Latent Variable Modelling and Statistical Analysis for High-Dimensional Data

Sze Ming Lee

Department of Statistics
London School of Economics and Political Science

This dissertation is submitted for the degree of

Doctor of Philosophy

October 2025

To my family and friends

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of 23864 words.

I confirm that Chapter 2 is jointly co-authored with Dr. Tony Sit (The Chinese University of Hong Kong) and my supervisor, Dr. Yunxiao Chen. Chapters 3 and 4 are jointly co-authored with Dr. Yunxiao Chen. I contributed 75% to each work. The contents in Chapters 2 and 4 have been submitted to peer-reviewed statistical journals, and we plan to submit Chapter 3 for publication soon.

Sze Ming Lee
October 2025

Acknowledgements

I would first like to thank my thesis supervisor, Dr. Yunxiao Chen, for his insightful guidance, continuous support, and constant encouragement. This thesis would not have been possible without his help, and it has been a great honour to work under his supervision. I am also grateful to my second supervisor, Professor Irini Moustaki, for her invaluable mentorship and encouragement throughout my academic development. I am also grateful to Professor Fiona Steel for her thoughtful feedback during the early stages of this work, which greatly contributed to its development. Special thanks go to Dr. Tony Sit for his generous advice and continued support over the course of my studies. Besides, I am grateful to Professor Tengyao Wang and Professor Yi Yu for serving as my examiners and for their valuable comments.

I would also like to thank the Professors, the classmates and the staff in the Department of Statistics who helped to create an enjoyable and fruitful environment during my stay in the London School of Economics and Political Science.

Lastly, I am deeply thankful to my parents for their unwavering support and care throughout my years of study. My heartfelt thanks go to my beloved wife, Carrie, for her enduring love, understanding, and encouragement during my graduate journey.

Abstract

In recent years, advances in technology have made it easier to collect and store high-dimensional data, creating a growing need for effective statistical tools. This thesis presents new approaches through three related studies to improve existing methods and enhance their practical applicability.

Chapter 2 proposes a novel latent variable model tailored for high-dimensional multivariate longitudinal data. This model accommodates mixed data types and missing observations by incorporating unobserved factors that capture dependence across variables and time points, facilitating both statistical inference and predictive performances. A central limit theorem is established for inference on regression coefficients, and an information criterion is developed to consistently determine the number of factors. The method is applied to grocery shopping data to predict and interpret consumer behaviour.

Chapter 3 introduces a stability-based method for selecting the number of latent factors in linear factor models, using principal angles between loading spaces obtained from data splitting. Consistency is established under weaker asymptotic requirements than existing approaches. Simulations and real data examples demonstrate the method's improved accuracy and robustness.

Chapter 4 develops a flexible statistical modelling framework for pairwise comparison data, relaxing the conventional stochastic transitivity assumptions in classical models. By imposing an approximately low-dimensional skew-symmetric structure, the method achieves minimax-optimal estimation rates and performs well with sparse data. Its superiority over the traditional Bradley-Terry model is supported by simulations and real-world applications.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent decades, advances in data collection have led to an abundance of high-dimensional data across diverse fields, including biomedical research, finance and social sciences, among many others (e.g., Fan et al. (2011); Rahnenführer et al. (2023)). High-dimensional datasets, characterized by having a large number of observed variables relative to the number of observations, present significant analytical challenges, including computational complexity, risks of overfitting, and difficulties in uncovering meaningful and interpretable structures. Traditional statistical techniques often fall short in managing these complexities effectively, motivating the development of advanced methodologies.

High-dimensional factor models offer an effective framework for capturing complex dependence structures by representing observed variables through a lower-dimensional set of latent factors. Both the latent factors and loading matrices are treated as fixed parameters during estimation. Early applications focused on linear settings, with estimation methods such as principal component analysis and maximum likelihood (e.g., Stock and Watson (2002); Bai (2003); Bai and Li (2012)). More recently, these models have been extended to broader contexts, including quantile factor models that target conditional quantiles of outcomes (Chen et al., 2021), and generalised latent factor models for data following exponential family distributions (Chen et al., 2019c, 2020).

Another important class of high-dimensional models imposes an approximately low-rank structure, offering greater flexibility for modelling noisy or heterogeneous data. Instead of enforcing an exact rank constraint often implied by latent variable formulations, these models regularize the nuclear norm or max-norm of the data matrix to encourage approximately low-rank solutions. This framework has been widely adopted in applications such as matrix completion, collaborative filtering, and signal recovery, with application in real-valued matrices (e.g., Candes and Plan (2010); Cai and Zhou (2016)) and binary matrices (e.g., Cai and Zhou (2013); Davenport et al. (2014)).

This thesis focuses on advancing latent variable modelling and statistical analysis for high-dimensional data by addressing three distinct yet interconnected problems. These problems are explored in detail in Chapters 2, 3, and 4. Below, we summarize the focus of each chapter.

Chapter 2 discusses the analysis of high-dimensional multivariate longitudinal data. A novel latent variable model for drawing statistical inferences on covariate effects and predicting future outcomes is proposed. This model introduces unobserved factors to account for the between-variable and across-time dependence and assist the prediction. Statistical inference and prediction tools are developed under a general setting that allows outcome variables to be of mixed types and possibly unobserved for certain time points, for example, due to right censoring. A central limit theorem is established for drawing statistical inferences on regression coefficients. Additionally, an information criterion is introduced to choose the number of factors. The proposed model is applied to customer grocery shopping records to predict and understand shopping behaviour.

Chapter 3 concerns the selection of the number of latent factors. We propose a novel method for determining the number of factors in linear factor models under stability considerations. An instability measure is proposed based on the principal angle between the estimated loading spaces obtained by data splitting. Based on this measure, criteria for determining the number of factors are proposed and shown to be consistent. This consistency is obtained using results from random matrix theory, especially the complete delocalization of non-outlier eigenvectors. The advantage of the proposed methods over the existing ones is shown via weaker asymptotic requirements for consistency, simulation studies and a real data example.

Chapter 4 focuses on the analysis of pairwise comparisons data. We propose a general family of statistical models for pairwise comparison data that relax the common stochastic transitivity assumption underlying many existing approaches, such as Bradley-Terry (BT) and Thurstone models. In this model, the pairwise probabilities are determined by an approximately low-dimensional skew-symmetric matrix. Likelihood-based estimation methods and computational algorithms are developed, which allow for sparse data with only a small proportion of observed pairs. Theoretical analysis shows that the proposed estimator achieves minimax-rate optimality, which adapts effectively to the sparsity level of the data. The spectral theory for skew-symmetric matrices plays a crucial role in the implementation and theoretical analysis. The proposed method's superiority against the BT model, along with its broad applicability across diverse scenarios, is further supported by simulations and real data analysis.

The ideas developed in Chapters 2 to 4 are interconnected, and together they open up many interesting directions for further exploration. For instance, in Chapter 2, the proposed latent variable model relies on a consistent estimate of the number of factors, which is currently determined by an information criterion. It would be of particular interest to apply the stability principle from Chapter 3 for selecting the number of factors, especially when interpretability of the latent variables is desired in the analysis of high-dimensional multivariate longitudinal data. Moreover, when prediction of longitudinal outcomes is the primary goal rather than inference on covariate effects, one might consider imposing an approximate low-rank structure, as developed in Chapter 4, to yield a more robust model. The connections also extend in the opposite direction: developing valid statistical theory for exact-rank models with rank greater than two would naturally generalize beyond the Bradley–Terry framework for the analysis of pairwise comparison data. This is especially valuable when the aim is to extract stable and interpretable latent factors that help explain stochastic intransitivity among players.

Chapter 2

A Latent Variable Approach to Learning High-dimensional Multivariate longitudinal Data

2.1 Introduction

High-dimensional multivariate longitudinal data are becoming increasingly common, especially in social, behavioural and health sciences, where many outcomes are measured repeatedly within individuals. Examples include ecological momentary assessment data collected by smartphones or wearable devices for understanding within-subject social, psychological, and physiological processes in everyday contexts (Bolger and Laurenceau, 2013; Wang et al., 2014), electronic health record data for predicting and understanding health-related conditions (Lian et al., 2015; Zhang et al., 2020b), computer logfile data for understanding human-computer interactions from solving complex computer-simulated tasks (Chen et al., 2019b; Chen, 2020), and grocery shopping data for market basket analysis (Wan et al., 2017, 2018). These data may involve outcome variables of different types. For example, for ecological momentary assessment, physiological processes are typically measured by continuous variables, such as blood pressure, while psychological processes are recorded by participants' responses to survey items that involve binary or categorical variables. In addition, many multivariate longitudinal data may be derived from multitype recurrent event history (Chapter 2.5, Cook and Lawless, 2007), for which an outcome variable records whether a specific type of event occurs (e.g., purchasing a merchandise item) or the count of its occurrences within a time interval (e.g., the number of purchases).

In this chapter, we study high-dimensional multivariate longitudinal data, aiming to (1) infer the effect of covariates on each outcome variable and (2) predict future outcomes based on covariates and historical data. These tasks involve three challenges. First, due to the nature of the data, there is a complex within-individual dependence structure which exists between outcome variables and across time. Valid statistical inference and accurate prediction become a challenge if one fails to account for the dependence properly. Second, the presence of many outcome variables implies a substantial number of item-specific parameters, bringing challenges to the statistical inference. The classical theory for M- or Z-estimators no longer applies, and new asymptotic results concerning the consistency and asymptotic normality under a high-dimensional regime are needed. Third, some observation units may be lost to follow-up or observed only intermittently, resulting in incomplete data. For example, grocery shopping records based on membership may be incomplete if customers occasionally shop without using their membership card.

To tackle these challenges, we propose a high-dimensional generalised latent factor model. In this model, low-dimensional factors are introduced within each observation unit to capture the between-item and across-time dependence that is not attributable to the covariates. The model is very flexible, allowing for many types of outcome variables, including binary, count, and continuous variables. In addition, a computationally efficient joint likelihood estimator is proposed that estimates the unobserved factors, loading parameters, and regression coefficients simultaneously, which treats the factors as fixed parameters. Asymptotic properties of this estimator are established, including a central limit theorem for drawing statistical inferences on regression coefficients and an information criterion for choosing the number of factors. Moreover, we introduce a missing indicator approach (see Chapter 26, Molenberghs and Verbeke, 2005) to account for data missingness. Under a Missing at Random (MAR) assumption, this approach can handle many missingness patterns, including right-censoring that is common to recurrent event data.

Various statistical methods have been proposed for analysing multivariate longitudinal data. Generalised estimating equation (GEE) methods (e.g., Liang and Zeger, 1986; Prentice, 1988; Carey et al., 1993; Gray and Brookmeyer, 2000) are widely used for drawing statistical inferences on regression parameters relating the means of outcome variables to a set of covariates and parameters characterizing the marginal association between outcome variables. These methods often provide valid statistical inferences on parameters of interest without a need to specify a full joint distribution for the outcome variables. On the other hand, many joint models have been proposed for multivariate longitudinal data that are better at making predictions while still capable of drawing statistical inferences on parameters of interest, though the latter may be jeopardized by model misspecification. Joint models for multivariate longitudinal data include transition models (Liang and Zeger, 1989; Zeng and Cook, 2007) that are specified through a sequence of conditional probabilities of outcome variables given historical outcome variables and covariates, copula-based models (Lambert and Vandenhende, 2002; Smith et al., 2010; Panagiotelis et al., 2012) that specify a joint distribution via copulas, and latent variable models (Ten Have and Morabia, 1999; Oort, 2001; Liu and Hedeker, 2006; Hsieh et al., 2010; Proust-Lima et al., 2013; Wang et al., 2016; Ounajim et al., 2023; Sørensen et al., 2023) that capture the complex dependence structure by introducing latent variables. Latent variable models are very popular, thanks to their flexibility and interpretability. However, the statistical inference for these traditional latent variable models is carried out based on a marginal likelihood, where the latent variables are treated as random variables and marginalized out. This approach can hardly be extended to the high-dimensional setting with many latent variables due to the high computational cost of optimizing the marginal likelihood. Our method extends the traditional latent variable models to the high-dimensional setting and further overcomes their computational challenge using the proposed joint likelihood estimator.

