 \leqslant \delta$ and $\mathbf{T} \in \overline{\mathcal{M}} \cap B_\epsilon(\mathcal{T}^*)^{\mathrm{C}}$,

$$\begin{aligned} g(\mathbf{A}, \mathbf{T}) - g(\mathbf{A}, \mathbf{T}^*) \geqslant & (g(\mathbf{A}, \mathbf{T}) - g(\mathbf{A}^*, \mathbf{T})) + (g(\mathbf{A}^*, \mathbf{T}) - g(\mathbf{A}^*, \mathbf{T}^*)) + \\ & (g(\mathbf{A}^*, \mathbf{T}^*) - g(\mathbf{A}, \mathbf{T}^*)) \\ \geqslant & -\frac{\delta_1}{3} + \delta_1 - \frac{\delta_1}{3} = \frac{\delta_1}{3} \end{aligned}$$

which means that $\mathbf{T}$ can not be the minimiser. □

In Lemma 5, we prove that after an orthogonal transformation, $\hat{\mathbf{A}}_N$ lies in a small neighborhood of $\mathbf{A}^*$ asymptotically with probability 1.

Lemma 5. *Under conditions C1 and C2, there exists a sequence of orthogonal matrices $\{\mathbf{O}_N\}$ such that $\hat{\mathbf{A}}_N \mathbf{O}_N \xrightarrow{pr} \mathbf{A}^*$.*

Proof. By condition C1, $\hat{\mathbf{A}}_N \hat{\mathbf{A}}'_N \xrightarrow{pr} \mathbf{A}^* \mathbf{A}^{*\prime}$. After multiplying both sides with $\mathbf{A}^*$ and rearranging, we get that

$$\hat{\mathbf{A}}_N \hat{\mathbf{A}}'_N \mathbf{A}^* - \mathbf{A}^* \mathbf{A}^{*\prime} \mathbf{A}^* \xrightarrow{pr} 0.$$

By condition C2, $\mathrm{rank}(\mathbf{A}^{*'}\mathbf{A}^*) = \mathrm{rank}(\mathbf{A}^*\mathbf{A}^{*'}) = K$, so $(\mathbf{A}^{*'}\mathbf{A}^*)^{-1}$ exists. Thus,

$$\hat{\mathbf{A}}_N \hat{\mathbf{A}}'_N \mathbf{A}^* (\mathbf{A}^{*'} \mathbf{A}^*)^{-1} - \mathbf{A}^* \xrightarrow{pr} 0. \tag{2.22}$$

Define $\mathbf{B}_N = \hat{\mathbf{A}}'_N \mathbf{A}^* (\mathbf{A}^{*'} \mathbf{A}^*)^{-1}$, and $\mathbf{O}_N = \mathbf{B}_N (\mathbf{B}'_N \mathbf{B}_N)^{-1/2}$. Then $\mathbf{O}_N$ is an orthogonal matrix. Therefore, we only need to prove that $\mathbf{O}_N$ forms the desired sequence of matrices. Let $\Delta_N = \mathbf{O}_N - \mathbf{B}_N$. Then

$$\begin{aligned} ||\hat{\mathbf{A}}_N \mathbf{O}_N - \mathbf{A}^*||_F = & ||\hat{\mathbf{A}}_N (\mathbf{B}_N + \Delta_N) - \mathbf{A}^*||_F \\ \leqslant & ||\hat{\mathbf{A}}_N \mathbf{B}_N - \mathbf{A}^*||_F + ||\hat{\mathbf{A}}_N \Delta_N||_F, \end{aligned} \tag{2.23}$$

where $||\cdot||_F$ denotes the Frobenius norm and the first term on the right-hand side of the inequality converges to 0 in probability according to (2.22). For the second term, we have that $||\hat{\mathbf{A}}_N\Delta_N||_F \leqslant ||\hat{\mathbf{A}}_N||_F||\Delta_N||_F$ by the sub-multiplicativity of the Frobenius norm. To control $||\hat{\mathbf{A}}_N||_F$, we can show that under condition C1, $||\hat{\mathbf{A}}_N||_F = \sqrt{tr(\hat{\mathbf{A}}'_N\hat{\mathbf{A}}_N)} = \sqrt{tr(\hat{\mathbf{A}}_N\hat{\mathbf{A}}'_N)} \xrightarrow{pr} \sqrt{tr(\mathbf{A}^*\mathbf{A}^{*\prime})} = ||\mathbf{A}^*||_F$. It is thus bounded. To control Δ_N, we use Theorem 4.1 in Higham (1988), which states that

$$||\Delta_N||_F = \sqrt{\sum_{k=1}^{K}(1 - d_i(\mathbf{B}_N))^2} \tag{2.24}$$

where $d_k(\mathbf{B}_N)$ is the k:th largest singular value of $\mathbf{B}_N$. Define $\mathbf{A}^+ = \mathbf{A}^*(\mathbf{A}^{*'}\mathbf{A}^*)^{-1}$ and $\mathbf{E}_N = \hat{\mathbf{A}}_N\hat{\mathbf{A}}'_N - \mathbf{A}^*\mathbf{A}^{*'}$. Then

$$\mathbf{B}'_N\mathbf{B}_N = (\mathbf{A}^{*'}\mathbf{A}^*)^{-1}\mathbf{A}^{*'}(\mathbf{A}^*\mathbf{A}^{*'} + \mathbf{E}_N)\mathbf{A}^*(\mathbf{A}^{*'}\mathbf{A}^*)^{-1} = \mathbf{I} + (\mathbf{A}^{+'})\mathbf{E}_N\mathbf{A}^+, \tag{2.25}$$

and

$$\begin{aligned}\max_{1\leqslant i\leqslant K}|d_i(\mathbf{B}_N)^2 - 1| &= \max_{1\leqslant i\leqslant K}|d_i(\mathbf{B}'_N\mathbf{B}_N) - d_i(\mathbf{I})| \\ &\leqslant ||(\mathbf{A}^+)'\mathbf{E}_N\mathbf{A}^+||_2 \\ &\leqslant ||\mathbf{E}_N||||\mathbf{A}^+||_2^2 \xrightarrow{pr} 0\end{aligned} \tag{2.26}$$

where the first inequality is due to Lemma 2, the second inequality is due to the sub-multiplicativity of the matrix norm, and the convergence is due to $\mathbf{E}_N \xrightarrow{pr} 0$. Therefore, $d_k(\mathbf{B}_N)^2 \xrightarrow{pr} 1$ for $k = 1, 2, .., K$. By the continuous mapping theorem, $d_i(\mathbf{B}_N) \xrightarrow{pr} 1$ for $i = 1, 2, .., K$. Combined with (2.24), we therefore have that $||\Delta_N||_F \xrightarrow{pr} 0$ and $||\hat{\mathbf{A}}_N\mathbf{O}_N - \mathbf{A}^*||_F \xrightarrow{pr} 0$. □

A2.5. Proof of Theorem 2

Proof. For a certain threshold $c \in (0, c_0)$, define $\mathbf{\Lambda}^{*(N)} = \mathbf{\Lambda}^*\tilde{\mathbf{D}}_N^{-1}\mathbf{D}_N^{-1} = \{\lambda_{ij}^{*(N)}\}_{J\times K}$ and $E_N = \{||\hat{\mathbf{\Lambda}}_{N,p} - \mathbf{\Lambda}^{*(N)}||_{max} < \min(c, c_0 - c)\}$. By Theorem 1, under conditions C1-C3, $\hat{\mathbf{\Lambda}}_{N,p}\mathbf{D}_N\tilde{\mathbf{D}}_N \xrightarrow{pr} \mathbf{\Lambda}^*$. Therefore, for any $\epsilon > 0$, there exists a N_0 such that when $N > N_0$, $P(E_N) > 1 - \epsilon$. Denote the entries of $\hat{\mathbf{\Gamma}}_{N,p} = \left(\text{sgn}(\hat{\lambda}_{ij}^{(N,p)}) \times 1_{\{|\hat{\lambda}_{ij}^{(N,p)}|>c\}}\right)_{J\times K}$ on E_N as $\hat{\gamma}_{ij}^{(N,p)}$:

$$\hat{\gamma}_{ij}^{(N,p)} = \begin{cases} 0, & \text{if } \lambda_{ij}^{*(N)} = 0, \text{ since } |\hat{\lambda}_{ij}^{N,p}| - 0 < c \\ \text{sgn}(\hat{\lambda}_{ij}^{(N,p)}) = \text{sgn}(\lambda_{ij}^{*(N)}), & \text{if } \lambda_{ij}^{*(N)} \neq 0, \text{ since } |\hat{\lambda}_{ij}^{N,p}| > |\lambda_{ij}^{*(N)}| - (c_0 - c) \geqslant c \end{cases} \tag{2.27}$$

Therefore, when $N > N_0$, $\hat{\mathbf{\Gamma}}_{N,p} = (\text{sgn}(\lambda_{ij}^{*(N)}))_{J\times K} = \mathbf{\Gamma}^*\tilde{\mathbf{D}}_N^{-1}\mathbf{D}_N^{-1}$ with probability at least $1-\epsilon$. □

A2.6. Proof of Theorem 3

Proof. For a fixed s and k, let $A_N = \{\lambda_{sk}^{*(N)} \in (l_{sk}^{(N)}, u_{sk}^{(N)})\}$ be the event of interest, where $\lambda_{sk}^{*(N)}$ are the entries of $\mathbf{\Lambda}^{*(N)} = \mathbf{\Lambda}^* \tilde{\mathbf{D}}_N^{-1} \mathbf{D}_N^{-1}$. Let $B_N = \{\lambda_{sk}^* \in (l_{sk}^{*(N)}, u_{sk}^{*(N)})\}$ be the event of the confidence interval coverage based on the true sign pattern $\mathbf{\Gamma}^*$. Let $C_N = \{\hat{\mathbf{\Gamma}}_{N,p} \mathbf{D}_N \tilde{\mathbf{D}}_N = \mathbf{\Gamma}^*\}$ be the event that the selected sign pattern is consistent. Since $A_N \cap C_N = B_N \cap C_N$,

$$\mathbb{P}(A_N \cap C_N) = \mathbb{P}(B_N \cap C_N) = \mathbb{P}(B_N) - \mathbb{P}(B_N \cap C_N^C) \xrightarrow{pr} 1 - \alpha,$$

where the limit is due to $P(B_N) \xrightarrow{pr} 1 - \alpha$ by condition C6 and $0 \leqslant \mathbb{P}(B_N \cap C_N^C) \leqslant \mathbb{P}(C_N^C) \xrightarrow{pr} 0$ by condition C5. Therefore, combined with $\mathbb{P}(A_N \cap C_N^C) \xrightarrow{pr} 0$ by condition C5,

$$\mathbb{P}(A_N) = \mathbb{P}(A_N \cap C_N) + \mathbb{P}(A_N \cap C_N^C) \xrightarrow{pr} 1 - \alpha.$$

□

A2.7. Computational Complexity

For the computational complexity, we remind that we have a loading matrix $\mathbf{\Lambda}$ of dimension $J \times K$, a rotation matrix $\mathbf{T}$ of dimension $K \times K$, a weight matrix $\boldsymbol{W} = \{w_{jk}\}_{J \times K}$ of dimension $J \times K$, and the diagonal of the residual covariance matrix, denoted $\boldsymbol{v}$, which is a vector of dimension $K \times 1$. In order to calculate the computational complexity, we count the number of floating point operations, which includes addition, subtraction, multiplication and division. The following results are simplified by ignoring all terms except the highest order term. We use $O(n)$ to denote a computational complexity of order n, meaning there exists a constant $C > 0$, such that the total number of floating point operations can be controlled by Cn. For example, an $m \times n$ matrix $\mathbf{A}$, $n \times q$ matrix $\mathbf{B}$ and $n \times n$ matrix $\mathbf{C}$, the matrix multiplication operation $\mathbf{AB}$ is of computational complexity $O(mnq)$. By Gauss-Jordan elimination we can also conclude that the inversion of $\mathbf{C}$ is of computational complexity $O(n^3)$.

At iteration t of the proposed IRGP algorithm in Algorithm 3 in the main text, the computations and their complexity are as follows,here we define the approximation function of the objective function of $\boldsymbol{W}$ and $\mathbf{\Lambda}$, by $Q_W(\boldsymbol{W}, \mathbf{\Lambda}) = \sum_{j,k} w_{jk} \lambda_{jk}^2$

$$\mathbf{\Lambda}_t = \hat{\mathbf{A}} (\mathbf{T}_t')^{-1}, \qquad O(JK^2 + K^3)$$

$$
\begin{aligned}
w_{jk}^{(t)} =& \frac{1}{((\mathbf{\Lambda}_t)_{jk}^2 + \epsilon^2)^{1-p/2}}, && O(JK) \\
\frac{dQ_W(\mathbf{W}, \mathbf{\Lambda}_t)}{d\mathbf{\Lambda}_t} =& 2\mathbf{W} \odot \mathbf{\Lambda}_t, \text{ where } \odot \text{ means element-wise product}, && O(JK) \\
\nabla G_t(\mathbf{T})) =& -(\mathbf{\Lambda}_t' \frac{dQ_W(\mathbf{W}, \mathbf{\Lambda}_t)}{d\mathbf{\Lambda}_t} \mathbf{T}_t^{-1})', && O(JK^2 + K^3) \\
\mathbf{T}_{t+1} =& \text{Proj}(\mathbf{T}_t - \alpha \nabla G_t(\mathbf{T})), && O(K^2)
\end{aligned}
$$

Therefore, the per-iteration complexity for Algorithm 3 is $O(JK^2 + K^3)$

At iteration t of the proximal gradient descent algorithm in Algorithm 4 in the main text, the computations and their complexity is in the following chart

$$
\begin{aligned}
\mathbf{\Sigma}(\theta) =& \mathbf{\Lambda}_t \mathbf{T}_t' \mathbf{T}_t \mathbf{\Lambda}' + \text{diag}(\exp(\boldsymbol{v}_t)), && O(J^2K + JK^2 + K^3) \\
\mathbf{Q} =& \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1} \mathbf{S} \mathbf{\Sigma}^{-1}, && O(J^3) \\
\nabla L_{t,\mathbf{\Lambda}} =& 2 \cdot \mathbf{Q} \mathbf{\Lambda}_t \mathbf{T}_t' \mathbf{T}_t, && O(J^2K + JK^2 + K^3) \\
& \text{Prox}_{\alpha,\gamma}(\mathbf{\Lambda}_t - \alpha \nabla L_{t,\mathbf{\Lambda}}), && O(JK) \\
\nabla L_{t,\mathbf{T}_{ij}} =& (2 \cdot \mathbf{T}_t \mathbf{\Lambda}_t' \mathbf{Q} \mathbf{\Lambda}_t)_{ij} \mathbf{1}_{\{i \leqslant j\}}, && O(J^2K + JK^2) \\
& \text{Proj}(\mathbf{T}_t - \alpha \nabla L_{t,\mathbf{T}}), && O(K^2) \\
\nabla L_{t,v_{t,i}} =& Q_{ii} \cdot \exp(v_{t,i}), i = 1, ..., J, && O(J)
\end{aligned}
$$

Therefore, the per-iteration complexity for Algorithm 4 is $O(J^3 + J^2K + JK^2 + K^3)$.

A2.8. Comparison with Other Rotation Criteria

In the following, we demonstrate scenarios where some of the most popular traditional rotation criteria fail to recover the true sparse structure, unlike the proposed criterion. Consider first the Geomin criterion (Yates, 1987), defined as

$$
Q_{geo} = \sum_{j=1}^{J} \Big(\prod_{k=1}^{K} \lambda_{jk} \Big)^{\frac{2}{K}}. \tag{2.28}
$$

The Geomin criterion thus measures the row-wise complexity and equals zero if at least one entry λ_{jk} in the loading matrix $\mathbf{\Lambda}$, for all $j = 1, \ldots, J$, equals zero. To refrain from indeterminacy of the minimizer, the criterion is commonly modified by adding a small positive constant ϵ, such that

$$
Q_{geo}^{\epsilon} = \sum_{j=1}^{J} \Big(\prod_{k=1}^{K} \lambda_{jk}^2 + \epsilon \Big)^{\frac{1}{K}}. \tag{2.29}
$$

In the **GPArotation** R package (Bernaards and Jennrich, 2005), (2.29) is the rotation criterion being minimized when the Geomin function is called, with default value $\epsilon = 0.01$.

Consider an initial loading matrix $\boldsymbol{A}$ of dimension 21×3, given in the first three columns of in Table 2.8. Notice that the first 15 rows of $\boldsymbol{A}$ contain only one non-zero entry per row, and that the remaining rows contain at least two non-zero entries. Also notice that several of the non-zero entries in the dense part of $\boldsymbol{A}$ are small in magnitude. A majority of the matrix is thus sparse, but with a dense component. One possible solution for the original Geomin criterion in 2.28 is given by $\boldsymbol{A}'$, since $Q_{geo}(\boldsymbol{A}') = 0$. This solution is displayed in columns four to six in Table 2.8. We verify that $\boldsymbol{A}'$ contains 26 zero entries, whereas $\boldsymbol{A}$ contain 32 zero entries. The dense part of $\boldsymbol{A}$ thus dominates the sparse structure in $\boldsymbol{A}$, making the Geomin criterion unable to recover the true sparse structure. In columns seven to nine in Table 2.8, the solution to the adjusted Geomin criterion in (2.29) is presented, with $\epsilon = 0.01$. As displayed, the adjusted Geomin is not either able to recover the true structure of $\boldsymbol{A}$.

	$\mathbf{A}$			$\mathbf{A}'$			$\arg\min Q_{geo}^{\epsilon=0.01}(\mathbf{A})$		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
1	0.633	0.000	0.000	1.096	0.000	0.895	0.633	0.005	0.014
2	0.000	0.686	0.000	0.000	0.686	0.000	-0.022	0.741	0.317
3	0.000	0.000	0.786	0.000	0.786	1.112	0.014	-0.043	0.769
4	0.954	0.000	0.000	1.653	0.000	1.349	0.954	0.007	0.021
5	0.000	0.601	0.000	0.000	0.601	0.000	-0.019	0.649	0.277
6	0.000	0.000	0.949	0.000	0.949	1.342	0.017	-0.052	0.929
7	0.972	0.000	0.000	1.684	0.000	1.375	0.973	0.008	0.022
8	0.000	0.830	0.000	0.000	0.830	0.000	-0.027	0.897	0.383
9	0.000	0.000	0.815	0.000	0.815	1.152	0.015	-0.045	0.797
10	0.531	0.000	0.000	0.920	0.000	0.751	0.531	0.004	0.012
11	0.000	0.603	0.000	0.000	0.603	0.000	-0.019	0.652	0.278
12	0.000	0.000	0.588	0.000	0.588	0.832	0.011	-0.032	0.575
13	0.844	0.000	0.000	1.461	0.000	1.193	0.844	0.007	0.019
14	0.000	0.692	0.000	0.000	0.692	0.000	-0.022	0.748	0.320
15	0.000	0.000	0.885	0.000	0.885	1.251	0.016	-0.049	0.866
16	0.000	0.117	0.489	0.000	0.606	0.691	0.005	0.100	0.532
17	0.496	-0.165	0.165	0.859	0.000	0.935	0.504	-0.184	0.096
18	0.575	1.138	-0.575	0.996	0.563	0.000	0.528	1.266	-0.024
19	0.000	0.110	0.524	0.000	0.634	0.741	0.006	0.090	0.563
20	0.513	-0.052	0.052	0.889	0.000	0.800	0.516	-0.056	0.039
21	0.559	1.065	-0.559	0.967	0.507	0.000	0.515	1.186	-0.042

Table 2.8: The initial loading matrix $\boldsymbol{A}$, the transpose of $\boldsymbol{A}$ which is the Geomin solution, and the solution to the adjusted Geomin criterion in (2.29) for a counterexample when the Geomin criterion fails to recover the true sparse structure.

We apply the proposed family of rotation criteria, with both $p = 0.5$ and $p = 1$, to the matrix $\boldsymbol{A}$. We verify that the solution is given by $\mathbf{A}$ using grid search over the whole oblique rotation matrix space $\mathcal{M}$. When $\boldsymbol{A}$ is used as a starting point for the proposed IRGP algorithm, all of the minimizers of $L^{0.5}$ and L^1 differ at most by a sign flip or column permutation of $\boldsymbol{T} = \mathbf{I}_3$, where $\mathbf{I}_3$ is an identity matrix of dimension 3×3. The true loading matrix $\boldsymbol{A}$ is thus recovered, up to a sign flip and column permutation.

We compare the results with the Quartimin criterion in the Oblimin family and the Promax algorithm as well. The former is defined as

$$Q_{obl} = \sum_{k=1}^{K} \sum_{k' \neq k}^{K} \sum_{j=1}^{J} \lambda_{jk}^2 \lambda_{jk'}^2. \tag{2.30}$$

The oblimin criterion (Harman and Harman, 1976) could thus be understood as a weighted sum related to the complexity of each row of the factor loading matrix. The Promax algorithm (Hendrickson and White, 1964) takes the rotation matrix from Varimax rotation and raises it to powers of 4 in the **stats** R package (Finch, 2006) This has the effect of pushing small values down to zero while larger values are not reduced as much.

In Table 2.9, the results of the proposed method for both $p = 0.5$ and $p = 1$, the Geomin and Oblimin criteria, and the Promax algorithm are presented in terms of their MSE. The starting point for all of the rotation criteria is $\mathbf{A}$. The first column displays the entrywise MSE, calculated as

$$\sum_{ij} \frac{(\mathbf{A}_{ij} - \text{rot}(\mathbf{A})_{ij})^2}{JK},$$

where $\text{rot}(\mathbf{A})$ represents the rotated solution for each respective method. The second column presents the value of the objective function at $\mathbf{A}$, the third column shows the value of rotation criteria at the rotated loading matrix, and the last column the contains the number of zeros produced by the rotated matrix with a cut-off at 0.01. Since Promax is an algorithm that does not involve an objective function, we do not report the objective value for it.

As demonstrated in Table 2.9, the MSE equals zero for both choices of p for the proposed criterion. The Promax algorithm shows the second to best performance and the Oblimin and Geomin with an $\epsilon = 0.01$ perform similarly. None of the methods, except for the Geomin with $\epsilon = 0$ comes close to the proposed method in terms of identifying the zero elements in the loading matrix, with the proposed method being able to identify all of them for both choices of p.

Lastly, we present the results of the average MSE for each respective rotation method over 500 simulations. The true loading matrix is still $\boldsymbol{A}$ given in Table 2.8, and with generated latent factors

	MSE	Obj	Obj. rot	Number of zeros
L^1	0.000	18.523	18.523	32
$L^{0.5}$	0.000	22.898	22.898	32
Oblimin	0.021	0.896	0.265	2
Geomin($\epsilon = 0.01$)	0.018	1.789	1.354	7
Geomin($\epsilon = 0$)	0.251	1.070	0.000	26
Promax	0.013	-	-	4

Table 2.9: Comparison of the component-wise loss function for $p = 1$ and $p = 0.5$, the Oblimin, the Geomin for $\epsilon = 0.01$ and $\epsilon = 0$, and the Promax rotation methods.

that are orthogonal to each other. The unique variances of the items corresponds to Item 1-21 in Table 2.12 under the column of Item Unique Variance. Three settings are considered, including $N = 400$, 800, and 1600. For each setting, 500 independent simulations are conducted. Table 2.10 presents the resulting MSEs, averaged over the number of simulations, and demonstrates the superior performance of the proposed method for both choices of p over the traditional methods.

	$N = 400$	$N = 800$	$N = 1600$
L^1	0.007	0.003	0.002
$L^{0.5}$	0.007	0.003	0.002
Oblimin	0.027	0.024	0.022
Geomin($\epsilon = 0.01$)	0.021	0.019	0.018
Promax	0.018	0.015	0.014

Table 2.10: The average MSE for the component-wise loss function for $p = 1$ and $p = 0.5$, the Oblimin, the Geomin for $\epsilon = 0.01$, and the Promax rotation methods, for $N = \{400, 800, 1600\}$.

A2.9. True Parameters for Simulation Study I

In this part, the parameters used in Study I are displayed in Table 2.11 to Table 2.13, including the true loading matrices $\mathbf{\Lambda}^*$, item unique variances $\mathbf{\Omega}^*$ and the lower diagonal part of the true covariance matrices of latent variables $\mathbf{\Phi}^*$ (which are symmetric).

A2.10. True Parameters for Study II

The loading matrix $\mathbf{\Lambda}^*$ is shown in Table 2.14. The covariance matrix for latent variable is the same as the 15×3 setting in Study I, listed in the last three columns of Table 2.13.

	Loading Matrix Item 1-15			Item Unique Variances Item 1-15
	F1	F2	F3	
1	0.71	0	0	1.27
2	0	0.75	0	1.38
3	0	0	0.83	1.57
4	0.96	0	0	1.92
5	0	0.68	0	1.20
6	0	0	0.96	1.90
7	0.98	0	0	1.95
8	0	0.86	0	1.67
9	0	0	0.85	1.63
10	0.62	0.35	0	1.06
11	0	0.68	0.42	1.21
12	0.5	0	0.67	1.17
13	0.87	0	0.31	1.68
14	0.43	0.75	0	1.39
15	0	0.48	0.91	1.77

Table 2.11: 15×3 factor loading patterns $\mathbf{\Lambda}^*$ and item unique variances $\mathbf{\Omega}^*$ in Simulation Study I

	Loading Matrix											Item Unique Variances	
	Item 1-15						Item 16-30					Item 1-15	Item 16-30
	F1	F2	F3	F4	F5		F1	F2	F3	F4	F5		
1	0.71	0	0	0	0	16	0.8	0.34	0	0	0	1.27	1.49
2	0	0.75	0	0	0	17	0	0.89	0.38	0	0	1.38	1.72
3	0	0	0.83	0	0	18	0	0	1	0.35	0	1.57	1.99
4	0	0	0	0.96	0	19	0	0	0	0.75	0.26	1.92	1.38
5	0	0	0	0	0.68	20	0.45	0	0	0	0.91	1.20	1.79
6	0.96	0	0	0	0	21	0.97	0	0.4	0	0	1.90	1.93
7	0	0.98	0	0	0	22	0	0.68	0	0.44	0	1.95	1.21
8	0	0	0.86	0	0	23	0	0	0.86	0	0.23	1.67	1.65
9	0	0	0	0.85	0	24	0.42	0	0	0.65	0	1.63	1.13
10	0	0	0	0	0.62	25	0	0.32	0	0	0.71	1.06	1.27
11	0.68	0	0	0	0	26	0.75	0.45	0.39	0	0	1.21	1.39
12	0	0.67	0	0	0	27	0	0.61	0.43	0.37	0	1.17	1.01
13	0	0	0.87	0	0	28	0	0	0.75	0.36	0.44	1.68	1.38
14	0	0	0	0.75	0	29	0.34	0	0	0.95	0.21	1.39	1.88
15	0	0	0	0	0.91	30	0.42	0.41	0	0	0.74	1.77	1.34

Table 2.12: 30×5 factor loading patterns $\mathbf{\Lambda}^*$ and item unique variances $\mathbf{\Omega}^*$ in Simulation Study I

A2.11. Additional Results for the Big-Five Personality Test Application

Tables 2.15 through 2.17 show the estimated loading parameters and the corresponding 95% confidence intervals obtained from the L^1 rotation.

	30 × 5 setting					15 × 3 setting		
	F1	F2	F3	F4	F5	F1	F2	F3
F1	1					1		
F2	0.085	1				0.021	1	
F3	0.429	0.042	1			0.502	0.274	1
F4	0.148	0.149	0.496	1				
F5	0.249	0.309	0.121	0.19	1			

Table 2.13: The true covariance matrices for latent variables in Simulation Study I.

	Loading Matrix							Item Unique Variances	
	Item 1-9				Item 10-18			Item 1-9	Item 10-18
	F1	F2	F3		F1	F2	F3		
1	0.531	0.760	0	10	0.124	0.765	0	1.27	1.06
2	0.744	0	0.216	11	0.412	0	0.047	1.38	1.21
3	0	1.870	1.447	12	0	0.681	0.954	1.57	1.17
4	1.816	0.424	0	13	1.374	0.964	0	1.92	1.68
5	0.403	0	1.642	14	0.768	0	1.385	1.20	1.39
6	0	0.251	1.294	15	0	0.987	0.955	1.90	1.77
7	1.889	0.534	0	16	0.995	0.372	0	1.95	1.49
8	1.322	0	1.106	17	1.435	0	0.876	1.67	1.72
9	0	0.027	1.059	18	0	1.337	0.490	1.63	1.99

Table 2.14: 18 × 3 true loading matrix and item unique variances in Simulation Study II

A2.12. Sensitivity Analysis of p

In this study, we conduct a sensitivity analysis to provide insights into the choice of p in L^p rotation for loading matrices with varying degrees of sparsity. Specifically, we simulate a toy example using a 60×3 loading matrix, where the proportion of simple items varies across five settings, ranging from 95% to 5%. Each simple item has a single dominant loading generated uniformly from $U[1, 2]$, while non-simple items are assigned an additional cross-loading generated from $U[0.2, 0.5]$. Main loadings are distributed evenly across all factors to ensure balanced representation.

To evaluate performance, we use the `GPArotation` package with 50 random starts for each value of p, ranging from 1.0 to 0.4. The algorithm is initialized with the true sparse loading matrix, ensuring that no estimation error is introduced at the initialization stage. For each setting, we record both the mean squared error (MSE) between the estimated and true loading matrices and the average computational time per random start. The lower bound of p is set to 0.4, as smaller values result in a substantial increase in computational time, reflecting the growing difficulty of the optimization problem. This effect is also illustrated in Figure 2.2, where the contour of the

$L^{0.5}$ criterion exhibits sharper corners and a more pronounced non-convexity.

Results. Figure 2.5(a) shows that, in general, smaller values of p lead to lower MSE across all sparsity levels. When the proportion of simple items exceeds 50%, MSE decreases steadily as p becomes smaller, and $p = 1$ is sufficient to recover the sparse loading structure. In the setting with 25% simple items, a sharp decline in MSE is observed when p decreases from 0.6 to 0.55, suggesting that this threshold marks the point at which L^p rotation begins to successfully identify the true sparse structure. In contrast, for highly dense loading matrices (e.g., with only 5% simple items), none of the tested p values yield low MSE, indicating the challenge of sparse recovery in such settings.

Figure 2.5(b) demonstrates that the average computational time per random start increases as p decreases. This trend highlights the increasing complexity of the non-convex optimization problem associated with smaller values of p. Overall, these results underscore the statistical–computational trade-off in the choice of p: smaller values tend to improve estimation accuracy but incur greater computational cost.

	E	ES	A	C	I
E1(+)	0.878*	-0.065	-0.069*	-0.005	0.082*
	(0.793, 0.983)	(-0.158, 0.011)	(-0.180,-0.014)	(-0.134, 0.038)	(0.005, 0.171)
E2(-)	-0.852*	0.127*	0.004	0.048	-0.014
	(-0.975,-0.770)	(0.047, 0.232)	(-0.056, 0.126)	(-0.028, 0.163)	(-0.103, 0.082)
E3(+)	0.785*	0.278*	0.202*	0.099	-0.095*
	(0.692, 0.868)	(0.197, 0.356)	(0.118, 0.274)	(-0.013, 0.148)	(-0.173,-0.018)
E4(-)	-0.922*	-0.063	-0.020	0.022	0.091*
	(-1.026,-0.847)	(-0.144, 0.011)	(-0.072, 0.079)	(-0.030, 0.128)	(0.034, 0.187)
E5(+)	0.889*	-0.024	0.153*	0.083	0.080*
	(0.810, 0.988)	(-0.117, 0.038)	(0.061, 0.212)	(-0.026, 0.131)	(0.022, 0.173)
E6(-)	-0.736*	-0.003	-0.087	-0.043	-0.137*
	(-0.854,-0.661)	(-0.065, 0.113)	(-0.160, 0.012)	(-0.132, 0.051)	(-0.225,-0.050)
E7(+)	1.125*	-0.077*	0.081*	0.081	-0.023
	(1.025, 1.229)	(-0.189,-0.014)	(0.003, 0.174)	(-0.032, 0.143)	(-0.119, 0.052)
E8(-)	-0.710*	-0.095*	0.029	0.138*	-0.064
	(-0.824,-0.628)	(-0.184,-0.002)	(-0.035, 0.143)	(0.059, 0.246)	(-0.156, 0.025)
E9(+)	0.827*	0.057	-0.002	-0.037	0.243*
	(0.737, 0.945)	(-0.045, 0.146)	(-0.108, 0.078)	(-0.174, 0.022)	(0.146, 0.334)
E10(-)	-0.826*	-0.119*	-0.047	-0.099	0.006
	(-0.930,-0.739)	(-0.192,-0.020)	(-0.121, 0.049)	(-0.169, 0.009)	(-0.078, 0.093)
ES1(-)	-0.085*	-0.988*	0.008	0.110*	-0.104*
	(-0.187,-0.003)	(-1.100,-0.895)	(-0.117, 0.079)	(0.067, 0.260)	(-0.195,-0.019)
ES2(+)	0.113*	0.684*	-0.000	-0.106*	0.085*
	(0.001, 0.178)	(0.614, 0.804)	(-0.065, 0.112)	(-0.259,-0.074)	(0.020, 0.194)
ES3(-)	-0.164*	-0.796*	0.233*	0.146*	0.044
	(-0.232,-0.056)	(-0.919,-0.726)	(0.131, 0.308)	(0.109, 0.296)	(-0.039, 0.135)
ES4(+)	0.206*	0.571*	0.000	0.046	0.010
	(0.089, 0.286)	(0.486, 0.688)	(-0.075, 0.118)	(-0.112, 0.090)	(-0.077, 0.116)
ES5(-)	0.056	-0.475*	-0.040	-0.096	-0.228*
	(-0.046, 0.167)	(-0.577,-0.361)	(-0.159, 0.049)	(-0.187, 0.032)	(-0.349,-0.137)
ES6(-)	-0.087	-0.817*	0.259*	-0.001	-0.133*
	(-0.172, 0.007)	(-0.930,-0.736)	(0.154, 0.334)	(-0.030, 0.159)	(-0.231,-0.056)
ES7(-)	0.052	-0.973*	-0.110*	-0.020	0.004
	(-0.008, 0.157)	(-1.077,-0.887)	(-0.213,-0.041)	(-0.059, 0.112)	(-0.090, 0.072)
ES8(-)	0.036	-1.142*	-0.133*	-0.047	0.001
	(-0.021, 0.154)	(-1.259,-1.055)	(-0.247,-0.066)	(-0.076, 0.104)	(-0.104, 0.065)
ES9(-)	0.001	-0.879*	-0.292*	0.195*	-0.016
	(-0.086, 0.092)	(-0.990,-0.795)	(-0.388,-0.207)	(0.145, 0.329)	(-0.110, 0.066)
ES10(-)	-0.332*	-0.846*	0.071	-0.081	0.100*
	(-0.399,-0.220)	(-0.957,-0.765)	(-0.019, 0.163)	(-0.116, 0.068)	(0.011, 0.184)

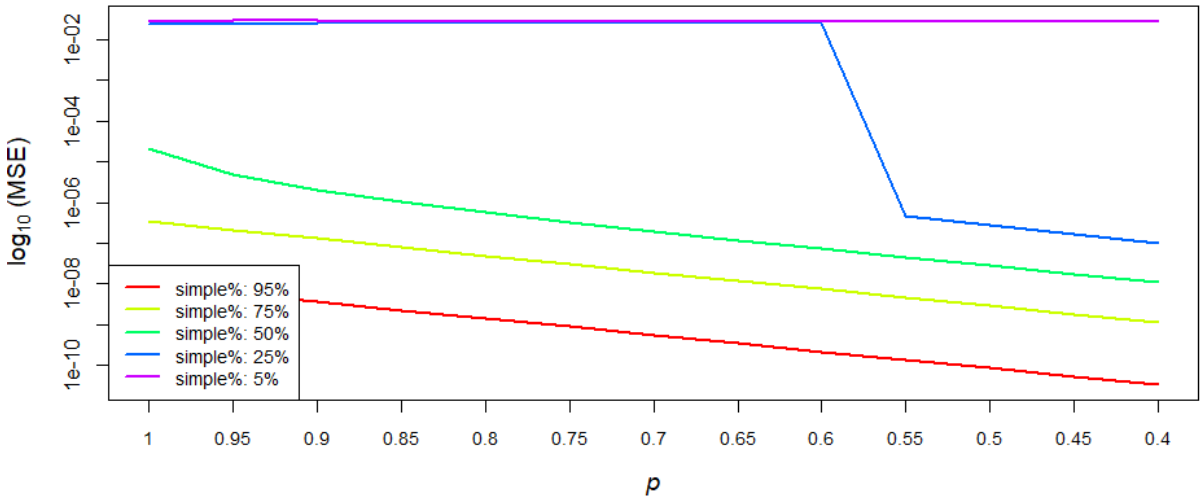
Table 2.15: Part I: Point estimates and confidence intervals constructed by L^1, big-five personality test. The loadings that are significantly different from zero according to the 95% confidence intervals are indicated by asterisks.

	E	ES	A	C	I
A1(-)	0.002	-0.128*	-0.779*	0.035	0.046
	(-0.118, 0.085)	(-0.209,-0.011)	(-0.872,-0.666)	(-0.081, 0.123)	(-0.059, 0.136)
A2(+)	0.433*	-0.004	0.557*	-0.054	0.042
	(0.362, 0.528)	(-0.092, 0.065)	(0.465, 0.627)	(-0.157, 0.003)	(-0.025, 0.129)
A3(-)	0.192*	-0.577*	-0.566*	-0.067	0.140*
	(0.090, 0.299)	(-0.679,-0.465)	(-0.682,-0.471)	(-0.166, 0.048)	(0.029, 0.232)
A4(+)	0.013	-0.001	0.980*	-0.018	-0.002
	(-0.008, 0.168)	(-0.097, 0.047)	(0.892, 1.045)	(-0.093, 0.039)	(-0.062, 0.070)
A5(-)	-0.165*	-0.038	-0.815*	0.006	0.089*
	(-0.258,-0.097)	(-0.108, 0.049)	(-0.892,-0.723)	(-0.074, 0.084)	(0.012, 0.164)
A6(+)	-0.054	-0.186*	0.718*	0.011	0.013
	(-0.136, 0.042)	(-0.278,-0.104)	(0.628, 0.810)	(-0.082, 0.096)	(-0.077, 0.095)
A7(-)	-0.366*	-0.093*	-0.732*	0.070*	0.031
	(-0.458,-0.300)	(-0.169,-0.020)	(-0.798,-0.637)	(0.006, 0.157)	(-0.047, 0.099)
A8(+)	0.110*	-0.042	0.692*	0.076*	0.027
	(0.044, 0.190)	(-0.130, 0.013)	(0.618, 0.771)	(0.001, 0.147)	(-0.034, 0.107)
A9(+)	0.113*	-0.115*	0.752*	0.062	0.113*
	(0.047, 0.207)	(-0.212,-0.056)	(0.669, 0.837)	(-0.010, 0.150)	(0.041, 0.195)
A10(+)	0.432*	0.069	0.320*	0.112	0.053
	(0.348, 0.513)	(-0.007, 0.151)	(0.245, 0.402)	(-0.004, 0.158)	(-0.019, 0.138)
C1(+)	0.096	0.089	-0.039	0.682*	0.133*
	(-0.004, 0.178)	(-0.005, 0.181)	(-0.098, 0.089)	(0.563, 0.754)	(0.064, 0.246)
C2(-)	0.056	-0.180*	0.110	-0.658*	0.145*
	(-0.000, 0.206)	(-0.262,-0.050)	(-0.022, 0.181)	(-0.798,-0.578)	(0.009, 0.212)
C3(+)	-0.007	-0.007	0.112*	0.399*	0.284*
	(-0.091, 0.071)	(-0.111, 0.052)	(0.050, 0.210)	(0.302, 0.473)	(0.218, 0.382)
C4(-)	-0.107*	-0.604*	0.051	-0.478*	-0.041*
	(-0.169,-0.005)	(-0.670,-0.496)	(-0.048, 0.123)	(-0.544,-0.371)	(-0.174,-0.008)
C5(+)	0.093	0.030	-0.002	0.779*	-0.051
	(-0.020, 0.169)	(-0.091, 0.113)	(-0.048, 0.154)	(0.679, 0.881)	(-0.122, 0.072)
C6(-)	0.003	-0.172*	0.048	-0.704*	0.088
	(-0.074, 0.139)	(-0.255,-0.035)	(-0.081, 0.136)	(-0.837,-0.608)	(-0.028, 0.187)
C7(+)	-0.121*	-0.150*	0.109*	0.535*	0.040
	(-0.219,-0.041)	(-0.267,-0.085)	(0.038, 0.216)	(0.464, 0.653)	(-0.022, 0.158)
C8(-)	-0.000	-0.268*	-0.240*	-0.518*	-0.000
	(-0.073, 0.109)	(-0.340,-0.155)	(-0.355,-0.173)	(-0.604,-0.413)	(-0.123, 0.058)
C9(+)	0.053	-0.029	0.121*	0.725*	-0.076
	(-0.062, 0.125)	(-0.177, 0.024)	(0.055, 0.243)	(0.639, 0.841)	(-0.149, 0.040)
C10(+)	-0.022	-0.025	0.126*	0.523*	0.238*
	(-0.116, 0.050)	(-0.146, 0.025)	(0.068, 0.234)	(0.431, 0.609)	(0.172, 0.340)

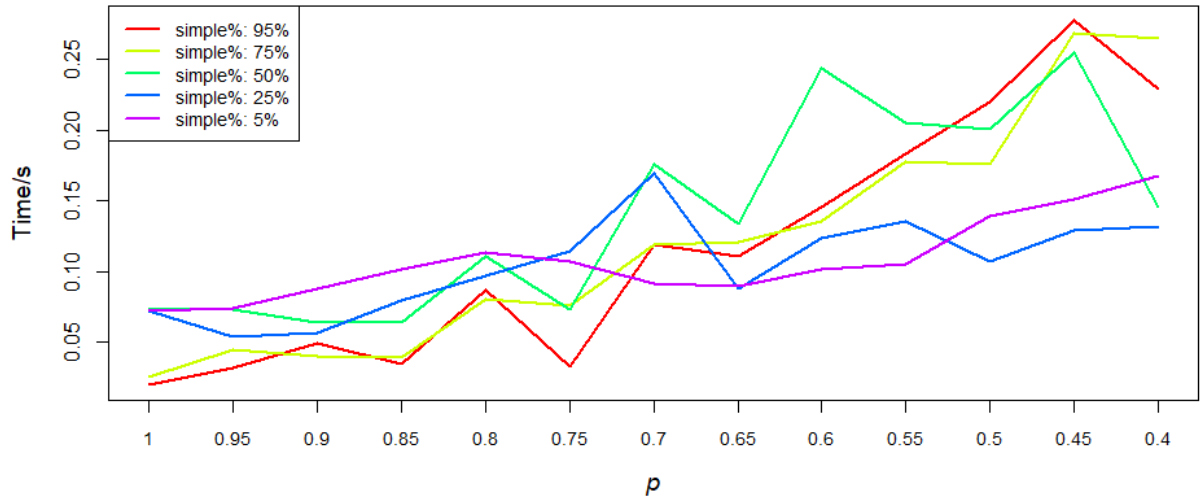
Table 2.16: Part II: Point estimates and confidence intervals constructed by L^1, big-five personality test.

	E	ES	A	C	I
I1(+)	0.003	0.002	-0.047	-0.007	0.630*
	(-0.037, 0.131)	(-0.113, 0.060)	(-0.147, 0.015)	(-0.106, 0.062)	(0.539, 0.716)
I2(-)	0.083	-0.226*	-0.086*	0.020	-0.588*
	(-0.020, 0.157)	(-0.297,-0.121)	(-0.185,-0.017)	(-0.047, 0.128)	(-0.683,-0.505)
I3(+)	0.004	-0.152*	0.023	-0.001	0.595*
	(-0.038, 0.131)	(-0.254,-0.092)	(-0.058, 0.099)	(-0.085, 0.080)	(0.501, 0.668)
I4(-)	0.105	-0.209*	-0.153*	0.046	-0.578*
	(-0.020, 0.153)	(-0.273,-0.098)	(-0.225,-0.059)	(-0.028, 0.146)	(-0.660,-0.484)
I5(+)	0.165*	0.065	-0.060	0.164*	0.586*
	(0.109, 0.252)	(-0.003, 0.138)	(-0.127, 0.006)	(0.054, 0.194)	(0.509, 0.657)
I6(-)	-0.149*	-0.004	-0.046	0.038	-0.515*
	(-0.264,-0.090)	(-0.065, 0.108)	(-0.122, 0.043)	(-0.034, 0.140)	(-0.608,-0.432)
I7(+)	0.013	0.168*	-0.036	0.087	0.455*
	(-0.044, 0.099)	(0.088, 0.229)	(-0.099, 0.037)	(-0.010, 0.135)	(0.384, 0.528)
I8(+)	-0.095	-0.164*	-0.108*	-0.001	0.664*
	(-0.177, 0.011)	(-0.276,-0.091)	(-0.194,-0.017)	(-0.097, 0.091)	(0.572, 0.768)
I9(+)	-0.081	-0.220*	0.239*	0.111*	0.262*
	(-0.149, 0.014)	(-0.321,-0.159)	(0.159, 0.318)	(0.042, 0.208)	(0.182, 0.343)
I10(+)	0.158*	-0.005	-0.002	0.086*	0.692*
	(0.110, 0.259)	(-0.114, 0.038)	(-0.068, 0.070)	(0.006, 0.158)	(0.613, 0.769)

Table 2.17: Part III: Point estimates and confidence intervals constructed by L^1, big-five personality test.



(a) Mean squared error (MSE) of L^p rotation under varying values of p



(b) Computational time per random start for L^p rotation

Figure 2.5: Statistical and computational trade-offs for L^p rotation across different values of p. The legend *Simple%* indicates the proportion of simple items in the loading matrix.

Chapter 3

Identifiability Conditions for Sparse Loading Matrices with L^1 Rotation

3.1. Introduction

Thurstone (1947) proposed leveraging sparsity to resolve the rotational indeterminacy inherent in factor loading matrices and factor covariances. In his words, "In numerical terms, this is a demand for the rotation which provides the smallest number of non-vanishing entries in each row of the oblique factor matrix." The interpretability enabled by sparsity is crucial for understanding how latent constructs are manifested in individuals' behaviors or perceptions. When an item loads exclusively on a single latent factor, it provides a clear and intuitive representation of that construct, thereby enhancing the interpretability of the model, which is called a simple item. For example, an item such as "I sympathize with others' feelings" offers a straightforward representation of the latent trait of agreeableness if it is associated with that factor alone. Conversely, non-simple items—those that load onto multiple latent factors—can obscure the interpretability and complicate the theoretical understanding of the constructs

In Chapter 2, we introduced the L^p rotation criterion, which empirically outperforms existing rotation methods in recovering sparse loading structures—structures that commonly used criteria often fail to identify. However, a closer examination reveals that Theorem 1 only establishes the consistency of L^p rotation under Condition C3 (see 2.3.1), which requires the true loading matrix to be uniquely identifiable by the L^p criterion. Our current understanding of when this condition holds is limited, except in the idealized setting where all items are simple. The goal of this chapter is to deepen our understanding of Condition C3 by deriving sufficient conditions under which it

holds.

Although many theoretical studies attempt to justify traditional rotation methods, none provide identifiability guarantees for loading matrices with non-simple items. For instance, Theorem 4.1 in Rohe and Zeng (2022) shows that, under the assumption that the rows of the loading matrix are i.i.d. and follow a leptokurtic distribution, the varimax rotation can recover the underlying structure. However, this assumption has limited practical relevance for interpretability in applied settings

Several rotation criteria have been proposed in the literature (Browne, 2001); however, most lack rigorous theoretical guarantees regarding the types of loading matrices that can be recovered, except under the highly idealized assumption of a perfect simple structure. For example, rotation criteria in the Crawford-Ferguson family (Crawford and Ferguson, 1970; Crawford, 1975), such as Quartimin(Carroll, 1953), Varimax(Kaiser, 1958, 1959), Equamax, Parsimax, and Factor Parsimony, are based on a weighted sum of two components: a measure of row complexity (to simplify item-level interpretation) and a measure of column complexity (to discourage a general factor). The only available theoretical result for these criteria shows that their loss functions are minimized when each row or column contains only a single non-zero entry. However, for more complex loading matrices, there are no theoretical results offering insight into the behavior of the rotated solutions.

Yates' Geomin rotation(Yates, 1987), for instance, reaches its minimum when each row contains at least one zero. Yet, when the number of latent factors exceeds three, this sparsity condition becomes insufficient. A truly sparse loading matrix with a high proportion of simple items—alongside a few complex items that load on all factors—may be penalized more than a matrix in which each item loads on multiple factors but happens to have at least one zero. Similarly, entropy-based criteria such as Infomax (McKeon, 1968) achieve their minimum when the loading matrix exhibits a perfect simple structure and when the column-wise sum of squared loadings is uniform—a stricter condition than simple structure alone. In Jennrich (2006), it is shown that the componentwise loss function can also recover the loading matrix when it adheres to a perfect simple structure.

However, the assumption of a perfect simple structure is overly restrictive in exploratory factor analysis (EFA), as we typically lack prior knowledge of the loading matrix and cannot ensure that each item loads exclusively on a single latent factor. Thus, allowing for the existence of non-simple items is critical for the effectiveness and applicability of rotation-based identification methods.

In Chapter 3, we address this limitation by extending Proposition 2 from Chapter 2 to accommodate loading matrices that contain both simple and non-simple items. Specifically, we establish new theoretical conditions for L^1 rotation under which the identifiability of loading matrices can be

guaranteed, even in the presence of cross-loadings. This advancement broadens the applicability of sparse factor analysis and offers a more flexible and realistic framework for modeling and interpreting complex data structures. Our contributions not only enhance the theoretical understanding of factor rotation but also provide practical implications for algorithm design and methodological development in EFA.

3.2. Methodology

This chapter builds on the L^p rotation framework introduced in Chapter 2, using the same factor analysis model. Under the sparse loading matrix assumption, consistent estimation is challenging due to the non-identifiability arising from rotational indeterminacy. Let $\mathbf{\Lambda}^*$ denote the true loading matrix and $\mathbf{\Phi}^*$ the true factor variance-covariance matrix. To uniquely recover the sparse loading matrix, we aim to define a rotation criterion that selects the true structure from the equivalence class of an initial loading matrix $\mathbf{\Lambda}^0$.

We assume that the initial estimate has orthogonal latent factors with unit variance, i.e., $\mathbf{\Phi}^0 = \mathbf{I}$, such that

$$\mathbf{\Lambda}^0 \mathbf{\Lambda}^{0\prime} = \mathbf{\Lambda}^* \mathbf{\Phi}^* \mathbf{\Lambda}^{*\prime},$$

and that the rotated solution allows for correlated factors with unit variance, i.e., $(\mathbf{T}'\mathbf{T})_{kk} = 1$ for all $k = 1, \ldots, K$.

The non-smooth L^p rotation criterion (Liu et al., 2023),

$$Q_p(\mathbf{\Lambda}) = \sum_j \sum_k |\lambda_{jk}|^p, \tag{3.1}$$

is effective in promoting sparsity in the loading matrix when minimized over the class of rotated solutions. Specifically, the loading matrix is parameterized as

$$\mathbf{\Lambda} = \mathbf{\Lambda}^0 \mathbf{T}^{\prime -1},$$

where the rotation matrix $\mathbf{T}$ belongs to the set

$$\mathcal{M} = \left\{ \mathbf{T} \in \mathbb{R}^{K \times K} : \forall k, (\mathbf{T}'\mathbf{T})_{kk} = 1, \operatorname{rank}(\mathbf{T}) = K \right\}.$$

We investigate identifiability of the L^1-based criterion under relaxed conditions, aiming to find

sufficient assumptions under which the maximizer

$$\hat{\mathbf{\Lambda}} = \text{argmax}_{\mathbf{\Lambda}=\mathbf{\Lambda}^0\mathbf{T}'^{-1},\mathbf{T}\in\mathcal{M}} Q_1(\mathbf{\Lambda})$$

coincides with the ground truth $\mathbf{\Lambda}^*$. It is known that this holds when $\mathbf{\Lambda}^*$ has a "perfect simple structure", i.e., each item loads on only one factor. Our goal is to relax this assumption.

We define a "simple item" as one that loads on exactly one latent factor. Formally, for each simple item j, there exists a unique $k_j \in \{1, \ldots, K\}$ such that $\lambda^*_{jk_j} \neq 0$. Let $Sp \subseteq \{1, \ldots, J\}$ denote the set of simple items, and define $Sp_k = \{j \in Sp : \lambda^*_{jk} \neq 0\}$ for factor k. Let Sp^c be the set of "non-simple items", i.e., those loading on multiple factors. We assume that the smallest number of simple items across all factors is relatively large compared to the number of non-simple items $|Sp^c|$. Specifically, we require the ratio $\frac{\min_k |Sp_k|}{|Sp^c|}$ to be sufficiently large.

To account for label switching and sign indeterminacy, we define

$$\mathcal{T}^* := \left\{ \mathbf{T}^*\mathbf{D}\tilde{\mathbf{D}} : \mathbf{D} \in \mathcal{D}_1,\, \tilde{\mathbf{D}} \in \mathcal{D}_2 \right\},$$

where $\mathcal{D}_1$ and $\mathcal{D}_2$ denote the sets of permutation and sign-flip matrices, respectively. Thus, $\mathcal{T}^*$ has cardinality $2^K K!$, covering all permutations and sign changes of the columns of $\mathbf{T}^*$.

Finally, we assume that non-zero loadings are "doubly bounded"—distinguishable from zero and not excessively large:

C1 There exist constants $0 < c < C$ such that for all j, k with $\lambda^*_{jk} \neq 0$, we have $c < |\lambda^*_{jk}| < C$.

Theorem 4. *Identifiability Condition for L^1 Rotation Criteria. Under C1, there exists a constant C_0, if $\frac{\min_k |Sp_k|}{|Sp^C|} > C_0$, then $\mathcal{T}^*$ contains all global minimum points of*

$$||\mathbf{\Lambda}^0\mathbf{T}^{-1'}||_1, \tag{3.2}$$

where $\mathbf{\Lambda}^0 \in \mathbb{R}^{J\times K}$ is any orthogonal solution satisfying $\mathbf{\Lambda}^0\mathbf{\Lambda}^{0'} = \mathbf{\Lambda}^*\mathbf{\Phi}\mathbf{\Lambda}^{*'}$.

The proof is provided in the Appendix for Chapter 3. This theorem suggests that when the proportion of non-simple items is small, the L^1 rotation can effectively recover the loading matrix. The proof shows that for simple items, any rotated loading matrix different from the true one will result in a strictly higher L^1 norm, while for non-simple items, the difference can be negative. Therefore, if there are enough simple items, their positive contributions dominate, ensuring that the true sparse loading matrix uniquely minimizes the L^1 norm. This guarantees that the L^1

rotation criterion can correctly recover the true loading structure when the proportion of simple items is sufficiently large.

Remark 4. *We believe that a similar result may hold for the L^p rotation with $p < 1$, but the analysis is more involved and is left for future work. In the proof below, we rely on the fact that $\sum_i |x_i|$ defines a norm and that the triangle inequality applies to it—this is used in Step 3 of Lemma 6 in the Appendix. However, for $p < 1$, the quantity $(\sum_i |x_i|^p)^{1/p}$ does not satisfy the properties of a norm, and therefore the triangle inequality no longer holds. Additionally, Hölder's inequality, which underpins many key steps in the analysis, is only valid for $p \geqslant 1$.*

Remark 5. *Compared with Proposition 2 in Chapter 2 (Liu et al., 2023), this identifiability condition addresses a more complex scenario where a small proportion of the items are non-simple, as opposed to all items being simple, which would yield a perfectly simple structure in the loading matrix. The conditions presented here ensure the applicability of condition C3 for Theorem 1 in Chapter 2 (Liu et al., 2023). Specifically, if the loading matrix is doubly bounded and contains only a minor proportion of non-simple items, the L^1 rotation criteria are sufficient to consistently recover the loading matrix, up to column swaps and sign flips, provided the sample size is sufficiently large.*

3.3. Numerical Experiment

As stated in Theorem 5, reproducing the loading matrix requires that the proportion of simple items remains below a certain threshold. However, our theoretical results do not provide a precise characterization of this constant. To investigate this limitation, we conduct a simulation study. Although the result is not formally proven for L^p rotation, we conjecture that it holds and thus include empirical evidence in this thesis.