The proposed method is also related to high-dimensional factor models for multivariate cross-sectional data or panel data that do not directly apply to the current problem. These models are estimated by minimizing a loss function of both unobserved factors and loading parameters. In other words, although unobserved factors may be regarded as random variables in the model specification, they are conditioned upon and treated as unknown parameters at the estimation stage. In this direction, Stock and Watson (2002) and Bai and Li (2012) considered linear factor analysis and proposed estimation methods based on quadratic and likelihood-based loss functions, respectively. Chen et al. (2020) and Liu et al. (2023a) considered generalised latent factor models that allow for various data types and proposed likelihood-based estimation procedures. Moreover, Chen et al. (2021) introduced a quantile factor model for multivariate data and proposed estimators based on the check loss function for quantile estimation. Although Liu et al. (2023a) and Chen et al. (2021) established some asymptotic normality results, they focused on factor models without covariates, and their results are not directly applicable to the current setting.

To summarize, our main contributions are three-fold: First, we propose a flexible latent variable model framework for analysing high-dimensional multivariate longitudinal data with mixed outcome types and missing values, which is unmanageable by traditional approaches. This framework accommodates a range of correlation structures by allowing both time-invariant and time-varying regression parameters and factor loadings, as well as structured forms of time-dependent intercepts. Notably, this modelling flexibility is novel compared to the existing high-dimensional factor models that are not tailored to longitudinal settings. Second, we propose a method for drawing statistical inferences on covariate effects and establish its statistical theory. In particular, we give conditions on the latent variables that ensure the identifiability of the covariate effects and further establish a central limit theorem that ensures the proposed inference method is asymptotically valid and efficient. Third, our theoretical result can be applied to the statistical inference of generalised latent factor models with covariates, which, to our knowledge, have not been previously explored. It thus may be of independent theoretical interest.

The rest of the chapter is organized as follows. Section 2.2 introduces a factor model for high-dimensional longitudinal data and proposes a likelihood-based estimator, along with several extensions and variants. Section 2.3 establishes the theoretical properties of the proposed estimator. Specifically, a central limit theorem is established for statistical inference on regression coefficients, and an information criterion is introduced for choosing the number of factors. The proposed method is evaluated by simulation studies in Section 2.4 regarding its finite sample performance and is further applied to a grocery shopping dataset in Section 2.5 for understanding and predicting customers' shopping behaviour. The chapter is concluded with discussions in Section 2.6. A software implementation for R is available at https://github.com/Arthurlee51/LVHML. Further details about the computation and proofs of the theoretical results are given in the supplementary material in Appendix A.

2.2 Proposed Method

2.2.1 Setting and Proposed Model

Consider multivariate longitudinal data with N individuals and J outcome variables observed on discrete time points $t = 1, ..., T$. Let $Y_i = (y_{ijt})_{j=1,...,J,t=1,...,T}$ be a $J \times T$ data matrix for each individual i, where y_{ijt} is a random variable indicating the measurement of the jth outcome at time t. We further use the vector $\mathbf{y}_{it} = (y_{i1t}, \ldots, y_{iJt})^\top$ to denote all of the individual's outcomes at time t. Besides the measured outcomes, a set of p covariates are collected for each individual i, denoted by $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^\top$. For ease of exposition, we assume the covariates to be static, and an extension to time-dependent covariates will be discussed in Section 2.2.3. Furthermore, to account for missing observations, we let r_{it} be a missing indicator for individual i at time t, where $r_{it} = 1$ if $\mathbf{y}_{it}$ is observed and $r_{it} = 0$ otherwise. We then partition Y_i into Y_i^o and Y_i^m, where Y_i^o contains those $\mathbf{y}_{it}$ for which $r_{it} = 1$ and Y_i^m contains the remaining components. Let $\mathbf{r}_i = (r_{i1}, \ldots, r_{iT})^\top$

denote the vector of the individual's missing indicators. We observe independent and identically distributed (i.i.d.) copies of the triplet Y_i^o, $\mathbf{r}_i$ and $\mathbf{x}_i$, $i = 1, \ldots, N$.

In this context, we propose a high-dimensional factor model to achieve two goals. First, we hope to draw statistical inferences on how the covariates affect each outcome variable based on the trained model. Second, given the up-to-date information, we hope to train a model and use it to predict future outcome variables y_{ijt} at time $(T+1)$, $i = 1, ..., N$, $j = 1, ..., J$. We introduce individual-specific random variables $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{iK})^\top$, also known as the latent factors, to capture the within-individual data dependence unexplained by the covariates, where K is a pre-specified number of factors. The latent dimension K is assumed to be small relative to N and J, but can still be large in absolute value.

Suppose that each of $y_{ijt}, i = 1, \ldots, N, j = 1, \ldots, J, t = 1, \ldots, T$, follows an exponential family distribution with natural parameter $\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i$, and possibly a scale dispersion parameter ϕ_j. Here γ_{jt}, $\mathbf{a}_j = (a_{j1}, \ldots, a_{jK})^\top$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^\top$ are item-specific parameters. Specifically, γ_{jt} is a variable- and time-specific intercept capturing the baseline intensity, $\mathbf{a}_j$ is a vector of the loading parameters, and $\boldsymbol{\beta}_j$ contains the regression coefficients. More precisely, the probability density/mass function for y_{ijt} takes the form

$$\begin{aligned} & f(y_{ijt} \mid \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i, \phi_j) \\ = & \exp\left(\frac{y_{ijt}(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i) - b_j(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i)}{\phi_j} + c_j(y_{ijt}, \phi_j) \right), \end{aligned} \tag{2.1}$$

where $b_j(\cdot)$ and $c_j(\cdot)$ are pre-specified variable-specific functions that are determined by the choice of the exponential family distribution. This model assumption allows us to model outcome variables of mixed types, including binary, count and continuous data. For example, for a binary variable, (2.1) leads to a logistic model where

$$P(y_{ijt} = 1 \mid \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i) = \frac{\exp(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i)}{1 + \exp(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i)}, \tag{2.2}$$

for which $b_j(\cdot) = \log(1 + \exp(\cdot)), \phi_j = 1$ and $c_j(\cdot, \cdot) = 0$. For count data, (2.1) gives

$$P(y_{ijt} = y \mid \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i) = \frac{\exp(y(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i) - \exp(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i))}{y!}, \tag{2.3}$$

a Poisson model for which $b_j(\cdot) = \exp(\cdot), \phi_j = 1$ and $c_j(y, \phi_j) = -\log(y!)$.

For each individual i, we assume that y_{ijt}s are conditionally independent given the latent variables $\boldsymbol{\theta}_i$ and covariates $\mathbf{x}_i$. Furthermore, we assume the missing outcome variables to be MAR, such that the missing indicator $\mathbf{r}_i$ is conditionally independent of the unobserved data Y_i^m given the observed data Y_i^o. The MAR assumption is widely adopted in the longitudinal analysis literature, as it allows the likelihood to be expressed conveniently for estimation, while being less restrictive and generally more realistic than the missing completely at random (MCAR) assumption, which requires missingness to be independent of both observed and unobserved data. Nevertheless, the MAR assumption cannot be verified statistically, as it concerns unobserved outcomes, and violations may result in biased estimates.

We denote $A = (a_{jk})_{J\times K}$ as the loading matrix, $\Theta = (\theta_{ik})_{N\times K}$ as the matrix for factor scores, and $X = (x_{il})_{N\times p}$ as the covariate matrix. To ensure the identifiability of the regression coefficients $\boldsymbol{\beta}_j$, we impose the restriction

$$\Theta^\top X = 0_{K\times p}, \tag{2.4}$$

where $0_{K\times p}$ is a $K \times p$ matrix with all the entries being zero. This constraint requires the latent factors to be uncorrelated with the observed covariates, coinciding the assumption in traditional random effects models that random effects and regressors are orthogonal.

We provide several remarks on the model. First, it is assumed that the within-individual dependence, both between items and across time, is completely captured by covariates $\mathbf{x}_i$ and low-dimensional factors $\boldsymbol{\theta}_i$. Conditioning on these variables, the outcome variables are independent, and the right-hand side of (2.1) does not depend on the outcomes at other time points. Second, one can regard the latent variables $\boldsymbol{\theta}_i$ as random effects capturing unobserved within-individual heterogeneity. Traditional latent variable models for multivariate longitudinal data typically impose a parametric distributional assumption (e.g., normality) on the latent variables and then performs statistical inferences based on a marginal likelihood where the latent variables are marginalized out. While this approach works well for low-dimensional latent variable models, it becomes computationally challenging when the latent dimension becomes moderately large (Chapter 6, Skrondal and Rabe-Hesketh, 2004) and is not suitable for the current setting where K can be large. In this chapter, we adopt an approach commonly applied in high-dimensional factor models (Chen et al., 2020, 2021; Liu et al., 2023a), optimizing an objective function involving both the fixed parameters such as γ_{jt}, $\mathbf{a}_j$ and $\boldsymbol{\beta}_j$ and the latent variables $\boldsymbol{\theta}_i$, without requiring distributional assumption on the latent variables. Third, while the current approach assumes K to be fixed, it is of theoretical interest to relax this assumption and allow K to diverge with N and J, as in Chen and Li (2022). We leave this extension for future research. Fourth, except for $\boldsymbol{\beta}_j$, the rest of the unknown parameters in (2.1) are not identifiable without additional constraints. For example, one can add a constant to each entry of Θ and compensate it by adjusting the intercepts γ_{jt}, without changing the density/probability (2.1). We note that similar indeterminacy phenomena are common in factor analysis, and the identification of these parameters can be ensured by imposing additional constraints similar to those in Bai and Li (2012). This indeterminacy does not affect making predictions but affects the interpretation of the factors and the inference of the corresponding loading parameters. As we are mainly interested in drawing statistical inferences on the regression coefficients, we do not impose constraints to fix the rotational indeterminacies. As will be shown in Section 2.2.2 below, the estimate of $\boldsymbol{\beta}_j$ is consistent and asymptotically normal, regardless of the identification of the rest of the parameters.

Finally, although the covariates, latent variables, and most of the parameters are assumed to be time-independent in (2.1), we can extend our model to allow them to be time-dependent. Some of such extensions are discussed in Sections 2.2.3, 2.2.4 and 2.6, respectively. However, we should note that these extensions also introduce more model parameters, which may lead to a higher variance in prediction and additional challenges with interpretations.