In the simulation, we use a 60×3 loading matrix and vary the proportion of non-simple items from 0% to 95%. We evaluate the mean squared error (MSE) of the L^1 and $L^{0.5}$ rotation criteria, and compare their performance with two widely used oblique rotation methods: Oblimin and Geomin. The results are presented in Table 3.1.

To generate the initial sparse loading matrix, we first construct a 3×3 diagonal matrix for the primary loadings, where each diagonal element is independently sampled from the uniform distribution $U[1, 2]$. This process is repeated 20 times to create an initial block-diagonal structure for the loading matrix.

To evaluate the reproducibility of different rotation criteria under varying proportions of simple

items, cross-loadings are systematically added for non-simple items. For proportions of non-simple items ranging from 5% to 100%, secondary loadings are generated and assigned randomly to factors other than their primary factor. These cross-loadings are sampled from $U[0.2, 0.5]$, ensuring that they remain weaker than the primary loadings.

In the implementation, we utilize the `GPArotation` package (Bernaards and Jennrich, 2005) with 50 random rotation matrix starting points for the L^1, $L^{0.5}$, Geomin, and Oblimin rotation functions. The process is initialized using the true sparse loading matrix, assuming no initial error. To assess performance, we compute the Mean Squared Error (MSE) between the true and estimated loadings. Additionally, we evaluate the true positive rate (TPR) and true negative rate (TNR) to measure accuracy in identifying nonzero and zero entries, using a cutoff of $c = 0.01$. Loadings with a magnitude exceeding this threshold are classified as nonzero. This small cutoff is chosen because we start with the true sparse loading matrix rather than an estimated one based on a sample.

Proportion of	L^1			$L^{0.5}$		
Simple items	MSE	TNR	TPR	MSE	TNR	TPR
95%	1.27×10^{-8}	1	1	8.33×10^{-11}	1	1
90%	3.64×10^{-8}	1	1	2.96×10^{-10}	1	1
85%	7.67×10^{-8}	1	1	6.28×10^{-10}	1	1
80%	1.57×10^{-7}	1	1	1.21×10^{-9}	1	1
75%	2.76×10^{-7}	1	1	1.96×10^{-9}	1	1
70%	4.38×10^{-7}	1	1	3.02×10^{-9}	1	1
65%	5.72×10^{-7}	1	1	3.71×10^{-9}	1	1
60%	7.31×10^{-7}	1	1	4.90×10^{-9}	1	1
55%	1.00×10^{-6}	1	1	6.32×10^{-9}	1	1
50%	1.41×10^{-6}	1	1	8.20×10^{-9}	1	1
45%	1.66×10^{-6}	1	1	9.41×10^{-9}	1	1
40%	2.05×10^{-6}	1	1	1.11×10^{-8}	1	1
35%	2.83×10^{-6}	1	1	1.36×10^{-8}	1	1
30%	6.51×10^{-6}	1	1	1.95×10^{-8}	1	1
25%	6.93×10^{-6}	1	1	2.15×10^{-8}	1	1
20%	4.96×10^{-5}	0.778	1	2.88×10^{-8}	1	1
15%	4.99×10^{-5}	0.783	1	3.04×10^{-8}	1	1
10%	7.63×10^{-4}	0.455	1	3.76×10^{-8}	1	1
5%	1.22×10^{-2}	0.317	0.957	4.72×10^{-8}	1	1
0%	1.56×10^{-2}	0.333	0.958	1.53×10^{-2}	0.75	0.942

Table 3.1: The Power of Recovering Zeros with L^1 and $L^{0.5}$ Rotation with Varying Proportions of Simple Items

As shown in Table 3.1, by comparing the number of estimated zeros with the true zeros in the

Proportion of	Oblimin			GeominQ($\epsilon = 0.01$)		
Simple items	MSE	TNR	TPR	MSE	TNR	TPR
95%	1.13×10^{-4}	0.530	1	3.51×10^{-6}	1	1
90%	1.66×10^{-4}	0.509	1	8.61×10^{-6}	1	1
85%	3.53×10^{-4}	0.216	1	1.79×10^{-5}	0.919	1
80%	7.93×10^{-4}	0.185	1	3.70×10^{-5}	0.852	1
75%	1.30×10^{-3}	0.190	1	5.66×10^{-5}	0.771	1
70%	1.79×10^{-3}	0.196	1	8.61×10^{-5}	0.755	1
65%	2.14×10^{-3}	0	1	7.77×10^{-5}	0.626	1
60%	2.50×10^{-3}	0	1	8.97×10^{-5}	0.563	1
55%	3.15×10^{-3}	0	1	1.02×10^{-4}	0.516	1
50%	3.82×10^{-3}	0	1	1.33×10^{-4}	0.389	1
45%	4.71×10^{-3}	0	1	1.42×10^{-4}	0.322	1
40%	5.14×10^{-3}	0	1	1.63×10^{-4}	0.345	1
35%	6.00×10^{-3}	0	1	1.98×10^{-4}	0.296	1
30%	7.58×10^{-3}	0	1	3.04×10^{-4}	0.321	1
25%	8.14×10^{-3}	0	1	2.77×10^{-4}	0.173	1
20%	9.23×10^{-3}	0	0.991	3.70×10^{-4}	0.139	1
15%	9.42×10^{-3}	0	0.991	3.58×10^{-4}	0.087	1
10%	1.13×10^{-2}	0	0.974	4.37×10^{-4}	0.091	1
5%	1.31×10^{-2}	0	0.974	5.02×10^{-4}	0.063	1
0%	1.36×10^{-2}	0	0.975	4.97×10^{-4}	0.017	1

Table 3.2: Benchmark: The Power of Recovering Zeros with Oblimin and Geomin Rotation at Different Proportions of Simple Items

loading matrix, L^1 rotation accurately reproduces all zeros when the proportion of simple items exceeds 20%. Meanwhile, $L^{0.5}$ rotation demonstrates even greater power, successfully reproducing zeros even when the proportion of simple items is 5%. In contrast, benchmark methods such as Oblimin and GeominQ perform poorly, ceasing to work effectively when the proportion of simple items is high (e.g., 95% and 85%, respectively). These results demonstrate that L^p rotation is significantly more effective at recovering sparse loading matrices, even with a limited number of simple items.

3.4. Concluding Remarks

In this chapter, we have established that when the proportion of simple items is sufficiently large, the true sparse loading matrix can be recovered using the L^1 rotation criterion. This result fills a critical gap in the existing literature, which largely lacks theoretical guarantees for the identifiab-

ility of sparse loading matrices in exploratory factor analysis (EFA), except under the restrictive assumption that all items are simple.

We also generalize Proposition 2 from Chapter 2—originally formulated for L^1 rotation under a perfectly simple structure—to cases where the loading matrix includes non-simple items. This generalization is particularly valuable for applied settings, as EFA is commonly used in early stages of scale development or in contexts where the underlying loading structure is unknown. In practice, it is unrealistic to assume that every item loads exclusively on a single factor. For instance, ambiguous item wording can lead to cross-loadings, with items capturing multiple latent constructs simultaneously. A concrete example appears in Table 2.6, where item A2, "+I am interested in people," loads on both Extraversion and Agreeableness.

Despite these contributions, our work has certain limitations. Notably, we are unable to extend the theoretical guarantee to the case of L^p rotation with $p < 1$, as the function $(\sum_i |x_i|^p)^{1/p}$ is not a proper norm in this regime. Future research may explore whether identifiability results can still be established under this setting, and whether the L^p criterion with $p < 1$ could allow for recovery with a smaller proportion of simple items compared to the L^1 case as found in the simulation. Furthermore, it is known that component loss functions can recover a loading matrix composed entirely of simple items. Future research could investigate whether this result can be extended to cases involving a mix of simple and non-simple items, potentially broadening the applicability of component-based methods in practical EFA scenarios.

Lastly, another promising direction is to refine the theoretical conditions required for identifiability under L^1 rotation. For example, the required proportion of simple items may increase in scenarios with a large ratio between the upper and lower bounds of the nonzero loadings, a higher number of latent factors, or when the smallest eigenvalue of the factor covariance matrix is low. A more explicit relationship between these parameters and the required proportion of simple items could inform both theory and practice. Simulation studies could help illuminate this dependency and validate theoretical insights.

Appendix for Chapter 3

A3.1. Proof of Theorem 4

Notations

Norms

- $\|\cdot\|_1$: Entry-wise 1-norm (for matrices/vectors, not induced)
- $\|\cdot\|_2$: Matrix norm induced by the ℓ^2 vector norm
- $\|\cdot\|_F$: Frobenius norm

Subscript Conventions

- $\boldsymbol{\Lambda}_k$: k-th row of matrix $\boldsymbol{\Lambda}$
- $\boldsymbol{t}_k$, $\boldsymbol{\Phi}_k$, $\boldsymbol{A}_j$: k-th columns of matrices $\mathbf{T}$, $\boldsymbol{\Phi}$, and $\boldsymbol{A}$ respectively

Key Definitions

1. Initial loading matrix $\boldsymbol{\Lambda}^0$ is an arbitrary orthogonal solution satisfying:

$$\boldsymbol{\Lambda}^0(\boldsymbol{\Lambda}^0)' = \boldsymbol{\Lambda}^*\boldsymbol{\Phi}(\boldsymbol{\Lambda}^*)'$$

2. For positive definite $\boldsymbol{\Phi}$ with unit diagonal, there exists a unique rotation matrix $\boldsymbol{T}^*$ with unit columns ($\|\boldsymbol{t}_k^*\|_2 = 1\ \forall k$) such that:

$$\boldsymbol{\Lambda}^0\boldsymbol{T}^{*-1} = \boldsymbol{\Lambda}^*$$

The oblique rotation matrix space is:

$$\mathcal{M} = \{\boldsymbol{T} \in \mathbb{R}^{K\times K} : \|\boldsymbol{t}_i\|_2 = 1, \text{rank}(\boldsymbol{T}) = K\}$$

With properties:

$$\boldsymbol{\Phi} = \boldsymbol{T}^{*\prime}\boldsymbol{T}^*, \quad \phi_{ij} = \langle \boldsymbol{t}_i^*, \boldsymbol{t}_j^* \rangle$$

Define singular values $\lambda_1(\boldsymbol{\Phi})$ (largest) and $\lambda_K(\boldsymbol{\Phi})$ (smallest), and:

$$\Phi_{\text{offmax}} = \max_{i \neq j} |\phi_{ij}| = \max_{i \neq j} |\langle \boldsymbol{t}_i^*, \boldsymbol{t}_j^* \rangle| \in [0, 1)$$

3. S: Index set of simple items in $\boldsymbol{\Lambda}^*$ (loading on exactly one factor)
   - $S_k = \{j \in S : \lambda_{jk}^* \neq 0\}$ (simple items for factor k)
   - S^c: Non-simple items (loading on $\geqslant 2$ factors)
4. $\mathcal{D}_1$: Set of $K \times K$ permutation matrices ($K!$ elements)
5. $\mathcal{D}_2$: Set of $K \times K$ sign-flip matrices (2^K elements)
6. $\mathcal{T}^*$: Solution set with $2^K K!$ elements:

$$\mathcal{T}^* := \{\boldsymbol{T}^* \boldsymbol{D} \tilde{\boldsymbol{D}} : \boldsymbol{D} \in \mathcal{D}_1, \tilde{\boldsymbol{D}} \in \mathcal{D}_2\}$$

   Labeled as $\boldsymbol{T}^{*(i)}$ for $i = 1, \ldots, 2^K K!$
7. Voronoi cell for $\boldsymbol{T}^{*(i)}$:

$$V_i := \{\boldsymbol{T} \in \mathcal{M} : \min_{\tilde{\boldsymbol{T}}^* \in \mathcal{T}^*} \|\boldsymbol{T} - \tilde{\boldsymbol{T}}^*\|_F^2 = \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F^2\}$$

The structure of our proof follows a clear framework. In Lemma 8, we show that for simple items, the difference in the L_1 norm between the rotated loading and the true sparse loading is lower bounded by a positive value. In Lemma 9, we prove that for non-simple items, the same difference is lower bounded by a negative value. By combining these two results, we demonstrate that as long as the proportion of simple items is sufficiently large, the overall difference remains positive. Consequently, any rotated matrix other than the true sparse one cannot achieve the minimum.

To prove Lemma 8, we further rely on Lemma 6, which establishes that the result holds in a local neighborhood of the true parameter, and Lemma 7, which extends the result to regions that are at a distance from the true parameter.

Proof. Combining Lemmas 8 and 9, for any $\boldsymbol{T} \in V_i$ satisfying

$$\frac{\min_k |S_k|}{|S^c|} > \frac{2CK}{d_K(\boldsymbol{T}^*) C_{\text{sim}}} \stackrel{\text{def}}{=} C_p,$$

we analyze the norm difference:

$$\|\boldsymbol{\Lambda}\|_1 - \|\boldsymbol{\Lambda}^*\|_1 = \underbrace{\|\boldsymbol{\Lambda}_S\|_1 - \|\boldsymbol{\Lambda}^*_S\|_1}_{\text{Simple items}} + \underbrace{\|\boldsymbol{\Lambda}_{S^c}\|_1 - \|\boldsymbol{\Lambda}^*_{S^c}\|_1}_{\text{Non-simple items}} \tag{3.3}$$

$$\geqslant \left(\min_k |S_k| \cdot C_{\text{sim}} - |S^c| \cdot \frac{2CK}{d_K(\boldsymbol{T}^*)} \right) \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F \tag{3.4}$$

$$\geqslant 0 \tag{3.5}$$

The final equality holds if and only if $\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F = 0$, which implies $\boldsymbol{T} \in \mathcal{T}^*$. □

In Lemma 1 below, we show that for any simple item, the difference in the L^1 norm between a rotated loading matrix and the true sparse loading matrix is locally lower bounded by a positive value. This result strengthens Lemma 1 in Jennrich (2006) by improving the lower bound from zero to a strictly positive value, which is essential to compensate for the presence of non-simple items in the true loading matrix. We obtain this result by applying a Taylor expansion to the inverse of the true rotation matrix, T^{*-1}.

Lemma 6. *There exists $\epsilon_0 > 0$ such that for all $\boldsymbol{T} \in \mathcal{M}$ with $\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F \leqslant \epsilon_0$, the following inequality holds:*

$$\|\boldsymbol{\Lambda}_S\|_1 - \|\boldsymbol{\Lambda}^*_S\|_1 \geqslant \min_k |S_k| \cdot C_{\text{loc}} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F,$$

where

$$C_{\text{loc}} = \frac{c(1 - \Phi_{\text{offmax}})}{2\sqrt{2\lambda_1(\boldsymbol{\Phi})}(1 + \Phi_{\text{offmax}})}.$$

Neighborhood Parameterization

For any $\boldsymbol{T} \in \mathcal{M}$ in the ϵ_0-neighborhood of $\boldsymbol{T}^{*(i)}$, we employ the following representation:

$$\boldsymbol{T} = \left[\sqrt{1 - \epsilon_1^2}\boldsymbol{t}_1^{*(i)} + \epsilon_1 \Delta \boldsymbol{t}_1, \ldots, \sqrt{1 - \epsilon_K^2}\boldsymbol{t}_K^{*(i)} + \epsilon_K \Delta \boldsymbol{t}_K \right] = \boldsymbol{T}^{*(i)} \boldsymbol{D}_{1-\epsilon^2} + \Delta \boldsymbol{T} \boldsymbol{D}_\epsilon$$

where the parameters satisfy:

1. **Magnitude parameters**: For each $k = 1, \ldots, K$,

$$\epsilon_k \in [0, \tfrac{1}{2}], \quad \epsilon = \sqrt{\sum_{k=1}^{K} \epsilon_k^2}$$

The Frobenius norm distance decomposes as:

$$\begin{aligned}\|\boldsymbol{T}-\boldsymbol{T}^{*(i)}\|_F^2 &= \sum_{k=1}^{K}\|\boldsymbol{t}_k-\boldsymbol{t}_k^{*(i)}\|_2^2 \\ &= \sum_{k=1}^{K}\left\|(\sqrt{1-\epsilon_k^2}-1)\boldsymbol{t}_k^{*(i)}+\epsilon_k\Delta\boldsymbol{t}_k\right\|_2^2 \\ &= \sum_{k=1}^{K}\left[(\sqrt{1-\epsilon_k^2}-1)^2+\epsilon_k^2\right]\end{aligned}$$

Using the inequality $0 \leqslant (1-\sqrt{1-\epsilon_k^2})^2 \leqslant \epsilon_k^2$ for $\epsilon_k \in [0, \frac{1}{2}]$, we obtain the key bounds:

$$\epsilon^2 \leqslant \|\boldsymbol{T}-\boldsymbol{T}^{*(i)}\|_F^2 \leqslant 2\epsilon^2 \tag{3.6}$$

This representation covers all matrices in the ϵ_0-neighborhood of $\boldsymbol{T}^{*(i)}$ in $\mathcal{M}$ for $\epsilon \in (0, \epsilon_0]$.

2. **Diagonal scaling matrices**:

$$\boldsymbol{D}_\epsilon = \operatorname{diag}(\epsilon_1,\ldots,\epsilon_K), \quad \boldsymbol{D}_{1-\epsilon^2} = \operatorname{diag}\left(\sqrt{1-\epsilon_1^2},\ldots,\sqrt{1-\epsilon_K^2}\right)$$

with inverse approximation:

$$\boldsymbol{D}_{1-\epsilon^2}^{-1} = \boldsymbol{I}+\boldsymbol{O}_K(\epsilon^2)$$

3. **Column properties**: For each column k,

$$\|\boldsymbol{t}_k^{*(i)}\|_2 = \|\Delta\boldsymbol{t}_k\|_2 = 1 \quad \text{and} \quad \langle\Delta\boldsymbol{t}_k, \boldsymbol{t}_k^{*(i)}\rangle = 0$$

(i) **Coefficient Matrix Analysis**:

Given the full-rank matrix $\boldsymbol{T}^{*(i)} = [\boldsymbol{t}_1^{*(i)},\ldots,\boldsymbol{t}_K^{*(i)}]$, any perturbation vector can be expressed as:

$$\Delta\boldsymbol{t}_j = \sum_{k=1}^{K} a_{kj}\boldsymbol{t}_k^{*(i)} = \boldsymbol{T}^{*(i)}\boldsymbol{A}_j$$

where $\boldsymbol{A}_j = (a_{1j},\ldots,a_{Kj})'$ is the j-th column of the coefficient matrix $\boldsymbol{A}$.

a) **Orthogonality Condition**: From the constraint $\langle\Delta\boldsymbol{t}_j, \boldsymbol{t}_j^{*(i)}\rangle = 0$, we derive:

$$0 = \langle\boldsymbol{T}^{*(i)}\boldsymbol{A}_j, \boldsymbol{t}_j^{*(i)}\rangle = \boldsymbol{A}_j'\boldsymbol{\Phi}_j^{(i)} \tag{3.7}$$

where $\boldsymbol{\Phi}^{(i)} = \boldsymbol{T}^{*(i)\prime}\boldsymbol{T}^{*(i)}$ with entries $\Phi_{kj}^{(i)} = \langle\boldsymbol{t}_k^{*(i)}, \boldsymbol{t}_j^{*(i)}\rangle$.

b) **Normalization Condition**: The unit norm requirement $\|\Delta \boldsymbol{t}_j\|_2 = 1$ yields:

$$1 = \boldsymbol{A}_j' \boldsymbol{\Phi}^{(i)} \boldsymbol{A}_j \tag{3.8}$$

(ii) **Spectral Bounds**: The quadratic form in (3.8) satisfies:

$$\lambda_K(\boldsymbol{\Phi}) \|\boldsymbol{A}_j\|_2^2 \leqslant \boldsymbol{A}_j' \boldsymbol{\Phi}^{(i)} \boldsymbol{A}_j \leqslant \lambda_1(\boldsymbol{\Phi}) \|\boldsymbol{A}_j\|_2^2$$

Combining with (3.8) gives the key inequality:

$$\frac{1}{\sqrt{\lambda_1(\boldsymbol{\Phi})}} \leqslant \|\boldsymbol{A}_j\|_2 \leqslant \frac{1}{\sqrt{\lambda_K(\boldsymbol{\Phi})}} \tag{3.9}$$

(iii) **Matrix Representation**: Collecting all columns yields the compact form:

$$\Delta \boldsymbol{T} = \boldsymbol{T}^{*(i)} \boldsymbol{A} \tag{3.10}$$

4. **Inverse Approximation**: Using Woodbury identity:

$$\begin{aligned}(\boldsymbol{I} + \boldsymbol{D}_\epsilon \boldsymbol{D}_{1-\epsilon^2}^{-1} \boldsymbol{A})^{-1} &= \boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{D}_\epsilon \boldsymbol{D}_{1-\epsilon^2}^{-1} \boldsymbol{A})^{-1} \boldsymbol{D}_\epsilon \boldsymbol{D}_{1-\epsilon^2}^{-1} \boldsymbol{A} \\ &= \boldsymbol{I} + \boldsymbol{O}_K(\epsilon)\end{aligned}$$

where we use:

- $\|\boldsymbol{D}_\epsilon \boldsymbol{D}_{1-\epsilon^2}^{-1} \boldsymbol{A}\|_F = \mathcal{O}(\epsilon)$
- Submultiplicativity of $\|\cdot\|_F$
- The bound $\|\boldsymbol{A}\|_F \leqslant \sqrt{K/\lambda_K(\boldsymbol{\Phi})}$ from (3.9)

Detailed Derivation. The proof proceeds through several key steps:

Step 1: Inverse Matrix Decomposition Applying the Woodbury matrix identity to $\boldsymbol{T}^{-1}$, we obtain:

$$\begin{aligned}\boldsymbol{T}^{-1} &= \left(\boldsymbol{T}^{*(i)} \boldsymbol{D}_{1-\epsilon^2} + \Delta \boldsymbol{T} \boldsymbol{D}_\epsilon\right)^{-1} \\ &= \left[\boldsymbol{I} - \boldsymbol{D}_{1-\epsilon^2}^{-1} \boldsymbol{A} \left(\boldsymbol{I} + \boldsymbol{D}_\epsilon \boldsymbol{D}_{1-\epsilon^2}^{-1} \boldsymbol{A}\right)^{-1} \boldsymbol{D}_\epsilon\right] \boldsymbol{D}_{1-\epsilon^2}^{-1} \boldsymbol{T}^{*(i)-1} \\ &= \left[\boldsymbol{I} - \boldsymbol{A} \boldsymbol{D}_\epsilon + \boldsymbol{O}_K(\epsilon^2)\right] \boldsymbol{T}^{*(i)-1}\end{aligned}$$

where $\boldsymbol{O}_K(\epsilon^2)$ denotes a $K \times K$ matrix with all entries of order $\mathcal{O}(\epsilon^2)$, with constants independent of J and ϵ.

Step 2: Norm Difference Decomposition For each loading vector $\boldsymbol{\Lambda}_j$, we analyze:

$$\begin{aligned}\|\boldsymbol{\Lambda}_j\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 &= \|\boldsymbol{\Lambda}_j^0 \boldsymbol{T}'^{-1}\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 \\ &\geqslant \|\boldsymbol{\Lambda}_j^* - \boldsymbol{\Lambda}_j^* \boldsymbol{D}_\epsilon \boldsymbol{A}'\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 - C_1\epsilon^2\end{aligned}$$

The constant C_1 bounds the residual term:

$$\|\boldsymbol{\Lambda}_j^* \boldsymbol{O}_K(\epsilon^2)\|_1 \leqslant C_1\epsilon^2$$

Step 3: Simple Items ($j \in S$) For items loading on a single factor k_j:

$$\begin{aligned}&\sum_{k=1}^{K} \left(|\lambda_{jk}^* - \lambda_{jk_j}^* \epsilon_{k_j} a_{kk_j}| - |\lambda_{jk}^*| \right) \\ &= \underbrace{|\lambda_{jk_j}^* - \lambda_{jk_j}^* \epsilon_{k_j} a_{k_j k_j}| - |\lambda_{jk_j}^*|}_{\text{Main Loading term}} + \underbrace{\sum_{k \neq k_j} |\lambda_{jk_j}^* \epsilon_{k_j} a_{kk_j}|}_{\text{Zero Loading terms}} \\ &\geqslant |\lambda_{jk_j}^* \epsilon_{k_j}| \left(\sum_{k \neq k_j} |a_{kk_j}| - |a_{k_j k_j}| \right)\end{aligned}$$

From the orthogonality condition (3.7), we derive the key inequality:

$$|a_{k_j k_j}| = \left| \sum_{k \neq k_j} a_{kk_j} \boldsymbol{\Phi}_{kk_j}^{(i)} \right| \leqslant \boldsymbol{\Phi}_{\text{offmax}} \sum_{k \neq k_j} |a_{kk_j}| \tag{3.11}$$

Using norm relationships and (3.9), we obtain:

$$\frac{1}{\sqrt{\lambda_1(\boldsymbol{\Phi})}} \leqslant \|\boldsymbol{A}_{k_j}\|_2 \leqslant (1 + \boldsymbol{\Phi}_{\text{offmax}}) \sum_{k \neq k_j} |a_{kk_j}| \tag{3.12}$$

Combining these results yields:

$$\|\boldsymbol{\Lambda}_j\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 \overset{(3.11)}{\geqslant} \underbrace{\mathcal{O}(\epsilon^2)}_{\text{Higher order}} + \underbrace{c\epsilon_{k_j}(1 - \boldsymbol{\Phi}_{\text{offmax}}) \sum_{k \neq k_j} |a_{kk_j}|}_{\text{Main term}}$$

$$\overset{(3.12)}{\geqslant} -C_1\epsilon^2 + C_2\epsilon_{k_j}$$

where the crucial constant is:

$$C_2 = c\frac{1 - \mathbf{\Phi}_{\text{offmax}}}{\sqrt{\lambda_1(\mathbf{\Phi})}(1 + \mathbf{\Phi}_{\text{offmax}})}$$

Note that when $\mathbf{\Phi} = \boldsymbol{I}$ (orthogonal case), $\mathbf{\Phi}_{\text{offmax}} = 0$ simplifies C_2 to 1.

Step 4: Final Bound Choosing $\epsilon_0 = \frac{C_2}{2C_1K}$ ensures:

$$\begin{aligned}\|\mathbf{\Lambda}_S\|_1 - \|\mathbf{\Lambda}^*_S\|_1 &\geqslant \min_k |S_k| \left(C_2 - C_1K\epsilon\right)\epsilon \\ &\geqslant \min_k |S_k| \frac{C_2}{2}\frac{\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F}{\sqrt{2}}\end{aligned}$$

completing the proof with the stated constant $C_{\text{loc}} = \frac{C_2}{2\sqrt{2}}$. □

In Lemma 2 below, we show that for any simple item, the difference in the L^1 norm between a rotated loading matrix and the true sparse loading matrix is lower bounded by a positive value within a region that is at a nonzero distance from the true parameter.

Lemma 7. *For any $\boldsymbol{T} \in V_i$, we have the global lower bound:*

$$\|\mathbf{\Lambda}_S\|_1 - \|\mathbf{\Lambda}^*_S\|_1 \geqslant \min_k |S_k| \cdot C_{\text{glo}}\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F^2, \tag{3.13}$$

where the global constant is:

$$C_{\text{glo}} = \frac{c(1 - \mathbf{\Phi}_{\text{offmax}})}{4K}.$$

Proof. We analyze the rotation relationship $\mathbf{\Lambda}^0 = \mathbf{\Lambda}^*\boldsymbol{T}^{*\prime} = \mathbf{\Lambda}\boldsymbol{T}'$, which implies for each item j:

$$\mathbf{\Lambda}^0_j = \sum_k \lambda^*_{jk}\boldsymbol{t}^{*\prime}_k = \sum_k \lambda_{jk}\boldsymbol{t}'_k. \tag{3.14}$$

Simple Items Analysis For $j \in S$ with unique non-zero entry at k_j:

$$\begin{aligned}\|\mathbf{\Lambda}_j\|_1 - \|\mathbf{\Lambda}^*_j\|_1 &= \sum_k |\lambda_{jk}| - |\lambda^*_{jk_j}| \\ &= \sum_k |\lambda_{jk}| - \left|\sum_k \lambda_{jk}\langle \boldsymbol{t}_k, \boldsymbol{t}^*_{k_j}\rangle\right| \\ &\geqslant \sum_k |\lambda_{jk}|(1 - |\langle \boldsymbol{t}_k, \boldsymbol{t}^*_{k_j}\rangle|).\end{aligned}$$

From the norm identity:

$$\min_{a=\pm 1} \|\boldsymbol{t}_k - a\boldsymbol{t}_{k_j}^*\|_2^2 = 2(1 - |\langle \boldsymbol{t}_k, \boldsymbol{t}_{k_j}^* \rangle|), \tag{3.15}$$

and using $\sum_k |\lambda_{jk}| \geqslant |\lambda_{jk_j}^*| \geqslant c$ from Jennrich (2006), we obtain:

$$\|\boldsymbol{\Lambda}_j\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 \geqslant \frac{c}{2} \min_{a=\pm 1, k} \|\boldsymbol{t}_k - a\boldsymbol{t}_{k_j}^*\|_2^2.$$

Aggregate Bound Summing over all simple items:

$$\|\boldsymbol{\Lambda}_S\|_1 - \|\boldsymbol{\Lambda}_S^*\|_1 \geqslant \frac{c}{2} \sum_k |S_k| \min_{a=\pm 1, k'} \|\boldsymbol{t}_{k'} - a\boldsymbol{t}_k^*\|_2^2 \tag{3.16}$$

Case Analysis Let $m_k = \arg\min_{k'} \min_{a=\pm 1} \|\boldsymbol{t}_{k'} - a\boldsymbol{t}_k^*\|_2^2$ and $a_k = \arg\min_{a=\pm 1} \|\boldsymbol{t}_{m_k} - a\boldsymbol{t}_k^*\|_2^2$.

Case 1: **Non-permutation case**: If $\exists k_1 \neq k_2$ with $m_{k_1} = m_{k_2}$:

$$\begin{aligned}
\sum_k \|\boldsymbol{t}_{m_k} - a_k \boldsymbol{t}_k^*\|_2^2 &\geqslant ||\boldsymbol{t}_{m_{k_1}} - a_{k_1}\boldsymbol{t}_{k_1}^*||_2^2 + ||\boldsymbol{t}_{m_{k_2}} - a_{k_2}\boldsymbol{t}_{k_2}^*||_2^2 \\
&\geqslant \frac{(||\boldsymbol{t}_{m_{k_1}} - a_{k_1}\boldsymbol{t}_{k_1}^*||_2 + ||\boldsymbol{t}_{m_{k_2}} - a_{k_2}\boldsymbol{t}_{k_2}^*||_2)^2}{2} \\
&\geqslant \frac{\|\boldsymbol{t}_{k_1}^* - \boldsymbol{t}_{k_2}^*\|_2^2}{2} \\
&\geqslant 1 - \boldsymbol{\Phi}_{\text{offmax}} \\
&> 0
\end{aligned}$$

Using $\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F^2 \leqslant 2K$, we get:

$$\sum_k \|\boldsymbol{t}_{m_k} - a_k \boldsymbol{t}_k^*\|_2^2 \geqslant \frac{1 - \boldsymbol{\Phi}_{\text{offmax}}}{2K} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F^2 \tag{3.17}$$

Case 2: **Permutation case**: When $\{m_k\}$ forms a permutation,then by the definition of V_i,:

$$\sum_k \|\boldsymbol{t}_{m_k} - a_k \boldsymbol{t}_k^*\|_2^2 \geqslant \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F^2 \geqslant \frac{1 - \boldsymbol{\Phi}_{\text{offmax}}}{2K} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F^2 \tag{3.18}$$

Substituting (3.17) or (3.18) into (3.16) yields the desired bound. □

Lemma 8. *For any rotation matrix $\boldsymbol{T} \in V_i$, there exists a constant $C_{\text{sim}} > 0$ such that:*

$$\|\boldsymbol{\Lambda}_S\|_1 - \|\boldsymbol{\Lambda}_S^*\|_1 \geqslant \min_k |S_k| \cdot C_{\text{sim}} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F$$

Proof. We consider two cases based on the distance to the optimal rotation:

Case 1: Local Neighborhood ($\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F \leqslant \epsilon_0$) By Lemma 6, we have the local lower bound:

$$\|\boldsymbol{\Lambda}_S\|_1 - \|\boldsymbol{\Lambda}_S^*\|_1 \geqslant \min_k |S_k| \cdot C_{\text{loc}} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F$$

Case 2: Global Region ($\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F \geqslant \epsilon_0$) By Lemma 7, we obtain the global lower bound:

$$\|\boldsymbol{\Lambda}_S\|_1 - \|\boldsymbol{\Lambda}_S^*\|_1 \geqslant \min_k |S_k| \cdot C_{\text{glo}} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F^2 \geqslant \min_k |S_k| \cdot C_{\text{glo}} \epsilon_0 \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F$$

Combining Both Cases Taking $C_{\text{sim}} = \min(C_{\text{loc}}, C_{\text{glo}}\epsilon_0)$ yields the unified bound:

$$\|\boldsymbol{\Lambda}_S\|_1 - \|\boldsymbol{\Lambda}_S^*\|_1 \geqslant \min_k |S_k| \cdot C_{\text{sim}} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F$$

for all $\boldsymbol{T} \in V_i$, completing the proof. □

In Lemma 4 below, we show that for any non-simple item, the difference in the L^1 norm between a rotated loading matrix and the true sparse loading matrix is lower bounded by a negative value. This result holds due to the boundedness of the loadings and the fact that the true rotation matrix T^* is not rank-deficient.

Lemma 9. *For any $\boldsymbol{T} \in V_i$, the non-simple items satisfy:*

$$\|\boldsymbol{\Lambda}_{S^c}\|_1 - \|\boldsymbol{\Lambda}_{S^c}^*\|_1 \geqslant -|S^c| \frac{2CK}{d_K(\boldsymbol{T}^*)} \|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F \tag{3.19}$$

Proof. We analyze the non-simple items through careful estimation:

Setup For $\boldsymbol{T} \in V_i$, we have:

$$\begin{aligned}
\|\boldsymbol{\Lambda}_{S^c}\|_1 - \|\boldsymbol{\Lambda}_{S^c}^*\|_1 &= \|\boldsymbol{\Lambda}_{S^c}^0 \boldsymbol{T}'^{-1}\|_1 - \|\boldsymbol{\Lambda}_{S^c}^*\|_1 \\
&= ||(\boldsymbol{\Lambda}_{S^c}^* \boldsymbol{T}^{*(i)'}) \boldsymbol{T}'^{-1}||_1 - ||\boldsymbol{\Lambda}_{S^c}^*||_1 \\
&= \|\boldsymbol{\Lambda}_{S^c}^* (\boldsymbol{T} + \boldsymbol{T}^{*(i)} - \boldsymbol{T})' \boldsymbol{T}'^{-1}\|_1 - \|\boldsymbol{\Lambda}_{S^c}^*\|_1 \\
&= \sum_{j \in S^c} \left(\|\boldsymbol{\Lambda}_j^* + \boldsymbol{\Lambda}_j^* \boldsymbol{P}\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 \right)
\end{aligned}$$

where $\boldsymbol{P} = (\boldsymbol{T}^{*(i)} - \boldsymbol{T})' \boldsymbol{T}'^{-1}$.

Case Analysis For each non-simple item $j \in S^c$, consider:

Case 1: $\|\boldsymbol{T} - \boldsymbol{T}^*\|_F \leqslant \frac{d_K(\boldsymbol{T}^*)}{2}$

By Weyl's inequality:

$$|d_K(\boldsymbol{T}) - d_K(\boldsymbol{T}^*)| \leqslant \|\boldsymbol{T} - \boldsymbol{T}^*\|_2 \leqslant \|\boldsymbol{T} - \boldsymbol{T}^*\|_F \leqslant \frac{d_K(\boldsymbol{T}^*)}{2}$$

Thus $d_K(\boldsymbol{T}) \geqslant \frac{d_K(\boldsymbol{T}^*)}{2}$.

For the norm difference:

$$\begin{aligned}\|\boldsymbol{\Lambda}_j\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 &\geqslant -\sum_k |\boldsymbol{\Lambda}_j^* \boldsymbol{P}_k| \\ &\geqslant -C\|\boldsymbol{P}\|_1 \quad \text{(by Cauchy-Schwarz)}\end{aligned}$$

Estimating $\|\boldsymbol{P}\|_1$:

$$\begin{aligned}\|\boldsymbol{P}'\|_1 &= \|\boldsymbol{T}^{-1}(\boldsymbol{T} - \boldsymbol{T}^{*(i)})\|_1 \\ &\leqslant \sum_k \sqrt{K}||\boldsymbol{T}^{-1}||_2||\boldsymbol{t}_k - \boldsymbol{t}_k^{*(i)}||_2 \\ &\leqslant \frac{2K}{d_K(\boldsymbol{T}^*)}\|\boldsymbol{T} - \boldsymbol{T}^{*(i)}\|_F\end{aligned}$$

Case 2: $\|\boldsymbol{T} - \boldsymbol{T}^*\|_F > \frac{d_K(\boldsymbol{T}^*)}{2}$

The trivial bound holds:

$$\|\boldsymbol{\Lambda}_j\|_1 - \|\boldsymbol{\Lambda}_j^*\|_1 \geqslant -KC \geqslant -\frac{2KC}{d_K(\boldsymbol{T}^*)}\|\boldsymbol{T} - \boldsymbol{T}^*\|_F$$

Conclusion Summing over all $j \in S^c$ yields the final result. □

Chapter 4

Controlling False Discovery Rate for Exploratory Factor Analysis

4.1. Introduction

In the previous chapters, we introduced L^p rotation, which has proven effective in identifying factors. Now, we aim to determine which indicators are truly associated with the latent variables and contribute to their interpretation (Bartholomew et al., 2008). To distinguish meaningful associations, we require hypothesis testing rather than relying solely on point estimates.

Although standard errors for the rotated loading matrix in EFA can be computed using smooth rotation criteria (Jennrich, 1974; Jennrich and Clarkson, 1980), these methods have limitations. As demonstrated in Chapter 3, smooth rotation criteria lack theoretical guarantees for recovering the loading matrix under sparse settings. While Theorem 4 in Chapter 3 establishes that L^p rotation can identify the true sparse loading matrix when the proportion of simple items is sufficiently large, the non-smooth nature of the rotation criteria makes inference challenging. To the best of our knowledge, the only prior attempt to quantify uncertainty in EFA under sparsity was Algorithm 2, discussed in Chapter 2. That approach leveraged the consistency of L^p rotation to recover the full loading structure for all J items, treating it as the ground truth—an assumption that is difficult to achieve with a small sample size. In contrast, in Section 4.3, we propose a new method that only requires K simple items in the loading structure to be correctly identified, significantly relaxing the identifiability assumptions.

After constructing p-values for individual loadings, we face the challenge of multiple hypothesis testing, where we must control the overall error rate while accounting for dependencies among

p-values. This introduces a statistically rigorous approach to item selection. Traditionally, psychometricians have relied on ad hoc methods, such as hard thresholds on factor loadings (e.g., above 0.3 or 0.4) or removing items with low communalities (Hair et al., 2019). However, these arbitrary cutoffs are often suboptimal across datasets. A data-driven statistical approach is needed to ensure more objective and reliable item selection.

To ensure the quality of multiple hypothesis testing, we require well-defined criteria to quantify uncertainty. In personality assessment, the goal is to select items that genuinely measure a given latent trait for subsequent analysis. A natural approach might be to strictly control false selections, ensuring that no irrelevant items are included—an objective analogous to controlling the familywise error rate (FWER) (Holm, 1979). However, such a stringent criterion is often overly conservative, leading to the omission of many meaningful items. Instead, an alternative approach is to control the false discovery rate (FDR) (Benjamini and Hochberg, 1995), which offers a balance between selecting relevant items and maintaining statistical rigor. This chapter explores statistical methods that achieve this balance, thereby improving item selection in factor analysis.

While the FDR control problem has garnered significant attention in statistical literature over the last two decades, many of the available methods are applicable in the common regression setting, where latent variables are not present. For example, Barber and Candès (2015) developed the knockoff filter to control FDR with high power by constructing copies of exploratory variables as a control group. However, their approach requires accurately estimating the distribution of these variables (Barber et al., 2020), which is infeasible in EFA since the variables are latent. A natural alternative is mirror statistics (Dai et al., 2023), which bypasses the need for data copies. However, the mirror statistic procedure relies on weak dependence assumptions that do not hold in EFA, making it inapplicable. Thus, existing methods fail to provide a valid FDR control framework for EFA.

In settings where the estimated coefficients are either independent or exhibit positive regression dependence on a subset of zero loadings (PRDS), the traditional Benjamini and Hochberg (BH) procedure (Benjamini and Hochberg, 1995) effectively controls the false discovery rate (FDR). However, this assumption does not hold in the context of exploratory factor analysis (EFA), where estimated loadings are inherently correlated. To address this challenge, Benjamini and Yekutieli (2001) introduced a modified BH procedure incorporating a logarithmic correction to account for dependence among estimated coefficients. Additionally, recent developments in statistical inference have explored the use of e-values as an alternative to p-values, with Wang and Ramdas (2022) extending the BH procedure to the e-value framework.

In this chapter, we propose a data-driven method to control the FDR at a prespecified level, enabling the selection of significant loadings at both the test and factor levels. Our approach resolves several previously unaddressed issues in uncertainty quantification and item selection for EFA models. First, as a prerequisite for implementing the FDR control procedures, we introduce a method that establishes the minimal conditions required to construct valid p-values for the EFA model under the assumption of a sparse loading matrix. We provide a theoretical proof demonstrating that this condition is sufficient for model identification. Second, we develop a novel effective latent variable selection procedure that ensures FDR control in the EFA model. Specifically, we adapt three widely used FDR control procedures to latent variable models.

The remainder of this chapter is organized as follows. Section 4.3 presents our proposed method for constructing p-values and e-values for EFA, along with theoretical validation. Section 4.4 reviews three standard FDR control methods and their adaptation for factor-level and test-level FDR control in EFA. Section 4.5 evaluates our approach through simulation studies, while Section 4.6 applies our method to the Big Five personality assessment. Finally, Section 4.7 discusses limitations and potential future research directions.

4.2. Problem Setup

As in previous chapters, we consider the following exploratory linear factor model with a J-dimensional vector of manifest variables $\boldsymbol{X}$ and a K-dimensional vector of common factors $\boldsymbol{\xi}$.

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

where $\boldsymbol{\Lambda} = (\lambda_{jk})_{J\times K}$ is the loading matrix, $\boldsymbol{\Phi} = (\phi_{ij})_{K\times K}$ and $\boldsymbol{\Omega} = (\omega_{ij})_{J\times J}$ are the covariance matrices for latent factors and residuals. We further require both covariance matrices being positive-definite and all latent factors having unit variances (i.e. $\phi_{kk} = 1$). Moreover, the manifest variables are assumed to be conditionally independent given $\boldsymbol{\xi}$, i.e., $\boldsymbol{\Omega}$ is a diagonal matrix. For simplifying the notation, we use $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$ to denote the set of all unknown parameters. The marginal distribution of X given $\boldsymbol{\theta}$ is

$$\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Omega}) \tag{4.1}$$

A non-zero loading λ_{jk} means the k-th latent common factor ξ_k contributes to the explanation

of j-th manifest variable X_j. The goal of this chapter is to select these non-zero loadings under the sparse loading matrix assumption while controlling the FDR. Based on our purpose, we can choose to control the FDR for the full or a single column of the loading matrix.

Factor-level FDR Control: In some cases, it may be desirable to control the proportion of incorrectly selected manifest variables for each common factor k . For example, a psychometrician might seek to include relevant items for a specific latent factor while allowing a small proportion of irrelevant items, within a pre-specified threshold, to shorten the test. Consequently, there is interest in managing the False Discovery Rate (FDR) for each column in the loading matrix.

For the k-th common factor, let $S_k \subseteq \{j : 1 \leqslant j \leqslant J\}$ denote the selection set of items j for which we hypothesize that the true factor loading $\lambda^*_{jk} \neq 0$. Here, λ^*_{jk} represents the (j,k)-th element in the true factor loading matrix, indicating the strength of the relationship between item j and factor k.

The false discovery rate (FDR) for the k-th factor is then defined as:

$$\text{FDR}_k = \mathbb{E}\left(\frac{|\{j \in S_k, \lambda^*_{jk} = 0\}|}{|S_k| \vee 1}\right), k = 1, ..., K. \tag{4.2}$$

where $a \vee b = \max(a, b)$. The denominator above ensures that FDR_k is zero when no items are selected. Suppose there are 100 items selected for latent factor k. Controlling the FDR at level $q = 0.1$ means that we can expect at most 10 of the selected items to be irrelevant to factor k, and at least 90 to be relevant.

test-level FDR Control: When our objective is to control the False Discovery Rate (FDR) for the entire matrix rather than focusing on a specific latent factor, we employ test-level FDR control. Let $S \subseteq \{(j,k) : 1 \leqslant j \leqslant J, 1 \leqslant k \leqslant K\}$ be the set of selected loadings. The goal here is to include as many as none zero loadings as possible while controlling FDR, the expected proportion of selected zero loadings, in the full matrix $\mathbf{\Lambda}$:

$$\text{FDR} = \mathbb{E}\left(\frac{|\{(j,k) \in S, \lambda^*_{jk} = 0\}|}{|S| \vee 1}\right). \tag{4.3}$$

Note that the maximum in the denominator is to ensure the division by zero does not occur. Suppose that 100 item loadings are selected for 3 latent factors. Controlling the FDR at level $q = 0.1$ implies that we can expect at most 10 of the selected loadings to be zero, and at least 90 to be non-zero.

4.3. Single hypothesize testing for Exploratory Factor Analysis Models

4.3.1 p-values and e-values for Hypothesis Testing

Before addressing false discovery rate (FDR) control—a multiple hypothesis testing problem—we begin by discussing valid procedures for single hypothesis testing in the context of sparse loading matrices.

Traditional hypothesis testing commonly uses a p-value, which is a random variable $p \in [0, 1]$ satisfying $\mathbb{P}(p \leqslant t) \leqslant t$ for all $t \in [0, 1]$ under the null hypothesis. A p-value quantifies how extreme the test statistic, derived from the data, is relative to the null distribution. Given a significance level q, we reject the null hypothesis if $p < q$.

However, p-values have notable limitations in the multiple testing setting. Aggregating them across tests often relies on strong independence assumptions. Furthermore, combining p-values analytically—such as through convolution to evaluate joint distributions—is computationally challenging. In contrast, the linearity of expectation makes e-values more tractable: the expectation of a sum of random variables equals the sum of their expectations, allowing for easier aggregation across tests.

An alternative is the use of e-values, which are based on expectations rather than probabilities. An e-value is a non-negative random variable e such that $\mathbb{E}[e] \leqslant 1$ under the null hypothesis (Vovk and Wang, 2021). A large e-value indicates evidence against the null. To control the Type-I error at level q, we reject the null if $e \geqslant 1/q$. Since,

$$\mathbb{P}(e \geqslant 1/q) \leqslant q\mathbb{E}[e] \leqslant q, \tag{4.4}$$

where the first inequality is due to Markov's inequality and the second follows from the definition of the e-value. e-values can be constructed from p-values using p-to-e calibrators, such as:

$$e_{jk} = p_{jk}^{-1/2} - 1, \quad \text{with} \quad \mathbb{E}[e_{jk}] = \int_0^1 (p^{-1/2} - 1)dp = 1, \tag{4.5}$$

$$e_{jk} = \kappa p_{jk}^{\kappa-1}, \quad \text{where } \kappa \in (0, 1), \quad \text{since } \mathbb{E}[e_{jk}] = \int_0^1 \kappa p^{\kappa-1} dp = 1. \tag{4.6}$$

4.3.2 Inference Procedure for EFA Models

Constructing valid p-values or e-values for exploratory factor analysis (EFA) models is challenging due to the presence of latent variables. To perform inference, the model must first be identifiable, which requires fixing certain parameters. As demonstrated in earlier chapters, L^1 rotation can identify a sparse loading matrix when there are sufficiently many simple items, and L^p rotation achieves superior mean squared error (MSE) performance under this setting compared to widely used rotation methods. However, as shown in Study A.I in the Appendix, although L^p rotation yields consistent estimators, the sampling distribution of zero loadings deviates significantly from normality when the number of items is small. As a result, the central limit theorem does not apply, and standard p-value construction methods based directly on the L^p-rotated loadings become unreliable.

A natural solution is to leverage the consistency of the L^p-rotated loading matrix for sparse structures to perform item selection. Once the structure is identified, a confirmatory factor analysis (CFA) model can be fitted to construct valid p-values. Initially, Algorithm 2 in Chapter 2 was proposed for this purpose. However, as discussed in the concluding remarks of that chapter, obtaining well-calibrated confidence intervals requires two strong conditions: accurate model selection for all J items (Condition C5) and identifiability of the loading matrix after the removal of each row (Condition C6). Furthermore, the original procedure does not account for the uncertainty introduced by model selection.

In this section, we introduce a more powerful inference procedure that relaxes previous assumptions, enhances computational efficiency, and enables inference for both factor covariances and unique variances. Specifically, instead of requiring accurate identification of the full loading pattern (Condition C5), we demonstrate in Subsection 4.3.3 that it suffices to identify K simple items using rank statistics, as defined in Subsection 4.3.4, to serve as *reference* variables. The procedure for selecting these reference items is outlined in Algorithm 5.

Finally, in subsection 4.3.5, we address the challenge of inference after model selection—where the same dataset is used for both selection and estimation—by employing a data-splitting strategy, implemented in Algorithm 6, to ensure the validity of inference.

4.3.3 Minimal Information Condition for Identification

To quantify uncertainty in an EFA model, additional constraints must be imposed to ensure identifiability (Koopmans and Reiersol, 1950; Anderson and Rubin, 1956; Bai and Li, 2012). We adopt a

Algorithm 5 Anchor Item Selection Using L^p Rotation

Input: Observed data X and the value of p.

1. Estimate the initial loading matrix, $\hat{\mathbf{\Lambda}}_0$, assuming orthogonal factors ($\mathbf{\Omega} = \mathbf{I}$) and imposing additional constraints on the loading matrix to resolve rotational indeterminacy. In the numerical study presented in this chapter, the initial loading matrix is constrained to be lower triangular to ensure model identification.
2. Apply the L^p rotation criterion to $\hat{\mathbf{\Lambda}}_0$.
3. Use the rank statistic defined in Equation (4.7) (Subsection 4.3.4) to identify one *reference* item per factor. For the k-th latent factor, select the item j with the smallest L_{jk} among $j \in \{1, 2, \ldots, J\}$ as the reference item, and set $\Gamma_{jk'} = 0$ for all $k' \neq k$. Denote the selected item for the k-th factor as id_k.

$$L_{jk} = \max\left(|\hat{\lambda}_{jk'}| : k' \neq k\right), \tag{4.7}$$

Output: The loading structure Γ.

Algorithm 6 Constructing p-Values or e-Values for the EFA Model

Input: Observed data X.

1. Split the data into two subsets, $X^{(1)}$ and $X^{(2)}$.
2. Apply Algorithm 5 to $X^{(1)}$ to obtain the loading structure $\Gamma^{(1)}$.
3. Use $\Gamma^{(1)}$ on $X^{(2)}$ to define the CFA model and compute p-values.
4. To ensure proper reference item selection, set $p_{id_k k'} = 1$ for all $k' \neq k$, ensuring that the zero loadings of reference items are excluded from selection.

Output: The p-values for each loading. We can also construct e-values using (4.5) or (4.6).

widely used identification condition with a long history in the factor analysis literature, as outlined in Table 1c of Jöreskog (1969b). This condition specifies the minimal requirements for recovering the parameters $(\mathbf{\Lambda}, \mathbf{\Phi})$ from the observed product $\mathbf{\Lambda\Phi\Lambda}'$. It is particularly well-suited to sparse loading matrices, where K simple items can be selected as reference variables. Additionally, we fix the latent factors to have unit variance, a standard approach to resolving scale indeterminacy in latent variable models. Formally, the identification condition is defined as follows:

1. **Factor Covariance Constraint:** $\mathbf{\Phi}_{ii} = 1$, for all $1 \leqslant i \leqslant K$.
2. **Loading Matrix Constraint:** Let $\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \end{bmatrix}$, where $\mathbf{\Lambda}_1$ is a diagonal matrix in $\mathbb{R}^{K\times K}$. After removing any row from $\mathbf{\Lambda}$, two disjoint submatrices must remain, each of rank K.

This condition represents the minimal requirements for model identifiability. Due to the presence of rotational indeterminacy, at least K^2 constraints must be imposed on $(\mathbf{\Lambda}, \mathbf{\Phi})$. The K simple items each have $K-1$ zero loadings, contributing a total of $K(K-1)$ constraints. The unit variance constraints on the diagonal entries of Φ provide additional K constraints, ensuring the necessary total of K^2 constraints.

The following theorem establishes that these conditions uniquely determine the CFA model:

Theorem 5 (Identification Condition for CFA Model)**.** *Suppose the above identification conditions hold. Then, the solution for* $(\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Omega})$ *is unique up to a column sign flip of* $\mathbf{\Lambda}$*.*

The proof is provided in the Appendix for Chapter 4 for completeness.

4.3.4 Simple Item Selection and Identification

According to Theorem 5, selecting one item for each latent factor is sufficient for identification. While additional item selections may enhance estimation accuracy, they also increase the risk of incorrect choices. To balance identifiability with selection reliability, we opt to select only one item per factor, thereby satisfying the minimum identifiability condition while minimizing selection errors by avoiding unnecessary reference items.

To systematically identify a simple item for each factor, we establish a selection criterion based on loading magnitudes. Specifically, for a simple item j that loads exclusively on factor k, the second-largest loading magnitude should ideally be zero, with its corresponding estimates being

small. Accordingly, we identify simple items using the rank statistic in (4.7), where L_{jk} represents the largest non-target loading magnitude for item j on factor k.

For each factor k, let $L_{j_{(1)}^k k} \leqslant L_{j_{(2)}^k k} \leqslant \ldots \leqslant L_{j_{(J)}^k k}$ denote the order statistics of $L_{1k}, \ldots, L_{Jk}$. We select the item $id_k = j_{(1)}^k$ as the reference item for the k-th factor and constrain all non-target loadings of the $j_{(1)}^k$-th item to zero in the confirmatory factor analysis (CFA) loading structure, thereby enabling the construction of a p-value or e-value.

4.3.5 Data Splitting

Although we propose a method to select a loading structure under minimal requirements, a common issue arises: performing inference on the same dataset after selection leads to biased results. To address this, we introduce Algorithm 6, which employs data splitting to mitigate this issue. As demonstrated in Study A.IV in Appendix, when the same dataset is used for both loading estimation and model fitting (1DS), the sampling distribution of each loading is non-normal, even asymptotically. However, when data splitting is applied (2DS), the distribution approximately follows a normal pattern.

4.4. FDR control for EFA model

In this section, we briefly review three popular FDR control methods developed in the regression setting: the Benjamini-Hochberg (BH) procedure, the Benjamini-Yekutieli (BY) procedure, and the BH procedure using e-values (eBH procedure). We then outline the steps to adapt these methods for factor- or test-level FDR control in the EFA model.

4.4.1 Introduction to the Benjamini-Hochberg Procedure

One of the challenges in developing methods to control the False Discovery Rate (FDR) is that the number of rejections, $|S|$, is not directly observable. The following ingenious method addresses this issue by determining a rejection threshold that is linked to the number of rejections. The Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) aims to control the FDR at level α at test-level by defining the rejection set:

$$S_{BH} = \left\{(j,k) : p_{jk} \leqslant \frac{\hat{s}q}{JK}\right\}, \text{ where } \hat{s} = \max\left\{s \in [JK] : p_{(s)} \leqslant \frac{qs}{JK}\right\}, \tag{4.8}$$

where q denotes the level at which the FDR is controlled, p_{jk} is the p-value associated with the null hypothesis $H_{jk} : \lambda^*_{jk} = 0$. Let $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(JK)}$ denote the order statistics of the p-values. This approach is equivalent to rejecting the $\hat{s}$-smallest p-values that are below the threshold $\frac{\hat{s}q}{JK}$, thereby rejecting as many loadings as possible while satisfying the constraint. This is done under the assumption:

C1 The p-value for any zero loading is independent of all other loadings, including both zero and non-zero loadings.