2.2.2 Estimation

We consider the estimation of the proposed model based on the joint log-likelihood function

$$l(\Xi) = \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T} r_{it}\left\{y_{ijt}(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i) - b_j(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i)\right\}, \qquad (2.5)$$

where Ξ is a vector containing unknown quantities including $\gamma_{jt}, \mathbf{a}_j, \boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_i$, for $i = 1, \ldots, N, j = 1, \ldots, J, t = 1, \ldots, T$. To estimate the regression coefficients $\boldsymbol{\beta}_j$, we maximize $l(\Xi)$ with respect to Ξ under certain compactness constraints on the model parameters. More specifically, let $\|\cdot\|$ denote the Euclidean norm, we solve the optimization problem

$$\begin{gathered}\hat{\Xi} = \arg\max_{\Xi} l(\Xi) \text{ such that for } i = 1, \ldots, N, j = 1, \ldots, J, \\ \|\boldsymbol{\theta}_i\| \leq c_1\sqrt{K}, \text{and } \|(\mathbf{a}_j^\top, \boldsymbol{\gamma}_j^\top, \boldsymbol{\beta}_j^\top)^\top\| \leq c_2\sqrt{T+p+K},\end{gathered} \qquad (2.6)$$

where c_1 and c_2 are two constraint parameters, and $\boldsymbol{\gamma}_j = (\gamma_{j1}, ..., \gamma_{jT})^\top$. Numerically, the compactness constraints prevent parameters from taking extreme values, which may happen when observed variables are discrete and certain categories are rarely observed. Theoretically, this constraint plays a crucial role in establishing the estimation consistency; see Section 2.3 for the details. This optimization problem is solved by a projected gradient descent algorithm (Chapter 4, Bertsekas, 1999) that is guaranteed to converge to a critical point. See Section A.1 for the computational details.

We call (2.5) a joint log-likelihood function to distinguish it from the marginal log-likelihood function used in the traditional latent variable models, as the current log-likelihood function involves both the model parameters and the latent factors. We also note that strictly speaking, the full joint log-likelihood function takes a form slightly different from $l(\Xi)$ in (2.5), given by $\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T} r_{it}\log f(y_{ijt} \mid \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i, \phi_j)$, as $l(\Xi)$ ignores all the scale parameters ϕ_j. They coincide (up to a constant difference) when all the outcome variables are binary or count variables that follow the Bernoulli or Poisson models in (2.2) and (2.3), respectively. The proposed estimator is suitable when all the scale parameters are close to each other. When the scale parameters are heterogeneous, we may use the full joint log-likelihood function to jointly estimate Ξ and all the scale parameters, for which the asymptotic results established in Section 2.3 can be adapted accordingly.

2.2.3 Extension to Incorporating Time-dependent Covariates

In scenarios where each individual i is associated with a time-dependent covariate vector $\mathbf{z}_{it} = (z_{i1t}, z_{i2t}, \ldots, z_{ip_zt})^\top$ at each time point t, with corresponding regression parameters $\mathbf{v}_j = (v_{j1}, \ldots, v_{jp_z})^\top$, our model adapts accordingly. The natural parameter of the exponential family distribution can be modelled as $\gamma_{jt} + \mathbf{a}_j^\top\boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top\mathbf{x}_i + \mathbf{v}_j^\top\mathbf{z}_{it}$. The conditional probability density/mass function $f(y_{ijt} \mid \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i, \mathbf{v}_j, \mathbf{z}_{it}, \phi_j)$ then becomes $\exp\left(\phi_j^{-1}\{y_{ijt}(\gamma_{jt} + \mathbf{a}_j^\top\boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top\mathbf{x}_i + \mathbf{v}_j^\top\mathbf{z}_{it}) - b_j(\gamma_{jt} + \mathbf{a}_j^\top\boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top\mathbf{x}_i + \mathbf{v}_j^\top\mathbf{z}_{it})\} + c_j(y_{ijt}, \phi_j)\right)$.

This modification maintains the structure of the likelihood as in equation (2.5), and, thus, the estimation algorithm described in Section 2.2.2 can still be applied. By incorporating additional assumptions for time-dependent covariates, we can derive theorems akin to those established under the current model in Section 2.3. The specific assumptions and proofs of these theorems are elaborated in Sections A.3 and A.4 of the supplementary material.

2.2.4 Extension for Time-dependent Loadings and Coefficients

To better accommodate the effects of time in complex datasets, the loadings and the coefficients of the covariates may also be made time-dependent by modelling the natural parameter as $\gamma_{jt} + \mathbf{a}_{jt}^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_{jt}^\top \mathbf{x}_i$. Similar to the extension discussed in Section 2.2.3, we could also include time-dependent covariate Z_t, and the estimation procedure outlined in Section 2.2.2 can be adapted to incorporate this extension. Theoretical results analogous to those in Section 2.3 can be established. The required modifications, assumptions and proofs are detailed in Sections A.1, A.3 and A.4 of the supplementary material, respectively.

2.2.5 Imposing Dependence Structure on Intercepts

In practical data analysis, it is often desirable to impose structured dependence on $\boldsymbol{\gamma}_j$ to enhance estimation efficiency and predictive accuracy, or reflect prior knowledge. An important example is $\gamma_{jt} = t\gamma_j$, which fits naturally under the framework developed in Section 2.2.3, treating γ_j as the coefficient for the time-dependent covariate t. Consequently, the asymptotic results established for that framework continue to hold.

This structure can also be incorporated into the extension in Section 2.2.4, with additional assumptions required for valid asymptotic theory. These conditions and related proof adjustments are provided in Sections A.3.2 and A.4 of the supplementary material.

2.3 Theoretical Results

2.3.1 Consistency and Asymptotic Normality

We now establish the asymptotic properties for the estimated regression coefficients $\hat{\boldsymbol{\beta}}_j$. Since J tends to infinity, we assume that as J varies, the sequence is regression coefficient vectors is nested, in the sense that $(\boldsymbol{\beta}_1^*, \ldots \boldsymbol{\beta}_J^*)^\top$ agrees with the first J vectors in $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{J+1}^*)$. Let $\mathbf{u}_j^* = \left(\boldsymbol{\gamma}_j^{*\top}, \boldsymbol{\beta}_j^{*\top}, \mathbf{a}_j^{*\top}\right)^\top$ denote the vector of true values of item-specific parameters. Additionally, define $\mathbf{D}_{it} = (D_{it1}, \ldots, D_{itT})^\top$ as a vector of dummy variables indicating the time periods, where $D_{itt} = 1$ and $D_{itt'} = 0$ for $t \neq t', i = 1 \ldots N$. We further define $\mathbf{e}_{it}^* = (\mathbf{D}_{it}^\top, \mathbf{x}_i^\top, \boldsymbol{\theta}_i^{*\top})^\top$ as the vector of true and observed individual-specific quantities. Let K^* denote the true dimension of the latent variables $\boldsymbol{\theta}_i^*$, and $P = T + p + K^*$ denote the dimension of $\mathbf{u}_j^*$. P is assumed to be fixed that does not vary with N and J. Let $\boldsymbol{\Xi}^* = \left(\mathbf{u}_1^{*\top}, \ldots, \mathbf{u}_J^{*\top}, \boldsymbol{\theta}_1^{*\top}, \ldots, \boldsymbol{\theta}_N^{*\top}\right)^\top$ denote the vector of true parameters. Let $\mathcal{U} \subset \mathbb{R}^P$, $\boldsymbol{\Theta} \subset \mathbb{R}^{K^*}$ and define the space of possible parameters $\mathcal{H}^{K^*} = \left\{\boldsymbol{\Xi} \in \mathbb{R}^{NK^*+PJ} : \mathbf{u}_j \in \mathcal{U}, \boldsymbol{\theta}_i \in \boldsymbol{\Theta} \text{ for all } i, j,\ \Theta^\top X = 0_{K^* \times p}\right\}$. For positive sequences a_n and b_n, we write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some $C > 0$, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The following regularity conditions ensure consistency of $\hat{\boldsymbol{\beta}}_j$.

Assumption 2.1. *$\mathcal{U}$ and $\boldsymbol{\Theta}$ are compact sets and $\boldsymbol{\Xi}^* \in \mathcal{H}^{K^*}$. Moreover, $\mathbf{x}_i \in \mathcal{X}$ for all i, where $\mathcal{X} \subset \mathbb{R}^p$ is a compact set.*

Assumption 2.2. *For any compact set $\mathcal{C} \subset \mathbb{R}$, there exists $\bar{b} > \underline{b} > 0$ (depending on $\mathcal{C}$) such that $\bar{b} \geq b_j^{''}(s) \geq \underline{b}$ and $|b_j^{'''}(s)| \leq \bar{b}$ for all $s \in \mathcal{C}, j = 1, \ldots, J$. Moreover, $\{\phi_j\} \lesssim 1$.*

Assumption 2.3. *$J^{-1} A^{*\top} A^*$ converges to a positive definite matrix as J tends to infinity. Also, $N^{-1} \Theta^{*\top} \Theta^*$ converge to a positive definite matrix as N tends to infinity.*

Assumption 2.4. *There exists $\kappa_1 > 0$ such that $\inf_{i=1,\ldots,N,t=1,\ldots,T} P(r_{it} = 1) \geq \kappa_1$.*

Assumption 2.5. *There exists $\kappa_2 > 0$ such that*

$$\liminf_{N\to\infty} \pi_{\min}\left((X, \mathbf{1_N})^\top (X, \mathbf{1_N}) \right) / N \geq \kappa_2,$$

where $\pi_{\min}(\cdot)$ is the minimum eigenvalue of a matrix, $\mathbf{1_N}$ is length-N vector of ones.

The following theorem establishes the consistency of $\hat{\boldsymbol{\beta}}_j$:

Theorem 2.1. *Under Assumptions 2.1 to 2.5, for each fixed j, $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\| = o_P(1)$, as N and J grow to infinity.*

Define $B^* = (\beta_{jl}^*)_{J\times p}$ and $\mathit{\Gamma}_t^* = (\gamma_{1t}^*, \ldots, \gamma_{Jt}^*)^\top$ for $t = 1, \ldots, T$. Furthermore, let $\hat{\Theta} = (\hat{\theta}_{ik})_{N\times K^*}, \hat{A} = (\hat{a}_{jk})_{J\times K^*}$, $\hat{B} = (\hat{\beta}_{jl})_{J\times p}$ and $\hat{\mathit{\Gamma}}_t = (\hat{\gamma}_{1t}, \ldots, \hat{\gamma}_{Jt})^\top, t = 1, \ldots, T$ be the estimated parameters from $\hat{\boldsymbol{\Xi}}$. The following theorem provides the average rate of convergence of $\hat{\boldsymbol{\Xi}}$ and $\hat{B}$:

Theorem 2.2. *Under Assumptions 2.1 to 2.5, we have*

$$\max_{t=1,\ldots,T} \frac{\left\| \hat{\Theta}\hat{A}^\top - \Theta^* A^{*\top} + X(\hat{B} - B^*)^\top + \mathbf{1_N}(\hat{\mathit{\Gamma}}_t - \mathit{\Gamma}_t^*)^\top \right\|_F}{\sqrt{NJ}} = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}), \tag{2.7}$$

$$\frac{1}{\sqrt{J}} \left\| \hat{B}^\top - B^{*\top} \right\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}). \tag{2.8}$$

We comment on the rate of convergence for $J^{-1/2}\|\hat{B}^\top - B^{*\top}\|_F$. One might expect a $N^{-1/2}$ rate since B corresponds to the regression component for observed variables. However, the latent variables are also estimated here, which introduces measurement error. Specifically, the estimated latent component $\hat{\Theta}\hat{A}^\top$ has an estimation error rate of $(NJ)^{-1/2} \left\| \hat{\Theta}\hat{A}^\top - \Theta^* A^{*\top} \right\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1})$. This measurement error dominates the estimation error of the regression component, resulting in the convergence rate stated in Theorem 2.2. To establish the asymptotic normality for each $\hat{\boldsymbol{\beta}}_j$, we need two additional assumptions.