This assumption implies that all zero loadings are mutually independent and independent of all non-zero loadings.

Theorem 6 (Benjamini and Hochberg (1995), Theorem 1)**.** *Under condition C1, the Benjamini–Hochberg procedure controls the false discovery rate (FDR) at level q.*

The proof is provided in the Appendix for Chapter 4 for completeness.

Remark 6. *To control the FDR at the factor level for a given factor k, we restrict the procedure to the J p-values associated with factor k by defining the rejection set as:*

$$S^k_{BH} = \left\{ j : p_{jk} \leqslant \frac{\hat{s}q}{J} \right\}, \quad \textit{where} \quad \hat{s} = \max \left\{ s \in [J] : p^k_{(s)} \leqslant \frac{qs}{J} \right\}.$$

Here, $p^k_{(1)} \leqslant p^k_{(2)} \leqslant \cdots \leqslant p^k_{(J)}$ denotes the order statistics of the J p-values associated with factor k.

Remark 7. *In general, Condition C1 is not satisfied by the EFA model. When two items load on the same factor, their loadings are correlated because their estimation depends on the same latent variable. This violation of independence undermines the assumptions required by the Benjamini–Hochberg (BH) procedure. As demonstrated numerically in Study A.IV in the Appendix, Condition C1 does not hold in practice. Therefore, the use of the BH procedure is not theoretically guaranteed to control the FDR in this context.*

We also present a relaxed version of Condition C1 under which the BH procedure may still be valid, although this condition is generally difficult to verify in practice.

Remark 8. *We can relax the independence assumption by requiring only positive regression dependency on a subset $S_0 = \{(j,k) : \lambda^*_{jk} = 0\}$ (PRDS, Benjamini and Yekutieli (2001)). This means that instead of assuming full independence, we assume that within the subset S_0, a form of positive dependency holds.*

C2 ***PRDS on*** S_0***:*** *For any increasing set D, meaning that if $x \in D$, then for any $y \geqslant x$ (coordinatewise), we have $y \in D$. Moreover, for each $i \in S_0$, the conditional probability $\mathbb{P}(\boldsymbol{X} \in D \mid X_i = x)$ is nondecreasing in x.*

Theorem 7 (Benjamini and Yekutieli (2001), Theorem 1.2)**.** *Under condition C2, the Benjamini–Hochberg procedure controls the false discovery rate (FDR) at level q.*

The proof is provided in the Appendix for Chapter 4 for completeness.

Remark 9. *PRDS ensures that within the subset S_0, larger values of a random variable do not decrease the probability of belonging to an increasing set. Intuitively, if a variable X_i increases, the likelihood of the entire vector $\boldsymbol{X}$ belonging to D should also increase or remain the same. This assumption is weaker than full independence but still provides sufficient structure to control false discovery rates effectively. However, this assumption is hard to verify.*

4.4.2 Introduction to the Benjamini-Yekutieli (BY) procedure

The independence assumption (**C1**) in the Benjamini-Hochberg (BH) procedure is not always valid. This is particularly evident in the loading matrix of factor analysis. In the context of test-level FDR control, the estimations of zero and non-zero loadings for the same item are dependent. Similarly, for factor-level FDR control, as discussed in Study A.IV in Appendix in the next section, the asymptotic behavior shows that the estimations of loadings for the same factor across different items are also dependent.

To address the violation of assumption **C1**, we apply the Benjamini-Hochberg procedure with a correction, which is also called BY procedure. To control the FDR at level q, the BH procedure is conducted with an adjusted significance level:

$$q' = \frac{q}{\sum_{m=1}^{JK} \frac{1}{m}} \approx \frac{q}{\log(JK) + 0.577},$$

where 0.577 is the Euler-Mascheroni constant.

Theorem 8 (Benjamini and Yekutieli (2001), Theorem 1.3)**.** *Without any additional conditions, the Benjamini–Hochberg procedure with q' in place of q controls the false discovery rate (FDR) at level q.*

The proof is provided in the Appendix for Chapter 4 for completeness.

Remark 10. *To control the FDR at the factor level for a given factor k, we modify the significance level q in the BH procedure from (6) using*

$$q'_k = \frac{q}{\sum_{m=1}^{J} \frac{1}{m}} \approx \frac{q}{\log(J) + 0.577}, \tag{4.9}$$

since we only have J items to select.

4.4.3 False discovery rate (FDR) control that utilizes e-values

However, the correction factor $\sum_{m=1}^{J} \frac{1}{m} \approx \log(J) + 0.577$ grows large as the number of items increases, making it ineffective as $J \to \infty$. To address FDR control under arbitrary dependence among estimated loadings, we propose using an e-value-based approach as a constraint-free alternative to traditional p-value-based methods.

In this setting, the e-BH procedure serves as the analogue of the Benjamini–Hochberg (BH) method, replacing p-values with e-values. It proceeds as follows: Suppose e_{jk} is the e-value associated with the null hypothesis $H_{jk} : \lambda^*_{jk} = 0$. Let $e_{(1)} \geqslant e_{(2)} \geqslant \ldots \geqslant e_{(JK)}$ denote the order statistics of the e-values. The rejection set is defined as

$$S_{ebh} = \left\{ (j,k) : e_{jk} \geqslant \frac{JK}{q\hat{s}} \right\}, \text{ where } \hat{s} = \max \left\{ s \in [JK] : e_{(s)} \geqslant \frac{JK}{qs} \right\} \tag{4.10}$$

If the latter set is empty, then we set $\hat{s} = 0$ and do not select any hypotheses. In other words, we want to reject $\hat{s}$ hypotheses with the largest e-values, and each of their corresponding e-values should be above the threshold $\frac{JK}{q\hat{s}}$. We reject as many hypotheses as we can as long as this criterion is satisfied, to maximize the power.

Theorem 9 (Wang and Ramdas (2022), Theorem 1)**.** *Without any additional conditions, the eBH procedure controls the false discovery rate (FDR) at level q.*

The proof is provided in the Appendix for Chapter 4 for completeness.

Remark 11. *To control the FDR at the factor level for a given factor k, we modify the eBH procedure in (4.10) as*

$$S_{ebh} = \left\{ j : e_{jk} \geqslant \frac{J}{q\hat{s}} \right\}, \quad \textit{where} \quad \hat{s} = \max \left\{ s \in [J] : e^k_{(s)} \geqslant \frac{J}{qs} \right\}. \tag{4.11}$$

Here, $e^k_{(1)} \geqslant e^k_{(2)} \geqslant \cdots \geqslant e^k_{(J)}$ denote the order statistics of the J item loadings on factor k.

In summary, the properties of these methods are listed in Table 4.1 below.

	Assumptions
BH	Independence or PRDS
BH with Correction	No
eBH	No

Table 4.1: Properties of BH, BH with Correction, and eBH

4.5. Result

4.5.1 Study I: BH, BY, and eBH Procedures

In this section, we compare the traditional Benjamini-Hochberg (BH) procedure with two variations: BY (BH procedure with correction) and eBH (BH procedure utilizing e-values). The e-values are constructed using equation (4.5).

To control the false discovery rate (FDR) at $q = 0.1$, we apply the proposed procedures to a dataset with a loading matrix of dimensions 60×3. Each factor comprises 10 simple items, along with 10 additional items that primarily load onto the factor but also exhibit secondary cross-loadings. The primary loadings and cross-loadings are uniformly sampled from $U[1, 2]$ and $U[0.2, 0.5]$, respectively.

We consider four different sample sizes: $N = 200, 500, 1000$, and 5000. For each sample size and experimental condition, we conduct $B = 1000$ independent replications. In terms of factor-level correction, the adjusted significance level for the BY procedure is determined using $J = 60$, whereas for test-level FDR control, it is calculated using $JK = 180$.

To examine the influence of initial item screening on FDR control, we compare the following two settings:

1. **Oracle setting**: In this scenario, the first three simple items are designated as reference items, ensuring zero selection error. Consequently, the resulting p-values are expected to be symmetrically distributed around zero.

2. **Data-driven setting**: Here, p-values and e-values are computed using Algorithm 6. Additionally, we record selection accuracy during the initial item screening stage to assess its impact on FDR control.

Evaluation Criteria:

We evaluate the proposed method based on three criteria under both test-level and factor-level

FDR control:

1. **Selection Accuracy:** For factor-level FDR control, the selection accuracy for the k-th latent factor is defined as:
$$\mathrm{ACC}_k = \sum_{b=1}^{B} \frac{\mathbf{1}_{\{id_k^{(b)} \text{ is a simple item for factor } k\}}}{B}, \tag{4.12}$$
where $id_k^{(b)}$ denotes the selected reference item for the k-th factor in the b-th replication.

   The overall reference item selection accuracy for test-level FDR control is given by the probability that the algorithm correctly identifies all K simple items, with one item corresponding to each factor, as reference items:
$$\mathrm{ACC} = \sum_{b=1}^{B} \frac{\prod_{k=1}^{K} \mathbf{1}_{\{id_k^{(b)} \text{ is a simple item for factor } k\}}}{B}. \tag{4.13}$$

2. **FDR Control:** The factor-level FDR for the k-th factor is defined as the proportion of incorrectly selected irrelevant items for factor k:
$$\widehat{\mathrm{FDR}}_k = \sum_{b=1}^{B} \frac{|\{j \in S_k^{(b)} : \lambda_{jk}^* = 0\}|}{|S_k^{(b)}| \vee 1}, \tag{4.14}$$
where $S_k^{(b)}$ represents the item selection set for factor k in the b-th replication.

   The test-level FDR is estimated as the proportion of mistakenly selected zero-loadings across the entire loading matrix:
$$\widehat{\mathrm{FDR}} = \sum_{b=1}^{B} \frac{|\{(j,k) \in S^{(b)} : \lambda_{jk}^* = 0\}|}{|S^{(b)}| \vee 1}, \tag{4.15}$$
where $S^{(b)}$ denotes the set of selected loadings in the b-th replication.

3. **Power:** The factor-level power for the k-th factor is computed as the proportion of correctly identified relevant items for factor k:
$$\widehat{\mathrm{Power}}_k = \sum_{b=1}^{B} \frac{|\{j \in S_k^{(b)} : \lambda_{jk}^* \neq 0\}|}{|\{j : \lambda_{jk}^* \neq 0\}| \vee 1}, \tag{4.16}$$
where $S_k^{(b)}$ denotes the item selection set for the k-th factor in the b-th replication.

4. **Power:** The factor-level power for the k-th factor is defined as the proportion of relevant

items correctly selected for that factor:

$$\widehat{\text{Power}}_k = \sum_{b=1}^{B} \frac{|\{j \in S_k^{(b)} : \lambda_{jk}^* \neq 0\}|}{|\{j : \lambda_{jk}^* \neq 0\}| \vee 1}, \tag{4.17}$$

where $S_k^{(b)}$ denotes the set of selected items for the k-th factor in the b-th replication.

Similarly, the test-level power is computed as the proportion of nonzero loadings correctly identified across the entire loading matrix:

$$\widehat{\text{Power}} = \sum_{b=1}^{B} \frac{|\{(j,k) \in S^{(b)} : \lambda_{jk}^* \neq 0\}|}{|\{(j,k) : \lambda_{jk}^* \neq 0\}| \vee 1}, \tag{4.18}$$

where $S^{(b)}$ represents the set of selected loadings in the b-th replication.

	BH		BY		eBH	
N	FDR	power	FDR	power	FDR	power
200	0.047	0.928	0.009	0.862	0.002	0.810
500	0.049	0.991	0.009	0.973	0.002	0.952
1000	0.049	1.000	0.008	0.998	0.002	0.995
5000	0.048	1.000	0.008	1.000	0.002	1.000

Table 4.2: Results of test-level FDR control using BH, BY, and eBH with oracle item screening information.

		BH		BY		eBH	
N	ACC	FDR	power	FDR	power	FDR	power
200	0.917	0.048	0.844	0.009	0.768	0.002	0.722
500	0.999	0.046	0.954	0.007	0.902	0.002	0.858
1000	1.000	0.044	0.992	0.007	0.976	0.002	0.957
5000	1.000	0.046	1.000	0.008	1.000	0.002	1.000

Table 4.3: Results of test-level FDR control using BH, BY, and eBH using selected reference items.

Test-Level FDR Control: From Table 4.2 and Table 4.3, we observe that although the assumptions required for the BH procedure to control FDR are not fully satisfied, it still maintains the FDR at the target level of 0.1. However, we believe this is a coincidence rather than a reliable property of the method.

Comparing the two settings, we find that the controlled FDR does not differ significantly between them. However, the power is lower in the data-driven setting than in the oracle setting.

This result is expected, as the data-driven setting introduces item screening errors, and the p-values are constructed primarily based on only half of the available data points.

Furthermore, we observe that the eBH procedure is more conservative than the BY procedure in controlling FDR. While eBH achieves an extremely low FDR, this comes at the cost of reduced power compared to the BY procedure. To achieve at least 95% power while maintaining effective FDR control, a sample size of $N = 1000$ is required for both the BY and eBH procedures.

		BH		BY		eBH	
k	N	FDR	power	FDR	power	FDR	power
$k = 1$	200	0.052	0.973	0.011	0.938	0.002	0.892
	500	0.053	0.999	0.012	0.997	0.002	0.992
	1000	0.054	1.000	0.012	1.000	0.002	1.000
	5000	0.055	1.000	0.012	1.000	0.002	1.000
$k = 2$	200	0.043	0.916	0.010	0.852	0.002	0.787
	500	0.043	0.991	0.009	0.973	0.002	0.943
	1000	0.046	1.000	0.010	0.999	0.003	0.996
	5000	0.043	1.000	0.008	1.000	0.002	1.000
$k = 3$	200	0.045	0.899	0.010	0.830	0.002	0.766
	500	0.046	0.985	0.010	0.962	0.002	0.929
	1000	0.045	0.999	0.009	0.996	0.002	0.990
	5000	0.043	1.000	0.010	1.000	0.002	1.000

Table 4.4: Comparison of factor-level FDR control using BH, BY, and eBH with oracle item screening information.

			BH		BY		eBH	
	N	ACC	FDR	power	FDR	power	FDR	power
$k = 1$	200	0.974	0.049	0.908	0.011	0.851	0.002	0.801
	500	1	0.050	0.985	0.010	0.962	0.002	0.929
	1000	1	0.047	0.999	0.010	0.996	0.002	0.990
	5000	1	0.050	1.000	0.011	1.000	0.003	1.000
$k = 2$	200	0.973	0.046	0.818	0.011	0.743	0.002	0.687
	500	0.999	0.040	0.944	0.007	0.892	0.001	0.834
	1000	1	0.039	0.990	0.008	0.973	0.002	0.946
	5000	1	0.043	1.000	0.010	1.000	0.002	1.000
$k = 3$	200	0.969	0.045	0.816	0.010	0.745	0.002	0.696
	500	1	0.045	0.936	0.009	0.880	0.001	0.823
	1000	1	0.044	0.989	0.009	0.970	0.002	0.941
	5000	1	0.045	1.000	0.009	1.000	0.002	1.000

Table 4.5: Comparison of factor-level FDR control using BH, BY, and eBH with selected reference items.

Results for Factor-Level FDR Control: From Table 4.4 and Table 4.5, we observe a similar pattern in factor-level FDR control, where the eBH procedure is more conservative than the BY procedure. To achieve at least 95% power, a minimum sample size of $N = 1000$ is required.

4.5.2 Study II: selection stability for BH, BY, and eBH

We evaluate the selection stability across 1000 different data splits. The true loading matrix is the same as in Study I, with dimensions 60×3, where 30 rows contain cross-loadings.

Figures 4.1 and 4.2 present two 4×3 grids of histograms, illustrating the stability of selection proportions across varying sample sizes ($N = 200, 500, 1000, 5000$) and selection methods (BH, BY, eBH), where e-values are constructed using equation (4.5). Each histogram represents the distribution of selection probabilities across $B = 1000$ data splits.

Evaluation Criteria For the (j, k)-th entry in the loading matrix, the selection probability is defined as follows:

For test-level FDR control:

$$p_{jk} = \frac{\sum_{b=1}^{B} \mathbf{1}\{(j,k) \in S^{(b)}\}}{B} \tag{4.19}$$

For factor-level FDR control:

$$p_{jk} = \frac{\sum_{b=1}^{B} \mathbf{1}\{j \in S_k^{(b)}\}}{B} \tag{4.20}$$

Shannon Entropy We measure the uncertainty of selection for each loading using average Shannon entropy. The maximum entropy value is 1, occurring when the selection probability is 0.5, while the minimum entropy is 0, achieved when the selection probability is either 0 or 1. The entropy is computed as:

$$\text{Shannon Entropy} = \frac{\sum_j \sum_k \left[-p_{jk} \log_2 p_{jk} - (1 - p_{jk}) \log_2 (1 - p_{jk})\right]}{J \cdot K}. \tag{4.21}$$

Results. As the sample size increases, the histograms exhibit more concentrated distributions at 0 and 1, indicating greater selection stability. For extremely large samples ($N = 5000$), the BY and eBH procedures achieve stable selection. It is not surprising that the BH method is less stable, as it selects more false positives, making these selections more susceptible to noise.

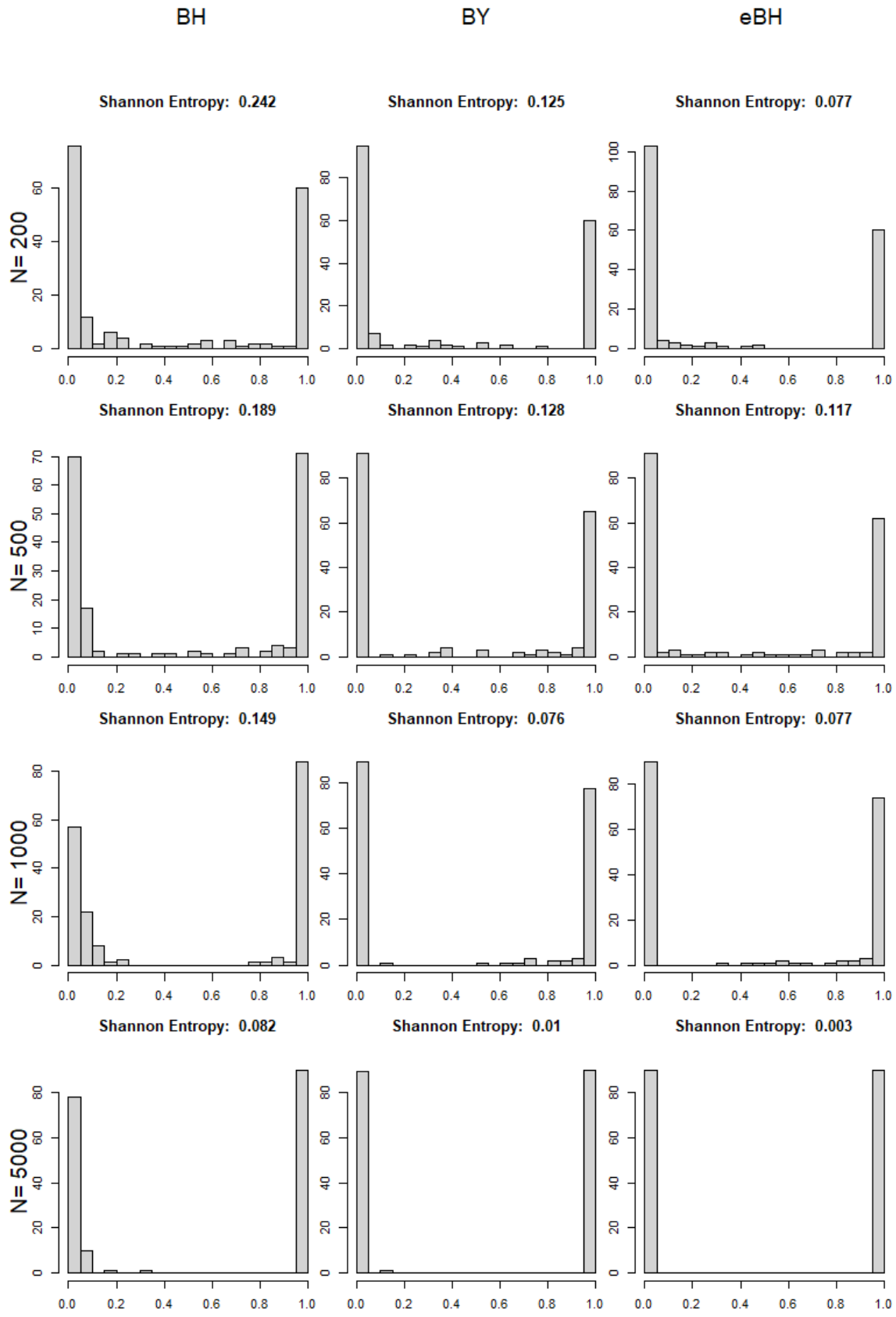


Figure 4.1: Histograms of selection probabilities for different test-level FDR control methods (BH, BY, eBH) across varying sample sizes ($N = 200, 500, 1000, 5000$). Each subplot shows the distribution of selection proportions, highlighting the proportion of stable selections (values close to 0 or 1). Stability increases with larger sample sizes, particularly for the BY and eBH methods.

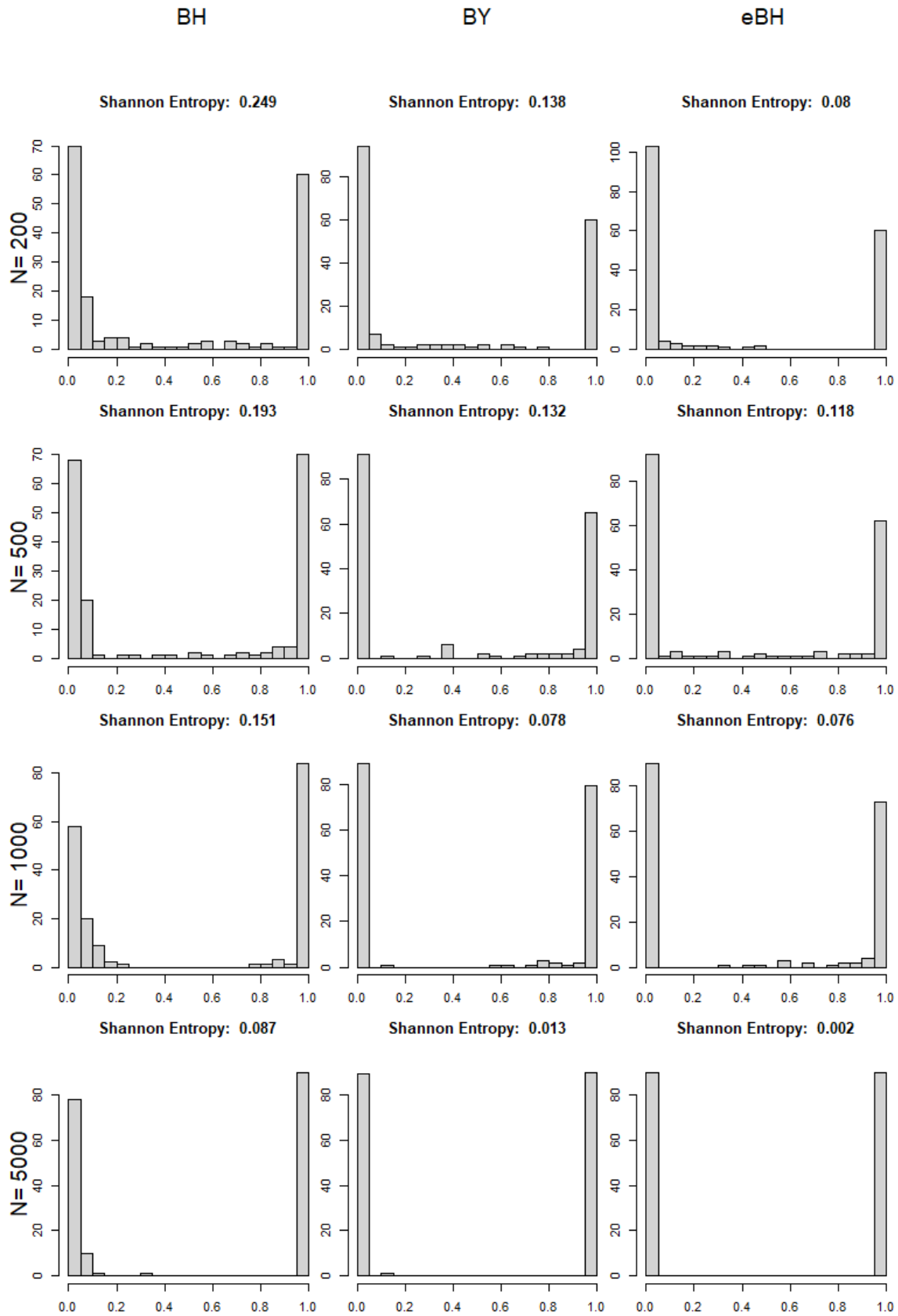


Figure 4.2: Histograms of selection probabilities for different factor-level FDR control methods (BH, BY, eBH) across varying sample sizes ($N = 200, 500, 1000, 5000$). Each subplot shows the distribution of selection proportions, highlighting the proportion of stable selections (values close to 0 or 1). Stability increases with larger sample sizes, particularly for the BY and eBH methods.

4.6. An Application to the Big Five Personality Test

We use the same dataset as in section 2.6, applying our item selection method to the Big Five Factor Markers from the International Personality Item Pool. This dataset consists of 50 items, with 10 per personality factor: Extraversion (E), Emotional Stability (ES), Agreeableness (A), Conscientiousness (C), and Intellect/Imagination (I). Each item is rated on a five-point Likert scale, treated as continuous. To reduce measurement noninvariance, we focus on British males ($N = 609$) and set the number of factors to $K = 5$.

Although the Big Five Personality Test is designed to measure a single personality trait per item, empirical data often report cross-loadings due to ambiguous item wording. To better understand the loading structure, we apply our proposed item selection method to control both factor-level and test-level false discovery rates (FDR) at $q = 0.1$. Specifically, for factor-level FDR control at $q = 0.1$, with 10 simple items intended to load primarily on each factor, if no additional items exhibit cross-loadings, we expect around 1 of the selected items per factor to be insignificant. This results in approximately 5 selected insignificant loadings for the overall matrix.

Results Table 4.6 provides an overview of the item selection results for Study II. Among the 50 loadings suggested by the answer key, only the eBH procedure at the factor-level FDR control excluded one item. Specifically, for the factor Openness to Experience, the item 'O9+ I spend time reflecting on things' was left out.

Number of Selection	BH	BY	eBH
Text-level FDR control	99	77	66
Factor-level FDR control	100	82	62(-1)

Table 4.6: Overview of Results for Study IV. If any of the 50 loadings in the answer key, which were designed by the researcher, are omitted, they are included in brackets with a negative number.

Comparing the selection tendencies of different procedures, we observe that the BH procedure selects the largest number of items, followed by the BY procedure, while the eBH procedure selects the fewest. Since we control the FDR at $q = 0.1$, we expect approximately 8 falsely selected loadings under the BY procedure and around 7 under the eBH procedure. If all 50 items are simple items, as specified in the answer key, the maximum number of loadings we should select is 58. However, all methods exceed this number, indicating the presence of meaningful cross-loadings in the Big Five questionnaire.

Tables 4.7 through 4.12 list all items selected by each FDR control method, in addition to those

listed in the answer key. For items that appear in the answer key but are not selected, we place them in brackets to indicate missed selections.

In Table 4.7, we observe that many items selected by the BH procedure do not intuitively reflect meaningful relationships. For instance, the BH procedure selects two Conscientiousness items under Extraversion: 'C2 – I leave my belongings around.' and 'C7+ – I like order.'

Furthermore, the BH procedure selects a total of 99 and 100 items under test-level and factor-level FDR control, respectively. On average, each item has two selected loadings, suggesting that the underlying assumptions of the procedure may not hold in real-world settings.

Although the BY procedure selects fewer items than the BH procedure, some selected items appear conceptually irrelevant. For example, in Table 4.8, the item 'O4 – I am not interested in abstract ideas' is selected under Agreeableness, and in Table 4.11, the item 'A7 – I am not really interested in others' is selected under Conscientiousness.

In contrast, the eBH procedure, shown in Tables 4.9 and 4.12, is highly conservative and selects the fewest items, yet it achieves the highest selection quality with minimal false positives. All selected items align well with theoretical expectations, making eBH the most precise procedure among the three.

	E	N	A	C	O
1	N2+I am relaxed most of the time.	E2-I don't talk a lot.	E3+I feel comfortable around people.	E8-I don't like to draw attention to myself.	E6-I have little to say.
2	N3-I worry about things.	E3+I feel comfortable around people.	E5+I start conversations.	N5-I am easily disturbed.	N5-I am easily disturbed.
3	N4+I seldom feel blue.	A1-I feel little concern for others.	E8-I don't like to draw attention to myself.	O6-I do not have a good imagination.	N10-I often feel blue.
4	N10-I often feel blue.	A3-I insult people.	N1-I get stressed out easily.		A3-I insult people.

Table 4.7: Result for test-level FDR control by BH

	E	N	A	C	O
5	A2+I am interested in people.	C4-I make a mess of things.	N3-I worry about things.		C1+I am always prepared.
6	A3-I insult people.	C8-I shirk my duties.	N6-I get upset easily.		C3+I pay attention to details.
7	A5-I am not interested in other people's problems.	O2-I have difficulty understanding abstract ideas.	C2-I leave my belongings around.		C7+I like order.
8	A7-I am not really interested in others.	O4-I am not interested in abstract ideas.	C3+I pay attention to details.		C10+I am exacting in my work.
9	A10+I make people feel at ease.	O7+I am quick to understand things.	C7+I like order.		
10	C2-I leave my belongings around.	O9+I spend time reflecting on things.	C8-I shirk my duties.		
11	C7+I like order.		C9+I follow a schedule.		
12	O6-I do not have a good imagination.		C10+I am exacting in my work.		
13			O3+I have a vivid imagination.		
14			O4-I am not interested in abstract ideas.		

Table 4.7: Result for test-level FDR control by BH

	E	N	A	C	O
15			O9+I spend time reflecting on things.		
16			O10+I am full of ideas.		

Table 4.7: (continued) Result for test-level FDR control by BH

	E	N	A	C	O
1	N2+I am relaxed most of the time.	E2-I don't talk a lot.	N3-I worry about things.		C1+I am always prepared.
2	N4+I seldom feel blue.	A1-I feel little concern for others.	N6-I get upset easily.		C3+I pay attention to details.
3	N10-I often feel blue.	A3-I insult people.	C8-I shirk my duties.		C10+I am exacting in my work.
4	A2+I am interested in people.	C4-I make a mess of things.	C9+I follow a schedule.		
5	A7-I am not really interested in others.	C8-I shirk my duties.	O4-I am not interested in abstract ideas.		
6	A10+I make people feel at ease.	O2-I have difficulty understanding abstract ideas.	O9+I spend time reflecting on things.		
7	C2-I leave my belongings around.	O4-I am not interested in abstract ideas.			
8	C7+I like order.	O7+I am quick to understand things.			
9	O6-I do not have a good imagination.	O9+I spend time reflecting on things.			

Table 4.8: Result for test-level FDR control by BY

	E	N	A	C	O
1	N10-I often feel blue.	E2-I don't talk a lot.	N3-I worry about things.		C3+I pay attention to details.
2	A2+I am interested in people.	A3-I insult people.	N6-I get upset easily.		C10+I am exacting in my work.
3	A7-I am not really interested in others.	C4-I make a mess of things.	O9+I spend time reflecting on things.		
4	A10+I make people feel at ease.	C8-I shirk my duties.			
5	C7+I like order.	O2-I have difficulty understanding abstract ideas.			
6		O4-I am not interested in abstract ideas.			

Table 4.9: Result for test-level FDR control by eBH

	E	N	A	C	O
1	N1-I get stressed out easily.	E2-I don't talk a lot.	E4-I keep in the background.	E8-I don't like to draw attention to myself.	E3+I feel comfortable around people.
2	N2+I am relaxed most of the time.	E3+I feel comfortable around people.	N2+I am relaxed most of the time.	A7-I am not really interested in others.	E9+I don't mind being the center of attention.
3	N4+I seldom feel blue.	A3-I insult people.	N3-I worry about things.	O5+I have excellent ideas.	N5-I am easily disturbed.
4	N6-I get upset easily.	A7-I am not really interested in others.	N6-I get upset easily.		N6-I get upset easily.
5	N10-I often feel blue.	C4-I make a mess of things.	N10-I often feel blue.		N9-I get irritated easily.

Table 4.10: Result for factor-level FDR control by BH

	E	N	A	C	O
6	A2+I am interested in people.	C7+I like order.	C7+I like order.		A3-I insult people.
7	A3-I insult people.	C8-I shirk my duties.	C8-I shirk my duties.		A5-I am not interested in other people's problems.
8	A5-I am not interested in other people's problems.	O2-I have difficulty understanding abstract ideas.	C10+I am exacting in my work.		C1+I am always prepared.
9	A7-I am not really interested in others.	O3+I have a vivid imagination.	O4-I am not interested in abstract ideas.		C3+I pay attention to details.
10	A9+I feel others' emotions.	O7+I am quick to understand things.	O9+I spend time reflecting on things.		C10+I am exacting in my work.
11	A10+I make people feel at ease.	O9+I spend time reflecting on things.			
12	C2-I leave my belongings around.				
13	C7+I like order.				
14	O5+I have excellent ideas.				
15	O6-I do not have a good imagination.				

Table 4.10: Result for factor-level FDR control by BH

	E	N	A	C	O
16	O10+I am full of ideas.				

Table 4.10: (continued) Result for factor-level FDR control by BH

	E	N	A	C	O
1	N10-I often feel blue.	E2-I don't talk a lot.	N3-I worry about things.	A7-I am not really interested in others.	N5-I am easily disturbed.
2	A2+I am interested in people.	E3+I feel comfortable around people.	N6-I get upset easily.	O5+I have excellent ideas.	A5-I am not interested in other people's problems.
3	A5-I am not interested in other people's problems.	A3-I insult people.	N10-I often feel blue.		C3+I pay attention to details.
4	A7-I am not really interested in others.	C4-I make a mess of things.	C7+I like order.		C10+I am exacting in my work.
5	A9+I feel others' emotions.	C8-I shirk my duties.	C8-I shirk my duties.		
6	A10+I make people feel at ease.	O2-I have difficulty understanding abstract ideas.	C10+I am exacting in my work.		
7	C2-I leave my belongings around.	O7+I am quick to understand things.	O9+I spend time reflecting on things.		
8	C7+I like order.	O9+I spend time reflecting on things.			
9	O5+I have excellent ideas.				
10	O6-I do not have a good imagination.				
11	O10+I am full of ideas.				

Table 4.11: Result for factor-level FDR control by BY

	E	N	A	C	O
1	N10-I often feel blue.	A3-I insult people.	N3-I worry about things.		C3+I pay attention to details.
2	A2+I am interested in people.	C4-I make a mess of things.	N6-I get upset easily.		(O9+I spend time reflecting on things.)
3	A5-I am not interested in other people's problems.	O9+I spend time reflecting on things.	C7+I like order.		
4	A7-I am not really interested in others.		O9+I spend time reflecting on things.		
5	A10+I make people feel at ease.				

Table 4.12: Result for factor-level FDR control by eBH

4.7. Concluding Remarks

In this chapter, we developed a method for computing p-values in EFA models by utilizing the L^p rotation criterion to achieve model identification under minimal requirements. Based on these p-values, we proposed a data-driven approach for controlling the False Discovery Rate (FDR), offering a statistically rigorous framework for selecting significant loadings at both the test and factor levels. These contributions address key challenges in uncertainty quantification and item selection, thereby bridging an important gap in the existing literature.

Among the three commonly used FDR control procedures, we find that only the BH with correction and the eBH procedure can be applied to EFA models with theoretical guarantees. In contrast, the standard BH procedure, which assumes independence among test statistics, is violated in the EFA setting, as demonstrated by our numerical simulations. Additionally, the positive regression dependency on each one from a subset (PRDS) assumption is difficult to verify, further limiting its applicability.

One limitation of our current work is that we only consider a finite number of items and do not extend our analysis to the high-dimensional factor analysis setting. A deeper investigation into the covariance structure of the loading matrix in high-dimensional scenarios, as explored in Bai and Li (2012), could provide a foundation for extending our theoretical framework.

One potential future direction is to explore methods for stabilizing the results. As demonstrated

in Study II, a large sample size is necessary to achieve consistent selection results. For smaller sample sizes, we could employ techniques such as multiple data splitting and aggregating the results. For example, we could apply the inclusion rate (Dai et al., 2023) or aggregate e-values simply by averaging them, as suggested in Ren and Barber (2024).

Finally, as with any simulation study, our analysis is inherently limited by the number of scenarios we can examine. While our results provide strong empirical support for the proposed methods, future research should explore a broader range of simulation settings to further validate the robustness of our approach.

Appendix for Chapter 4

A4.1. Proof of Theorem 5

Proof. Suppose there exist two solutions, $(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$ and $(\tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\Phi}}, \tilde{\boldsymbol{\Omega}})$, that satisfy the identification conditions, with the constraint that $\text{sign}(\boldsymbol{\Lambda}_1) = \text{sign}(\tilde{\boldsymbol{\Lambda}}_1)$. Given that

$$\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Omega} = \tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Lambda}}' + \tilde{\boldsymbol{\Omega}} \stackrel{\text{def}}{=} \boldsymbol{\Sigma},$$

we aim to prove that $(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega}) = (\tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\Phi}}, \tilde{\boldsymbol{\Omega}})$.

1. **Equality of $\boldsymbol{\Omega}$:** Following the proof of Theorem 5.1 in Anderson and Rubin (1956), we extend the result from the special case where $\boldsymbol{\Phi} = \boldsymbol{I}$ to the general condition. This generalizes the rotation from the orthogonal case to the oblique case. Since the diagonal entries of $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$ and $\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Lambda}}'$ correspond to the off-diagonal entries of $\boldsymbol{\Sigma}$, it suffices to show that

$$\text{diag}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}') = \text{diag}(\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Lambda}}').$$

   By the identification condition on $\boldsymbol{\Lambda}$, we have $J \geqslant 2K + 1$. Let

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{\lambda}_{K+1} \\ \boldsymbol{S}_2 \\ \boldsymbol{S}_3 \end{bmatrix},$$

   where $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are nonsingular matrices, and $\boldsymbol{\lambda}_{K+1}$ is the $(K+1)$-th row. We can rearrange the rows of $\boldsymbol{\Lambda}$ such that the first K rows and the $(K+2)$-th to $(2K+1)$-th rows are nonsingular. The corresponding partitioning holds for $\tilde{\boldsymbol{\Lambda}}$. Expanding $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$:

$$\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' = \begin{bmatrix} \boldsymbol{S}_1\boldsymbol{\Phi}\boldsymbol{S}_1' & \boldsymbol{S}_1\boldsymbol{\Phi}\boldsymbol{\lambda}_{K+1}' & \boldsymbol{S}_1\boldsymbol{\Phi}\boldsymbol{S}_2' & \boldsymbol{S}_1\boldsymbol{\Phi}\boldsymbol{S}_3' \\ \boldsymbol{\lambda}_{K+1}\boldsymbol{\Phi}\boldsymbol{S}_1' & \boldsymbol{\lambda}_{K+1}\boldsymbol{\Phi}\boldsymbol{\lambda}_{K+1}' & \boldsymbol{\lambda}_{K+1}\boldsymbol{\Phi}\boldsymbol{S}_2' & \boldsymbol{\lambda}_{K+1}\boldsymbol{\Phi}\boldsymbol{S}_3' \\ \boldsymbol{S}_2\boldsymbol{\Phi}\boldsymbol{S}_1' & \boldsymbol{S}_2\boldsymbol{\Phi}\boldsymbol{\lambda}_{K+1}' & \boldsymbol{S}_2\boldsymbol{\Phi}\boldsymbol{S}_2' & \boldsymbol{S}_2\boldsymbol{\Phi}\boldsymbol{S}_3' \\ \boldsymbol{S}_3\boldsymbol{\Phi}\boldsymbol{S}_1' & \boldsymbol{S}_3\boldsymbol{\Phi}\boldsymbol{\lambda}_{K+1}' & \boldsymbol{S}_3\boldsymbol{\Phi}\boldsymbol{S}_2' & \boldsymbol{S}_3\boldsymbol{\Phi}\boldsymbol{S}_3' \end{bmatrix}.$$

Define:

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{S}_1 \boldsymbol{\Phi} \boldsymbol{\lambda}'_{K+1} = \tilde{\boldsymbol{S}}_1 \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\lambda}}'_{K+1}, \\
\boldsymbol{B} &= \boldsymbol{S}_1 \boldsymbol{\Phi} \boldsymbol{S}'_2 = \tilde{\boldsymbol{S}}_1 \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{S}}'_2, \\
\boldsymbol{C} &= \boldsymbol{\lambda}_{K+1} \boldsymbol{\Phi} \boldsymbol{S}'_2 = \tilde{\boldsymbol{\lambda}}_{K+1} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{S}}'_2.
\end{aligned}$$

Since $\boldsymbol{B}$ is nonsingular, the determinant of a $(K+1) \times (K+1)$ submatrix must be zero:

$$0 = \begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{\lambda}_{K+1} \boldsymbol{\Phi} \boldsymbol{\lambda}'_{K+1} & \boldsymbol{C} \end{vmatrix} = (-1)^{(m+1)+1} \boldsymbol{\lambda}_{K+1} \boldsymbol{\Phi} \boldsymbol{\lambda}'_{K+1} |\boldsymbol{B}| + f(\boldsymbol{A}, \boldsymbol{C}).$$

Similarly,

$$0 = \begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \tilde{\boldsymbol{\lambda}}_{K+1} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\lambda}}'_{K+1} & \boldsymbol{C} \end{vmatrix} = (-1)^{(m+1)+1} \tilde{\boldsymbol{\lambda}}_{K+1} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\lambda}}'_{K+1} |\boldsymbol{B}| + f(\boldsymbol{A}, \boldsymbol{C}).$$

Since $\det(\boldsymbol{B}) \neq 0$, it follows that

$$\boldsymbol{\lambda}_{K+1} \boldsymbol{\Phi} \boldsymbol{\lambda}'_{K+1} = \tilde{\boldsymbol{\lambda}}_{K+1} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\lambda}}'_{K+1}.$$

Applying this argument to all diagonal entries yields

$$\operatorname{diag}(\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}') = \operatorname{diag}(\tilde{\boldsymbol{\Lambda}} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Lambda}}').$$

2. **Uniqueness of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$:** Given $\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}' = \tilde{\boldsymbol{\Lambda}} \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Lambda}}'$, consider the block structure:

$$\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}' = \begin{bmatrix} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}'_1 & \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}'_2 \\ \boldsymbol{\Lambda}_2 \boldsymbol{\Phi} \boldsymbol{\Lambda}'_1 & \boldsymbol{\Lambda}_2 \boldsymbol{\Phi} \boldsymbol{\Lambda}'_2 \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\Lambda}}_1 \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Lambda}}'_1 & \tilde{\boldsymbol{\Lambda}}_1 \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Lambda}}'_2 \\ \tilde{\boldsymbol{\Lambda}}_2 \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Lambda}}'_1 & \tilde{\boldsymbol{\Lambda}}_2 \tilde{\boldsymbol{\Phi}} \tilde{\boldsymbol{\Lambda}}'_2 \end{bmatrix}.$$

From the diagonal elements, we obtain

$$\lambda^2_{1,ii} \Phi_{ii} = \tilde{\lambda}^2_{1,ii} \tilde{\Phi}_{ii}, \quad i = 1, \ldots, K.$$

By the restrictions on $\boldsymbol{\Phi}$, we have $\lambda^2_{1,ii} = \tilde{\lambda}^2_{1,ii}$ for $i = 1, ..., K$. Considering $\operatorname{sign}(\boldsymbol{\Lambda}_1) = \operatorname{sign}(\tilde{\boldsymbol{\Lambda}}_1)$, we conclude that $\boldsymbol{\Lambda}_1 = \tilde{\boldsymbol{\Lambda}}_1$.

Since both $\boldsymbol{\Lambda}$ and $\tilde{\boldsymbol{\Lambda}}$ span the same column space, there exists a unique full-rank transformation matrix $\boldsymbol{T} \in \mathbb{R}^{K \times K}$ such that $\boldsymbol{\Lambda} = \tilde{\boldsymbol{\Lambda}} \boldsymbol{T}$. From $\boldsymbol{\Lambda}_1 = \tilde{\boldsymbol{\Lambda}}_1 \boldsymbol{T}$, we deduce that $\boldsymbol{T} = \boldsymbol{I}$,

implying $\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}$.

Finally, using $\mathbf{\Lambda}_1 \mathbf{\Phi} \mathbf{\Lambda}_1' = \tilde{\mathbf{\Lambda}}_1 \tilde{\mathbf{\Phi}} \tilde{\mathbf{\Lambda}}_1'$ and $\mathbf{\Lambda}_1 = \tilde{\mathbf{\Lambda}}_1$, we obtain:

$$\mathbf{\Phi} = \mathbf{\Lambda}_1^{-1} \tilde{\mathbf{\Lambda}}_1 \tilde{\mathbf{\Phi}} \tilde{\mathbf{\Lambda}}_1' \mathbf{\Lambda}_1^{-\top} = \tilde{\mathbf{\Phi}}.$$

□

A4.2. Proof of Theorem 6

The FDR is defined as:

$$FDR = \mathbb{E}\left(\frac{|\{(j,k) \in S : \lambda_{jk}^* = 0\}|}{|S| \vee 1} \right).$$

Expanding this expression, we have:

$$FDR = \sum_{s=1}^{JK} \mathbb{E}\left(\frac{|\{(j,k) \in S : \lambda_{jk}^* = 0\}|}{s} \mathbf{1}_{\{|S|=s\}} \right) = \sum_{s=1}^{JK} \frac{1}{s} \mathbb{E}\left(\sum_{\lambda_{jk}^*=0} \mathbf{1}_{\{p_{jk} \leqslant \frac{sq}{JK}\}} \mathbf{1}_{\{|S|=s\}} \right).$$

To further analyze the joint event $\mathbf{1}_{\{p_{jk} \leqslant \frac{sq}{JK}, |S|=s\}}$, consider its meaning within the BH procedure. It corresponds to rejecting the loading (j,k) as zero while rejecting s loadings in total. This event can be decomposed into two sub-events:

E1 The zero loading λ_{jk}^* is associated with a p-value $p_{jk} \leqslant \frac{sq}{JK}$.

E2 Among the remaining $JK - 1$ loadings, $s - 1$ have p-values below or equal to the threshold $\frac{sq}{JK}$, while $JK - s$ exceed the threshold. We denote this event as $\{|S^{-jk}| = s - 1\}$.

By assumption **C1**, events **E1** and **E2** are independent. Therefore, we can rewrite the probability as:

$$\mathbb{P}(p_{jk} \leqslant \frac{sq}{JK}, |S| = s) = \mathbb{P}(p_{jk} \leqslant \frac{sq}{JK}) \cdot \mathbb{P}(|S^{-jk}| = s - 1).$$

Substituting this back, the FDR becomes:

$$FDR = \sum_{s=1}^{JK} \frac{1}{s} \sum_{\lambda_{jk}^*=0} \mathbb{P}(p_{jk} \leqslant \frac{sq}{JK}) \cdot \mathbb{P}(|S^{-jk}| = s - 1).$$

Using the uniform distribution property of p-values for true null hypotheses:

$$\frac{1}{s} \mathbb{P}(p_{jk} \leqslant \frac{sq}{JK}) = \frac{q}{JK}.$$

Thus,

$$FDR = \sum_{s=1}^{JK} \sum_{\lambda_{jk}^*=0} \frac{q}{JK} \cdot \mathbb{P}(|S^{-jk}| = s-1).$$

Simplifying further:

$$FDR = \frac{q}{JK} \sum_{\lambda_{jk}^*=0} \left(\sum_{s=1}^{JK} \mathbb{P}(|S^{-jk}| = s-1) \right).$$

Since the probabilities sum to 1:

$$\sum_{s=1}^{JK} \mathbb{P}(|S^{-jk}| = s-1) = 1.$$

Finally,

$$FDR = \frac{q}{JK} \sum_{\lambda_{jk}^*=0} 1 \leqslant q.$$

This completes the proof that the BH procedure controls the FDR at level q.

A4.3. Proof of Theorem 7

Proof. Rearrange (A4.2):

$$\begin{aligned}
\text{FDR} &= \sum_{s=1}^{JK} \frac{1}{s} \mathbb{E} \left(\sum_{\lambda_{jk}^*=0} \mathbf{1}_{\{p_{jk} \leqslant \frac{sq}{JK}\}} \mathbf{1}_{\{|S|=s\}} \right) \\
&= \sum_{s=1}^{JK} \sum_{\lambda_{jk}^*=0} \frac{1}{s} \mathbb{P}\left(p_{jk} \leqslant \frac{sq}{JK}, |S| = s \right) \\
&= \sum_{s=1}^{JK} \sum_{\lambda_{jk}^*=0} \frac{1}{s} \mathbb{P}(p_{jk} \leqslant \frac{sq}{JK}) \mathbb{P}\left(|S| = s \mid p_{jk} \leqslant \frac{sq}{JK} \right) \\
&\leqslant \sum_{\lambda_{jk}^*=0} \frac{q}{JK} \sum_{s=1}^{JK} \mathbb{P}\left(|S^{-jk}| = s-1 \mid p_{jk} \leqslant \frac{sq}{JK} \right).
\end{aligned}$$

If we can guarantee that

$$\sum_{s=1}^{JK} \mathbb{P}\left(|S^{-jk}| = s-1 \mid p_{jk} \leqslant \frac{sq}{JK} \right) \leqslant 1, \tag{4.22}$$

then we obtain

$$\text{FDR} \leqslant \frac{\#\{\lambda_{jk}^* = 0\}}{JK} q \leqslant q.$$

Now we prove that Condition C2 implies (4.22). We proceed by induction. First, when $s = 1$, we have

$$\sum_{s=1} \mathbb{P}\left(|S^{-jk}| = s-1 \mid p_{jk} \leqslant \frac{sq}{JK}\right) \leqslant \mathbb{P}\left(|S^{-jk}| \leqslant s \mid p_{jk} \leqslant \frac{q}{JK}\right).$$

Define:

$$\begin{aligned}
&\mathbb{P}\left(|S^{-jk}| \leqslant s-1 \mid p_{jk} \leqslant \frac{sq}{JK}\right) + \mathbb{P}\left(|S^{-jk}| = s \mid p_{jk} \leqslant \frac{(s+1)q}{JK}\right) \\
\overset{C2}{\leqslant} &\mathbb{P}\left(|S^{-jk}| \leqslant s-1 \mid p_{jk} \leqslant \frac{(s+1)q}{JK}\right) + \mathbb{P}\left(|S^{-jk}| = s \mid p_{jk} \leqslant \frac{(s+1)q}{JK}\right) \\
\leqslant &\mathbb{P}\left(|S^{-jk}| \leqslant s \mid p_{jk} \leqslant \frac{(s+1)q}{JK}\right).
\end{aligned}$$

where the condition C2 satisfied for p-values due to the event

$$\{|S^{-jk}| \leqslant s-1\} = \{p_{(JK-1)}^{-jk} \geqslant q, p_{(JK-2)}^{-jk} \geqslant \frac{(JK-1)q}{JK}, ...p_{(s)}^{-jk} \geqslant \frac{(s+1)q}{JK}\}, \tag{4.23}$$

where $p_{(JK-1)}^{-jk} \geqslant p_{(JK-2)}^{-jk} \geqslant ... \geqslant p_{(s)}^{-jk}$ is the $JK - s$ largest vales among all JK p-values except for p_{jk} Since the set $[q, \infty) \times [\frac{(JK-1)q}{JK}, \infty) \times ... \times \times [\frac{(s+1)q}{JK}, \infty)$ is a increasing set, by C2, we have

$$P(|S^{-jk}| \leqslant s-1 | p_{jk} = q_1) \leqslant P(|S^{-jk}| \leqslant s-1 | p_{jk} = q_2), \text{ if } q_1 \leqslant q_2 \tag{4.24}$$

By Lehmann (1966), we have

$$\mathbb{P}(|S^{-jk}| \leqslant s-1 | p_{jk} \leqslant \frac{sq}{JK}) \leqslant \mathbb{P}(|S^{-jk}| \leqslant s-1 | p_{jk} \leqslant \frac{(s+1)q}{JK}) \tag{4.25}$$

□

A4.4. Proof of Theorem 8

Proof. Without assumption **C1**, events **E1** and **E2** are no longer independent. Therefore, the original FDR expression in (A4.2) must be modified as follows:

$$FDR = \sum_{s=1}^{JK} \frac{1}{s} \sum_{\lambda_{jk}^*=0} \mathbb{P}\left(p_{jk} \leqslant \frac{sq'}{JK}, |S^{-jk}| = s-1\right)$$

$$
\begin{aligned}
&= \sum_{s=1}^{JK} \frac{1}{s} \sum_{\lambda_{jk}^*=0} \sum_{m=1}^{s} \mathbb{P}\left(\frac{(m-1)q'}{JK} \leqslant p_{jk} \leqslant \frac{mq'}{JK}, |S^{-jk}| = s-1 \right) \\
&= \sum_{\lambda_{jk}^*=0} \sum_{m=1}^{JK} \sum_{s=m}^{JK} \frac{1}{s} \mathbb{P}\left(\frac{(m-1)q'}{JK} \leqslant p_{jk} \leqslant \frac{mq'}{JK}, |S^{-jk}| = s-1 \right) \\
&\leqslant \sum_{\lambda_{jk}^*=0} \sum_{m=1}^{JK} \sum_{s=1}^{JK} \frac{1}{m} \mathbb{P}\left(\frac{(m-1)q'}{JK} \leqslant p_{jk} \leqslant \frac{mq'}{JK}, |S^{-jk}| = s-1 \right) \\
&\leqslant \sum_{\lambda_{jk}^*=0} \sum_{m=1}^{JK} \frac{1}{m} \mathbb{P}\left(\frac{(m-1)q'}{JK} \leqslant p_{jk} \leqslant \frac{mq'}{JK} \right) \\
&= \frac{|\{\lambda_{jk}^* = 0\}|}{JK} \cdot q' \sum_{m=1}^{JK} \frac{1}{m} \\
&\approx \frac{|\{\lambda_{jk}^* = 0\}|}{JK} \cdot q.
\end{aligned}
$$

□

A4.5. Proof of Theorem 9

Proof. We notice that the false discovery proportion of the e-BH procedure satisfies:

$$
FDP \overset{def}{=} \frac{|\{(j,k) \in S, \lambda_{jk}^* = 0\}|}{|S| \vee 1} = \sum_{(j,k):\lambda_{jk}^*=0} \frac{\mathbf{1}\{e_{jk} \geqslant \frac{JK}{q\hat{s}}\}}{\hat{s} \vee 1} \leqslant \sum_{(j,k):\lambda_{jk}^*=0} \frac{e_{jk} q}{JK} \tag{4.26}
$$

where the inequality holds because when $\mathbf{1}\{e_{jk} \geqslant \frac{JK}{q\hat{s}}\}$ is true, we have $\frac{q\hat{s}e_{jk}}{JK} \geqslant 1$. Therefore, when C3 is satisfied, we have:

$$
FDR = \mathbb{E}FDP \leqslant (\frac{\sum_{(j,k):\lambda_{jk}^*=0} \mathbb{E}e_{jk}}{JK})q \leqslant q
$$

□

A4.6. Study A.I: The Sampling Distribution of Estimation Errors in L^1 and $L^{0.5}$ Rotation

In this study, we examine the sampling distribution of the L^1 and $L^{0.5}$ rotated loading matrices. Our primary objective is to demonstrate that, due to the non-smooth nature of the rotation criteria, the central limit theorem (CLT) may not always hold for L^p rotations.