Assumption 2.6. *The limits $\mathit{\Phi}_j = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T -\phi_j^{-1} E(r_{it})\, b_j^{''}(\mathbf{u}_{jt}^{*\top}\mathbf{e}_{it}^*)\mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}$ and $\mathit{\Psi}_i = \lim_{J\to\infty} \frac{1}{J} \sum_{j=1}^J \sum_{t=1}^T -\phi_j^{-1} E(r_{it})\, b_j^{''}(\mathbf{u}_{jt}^{*\top}\mathbf{e}_{it}^*)\mathbf{a}_j^*\mathbf{a}_j^{*\top}$ exist for $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Moreover, there exists $\kappa_3 > 0$ such that $\pi_{\min}(\mathit{\Phi}_j^\top \mathit{\Phi}_j) \geq \kappa_3$ and $\pi_{\min}(\mathit{\Psi}_i^\top \mathit{\Psi}_i) \geq \kappa_3$.*

Assumption 2.7. *As $N, J \to \infty$, $N \asymp J$.*

Theorem 2.3. *Under Assumptions 2.1 to 2.7 , for each fixed j, we have* $\sqrt{N}\left(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{E,j})$, *where the asymptotic variance* $\Sigma_{E,j} = (-\Phi_j^{-1})_{(T+1):(T+p),(T+1):(T+p)}$ *is a submatrix of* $-\Phi_j^{-1}$ *that corresponds to its* $T+1$ *to* $(T+p)$*th rows and columns.* $\Sigma_{E,j}$ *is uniquely determined by the true model without being affected by the indeterminacy of* γ_{jt}, $\mathbf{a}_j$, *and* $\boldsymbol{\theta}_i$.

Theorem 2.3 establishes the asymptotic normality of $\hat{\boldsymbol{\beta}}_j$, and specifies the form of the asymptotic variance. This shows that $\hat{\boldsymbol{\beta}}_j$ is efficient, as its asymptotic variance matches the maximum likelihood estimator in generalised linear model regression, where the latent factors $\boldsymbol{\theta}_i^*$ are directly observable. Assumptions 2.1, 2.3 and 2.6 are standard in the literature of factor analysis (see e.g., Bai and Ng, 2002 and Bai, 2003). Assumption 2.2 concerns the regularity conditions of the exponential family. The condition regarding the derivatives of $b_j(\cdot)$ is straightforward to verify and applies to a wide range of commonly used models under the exponential family, including the logistic model for binary data and Poisson model for count data. The condition for the scale parameter is a mild assumption ensuring the variance of y_{ijt} does not explode. Assumption 2.4 ensures that a sufficient proportion of outcomes are observed so that the effective sample size grows with N. In many applications, this condition is not restrictive: as long as each individual contributes at least one observed outcome, and there is no reason to assume zero probability of observation at other time points, the assumption is naturally satisfied. Nevertheless, it may be possible to relax it by instead requiring that the average observation rate $N^{-1}\sum_{i=1}^N r_{it}$ is bounded away from zero for each t. Such weaker conditions would also guarantee a growing effective sample size, though establishing the corresponding asymptotic theory would require a different proof strategy and is not pursed here. Assumption 2.5 is a condition that guarantee a degree of variability in the values of X. Assumption 2.7 ensures that N and T grow at the same rate. Similar assumptions are made when deriving asymptotic normality for high-dimensional factor models; see e.g., Bai and Li (2012), Galvao and Kato (2016a) and Chen et al. (2021).

Remark 2.1. *In practice, asymptotic variance is unknown and needs to be estimated. Define* $\hat{\mathbf{e}}_{it} = (\mathbf{D}_{it}^\top, \mathbf{x}_i^\top, \hat{\boldsymbol{\theta}}_i^\top)^\top$. *We can estimate* Φ_j *by*

$$\hat{\Phi}_j = N^{-1}\sum_{i=1}^N\sum_{t=1}^T -\hat{\phi}_j^{-1} r_{it} b_j^{''}(\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^\top,$$

and estimate $\hat{\Sigma}_{E,j}$ *by the corresponding submatrix. We show in Section A.4.6 in the supplementary material that* $\hat{\Sigma}_{E,j}$ *is a consistent estimator for the true asymptotic variance* $\Sigma_{E,j}$, *where* $\hat{\phi}_j$ *is any consistent estimator of the scale parameter* ϕ_j.

Remark 2.2. *Under the conditions of Theorem 2.3, the established asymptotic normality for each* $j \in \{1, \ldots, J\}$ *implies a uniform convergence rate of* $\max_{1\le j\le J}\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\| = O_P\left(N^{-1/2}\sqrt{\log J}\right)$. *This result follows by applying a union bound over* $j = 1, \ldots, J$ *to the sub-Gaussian tail probabilities of* $\hat{\beta}_{jl} - \beta_{jl}^*$ *for* $l = 1, \ldots, p$, *noting that* p *is assumed fixed.*

Remark 2.3. *The asymptotic results in this section are derived under a double-diverging regime where both the number of individuals* N *and the number of outcomes per individual* J *tend to infinity. An interesting question, as pointed out by the examiner, is whether one* *can consistently estimate the regression coefficients $\boldsymbol{\beta}_j$ under weaker growth conditions, for example in the large N, fixed J regime, by treating the latent variables as nuisance parameters. Exploring such restricted settings would help clarify the extent to which the double-diverging assumption is essential, and we regard this as an important direction for future research.*

A related question is whether the assumption of fixed T can be relaxed to allow T to diverge. We conjecture that the main results would continue to hold in this setting, since having more time points provides additional information, while only the number of time-dependent intercepts increases with T. In fact, sharper convergence rates may be achieved in some of the theorems to reflect the diverging nature of T. A precise characterization of the rates and the corresponding proofs, however, are beyond the scope of this work and are left for future research.

2.3.2 Determining the Number of Factors

In real application, the true number of factors K^* is unknown and, thus, needs to be estimated. To do so, we consider a finite set $\mathcal{K}$, containing the candidate numbers of factors. For each value of $K \in \mathcal{K}$, we estimate the proposed model and obtain the estimate $\hat{\Xi}_K$ and the corresponding log-likelihood function value, $l(\hat{\Xi}_K)$. We then construct an information criterion taking the form $\mathrm{IC}(K) = -2l(\hat{\Xi}_K) + K\Lambda_{NJ}$, where Λ_{NJ} is a penalty term to be discussed in the sequel. We then set

$$\hat{K} = \underset{K\in\mathcal{K}}{\arg\min}\, \mathrm{IC}(K). \tag{2.9}$$

Theorem 2.4. *Suppose that Assumptions 2.1 to 2.5 hold and $K^* \in \mathcal{K}$. If the penalty term Λ_{NJ} satisfies $\max\{N, J\} \lesssim \Lambda_{NJ} \lesssim NJ$, then $\lim_{N,J\to\infty} P(\hat{K} = K^*) = 1$.*

This result is an extension of an information criterion for a generalised latent factor model proposed by Chen and Li (2022) to the current model. Following the choice in Chen and Li (2022). we set

$$\Lambda_{NJ} = \max\{N, J\} \times \log\left(\max\{N, J\}^{-1} J \sum_{i=1}^{N} \sum_{t=1}^{T} r_{it}\right)$$

in implementation, where $J\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}$ records the total number of data points being observed. It is easy to see that the requirement on Λ_{NJ} is satisfied with this choice. We note that this choice of Λ_{NJ} lies toward the softer end of the penalty range allowed by the theory. This reduces the risk of discarding factors with weaker signals, which might otherwise be overlooked under a stronger penalty in a practical setting. We adopt this choice since the main focus of this chapter is on the estimation of regression coefficients and prediction performance.

Remark 2.4. *There exists multiple approaches for determining the number of factors beyond the information criterion adopted here. In particular, Chapter 3 introduces a stability-based estimator designed for linear factor models, which focuses on selecting loading structures that are reproducible across data splits. However, the theoretical results based on random matrix theory cannot be directly applied to this setting. Extending such a stability-based approach to the current setting would be a promising future research direction.*

2.4 Simulation Study

2.4.1 Simulation Setting

We assess the finite sample performance of the proposed method via Monte Carlo simulations under a variety of settings. Specifically, we consider $J = 100, 200, 300, 400$ with $N = 5J$ or $10J$, yielding eight combinations. For each setting, we generate 100 replications with the number of time points $T = 4$ and true latent dimensions $K^* = 3$ and 8. Model selection is performed over the candidate set $\mathcal{K} = \{1, 2, \ldots, 10\}$.

We simulate data in a binary response setting to mimic the real data example. The simulated data follows the logistic model given in (2.2):

$$P(y_{ijt} = 1 \mid \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i) = \frac{\exp(\gamma_{jt} + \sum_{k=1}^{K^*} a_{jk}\theta_{ik} + \sum_{l=1}^{5} \beta_{jl}x_{il})}{1 + \exp(\gamma_{jt} + \sum_{k=1}^{K^*} a_{jk}\theta_{ik} + \sum_{l=1}^{5} \beta_{jl}x_{il})}. \tag{2.10}$$

The variables are generated as follows, slightly abusing notation by using the same symbols before and after normalisation. Intercepts γ_{jt} are independently sampled from a uniform distribution $U[-1, 1]$, and regression coefficients β_{jl} from $U[0.5, 1]$. The latent variables θ_{ik} and a_{jk} are sampled from truncated standard normal distributions on $[-3, 3]$. The covariates (x_{i1}, x_{i2}) and (x_{i3},x_{i4}) are two pairs of dummy variables, each derived independently from a binomial distribution Bin$(2, 0.5)$. The last covariate x_{i5} is sampled from $U[-1, 1]$. The normalisation procedures described in Section A.2 of the supplementary material is applied to ensure identifiability of regression coefficients. We further set half of normalised coefficient pairs (β_{j1}, β_{j2}) to zero. We independently repeat the same procedure for (β_{j3},β_{j4}), and separately set half of β_{j5} to zero. The missingness indicator $\mathbf{r}_i$ is sampled from all possible binary combinations of 0 and 1 with equal probability, excluding the all-zero case. It leads to approximately 47% of the values in r_{it} being 0.

2.4.2 Evaluation Criteria

The performance of the proposed estimator is assessed based on several performance metrics, as given in Table 2.1. Specifically, in each replication, at the true number of factors K^*, we compute the "Loss" metric defined on the left-hand side of (2.7) to evaluate the convergence of $\hat{\Xi}$ in finite sample. Additionally, we compute "Bloss" as defined on the left-hand side of (2.8) to quantify the convergence of $\hat{B}$. The mean "Loss" and "Bloss" across 100 simulations are reported in Table 2.1.

To further assess the estimator's performance on individual parameters, the mean squared error (MSE) for each β_{jl}, where $j = 1, \ldots, J$ and $l = 1, \ldots, 4$, is computed across all trials. The maximum of these MSE values is reported as "MMSE". Additionally, the proportion of instances where the correct number of factors is accurately identified is denoted by $P(\hat{K} = K^*)$. The asymptotic variance for each simulation is estimated following the methodology proposed in Remark 2.1, based on which 95% confidence intervals for β_{jl}s are constructed. The empirical coverage probability (ECP) is then determined by aggregating the coverage probabilities across all parameters and simulation repetitions.

In addition, the recovery of the coefficients $\boldsymbol{\beta}_j$s is evaluated against a baseline approach that assumes no factors, that is, $K = 0$. The corresponding likelihood is given by $\prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{t=1}^{T} \left\{\exp((\gamma_{jt} + \boldsymbol{\beta}_j^\top \mathbf{x}_i)y_{ijt})(1 + \exp(\gamma_{jt} + \boldsymbol{\beta}_j^\top \mathbf{x}_i))^{-1}\right\}^{r_{it}}$. The optimization is carried out using the `glm` function in R, leveraging a logistic regression (LR) approach.

Additionally, we compare the proposed method to a logistic regression model incorporating a random intercept α_{ij}, where for each j, the α_{ij} are assumed to be identically and independently distributed as normal random variables. The likelihood for this model is $\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{t=1}^{T}\left\{\exp((\gamma_{jt}+\alpha_{ij}+\boldsymbol{\beta}_j^\top\mathbf{x}_i)y_{ijt})(1+\exp(\gamma_{jt}+\alpha_{ij}+\boldsymbol{\beta}_j^\top\mathbf{x}_i))^{-1}\right\}^{r_{it}}$. This model is optimized using the `glmer` function from the `lme4` package in R. The results, including the metrics "Bloss" and "MMSE" from both approaches, are reported in Table 2.1.

Moreover, as we try to draw statistical inferences on a large number of regression coefficients, it is essential to control for multiple testing. We report the mean false discovery rate (FDR) over 100 simulations using the Benjamini–Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001), which is valid under arbitrary dependence among hypotheses. Since x_{i1}, x_{i2} are dummy variables for a single covariate, we test the hypotheses $H_{0j}:\beta_{j1}=\beta_{j2}=0$ for $j=1,\ldots,J$. The Wald test is applied using the estimator for asymptotic variance derived in Theorem 2.3, as detailed in Remark 2.1. We reject hypotheses at a significance level of 0.05 based on the BY-adjusted p-values. The same procedure is applied to the coefficients associated with x_{i3} and x_{i4}. For the continuous covariate x_{i5}, we test $H_{0j}:\beta_{j5}=0$ for each $j=1,\ldots,J$. For each covariate, we compute the mean FDR (MFDR) across the 100 replications and report the maximum as "MMFDR" in Table 2.1. In addition, we report the maximum of the mean false non-discovery rates (MMFNR), as the proportion of true alternative hypotheses that are incorrectly not rejected, among all hypotheses not rejected. The results in Table 2.1 are based on setting the constraint parameters in Equation (2.6) to $c_1=c_2=5$. To assess sensitivity, we repeat the simulations with $c_1=c_2=c$ for $c\in\{3,4,5,6,7\}$. Results are shown in Table A.2 of the supplementary material.

2.4.3 Results

The simulation results align with the theory for the proposed method. As N and J increase, the metrics "Loss", "Bloss", and "MMSE" under the proposed method show a decreasing trend, which occurs regardless of the number of factors and the ratio between N and J. Moreover, the information criterion for determining the true number of factors K^*, introduced in Section 2.3.2, proves to be effective. This is evidenced by $P(\hat{K}=K^*)$ achieving 1 in every scenario, which means our method always identifies the correct number of factors. Additionally, as N and J increase, the empirical coverage probability (ECP) approaches the nominal 95% confidence interval level. This validates our asymptotic normality results in Theorem 2.3 and the asymptotic variance estimator in Remark 2.1. Moreover, "MMFDR" remain below the 0.05 significance level across all settings, confirming that the BY procedure successfully controls the false discovery rate. The metric "MMFNR" decreases as N and J increases, suggesting that our approach yields high power asymptotically.

In comparison, the "Bloss" and "MMSE" metrics under the logistic regression (LR) method are consistently higher than those of the proposed method. In addition, they do not further improve as N and J increase when they are sufficiently large, suggesting that this simplified model suffers from a large bias. In contrast, the logistic regression model with a random intercept (LRRI) outperforms the basic LR model by accounting for the effects of unobserved random intercepts. However, our proposed method continues to demonstrate superior performance as J and N increase, highlighting the importance of considering correlations among different outcomes to achieve optimal results.

Finally, Table A.2 indicates that the proposed estimator has stable performance across

Table 2.1: Summary statistics for the simulation study. The results for the proposed method, logistic regression (LR), and logistic regression with a random intercept (LRRI) across different combinations of N, K^* and J are reported.

		Proposed							LR		LRRI	
N	J	Loss	$P(\hat{K} = K^*)$	ECP	MMFDR	MMFNR	Bloss	MMSE	Bloss	MMSE	Bloss	MMSE
$K^* = 3$												
5J	100	0.55	1	0.94	0.01	0.20	0.49	0.14	0.545	0.512	0.485	0.137
5J	200	0.36	1	0.95	0.01	0.05	0.32	0.05	0.453	0.426	0.338	0.099
5J	300	0.29	1	0.95	0.00	0.00	0.25	0.03	0.427	0.407	0.268	0.104
5J	400	0.24	1	0.95	0.00	0.02	0.22	0.03	0.398	0.328	0.237	0.056
10J	100	0.48	1	0.94	0.01	0.01	0.33	0.06	0.469	0.400	0.333	0.063
10J	200	0.32	1	0.95	0.01	0.00	0.22	0.03	0.409	0.380	0.233	0.047
10J	300	0.26	1	0.95	0.00	0.00	0.18	0.01	0.399	0.388	0.194	0.030
10J	400	0.22	1	0.95	0.00	0.00	0.16	0.01	0.418	0.379	0.185	0.059
$K^* = 8$												
5J	100	1.26	1	0.91	0.02	0.24	0.64	0.27	0.700	0.869	0.614	0.371
5J	200	0.68	1	0.94	0.01	0.10	0.39	0.09	0.671	0.787	0.432	0.245
5J	300	0.52	1	0.94	0.01	0.07	0.31	0.06	0.642	1.059	0.366	0.403
5J	400	0.44	1	0.94	0.01	0.02	0.26	0.04	0.629	0.788	0.317	0.434
10J	100	1.12	1	0.91	0.02	0.11	0.45	0.15	0.672	0.955	0.482	1.119
10J	200	0.63	1	0.94	0.01	0.03	0.28	0.05	0.636	0.653	0.319	0.168
10J	300	0.48	1	0.94	0.01	0.00	0.22	0.03	0.613	0.582	0.260	0.113
10J	400	0.41	1	0.94	0.01	0.00	0.18	0.02	0.592	0.492	0.238	0.123

Loss: Frobenius loss measuring the convergence of $\hat{\Xi}$.
P($\hat{K} = K^*$): Proportion of instances where the correct number of factors is identified.
ECP: Empirical coverage probability of the confidence intervals.
MMFDR: Maximum mean false discovery rate across all covariates.
MMFNR: Maximum mean false non-discovery rate across all covariates.
Bloss: Frobenius loss measuring convergence of $\hat{B}$.
MMSE: Maximum mean squared error across all estimated β_{jl}s.

different choices of the constraint parameter c. This suggests that the method is not overly sensitive to the choice of constraint values. Therefore, we adopt $c_1 = c_2 = 5$ in Section 2.5.

2.5 Application to Grocery Shopping Data

2.5.1 Background

We illustrate the proposed method via an application to a grocery shopping dataset. This dataset encompasses household-level transactions over a span of two years from approximately $2,000$ frequent shoppers. It includes purchases made by each household, recorded daily, alongside demographic information such as age groups, household sizes, and income levels (recorded as categorical variables) for around 800 households. We focus on the subset of customers with demographic information to understand how customers' shopping behaviour is associated with their demographic variables and evaluate prediction performance based on latent factors and demographic variables. The dataset is available at https://www.dunnhumby.com/source-files.

In this analysis, daily transaction data are aggregated into 25 four-week periods, using the first $T = 24$ intervals for statistical analysis and model training. The 25th interval is reserved for assessing the predictive performance of our proposed model. We focus on the transactions involving the most popular J items during the first T intervals, with N denoting the count of customers who purchased any of the J items within these periods. In each time period t, let y_{ijt} be a binary indicator of purchase such that $y_{ijt} = 1$ if individual i purchased item j and $y_{ijt} = 0$ otherwise. The missing indicator, r_{it}, is set to 0 when the ith customer did not purchase any item, including those outside the J item list.

We introduce a covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^\top$ capturing household sizes and income levels through dummy variables. Here, $x_{i1} = 1$ indicates two-member households,

$x_{i2} = 1$ for three or more members, $x_{i3} = 1$ for incomes between \$35,000 and \$74,999, and $x_{i4} = 1$ for incomes above \$75,000. The baseline level with $x_{i1} = x_{i2} = 0$ for size and $x_{i3} = x_{i4} = 0$ for income represents single-member households earning below \$35,000.

2.5.2 Statistical Inference

We first focus on inferring the effects of covariates on customers' shopping behaviour. In this analysis, we focus on the most popular 100 items, i.e., $J = 100$. The number of observations is $N = 800$. We set the candidate set $\mathcal{K} = \{1, 2, ..., 15\}$ when selecting the number of factors. Using the proposed information criterion, we obtain $\hat{K} = 8$. We perform statistical inference under the eight-factor model.

We start with an overall significance test for all the covariates to see if any of the covariates are associated with customers' shopping behaviour. That is, we test the null hypothesis of $B = 0_{J\times 4}$. We use $\|\hat{B}\|_{\mathrm{F}}$ as the test statistic and perform a permutation test to obtain its reference distribution under the null hypothesis. Specifically, we perform 500 random permutations indexed by $l = 1, ..., 500$. In each permutation l, we randomly shuffle customers' covariates and then estimate the model parameters. Let the estimate of B be denoted by $\hat{B}^{(l)}$. The reference distribution for the test statistic is then obtained by the empirical distribution for $\|\hat{B}^{(l)}\|_F$, $l = 1, ..., 500$. This results in a p-value < 0.001, suggesting that these covariates are significantly associated with customers' shopping behaviour.

We move on to assess the influences of covariates on individual items. For each item j, we calculate the p-values associated with the null hypotheses of $\beta_{j1} = \beta_{j2} = 0$ and $\beta_{j3} = \beta_{j4} = 0$, respectively, for all j. These hypotheses test the effects of household income and size on the likelihood of purchasing item j, respectively. P-values are derived through Wald tests, utilizing the estimated coefficients and the asymptotic variance $\hat{\Sigma}_{E,j}$, as elaborated in Remark 2.1. To account for multiple testing, we adjust the p-values using the BY procedure for FDR control, as discussed in Section 2.4. This adjustment is carried out separately for the covariates of household income and size, enabling the identification of items significantly associated with each at the predetermined FDR threshold of 5%.

Our analysis examines the influence of household income and size on the purchase patterns of grocery items, where the items are categorized into six groups – Vegetables, Dairy and Eggs, Beverages, Fruits, Bakery and Miscellaneous items. Table 2.2 gives the items selected by the BY procedure for the covariate household size, the corresponding regression coefficients, p-value, BY-adjusted p-value (adj p-value), category, subcategory, average price, and package size. Table 2.3 is similar to Table 2.2 but gives the results for household income. Item details absent in the original dataset are marked as NA. The subcategory column represents the lowest level classification available within the dataset, and the price column represents the average unit price derived from all recorded transactions.

We examine the results about household size. As presented in Table 2.2, the coefficients in columns $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$—particularly those corresponding to $\hat{\boldsymbol{\beta}}_2$, which denote the coefficients for households comprising three or more individuals—suggest an overall increase in the likelihood of purchasing items compared to the baseline scenario of single-person households. This trend aligns with the expectation that households with a greater number of occupants tend to have higher consumption needs.

We then explore the effect of household income on consumer behaviour, as revealed in Table 2.3. Recall that $\hat{\boldsymbol{\beta}}_3$ and $\hat{\boldsymbol{\beta}}_4$ are the estimated coefficients for the dummy variables of middle and high household incomes, respectively. Notably, most coefficients in the fruits category are positive, suggesting a heightened health consciousness among these households in comparison to their lower-income counterparts. This hypothesis is consistent with the data in the Beverages category, where most soft drinks are associated with negative coefficients. Although there are exceptions, the vegetable category mostly displays positive coefficients, further reinforcing the trend toward healthier dietary preferences. For the Bakery category, a consistent negative trend across coefficients suggests that higher-income households are generally accepted to be less inclined to consume breakfast at home. These observations match our knowledge about how income levels are associated with dietary choices and lifestyle habits (see, e.g. French et al., 2019).