Settings To investigate this, we conducted 10,000 simulations under the settings described in Table 4.13. For each setting, we consider five sample sizes: $N = 100, 200, 500, 1000$, and 5000. In each simulation, we generate a new dataset and compute the initial orthogonal solution using confirmatory factor analysis (CFA), where the loading matrix is constrained to an upper triangular structure with fixed zeros, and the factor covariance matrix is set to the identity matrix. We then apply either the L^1 or $L^{0.5}$ rotation criterion to transform the initial loading matrix. The estimation error is then analyzed by comparing the estimated loadings with their true values. If the CLT holds, the sampling distribution of these errors should be symmetric around zero.

	Loading Matrix Item 1-12			Item Unique Variances Item 1-12	Factor Covariance		
	F1	F2	F3		F1	F2	F3
1	1	0	0	1	1		
2	0	1	0	1	0.021	1	
3	0	0	1	1	0.502	0.274	1
4	1	0	0	1			
5	0	1	0	1			
6	0	0	1	1			
7	1	0	0	1			
8	0	1	0	1			
9	0	0	1	1			
10	1	0.2	0.3	1			
11	0.3	1	0.2	1			
12	0.2	0.3	1	1			

Table 4.13: Parameters in Simulation Study A.I.

Results Figures 4.3, 4.4, 4.5, and 4.6 present the results of the simulations. The sample size is indicated at the top of each column, while the row labels on the left correspond to the location of the loadings and their respective values. As expected, the estimation error decreases as sample size increases, consistent with Theorem 1 in Chapter 2. However, for many zero loadings, such as $L[1, 2]$ and $L[2, 1]$ in Figure 4.3 and $L[1, 2]$ and $L[3, 1]$ in Figure 4.5, the sampling distribution deviates significantly from normality—it is non-symmetric and highly concentrated around zero.

Notably, the L^1 rotation exhibits a more pronounced asymmetry, whereas the $L^{0.5}$ rotation yields a distribution that is heavily centered at zero, suggesting its stronger ability to recover zero loadings. In contrast, for nonzero loadings—including both primary loadings and smaller cross-loadings—the sampling distribution appears approximately normal.

Conclusion These findings indicate that while L^p rotations yield consistent estimates with increasing sample size, the central limit theorem does not hold for zero loadings. Consequently, standard inferential procedures based on normality assumptions, such as calculating p-values, cannot be directly applied to the loading matrix when using L^p rotations.

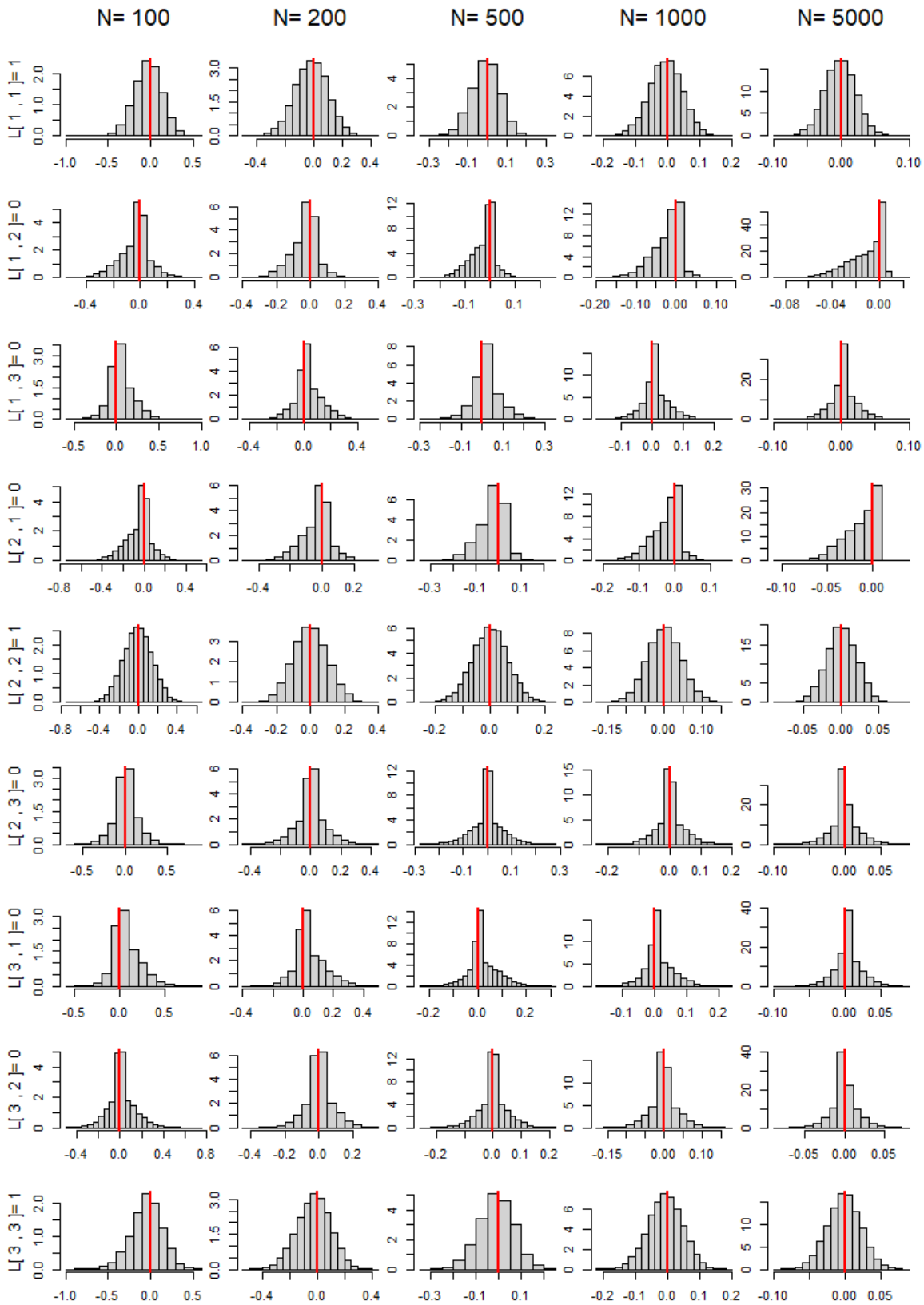


Figure 4.3: Sampling distribution of estimation errors of the top square submatrix in the loading matrix estimated using L^1 rotation.

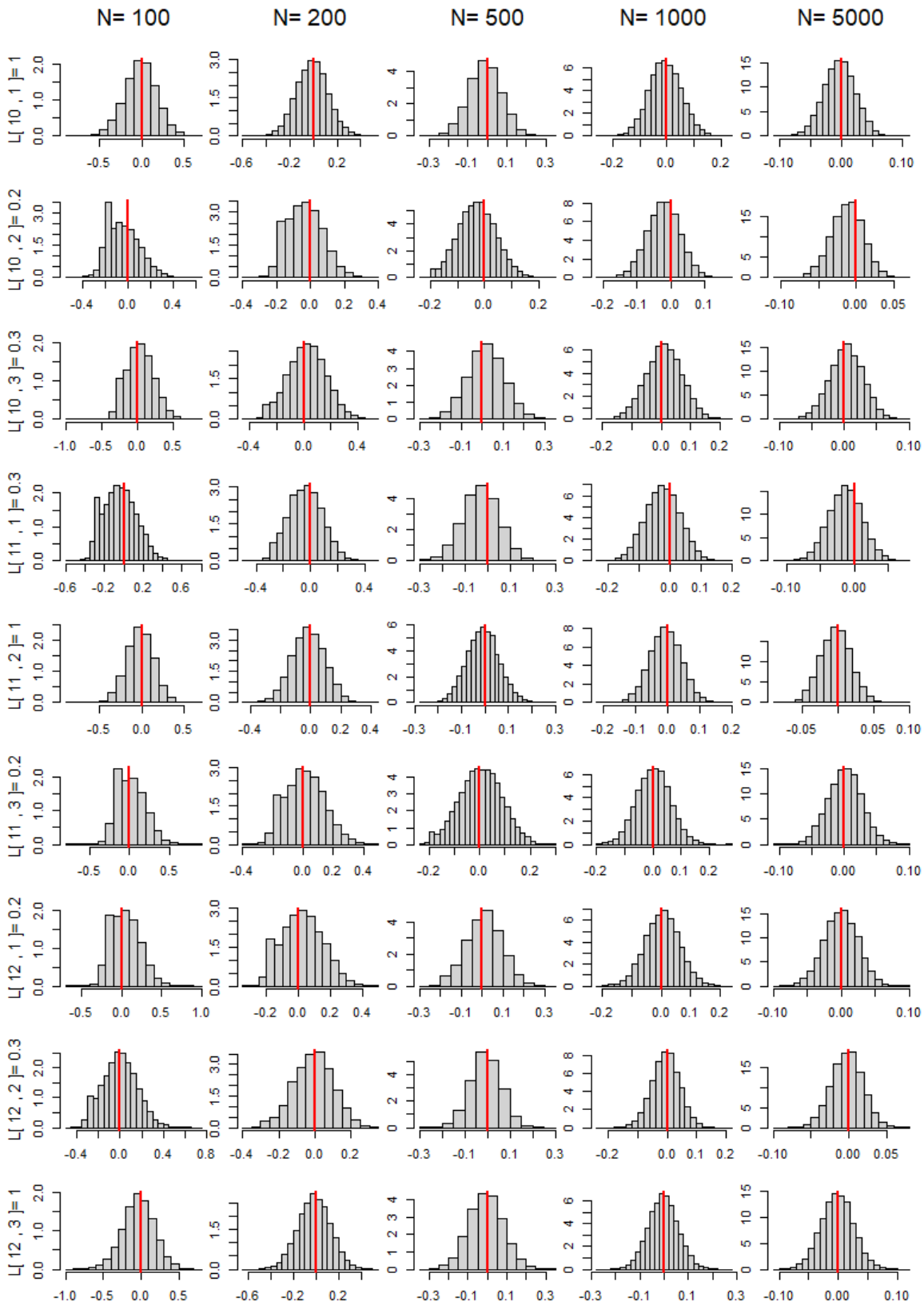


Figure 4.4: Sampling distribution of estimation errors of the bottom square submatrix in the loading matrix estimated using L^1 rotation.

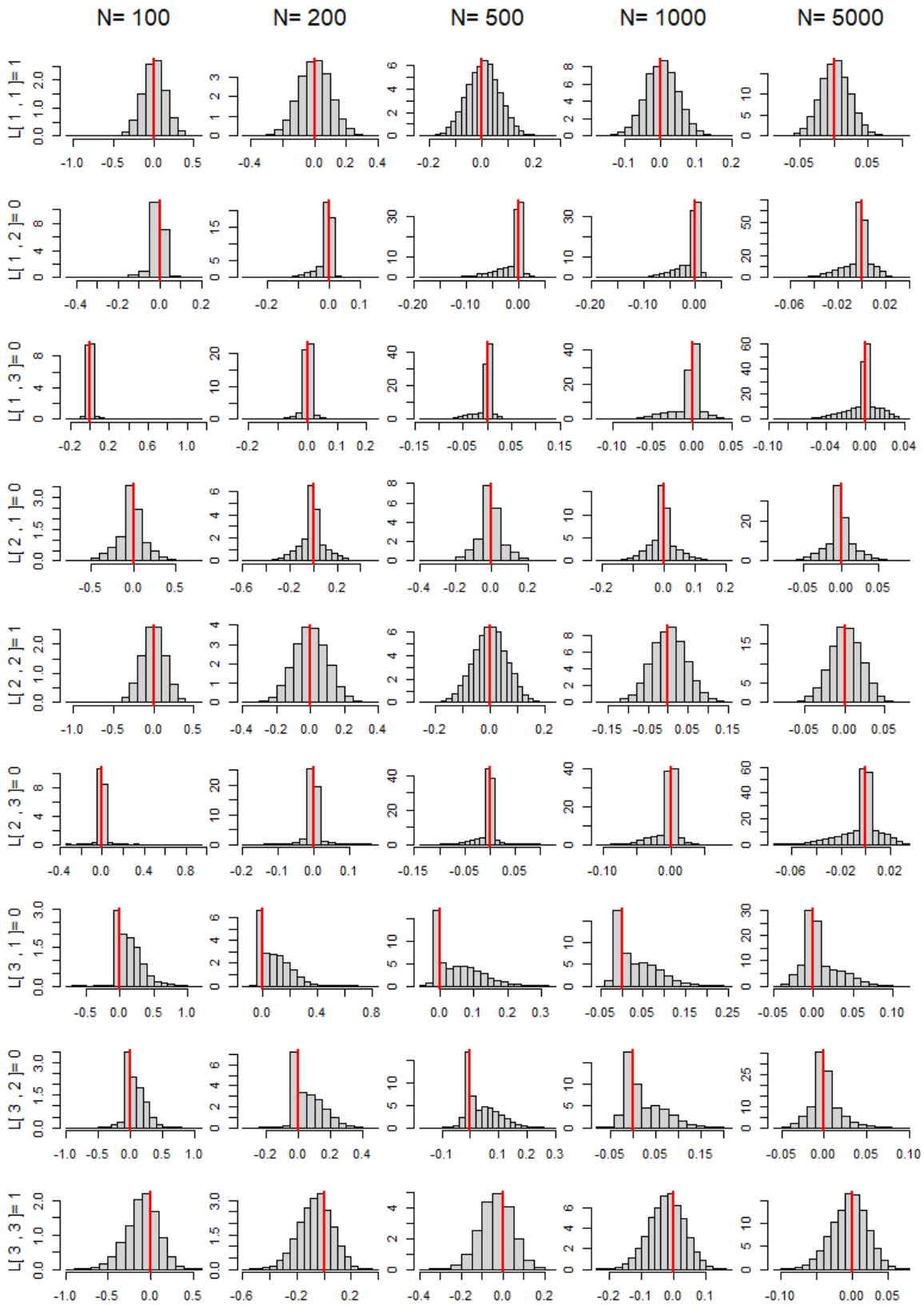


Figure 4.5: Sampling distribution of estimation errors of the top square submatrix in the loading matrix estimated using $L^{0.5}$ rotation.

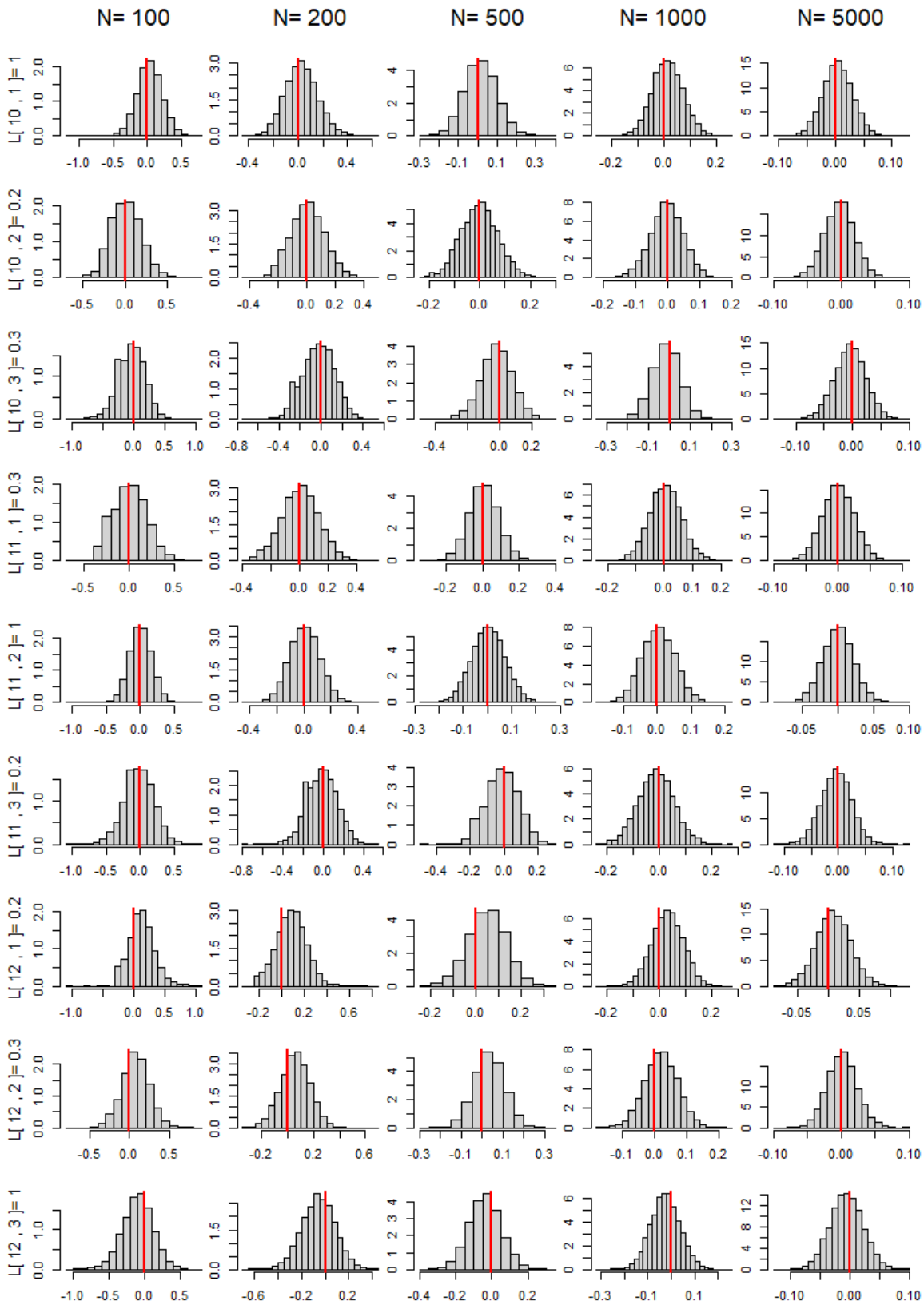


Figure 4.6: Sampling distribution of estimation errors of the bottom square submatrix in the loading matrix estimated using $L^{0.5}$ rotation.

A4.7. Study A.II: Evaluation of Anchor Item Selection Using Rank Statistics

To assess the accuracy of anchor item selection using Algorithm 5, we adopt the same simulation settings as in Section 3.4. Specifically, we generate a 60×3 loading matrix while varying the proportion of non-simple items from 0% to 95%. The main loadings are sampled from $U[1, 2]$, whereas cross-loadings are drawn from $U[0.2, 0.5]$. The factor covariance matrix is identical to that in Study A.I (see Table 4.13). We consider five sample sizes: $N = 100, 200, 500, 1000, 5000$, and conduct $B = 1000$ independent replications for both $p = 1$ and $p = 0.5$.

Evaluation Criteria We evaluate the performance of the proposed method using Selection Accuracy (ACC): The probability that the algorithm correctly identifies all K simple items as anchor items. Formally, we define:

$$ACC = \frac{1}{B} \sum_{b=1}^{B} \prod_{k=1}^{K} \mathbf{1}\{id_k \text{ is a simple item for factor } k\}. \tag{4.27}$$

Results From Table 4.14, we observe that both L^1 and $L^{0.5}$ rotations achieve anchor item selection accuracy exceeding 95% when the sample size is at least 200 and the proportion of simple items is above 40%. Moreover, L^1 rotation outperforms $L^{0.5}$ when the proportion of simple items exceeds 25%. This is due to $L^{0.5}$'s tendency to produce smaller cross-loadings, leading to the misclassification of non-simple items as anchor items.

However, when the proportion of simple items is extremely low (e.g., below 10%), we observe a contrasting trend. Specifically, for $N = 5000$, $L^{0.5}$ correctly identifies all anchor items, whereas L^1 achieves selection accuracies of only 0.427 and 0. This suggests that, under extremely sparse conditions, $L^{0.5}$ remains capable of recovering the loading matrix, while L^1 may converge to alternative local minima.

Conclusion The results demonstrate that Algorithm 5 performs well when the proportion of simple items exceeds 40%, for both $p = 1$ and $p = 0.5$. When the proportion of simple items is extremely low and the sample size is sufficiently large (e.g., $N > 500$), $L^{0.5}$ is a preferable choice, as it better preserves sparsity and improves selection accuracy under these conditions.

Proportion of Simple items	L^1					$L^{0.5}$				
N	100	200	500	1000	5000	100	200	500	1000	5000
95%	1	1	1	1	1	1	1	1	1	1
90%	0.994	1	1	1	1	0.993	0.999	1	1	1
85%	0.991	0.999	1	1	1	0.984	0.999	1	1	1
80%	0.98	1	1	1	1	0.965	0.998	1	1	1
75%	0.987	0.998	1	1	1	0.965	0.997	1	1	1
70%	0.971	1	1	1	1	0.94	0.996	1	1	1
65%	0.96	0.999	1	1	1	0.92	0.991	1	1	1
60%	0.951	0.997	1	1	1	0.905	0.984	1	1	1
55%	0.917	0.99	1	1	1	0.877	0.983	1	1	1
50%	0.927	0.99	1	1	1	0.857	0.974	1	1	1
45%	0.881	0.988	1	1	1	0.805	0.96	1	1	1
40%	0.818	0.974	1	1	1	0.737	0.95	1	1	1
35%	0.725	0.944	0.999	1	1	0.62	0.903	0.997	1	1
30%	0.599	0.872	0.992	1	1	0.532	0.869	0.997	1	1
25%	0.548	0.844	0.99	1	1	0.454	0.802	0.997	1	1
20%	0.351	0.627	0.933	0.996	1	0.337	0.697	0.991	1	1
15%	0.241	0.537	0.931	0.991	1	0.245	0.644	0.98	1	1
10%	**0.107**	**0.259**	**0.464**	**0.502**	**0.427**	**0.154**	**0.454**	**0.946**	**0.999**	**1**
5%	**0.016**	**0.054**	**0.048**	**0.013**	**0**	**0.056**	**0.21**	**0.801**	**0.984**	**1**

Table 4.14: The selection accuracy of anchor items using the L^1 and $L^{0.5}$ rotation criteria.

A4.8. Study A.III: The Sampling Distribution of 1DS and 2DS

In this study, we aim to demonstrate the necessity of data splitting for constructing valid p-values in Algorithm 6. To achieve this, we compare the 1DS with 2DS:

1. **1DS (One Data Set)**: The same dataset is used for both anchor item selection (Algorithm 5) and for refitting the CFA model.

2. **2DS (Two Data Sets)**: Algorithm 6 is applied, where the dataset is split into two equal subsets. Anchor item selection is performed on one subset to determine the loading structure $\Gamma^{(1)}$, and a CFA model is then fitted to the other subset.

Experimental Setting We use the same parameters as in Study A.I, as specified in Table 4.13. To highlight the asymptotic failure of 1DS, we consider an extremely large sample size of $N = 20,000$ and replicate the experiment $B = 1000$ times. As in Study A.I, estimation error is

analyzed by comparing the estimated loadings to their true values. If the central limit theorem (CLT) holds, the sampling distribution of these errors should be symmetric around zero. We present histograms that specify both the location and magnitude of each loading.

Results Given the large sample size, as shown at the top of each column, the item selection error for each factor is zero for both 1DS and 2DS. However, as illustrated in Figures 4.7 and 4.8, the 1DS method fails to produce normally distributed sampling distributions. Specifically, entries such as $L[1,3]$, $L[2,1]$, $L[4,2]$, $L[4,3]$, $L[5,1]$, $L[5,3]$, $L[6,1]$, $L[7,2]$, $L[7,3]$, $L[8,1]$, $L[9,1]$, $L[11,1]$, and $L[11,3]$ exhibit skewed distributions. Notably, all these loadings correspond either to zero loadings or small cross-loadings. In contrast, Figures 4.9 and 4.10 show that 2DS results in distributions that are approximately symmetric around zero.

The failure of 1DS arises from the selection process. In the item selection stage, the criterion ensures that the estimated loading has the smallest second-largest entry magnitude. As a result, aside from the primary loading, all other loadings are estimated to be values very close to zero. During the subsequent estimation stage, these non-primary loadings are set to zero, and a CFA model is fitted using the same dataset. However, this approach implicitly incorporates prior knowledge from the selection stage, leading to biased inference.

In contrast, when a separate dataset is used for selecting simple items, the covariance structure in the second dataset does not necessarily enforce small estimated values for non-primary loadings in the selected items determined by the first dataset. Since the two datasets are independent, the information in one does not dictate the properties of the other. This issue, known as *inference after selection*, necessitates data separation to avoid unintentional 'data peeking'.

Conclusion We demonstrate that the data-splitting procedure in Algorithm 6 is essential for valid inference. The problem of reusing data during the item selection stage cannot be mitigated, even with an extremely large sample size. Failure to account for prior exposure to the dataset leads to biased inference, reinforcing the necessity of data separation in this context.

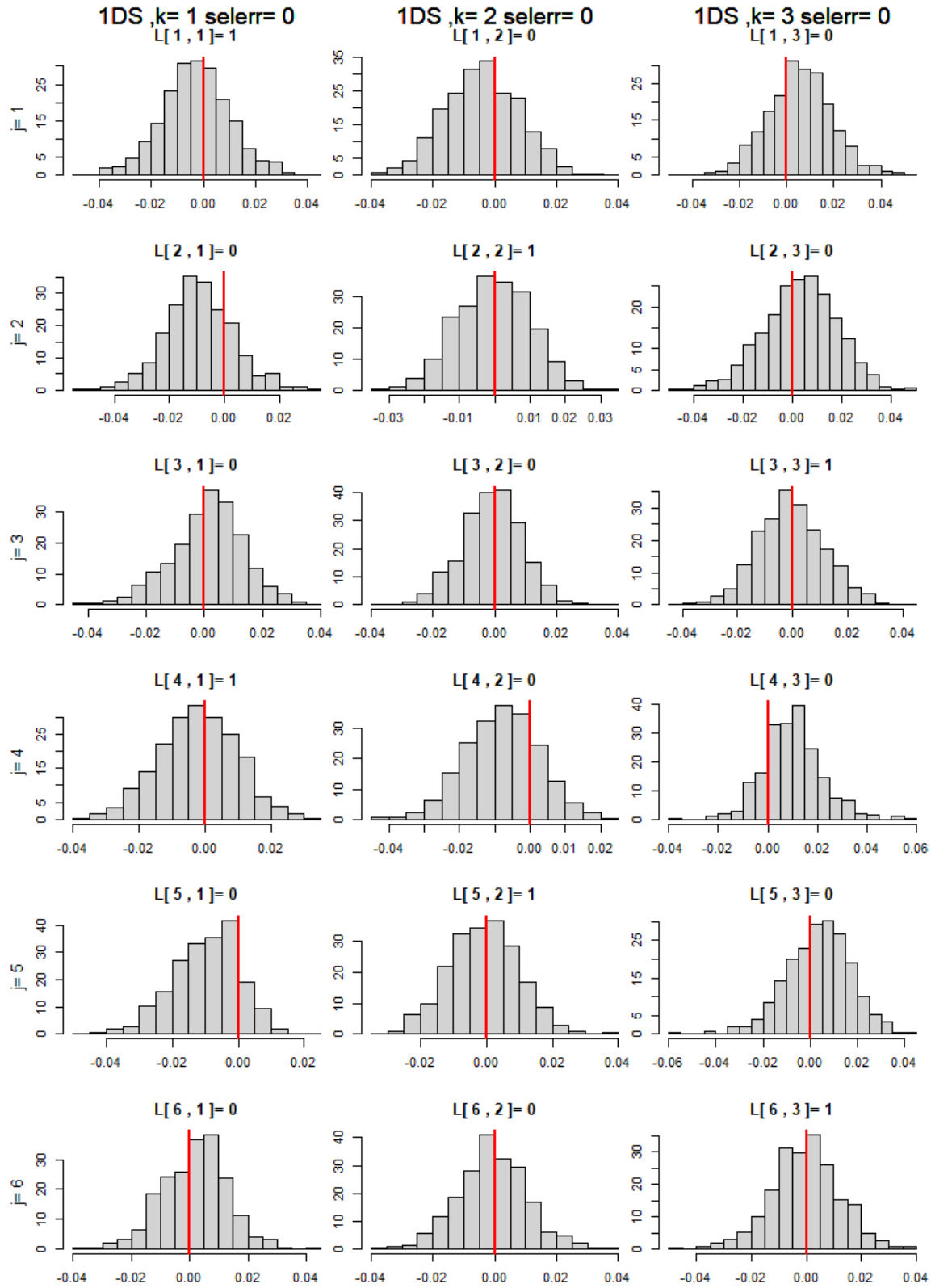


Figure 4.7: Sampling distribution of the top two squared submatrices in the loading matrix, estimated using 1DS.