On the other hand, divergent preferences across income levels are observed in the Dairy and Egg category. Specifically, we observe two opposite trends for milk: one subset exhibits positive and increasing coefficients across $\hat{\boldsymbol{\beta}}_3$ and $\hat{\boldsymbol{\beta}}_4$, signifying a preference among higher-income households, while the other shows negative and diminishing coefficients, indicating the contrary. Notably, price and size do not account for these trends, as evidenced by the table. Further investigation is needed to explore the cause, such as brand differentiation, that drives these preferences. These observations may offer insight for further investigations to explain the differences in preferences uncovered by the current exploratory model.

2.5.3 Prediction

Beyond inference for coefficients of covariates, a natural application of such models is for predictions and recommendations. In particular, we can estimate the probabilities for the outcome variables at time $T+1$ given $\hat{\boldsymbol{\Xi}}$, assuming that the model (2.2) still holds at $t = T+1$. Due to the absence of an estimate for the time-dependent intercept $\gamma_{j,T+1}$, we substitute $\hat{\gamma}_{j,T}$ in practice. More specifically, we predict the occurrence of outcome variable j at time $T+1$ based on the predicted probability $1/(1+\exp(-(\hat{\gamma}_{jT} + \hat{\mathbf{a}}_j^\top \hat{\boldsymbol{\theta}}_i + \hat{\boldsymbol{\beta}}_j^\top \mathbf{x}_i)))$. The same prediction approach applies to the extended model in Section 2.2.4, substituting $\hat{\boldsymbol{\beta}}_{j,T}$ and $\hat{\mathbf{a}}_{j,T}$ for $\boldsymbol{\beta}_{j,T+1}$ and $\mathbf{a}_{j,T+1}$, respectively. For both the original and extended models, we also assess performance under the restriction $\gamma_{jt} = t\gamma_j$.

To assess the performance of this approach under varied settings, besides the setting where $J = 100$, we also consider $J = 200, 300$ and 400, with all scenarios having $N = 800$.

Given the nature of the dataset, the focus of our evaluations is on recommendation performance. To evaluate the performance, we compute the sensitivity, namely, the number of actual purchases in the recommendations divided by the total number of actual purchases. We propose a comparison of seven approaches. Prop, Prop (2.2.4), Prop (2.2.5), and Prop (2.2.4 & 2.2.5) correspond to model variants using: (i) the main model, (ii) the extension with time-dependent loadings and regression coefficients in Section 2.2.4, (iii) the intercept constraint $\gamma_{jt} = t\gamma_j$ introduced in Section 2.2.5, and (iv) both modifications, respectively. The number of factors $\hat{K}$ selected by each method is reported in Table A.6 in Section A.5.3 of the supplementary material. Recommendations are ranked based on the sorted predicted probabilities of the J items from corresponding model estimates. Hist ranks recommendations by the sorted cumulative purchasing frequency for each individual, resorting to random selection when ties occur. Hist-Prop follows the ranking of Hist but employs sorted predicted probabilities from Prop to resolve ties. Lastly, Hist-Hist, like Hist, ranks recommendations but uses the overall cumulative

Table 2.2: Characteristics and Estimated coefficients of selected products based on household size

Category	Subcategory	Price	Size	$\hat{\beta}_1$	$\hat{\beta}_2$	Adj p-value	p-value
Beverages	DAIRY CASE 100% PURE JUICE - O	1.36	NA	-0.29	0.03	0.000	0.000
Beverages	DAIRY CASE TEA WITH SUGAR OR S	0.99	1 GA	0.56	0.80	0.000	0.000
Beverages	SFT DRNK 2 LITER BTL CARB INCL	1.16	2 LTR	0.08	0.46	0.000	0.000
Beverages	SFT DRNK 2 LITER BTL CARB INCL	1.23	2 LTR	0.15	0.63	0.000	0.000
Beverages	SFT DRNK 2 LITER BTL CARB INCL	1.16	2 LTR	0.04	0.86	0.000	0.000
Beverages	SFT DRNK SNGL SRV BTL CARB (EX	1.07	20 OZ	-0.89	0.19	0.000	0.000
Beverages	SOFT DRINKS 12/18&15PK CAN CAR	3.41	12 OZ	0.10	0.74	0.000	0.000
Beverages	SOFT DRINKS 12/18&15PK CAN CAR	3.33	12 OZ	0.38	0.31	0.000	0.000
Beverages	SOFT DRINKS 12/18&15PK CAN CAR	3.49	12 OZ	-0.05	0.49	0.000	0.000
Beverages	SOFT DRINKS 12/18&15PK CAN CAR	3.41	12 OZ	0.24	0.65	0.000	0.000
Breakfast	HAMBURGER BUNS	0.95	12 OZ	0.44	0.76	0.000	0.000
Breakfast	HOT DOG BUNS	0.95	11 OZ	0.40	0.88	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	0.97	20 OZ	-0.08	0.73	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	1.65	20 OZ	0.09	0.65	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	1.47	24 OZ	0.48	0.85	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	0.97	20 OZ	-0.17	1.03	0.000	0.000
Breakfast	SW GDS:DONUTS	0.49	NA	-0.17	0.10	0.023	0.003
Dairy and Eggs	CHOCOLATE MILK	1.28	NA	-0.18	0.50	0.000	0.000
Dairy and Eggs	CHOCOLATE MILK	2.36	1 GA	0.16	0.60	0.000	0.000
Dairy and Eggs	CREAM CHEESE	1.55	8 OZ	0.12	0.34	0.008	0.001
Dairy and Eggs	CREAM CHEESE	0.98	8 OZ	-0.09	0.47	0.000	0.000
Dairy and Eggs	EGGS - LARGE	1.02	1 DZ	0.01	0.26	0.000	0.000
Dairy and Eggs	EGGS - LARGE	1.40	18 CT	0.02	0.35	0.000	0.000
Dairy and Eggs	EGGS - MEDIUM	0.71	1 DZ	-0.16	0.28	0.000	0.000
Dairy and Eggs	EGGS - X-LARGE	1.06	1 DZ	0.31	0.29	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.35	NA	0.25	-0.10	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.35	NA	0.03	0.23	0.001	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.43	1 GA	0.37	1.00	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.37	NA	-0.35	-0.05	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.41	1 GA	0.47	0.77	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.43	1 GA	-0.16	0.64	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.43	1 GA	0.16	0.42	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.32	NA	-0.30	0.27	0.000	0.000
Dairy and Eggs	IWS SINGLE CHEESE	2.35	16 OZ	0.07	0.47	0.000	0.000
Dairy and Eggs	IWS SINGLE CHEESE	1.89	12 OZ	0.26	0.59	0.000	0.000
Dairy and Eggs	IWS SINGLE CHEESE	1.49	12 OZ	-0.13	0.45	0.000	0.000
Dairy and Eggs	SHREDDED CHEESE	1.57	8 OZ	0.03	0.65	0.000	0.000
Dairy and Eggs	SOUR CREAMS	1.11	16 OZ	0.28	0.68	0.000	0.000
Fruits	APPLES GRANNY SMITH (BULK&BAG)	2.65	NA	0.28	0.50	0.000	0.000
Fruits	CANTALOUPE WHOLE	2.13	NA	0.31	0.15	0.002	0.000
Fruits	GRAPES RED	3.19	18 LB	0.14	0.31	0.002	0.000
Fruits	GRAPES WHITE	3.63	18 LB	0.08	0.25	0.011	0.002
Fruits	STRAWBERRIES	2.61	16 OZ	0.29	0.42	0.000	0.000
Vegetables	BEANS GREEN: FS/WHL/CUT	0.52	14.5 OZ	0.02	0.48	0.000	0.000
Vegetables	BROCCOLI WHOLE&CROWNS	1.65	NA	0.24	0.39	0.000	0.000
Vegetables	CABBAGE	1.27	14-18 CT	0.35	0.15	0.001	0.000
Vegetables	CARROTS MINI PEELED	1.57	1 LB	0.03	0.28	0.000	0.000
Vegetables	CELERY	1.33	NA	0.24	0.17	0.009	0.001
Vegetables	CORN	0.51	15.25 OZ	0.03	0.42	0.000	0.000
Vegetables	CORN YELLOW	0.36	48 CT	0.25	0.02	0.046	0.007
Vegetables	CUCUMBERS	0.70	36 CT	0.28	0.30	0.000	0.000
Vegetables	GARDEN PLUS	2.31	10 OZ	0.46	0.37	0.000	0.000
Vegetables	GARDEN PLUS	2.26	12 OZ	0.26	0.27	0.048	0.007
Vegetables	HEAD LETTUCE	0.98	24 CT	0.25	0.40	0.000	0.000
Vegetables	HEAD LETTUCE	0.99	24 CT	0.23	0.43	0.001	0.000
Vegetables	MUSHROOMS WHITE SLICED PKG	1.86	8 OZ	-0.06	0.32	0.002	0.000
Vegetables	ONIONS OTHER	0.53	48 CT	0.19	0.36	0.000	0.000
Vegetables	ONIONS SWEET (BULK&BAG)	1.17	40 LB	0.28	0.14	0.011	0.002
Vegetables	POTATOES RUSSET (BULK&BAG)	3.48	10 LB	0.34	0.40	0.000	0.000
Vegetables	POTATOES RUSSET (BULK&BAG)	2.44	5 LB	0.11	0.31	0.001	0.000
Vegetables	POTATOES SWEET&YAMS	1.81	40 LB	0.40	0.20	0.000	0.000
Vegetables	ROMA TOMATOES (BULK/PKG)	1.93	25 LB	-0.39	-0.33	0.000	0.000
Vegetables	SALAD BAR FRESH FRUIT	2.37	NA	0.15	-0.12	0.007	0.001
Vegetables	TOMATOES HOTHOUSE ON THE VINE	2.57	13 LB	0.16	-0.05	0.017	0.003
Miscellaneous	CANDY BARS (SINGLES)(INCLUDING	0.42	1.6 OZ	-0.11	0.57	0.000	0.000
Miscellaneous	POTATO CHIPS	1.91	11.5 OZ	0.52	1.00	0.000	0.000
Miscellaneous	SOUP CRACKERS (SALTINE/OYSTER)	1.05	16 OZ	0.12	0.37	0.006	0.001
Miscellaneous	TORTILLA/NACHO CHIPS	2.33	12.5 OZ	0.10	0.59	0.000	0.000
Miscellaneous	LEAN	3.31	NA	0.00	0.28	0.001	0.000
Miscellaneous	MEAT: LUNCHMEAT BULK	2.75	NA	0.32	0.77	0.000	0.000
Miscellaneous	MEAT: SAUS DRY BULK	3.29	NA	0.15	0.73	0.000	0.000
Miscellaneous	PREMIUM - MEAT	2.50	1 LB	0.42	1.19	0.000	0.000
Miscellaneous	CIGARETTES	3.56	974246 PK	-1.00	-0.35	0.000	0.000
Miscellaneous	CONDENSED SOUP	0.64	10.5 OZ	0.05	0.49	0.000	0.000
Miscellaneous	GASOLINE-REG UNLEADED	0.00	NA	-0.64	-0.09	0.000	0.000
Miscellaneous	SUGAR	2.01	4 LB	-0.10	0.37	0.000	0.000
Miscellaneous	NA	NA	NA	0.26	0.43	0.000	0.000
Miscellaneous	NA	NA	NA	0.00	0.17	0.044	0.007
Miscellaneous	NEWSPAPER	1.42	NA	-0.19	0.52	0.000	0.000
Miscellaneous	TOILET TISSUE	1.02	83.5 SQ FT	-0.21	0.34	0.000	0.000

Table 2.3: Characteristics and Estimated coefficients of selected products based on household income

Category	Subcategory	Price	Size	$\hat{\beta}_3$	$\hat{\beta}_4$	Adj p-value	p-value
Beverages	DAIRY CASE 100% PURE JUICE - O	1.36	NA	-0.33	-0.97	0.000	0.000
Beverages	DAIRY CASE TEA WITH SUGAR OR S	0.99	1 GA	-0.71	-3.23	0.000	0.000
Beverages	SFT DRNK 2 LITER BTL CARB INCL	1.16	2 LTR	-0.54	-1.35	0.000	0.000
Beverages	SFT DRNK 2 LITER BTL CARB INCL	1.18	2 LTR	-0.49	-0.41	0.001	0.000
Beverages	SFT DRNK 2 LITER BTL CARB INCL	1.23	2 LTR	-0.21	-0.62	0.000	0.000
Beverages	SFT DRNK 2 LITER BTL CARB INCL	1.16	2 LTR	-0.48	-1.72	0.000	0.000
Beverages	SFT DRNK SNGL SRV BTL CARB (EX	1.07	20 OZ	-0.54	-1.57	0.000	0.000
Beverages	SOFT DRINKS 12/18&15PK CAN CAR	3.33	12 OZ	0.21	0.96	0.000	0.000
Beverages	SOFT DRINKS 12/18&15PK CAN CAR	3.49	12 OZ	-0.24	-0.95	0.000	0.000
Beverages	SOFT DRINKS 12/18&15PK CAN CAR	3.41	12 OZ	-0.28	-1.43	0.000	0.000
Breakfast	HAMBURGER BUNS	0.95	12 OZ	-0.09	-0.48	0.000	0.000
Breakfast	HOT DOG BUNS	0.95	11 OZ	-0.22	-0.49	0.000	0.000
Breakfast	MAINSTREAM WHEAT/MULTIGRAIN BR	0.96	20 OZ	-0.24	-0.52	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	0.97	20 OZ	-0.33	-0.66	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	1.65	20 OZ	0.06	-0.45	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	1.47	24 OZ	-0.09	-1.30	0.000	0.000
Breakfast	MAINSTREAM WHITE BREAD	0.97	20 OZ	-0.82	-1.40	0.000	0.000
Dairy and Eggs	CHOCOLATE MILK	1.28	NA	-0.05	-0.30	0.001	0.000
Dairy and Eggs	COTTAGE CHEESE	2.09	24 OZ	0.36	-0.17	0.000	0.000
Dairy and Eggs	CREAM CHEESE	1.55	8 OZ	0.26	0.35	0.011	0.002
Dairy and Eggs	CREAM CHEESE	0.98	8 OZ	0.55	-0.02	0.000	0.000
Dairy and Eggs	EGGS - LARGE	1.02	1 DZ	0.27	0.44	0.000	0.000
Dairy and Eggs	EGGS - LARGE	1.40	18 CT	-0.36	-0.25	0.000	0.000
Dairy and Eggs	EGGS - MEDIUM	0.71	1 DZ	-0.19	-0.73	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.35	NA	0.45	0.71	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.35	NA	-0.14	-0.28	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.43	1 GA	-0.20	-0.28	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.37	NA	0.37	1.00	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.41	1 GA	0.06	0.44	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.43	1 GA	-0.67	-0.98	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	2.43	1 GA	0.30	1.07	0.000	0.000
Dairy and Eggs	FLUID MILK WHITE ONLY	1.32	NA	-0.48	-1.11	0.000	0.000
Dairy and Eggs	IWS SINGLE CHEESE	2.35	16 OZ	0.17	-0.55	0.000	0.000
Dairy and Eggs	IWS SINGLE CHEESE	1.89	12 OZ	-0.06	-0.54	0.000	0.000
Dairy and Eggs	IWS SINGLE CHEESE	1.49	12 OZ	-0.32	-0.72	0.000	0.000
Dairy and Eggs	SHREDDED CHEESE	1.57	8 OZ	-0.02	-0.39	0.000	0.000
Dairy and Eggs	SOUR CREAMS	1.11	16 OZ	0.32	-0.41	0.000	0.000
Fruits	APPLES GALA (BULK&BAG)	2.36	NA	0.45	1.08	0.000	0.000
Fruits	APPLES GRANNY SMITH (BULK&BAG)	2.65	NA	0.35	1.23	0.000	0.000
Fruits	BANANAS	0.95	40 LB	0.43	1.08	0.000	0.000
Fruits	CANTALOUPE WHOLE	2.13	NA	0.26	0.64	0.000	0.000
Fruits	GRAPES RED	3.19	18 LB	0.22	0.53	0.000	0.000
Fruits	GRAPES WHITE	3.63	18 LB	-0.12	0.40	0.000	0.000
Fruits	LEMONS	0.61	NA	0.36	1.19	0.000	0.000
Fruits	ORANGES NAVELS ALL	0.51	NA	-0.06	0.54	0.000	0.000
Fruits	STRAWBERRIES	2.61	16 OZ	0.26	0.95	0.000	0.000
Vegetables	BROCCOLI WHOLE&CROWNS	1.65	NA	0.20	1.09	0.000	0.000
Vegetables	CABBAGE	1.27	14-18 CT	-0.18	-0.37	0.005	0.001
Vegetables	CARROTS MINI PEELED	1.57	1 LB	0.36	0.62	0.000	0.000
Vegetables	CELERY	1.33	NA	0.13	0.37	0.000	0.000
Vegetables	CORN	0.51	15.25 OZ	0.22	-0.32	0.000	0.000
Vegetables	CORN YELLOW	0.36	48 CT	0.34	0.57	0.000	0.000
Vegetables	CUCUMBERS	0.70	36 CT	0.16	0.35	0.000	0.000
Vegetables	GARDEN PLUS	2.31	10 OZ	0.52	1.19	0.000	0.000
Vegetables	HEAD LETTUCE	0.98	24 CT	0.03	-0.18	0.020	0.003
Vegetables	MUSHROOMS WHITE SLICED PKG	1.86	8 OZ	0.01	0.93	0.000	0.000
Vegetables	ONIONS OTHER	0.53	48 CT	0.23	0.58	0.000	0.000
Vegetables	ONIONS SWEET (BULK&BAG)	1.17	40 LB	0.38	0.83	0.000	0.000
Vegetables	ONIONS SWEET (BULK&BAG)	1.04	40 LB	0.34	0.59	0.000	0.000
Vegetables	ONIONS YELLOW (BULK&BAG)	1.91	3 LB	-0.30	-0.30	0.001	0.000
Vegetables	PEPPERS GREEN BELL	0.72	48-54 CT	0.17	0.47	0.000	0.000
Vegetables	POTATOES RUSSET (BULK&BAG)	3.48	10 LB	-0.33	-0.76	0.000	0.000
Vegetables	POTATOES RUSSET (BULK&BAG)	2.44	5 LB	-0.13	-0.42	0.000	0.000
Vegetables	POTATOES SWEET&YAMS	1.81	40 LB	0.35	0.80	0.000	0.000
Vegetables	REGULAR GARDEN	1.47	1 LB	-0.14	-0.30	0.020	0.003
Vegetables	ROMA TOMATOES (BULK/PKG)	1.93	25 LB	-0.07	0.46	0.000	0.000
Vegetables	SQUASH ZUCCHINI	1.38	18 LB	0.55	1.47	0.000	0.000
Vegetables	TOMATOES GRAPE	2.59	PINT	0.11	0.79	0.000	0.000
Vegetables	TOMATOES HOTHOUSE ON THE VINE	2.57	13 LB	0.27	0.69	0.000	0.000
Miscellaneous	CANDY BARS (SINGLES)(INCLUDING	0.42	1.6 OZ	-0.09	-1.37	0.000	0.000
Miscellaneous	POTATO CHIPS	1.91	11.5 OZ	-0.12	-0.35	0.020	0.003
Miscellaneous	SOUP CRACKERS (SALTINE/OYSTER)	1.05	16 OZ	-0.52	-0.68	0.000	0.000
Miscellaneous	TORTILLA/NACHO CHIPS	2.33	12.5 OZ	-0.07	-0.70	0.000	0.000
Miscellaneous	CHICKEN BREAST BONELESS	4.46	NA	0.25	0.50	0.000	0.000
Miscellaneous	LEAN	3.31	NA	-0.44	-1.27	0.000	0.000
Miscellaneous	MEAT: LUNCHMEAT BULK	2.75	NA	0.26	-0.77	0.000	0.000
Miscellaneous	MEAT: SAUS DRY BULK	3.29	NA	0.43	0.27	0.000	0.000
Miscellaneous	PREMIUM - MEAT	2.50	1 LB	-0.14	-0.54	0.000	0.000
Miscellaneous	PRIMAL	3.81	NA	-0.02	-0.53	0.000	0.000
Miscellaneous	CIGARETTES	3.56	974246 PK	0.85	0.83	0.000	0.000
Miscellaneous	CONDENSED SOUP	0.64	10.5 OZ	0.01	-0.59	0.000	0.000
Miscellaneous	GARLIC WHOLE CLOVES	0.46	10 LB	0.18	1.15	0.000	0.000
Miscellaneous	GASOLINE-REG UNLEADED	0.00	NA	0.61	1.08	0.000	0.000
Miscellaneous	NA	NA	NA	0.29	0.40	0.000	0.000
Miscellaneous	NA	NA	NA	0.27	0.43	0.000	0.000
Miscellaneous	PAPER TOWELS & HOLDERS	0.55	57 SQ FT	0.09	-0.59	0.000	0.000
Miscellaneous	TOILET TISSUE	1.02	83.5 SQ FT	-0.29	-1.10	0.000	0.000

Table 2.4: Sensitivity Based on Number of Recommendations

	10 recommendations				20 recommendations			
	J=100	J=200	J=300	J=400	J=100	J=200	J=300	J=400
Hist	0.448	0.348	0.297	0.264	0.652	0.518	0.447	0.404
Prop	0.352	0.256	0.223	0.199	0.533	0.394	0.340	0.304
Prop (2.2.4)	0.323	0.208	0.173	0.143	0.498	0.332	0.275	0.224
Prop (2.2.5)	0.337	0.247	0.208	0.189	0.489	0.360	0.303	0.278
Prop (2.2.4 & 2.2.5)	0.297	0.197	0.164	0.143	0.462	0.317	0.262	0.231
Hist-Hist	0.451	0.350	0.299	0.266	0.654	0.519	0.450	0.406
Hist-Prop	0.456	0.352	0.301	0.269	0.659	0.525	0.453	0.407
	30 recommendations				40 recommendations			
	J=100	J=200	J=300	J=400	J=100	J=200	J=300	J=400
Hist	0.774	0.627	0.546	0.496	0.848	0.705	0.618	0.565
Prop	0.658	0.490	0.425	0.382	0.752	0.570	0.496	0.447
Prop (2.2.4)	0.624	0.430	0.355	0.292	0.723	0.509	0.424	0.346
Prop (2.2.5)	0.612	0.456	0.382	0.348	0.712	0.529	0.445	0.406
Prop (2.2.4 & 2.2.5)	0.597	0.415	0.341	0.300	0.696	0.495	0.406	0.358
Hist-Hist	0.775	0.629	0.549	0.499	0.852	0.709	0.621	0.570
Hist-Prop	0.781	0.635	0.555	0.505	0.860	0.714	0.626	0.575

frequency of items across individuals to break ties. Table 2.4 displays the results for 10, 20, 30, and 40 recommendations across different values of J for all methods.