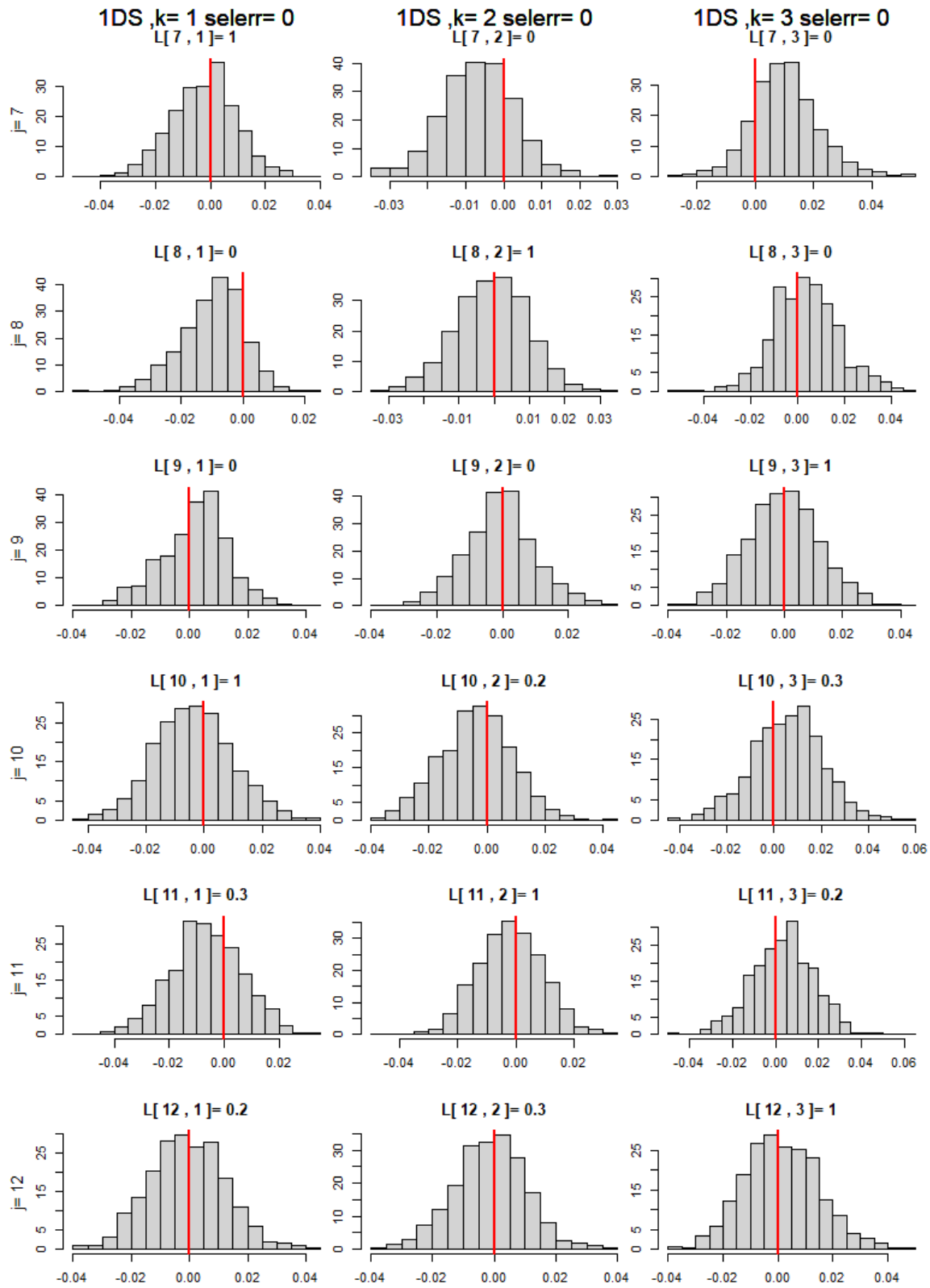


Figure 4.8: Sampling distribution of the bottom two squared submatrices in the loading matrix, estimated using 1DS.

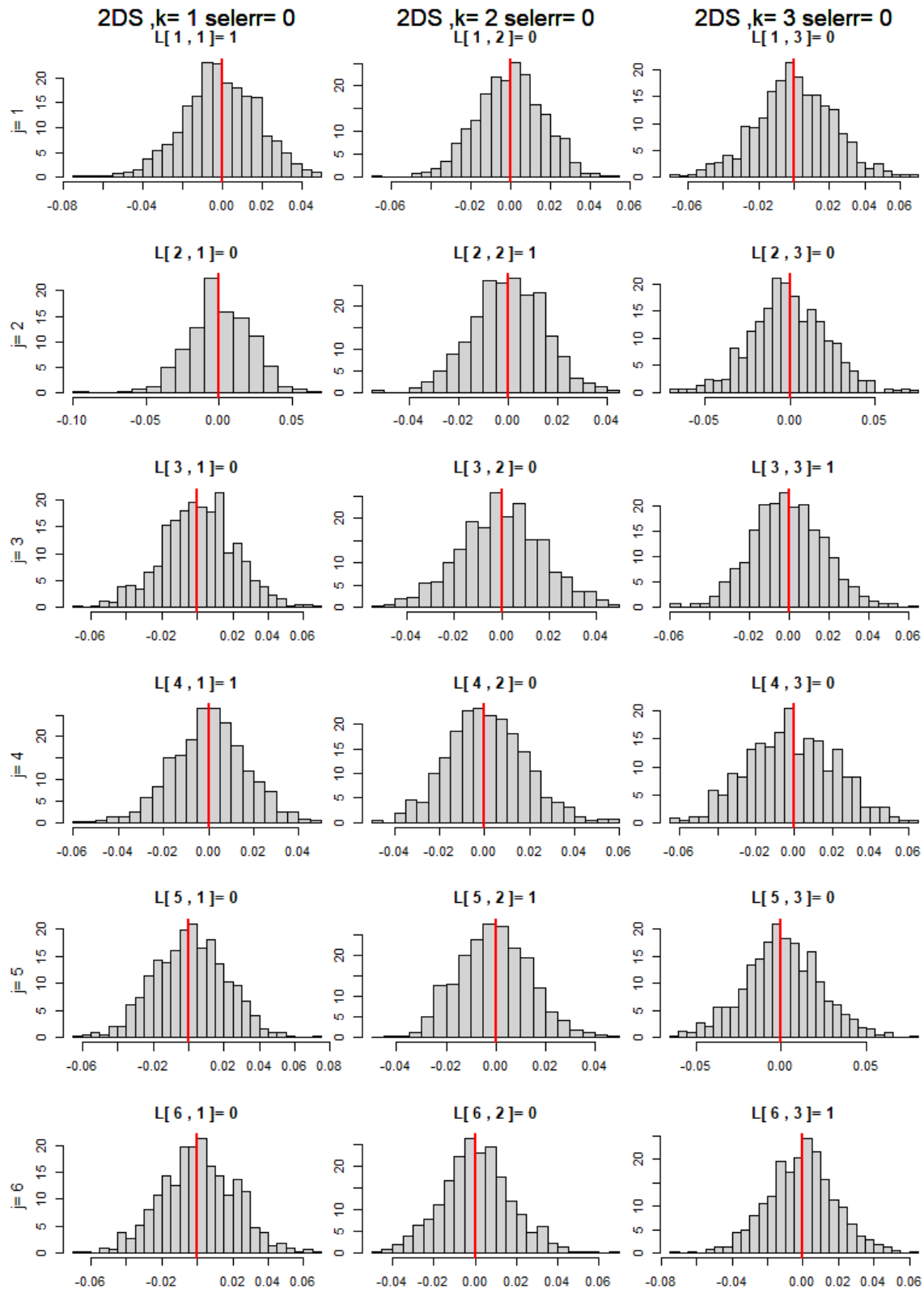


Figure 4.9: Sampling distribution of the top two squared submatrices in the loading matrix, estimated using 2DS.

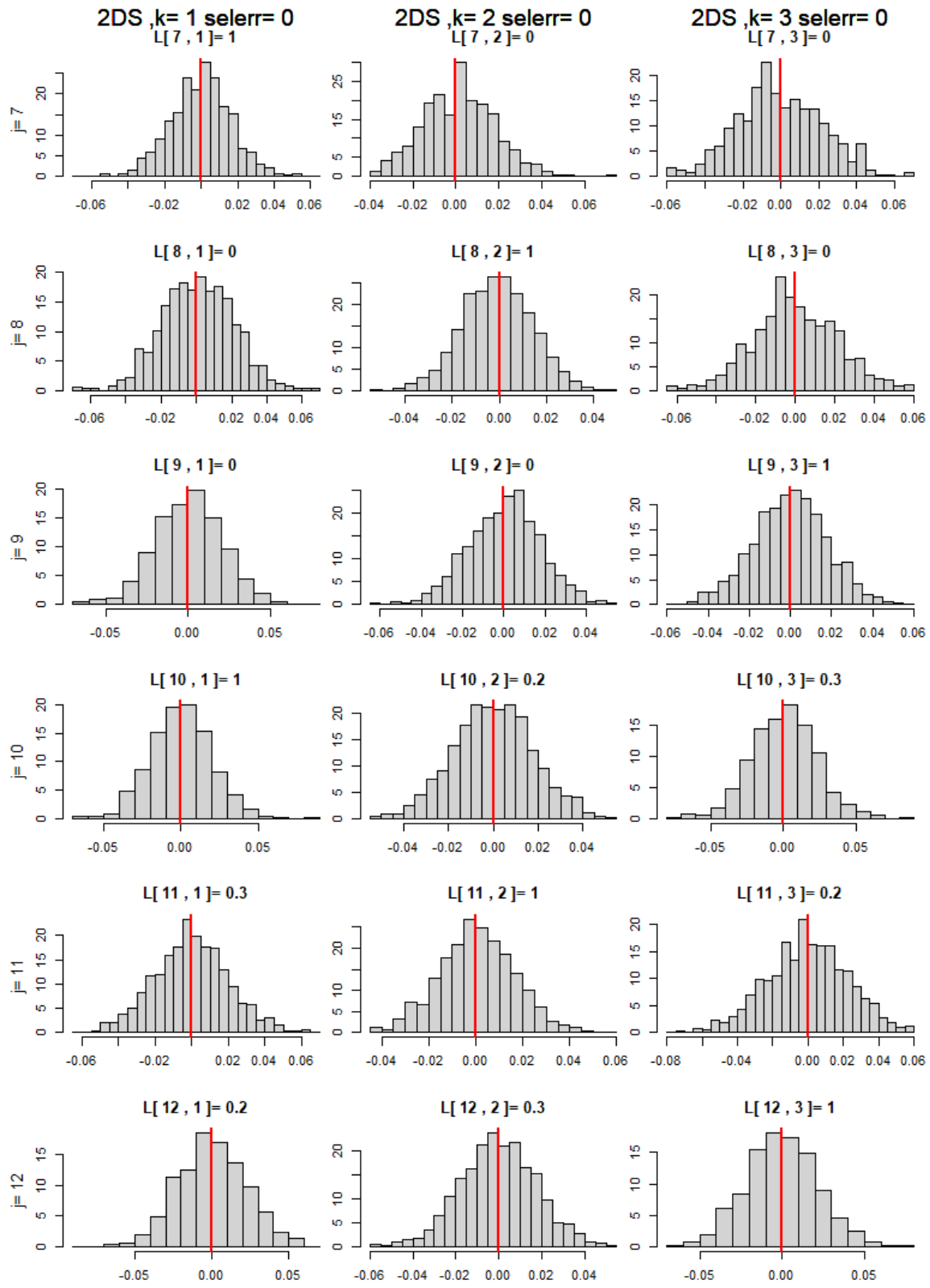


Figure 4.10: Sampling distribution of the bottom two squared submatrices in the loading matrix, estimated using 2DS.

A4.9. Study A.IV: Dependence in the CFA Model and its Implications for the BH Procedure

In this section, we demonstrate that the independence assumption required for the Benjamini-Hochberg (BH) procedure does not hold in the confirmatory factor analysis (CFA) model. This violation arises due to the correlation between zero loadings and nonzero loadings at both the test level and the factor level.

Experimental Setting To analyze the asymptotic variance of the loading matrix, we consider an extremely large sample size: $N = 100,000$. We run a CFA model on a dataset with a loading matrix of dimensions 60×3. Each factor consists of 10 simple items, along with 10 additional items that primarily load onto the factor but also exhibit secondary cross-loadings. The primary loadings and cross-loadings are uniformly sampled from $U[1, 2]$ and $U[0.2, 0.5]$, respectively. For each sample size and experimental condition, we conduct $B = 1000$ independent replications. We then plot the histogram of the covariance between zero loadings and nonzero loadings. Notably, the variances of all variables are excluded from the graph. To compute the sample covariance, we multiply it by the convergence rate N to account for the effect of sample size scaling.

Results As shown in Figure 4.11, the adjusted covariance values range from approximately -0.1 to 0.5. Given that the nonzero loadings are sampled from the range $[0.2, 2]$, the square root of these covariance values is of comparable magnitude to the true loading scale. This empirical evidence demonstrates that zero loadings are correlated with nonzero loadings at both the factor and test levels. Consequently, the independence assumption required for the BH procedure is violated, limiting its applicability in the CFA model.

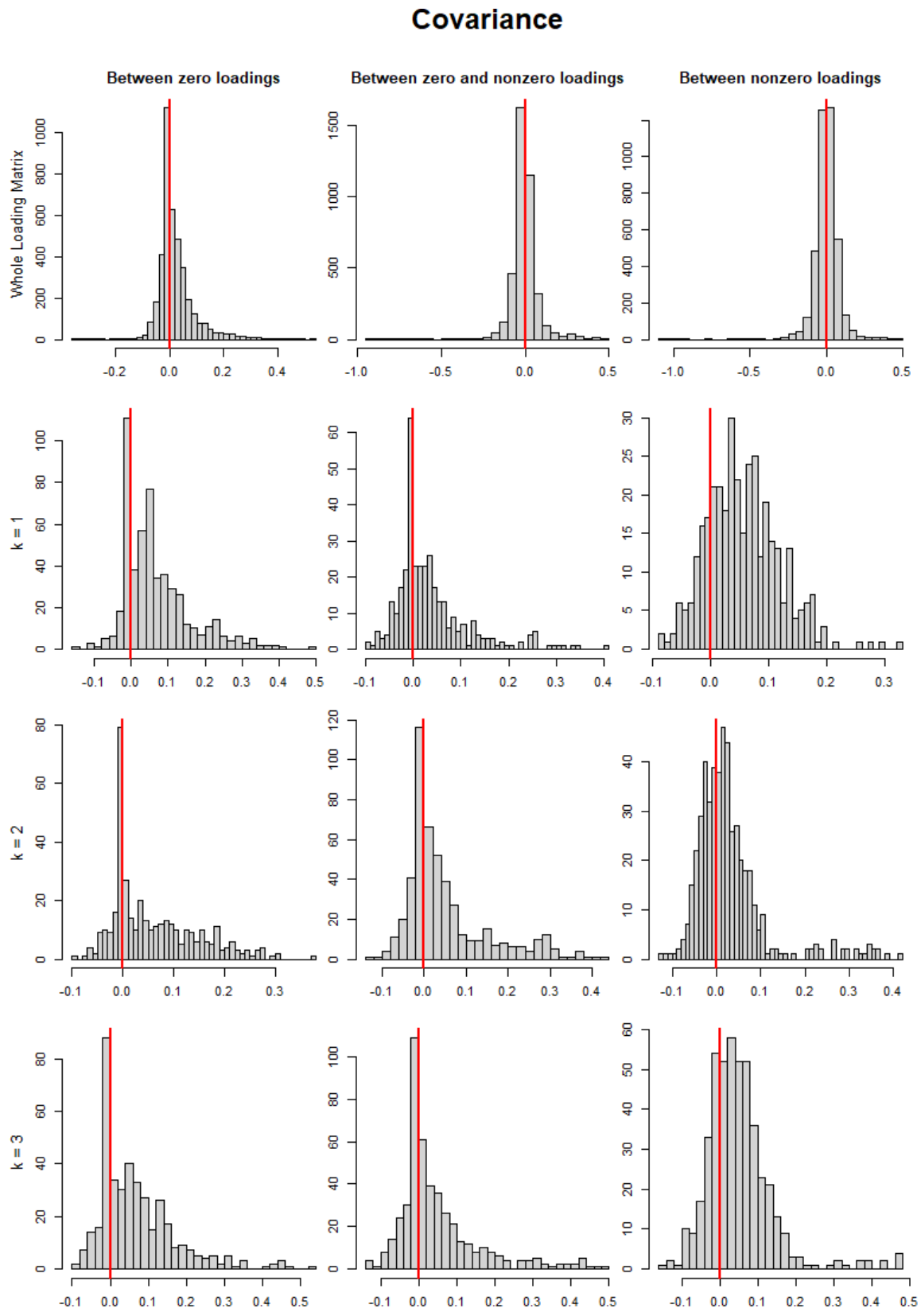


Figure 4.11: Sampling distribution of off-diagonal covariance values for different loading conditions. The first row represents the covariance between zero loadings, zero and nonzero loadings, and nonzero loadings in the full loading matrix. The subsequent rows show covariance distributions for each factor (k = 1, 2, 3).

Chapter 5

Discussions

This thesis introduces a novel rotation framework for Exploratory Factor Analysis (EFA), offering both theoretical guarantees and practical tools that enhance interpretability and support valid statistical inference.

Specifically, we developed the L^p rotation criterion to recover sparse loading matrices in EFA. This rotation method improves model interpretability by simplifying the associations between latent factors and manifest variables. Within the L^p framework, we introduced computational algorithms, identification theory, and variable selection techniques, with a particular focus on controlling the False Discovery Rate (FDR). The L^p rotation functionality has been implemented in the R package `GPArotation` (Bernaards and Jennrich, 2005), with the functions `lpT` and `lpQ` supporting orthogonal and oblique rotations, respectively.

In Chapter 2, we proposed a new family of oblique rotations based on component-wise L^p loss functions and developed an iteratively reweighted gradient projection algorithm to efficiently solve the resulting non-smooth optimization problem. Our results demonstrate that L^p rotation achieves accuracy comparable to l^p penalized estimation with a small tuning parameter, while significantly reducing computational cost—particularly when the number of items J is large.

In Chapter 3, we established theoretical results that confirm the identification conditions for L^p rotation estimation. These results address a gap in the literature, which has predominantly focused on loading matrices with only simple items. We extended the theory to accommodate cases where the loading matrix includes a small proportion of non-simple items—an assumption more appropriate for EFA, where the loading structure is typically unknown. Additionally, we showed that L^p rotation tends to outperform traditional rotation methods when the true loading matrix is sparse.

Chapter 4 further extends this methodology by applying L^p rotation to identify loading structures in Confirmatory Factor Analysis (CFA) models under less restrictive assumptions. Building upon this, we proposed a procedure for computing p-values in EFA models and adapted common FDR control methods for this context. We found that the Benjamini-Hochberg (BH) procedure violates the independence assumptions of EFA, whereas the Benjamini-Yekutieli (BY) and e-value-based Benjamini-Hochberg (eBH) procedures remain valid.

In many modern applications, we encounter settings where the number of items J is comparable to, or even exceeds, the sample size N; that is, $J \approx N$ or $J > N$. In such cases, it is more appropriate to consider a high-dimensional asymptotic regime when performing exploratory factor analysis (EFA). Despite its importance, relatively little theoretical work has been done in this area, with the notable exception of Rohe and Zeng (2022).

A promising direction for future research is to extend the existing consistency results to the high-dimensional setting, where $J \to \infty$. From a methodological perspective, we expect that the proposed L^p rotation method and the p-value computation procedure (Algorithm 6) can still operate effectively in this regime. However, for false discovery rate (FDR) control, the Benjamini–Yekutieli (BY) procedure becomes infeasible, as its correction factor diverges when $J \to \infty$. In contrast, the Benjamini–Hochberg (BH) procedure based on e-values remains applicable, making it a promising candidate for FDR control in high-dimensional settings and an important subject for further investigation.

From a theoretical standpoint, the consistency result for L^p rotation presented in Theorem 1 must be revisited in this regime, as the eigenvalues of the true loading matrix $\mathbf{\Lambda}^*$ may grow with J. Moreover, a central limit theorem (CLT) specifically tailored to high-dimensional factor analysis models—such as the one proposed in Bai and Li (2012)—is needed to justify the validity of p-value estimation. This high-dimensional CLT should be derived under identification conditions that involve simple items. Such theoretical developments are particularly important, as the classical CLT for maximum likelihood estimation no longer applies when the sample covariance matrix is not positive definite, which commonly occurs when $J > N$.

In this work, we treat the number of factors as given. However, in many applications of exploratory factor analysis (EFA), the true number of factors is unknown. We recommend the use of consistent and theoretically grounded selection methods—such as the Bayesian Information Criterion (BIC) or parallel analysis—to establish a valid foundation before applying factor rotation. This step is essential because rotation methods—whether orthogonal, oblique, or sparse—are post-processing procedures applied to an estimated factor loading matrix of fixed dimensionality. Among

various model selection criteria, BIC has been shown to consistently estimate the true number of factors under suitable regularity conditions (Bai and Ng, 2002). Parallel analysis (Horn, 1965) is another widely used method, which compares the observed eigenvalues to those generated under a null model. In contrast, methods such as the scree plot (Cattell, 1966) and the Kaiser criterion (eigenvalue > 1 rule) are based on the eigenvalues of the sample correlation matrix and thus derive from principal component analysis (PCA) rather than the factor model. While these techniques are often used as simple heuristics for dimensionality selection, they lack the statistical rigor required for EFA and should be interpreted with caution.

In practice, many datasets consist of categorical variables, for which linear factor models may be inappropriate. In such cases, nonlinear or generalized factor models—such as item response theory (IRT) models—are more suitable (Bartholomew et al., 2011; Skrondal and Rabe-Hesketh, 2004). A promising direction for future research is to extend the L^p rotation framework to these models. For instance, in IRT models where the manifest variables are categorical but the latent factors remain continuous, L^p rotation can be applied to produce sparser and more interpretable loading structures. More generally, for any model with multiple continuous latent factors, where the condition C1 on Chapter 2 are satisfied, a two-step procedure can be used: first, obtain an initial estimate of the loading matrix; second, apply L^p rotation to enhance sparsity and interpretability.

In models with a single latent factor, it is not common practice to apply rotation. This is because in the unidimensional case, any rotation would amount to a trivial rescaling or sign change, offering no new structural insight or interpretability benefits. Many widely used models fall under this category, including the classical one-factor model, the two-parameter logistic (2PL) IRT model, and the testlet model with a single general factor and structured residuals. In such settings, the primary goal is typically to estimate the latent trait or ability accurately rather than to uncover a sparse or interpretable factor loading structure. Rotation techniques are thus more relevant in multidimensional exploratory contexts, where identifying interpretable factor patterns is a central objective.

Bibliography

Anderson, T. and Rubin, H. (1956). Statistical inference in. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, page 111. University of California Press.

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

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436 – 465.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020). Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409–1431.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley & Sons, West Sussex, UK.

Bartholomew, D. J., Steele, F., Galbraith, J., and Moustaki, I. (2008). *Analysis of multivariate social science data*. CRC press.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Bernaards, C. A. and 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*. John Wiley & Sons.

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

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

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.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate behavioral research*, 36(1):111–150.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, 18(1):23–38.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Chen, Y. and 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., and Ying, Z. (2021). Item response theory–a statistical framework for educational and psychological measurement. *arXiv preprint arXiv:2108.08604*.

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

Chen, Y., Li, X., and 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., and Zou, H. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and its Interface*, 3(4):429–436.

Crawford, C. (1975). A comparison of the direct oblimin and primary parsimony methods of oblique rotation. *British Journal of Mathematical and Statistical Psychology*, 28(2):201–213.

Crawford, C. B. and Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, 35(3):321–332.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023). False discovery rate control via data splitting. *Journal of the American Statistical Association*, 118(544):2503–2520.

Daubechies, I., DeVore, R., Fornasier, M., and 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., and 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.

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement*, 43(1):39–52.

Fitzpatrick, P. (2009). *Advanced Calculus*, volume 5. American Mathematical Soc.

Geminiani, E., Marra, G., and 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.

Ghandwani, D. and Hastie, T. (2024). Scalable recommender system based on factor analysis. *arXiv preprint arXiv:2408.05896*.

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., and 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.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The annals of statistics*, 5(5):815–841.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2019). Multivariate data analysis.

Harerimana, A. and Mtshali, N. G. (2020). Using exploratory and confirmatory factor analysis to understand the role of technology in nursing education. *Nurse Education Today*, 92:104490.

Harman, H. H. and Harman, H. H. (1976). *Modern Factor Analysis*. University of Chicago press.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.

Hendrickson, A. E. and White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17(1):65–70.

Higham, N. J. (1988). Matrix nearness problems and applications. In Gover, M. and Barnett, S., editors, *Applications of Matrix Theory*, pages 1–27. Oxford University Press, Oxford.

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

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

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

Jennrich, R. I. (1974). Simplified formulae for standard errors in maximum-likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology*, 27(1):122–131.

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. and Clarkson, D. B. (1980). A feasible method for standard errors of estimate in maximum likelihood factor analysis. *Psychometrika*, 45(2):237–247.

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

Jin, S., Moustaki, I., and Yang-Wallentin, F. (2018). Approximated penalized maximum likelihood for exploratory factor analysis: An orthogonal case. *Psychometrika*, 83(3):628–649.

John, O. P. and Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In Pervin, L. A. and John, O. P., editors, *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press, New York.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482.

Jöreskog, K. G. and Goldberger, A. S. (1972). Factor analysis by generalized least squares. *Psychometrika*, 37(3):243–260.

Jöreskog, K. G. (1969a). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202.

Jöreskog, K. G. (1969b). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kaiser, H. F. (1959). Computer program for varimax rotation in factor analysis. *Educational and psychological measurement*, 19(3):413–420.

Kano, Y. (1986). Conditions on consistency of estimators in covariance structure model. *Journal of the Japan Statistical Society, Japanese Issue*, 16(1):75–80.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

Kiers, H. A. (1994). Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika*, 59:567–579.

Koopmans, T. C. and Reiersol, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181.

Lai, M.-J. and Wang, J. (2011). An unconstrained ℓ_q minimization with $0 < q \leqslant 1$ for sparse solution of underdetermined linear systems. *SIAM Journal on Optimization*, 21(1):82–101.

Lehmann, E. L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, 37:1137–1153.

Liu, X., Wallin, G., Chen, Y., and Moustaki, I. (2023). Rotation to sparse loadings using l^p losses and related inference problems. *Psychometrika*, 88(2):527–553.

Machado, J. A. (1993). Robust model selection and m-estimation. *Econometric Theory*, 9(3):478–493.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

McKeon, J. (1968). Rotation for maximum association between factors and tests. *Unpublished manuscript, Biometric Laboratory, George Washington University*.

Meinshausen, N. and 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*, pages 758–765.

O'Rourke, S., Vu, V., and Wang, K. (2018). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook. Version 20121115.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory*. Springer.

Ren, Z. and Barber, R. F. (2024). Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):122–154.

Rohe, K. and Zeng, M. (2022). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society - Series B*, page to appear.

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

Sellbom, M. and Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological assessment*, 31(12):1428.

Sen, B., Banerjee, M., and 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*, pages 221–242.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Sohrab, H. H. (2003). *Basic Real Analysis*, volume 231. Springer, New York.

Stock, J. H. and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, 4:351–394.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335.

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.

Vallerand, R. J., Blais, M. R., Brière, N. M., and Pelletier, L. G. (1992). The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement*, 52(4):1003–1017.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

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.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852.

Weyl, H. (1912). Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479.

Williams, B., Onsman, A., and Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian journal of paramedicine*, 8:1–13.

Yamamoto, M., Hirose, K., and 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., and Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2):358–372.

Zhao, S., Witten, D., and 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., and Long, T. (2017). Does l_p-minimization outperform l_1-minimization? *IEEE Transactions on Information Theory*, 63(11):6896–6935.