Among the proposed models, the main model achieves the best performance. The weaker performance under the intercept constraint $\gamma_{jt} = t\gamma_j$ implies the absence of a linear trend in consumer preferences over time, which is reasonable in the context of grocery shopping data. The weaker performance of the extension with time-dependent regression coefficients and factor loadings may be due to overparameterisation. For instance, at $J = 100$, with 24 time points and 4 estimated factors, the model introduces 192 parameters per item from the loadings and regression coefficients, which may lead to overfitting. We observe that Hist generally outperforms Prop. This is not surprising as there is strong tendency for consumers to purchase the same products repeatedly in grocery shopping data (see e.g. Wan et al., 2018), which is captured effectively by the Hist method. Nevertheless, Hist-Prop emerges as the most proficient approach, indicating that our model is beneficial for improving recommendations, especially when customer information is sparse. This shows the capability of our method to borrow information from similar customers and reflect their preference for previously not purchased products. By offering personalized recommendations, this method outperforms Hist-Hist, which merely suggests the most popular items to individuals when there are insufficient individual history data. Finally, we highlight that it is possible to devise more advanced approaches based on our method to further enhance recommendations performances, especially for suggesting relevant new products to customers. For example, instead of using the proposed method to resolve ties only, we could develop more sophisticated criteria to allocate the proportions of recommendations using individual cumulative frequency and sorted predicted probability, respectively.

2.6 Discussions

This chapter concerns the analysis of high-dimensional multivariate longitudinal data. A flexible modelling framework is proposed to account for between-variable and across-time dependence by latent variables. Statistical inference procedures are developed for parameter estimation and model selection, with statistical consistency and asymptotic normality results established. The method's application to customer grocery shopping records demonstrates its ability to identify demographic influences on purchasing patterns and improve recommendation precision, revealing its value for analytical and predictive uses in practical contexts. In particular, we find a positive association between household size and the likelihood of most purchases, whereas income level is positively associated with the consumption probabilities of healthy food and inversely with soft drinks. Moreover, our model's ability to capture information from other customers' purchase behaviour allows improved recommendation performance, when combined with the information from one's purchase history.

The current research may be extended in several directions besides the extensions discussed in Sections 2.2.3, 2.2.4 and 2.2.5. First, the current analysis focuses on the regression coefficients. In many applications, especially in applications of social sciences, the substantive interpretation of the factors may be of interest. Section A.2 of the supplementary material presents normalisation criteria that allow identification of the latent factors. These results are further supported by additional simulation studies presented in Section A.5.1. These identification criteria are not unique; rotation methods (e.g., Liu et al., 2023b and Rohe and Zeng, 2023) and regularized estimation methods (e.g., Zhu et al., 2016) may be used to obtain more interpretable factors. Theoretical analysis of these method-specific criteria under our model is beyond the current scope and represents a promising direction for future research. Second, as in the extension in Section 2.2.4, which allows the loadings $\mathbf{a}_j$ to vary over time, the static factors $\boldsymbol{\theta}_i$ can be modified to be time-variant, becoming $\boldsymbol{\theta}_{it}$. This alteration would not significantly change the estimation method but would require adjustments to the normalisation criteria and assumptions to ensure the identification of the parameters $\boldsymbol{\beta}_j$, as well as $\mathbf{v}_j$ in the model with time-dependent covariates. In this direction, it is of particular interest to consider a change-point setting that assumes the time-dependent factors $\boldsymbol{\theta}_{it}$ to have a piece-wise constant structure, allowing for individual-specific change points. This model allows us to detect structural changes within each individual, based on which adaptive interventions may be made (e.g., individualized marketing strategies). In addition, by controlling for the maximum number of change points, this change point model enables us to find a balance between model flexibility and parsimony, which leads to high prediction accuracy. Finally, the computational cost for the proposed estimator becomes high or even infeasible when some or all of N, J, T and p are large. In such cases, stochastic optimization algorithms may be developed to efficiently obtain approximation solutions, and further, central limit theorems may be established for the approximate solutions to facilitate statistical inference.

Supplementary material

Appendix A presents the estimation procedure, normalisation algorithm, additional conditions and theorems for extension, and the technical proofs for main theorems.

Chapter 3

Determining number of factors under stability considerations

3.1 Introduction

Factor analysis is a widely used technique for uncovering the latent structure of multivariate data. An important problem in factor analysis is to determine the number of factors in the model, which has been studied extensively in the literature (see e.g. Bai and Ng, 2002, Onatski, 2009, Ahn and Horenstein, 2013, Bai et al., 2018, Dobriban and Owen, 2019 and Ke et al., 2023). Most of the existing methods determine the number of factors by identifying a gap in the eigenvalues of the sample covariance matrix, as the factor model structure leads to some outlier eigenvalues, where the number of such eigenvalues equals the number of factors.

This chapter proposes a new method for determining the number of factors. This method is based on the concept of loading instability, which concerns the instability of the estimated loading matrix when there are multiple copies of data. As factors receive their interpretation based on the loading matrix (see e.g. Liu et al., 2023b and Rohe and Zeng, 2023), this concept is important to factor analysis, especially the reproducibility of the learned factors. However, given the symmetric roles that loadings and factor scores play in a factor model, the instability of factor scores can be defined similarly, based on which the proposed method can be adapted accordingly. More specifically, loading instability is defined by the principal angle of two independent loading matrix estimates. Intuitively, the loading instability tends to be high when the number of factors is over-specified, as the estimated loading space contains spurious directions that correspond to singular vectors of a noise matrix. On the other hand, due to the presence of the eigengap, the loading instability tends to be low when the number of factors is correctly specified, as implied, for example, by the Davis-Kahan theorem (see e.g. Yu et al., 2015 and O'Rourke et al., 2018). Making use of this property, we introduce several statistical criteria for determining the number of factors and prove that they are consistent under an asymptotic regime that is not covered by those adopted in many existing methods, including Bai and Ng, 2002 and Bai et al., 2018. The consistency is obtained using results from random matrix theory, especially the complete delocalization of non-outlier eigenvectors (Bloemendal et al., 2016). The superiority of the proposed criteria in selecting the correct number of factors, compared to existing selection criteria, is demonstrated through simulations. Additionally, the ability of the proposed criteria to preserve loading stability is illustrated through a real data example.

Stability is a core principle of data science (Yu and Kumbier, 2020), which is important to the reproducibility of scientific discoveries. This principle has been applied to several settings of statistical learning. Sun et al. (2016) defined classification instability to quantify the sampling variability of the prediction made by classification methods and proposed a stabilized nearest neighbour classifier. Liu et al. (2010), Sun et al. (2013), Yu (2013), and Lim and Yu (2016) introduced stability measures for selecting tuning parameters across various statistical models, including penalized regression and high-dimensional graphical models. Pfister et al. (2021) introduced a stabilized regression algorithm designed to identify an optimal subset of predictors that generalizes across different environments. Wang (2010) and Fang and Wang (2012) defined stability measures for cluster analysis and used them for choosing the number of clusters. However, to our knowledge, the stability principle has not been applied in factor analysis. The definition of stability in factor analysis is less straightforward than that in regression due to the rotational indeterminacy of the factor model (Bai, 2003; Anderson and Rubin, 1956).

3.2 Stability-based Approach

3.2.1 Factor model

We observe an $n \times p$ data matrix $X = (x_{ij})_{n\times p}$, which contains p features of n observations. We work with the linear factor model, satisfying

$$x_{ij} = \boldsymbol{\lambda}_j^\top \boldsymbol{\gamma}_i + \epsilon_{ij}, 1 \le i \le n, 1 \le j \le p. \tag{3.1}$$

Here, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \dots, \gamma_{iK})^\top$ is a K-dimensional vector of factor scores of the i-th object, with γ_{ik} being independent variables with zero mean and unit variance. $\boldsymbol{\lambda}_j = (\lambda_{j1}, \dots, \lambda_{jK})^\top$ is a deterministic K-dimensional vector of factor loadings, and ϵ_{ij} is an independent noise term with mean 0 and variance ψ. In matrix form, the model specified in (3.1) can be written as $X = \Gamma\Lambda^\top + \mathcal{E}$, where $\Gamma = (\gamma_{ik})_{n\times K}$, $\Lambda = (\lambda_{jk})_{p\times K}$ and $\mathcal{E} = (\epsilon_{ij})_{n\times p}$. An intercept term can be included if entries of X are not mean zero. Since our primary goal is to determine the number of factors, and the covariance matrices are invariant to the addition of an intercept, we proceed with model (3.1) without loss of generality. The same setting has been considered in Bai et al. (2018), under which information criteria are proposed for determining K.

Let $\mathbf{x}_i$ denote the i-th row of X. The population covariance matrix of $\mathbf{x}_i$ is $\Sigma = \Lambda\Lambda^\top + \psi I_p$. By eigenvalue decomposition, we can write $\Sigma = \sum_{j=1}^p \sigma_j \mathbf{v}_j \mathbf{v}_j^\top$, where $\sigma_1 \ge \cdots \ge \sigma_p$ are the eigenvalues of Σ, and $\{\mathbf{v}_j\}_{j=1}^p$ is the set of orthonormal eigenvectors.

Remark 3.1. *In this work, we focus on detecting factors whose loadings can be consistently estimated. As discussed in Section 3.3, when n and p grow at the same order, this requires $\sigma_K \to \infty$. By contrast, the literature sometimes considers detecting spikes separated from the bulk even when $\sigma_i = O(1)$, in which case consistent estimation of the associated eigenvectors is not possible. The detection of such spikes lies beyond the scope of this work.*

3.2.2 Instability Measure

We define an instability measure using the principal angle between loading matrices. Ideally, let U_k and V_k be subspaces spanned by the leading k eigenvectors of the sample covariance matrices obtained from two independent and identically distributed samples. The between-sample loading instability at k is defined as

$$\sin\angle(U_k, V_k) = \max_{u\in U_k, u\neq 0} \min_{v\in V_k, v\neq 0} \sin\angle(u, v). \tag{3.2}$$

In practice, we only observe data from a single sample. To obtain an instability measure, we use data splitting. Let $[w]$ denote the integer part of any real number w. We randomly sample the rows of $X = (\mathbf{x}_1, \cdots \mathbf{x}_n)^\top$ without replacement to form a new permuted data matrix $(\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_n^{(s)})^\top$. This data matrix is further split into two halves, $X^{(1)} = (\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_{n_1}^{(s)})^\top$ and $X^{(2)} = (\mathbf{x}_{n_1+1}^{(s)}, \ldots, \mathbf{x}_n^{(s)})^\top$, where $n_1 = [n/2]$ and $n_2 = n - n_1$. For $l = 1, 2$, we perform eigenvalue decomposition such that $n_l^{-1}(X^{(l)})^\top(X^{(l)}) = \sum_{j=1}^p \tilde{\sigma}_j^{(l)} \tilde{\mathbf{v}}_j^{(l)} (\tilde{\mathbf{v}}_j^{(l)})^\top$, where $\tilde{\sigma}_1^{(l)} \geq \cdots \geq \tilde{\sigma}_p^{(l)}$ are the eigenvalues, and $\tilde{\mathbf{v}}_1^{(l)}, \ldots, \tilde{\mathbf{v}}_p^{(l)}$ are the corresponding eigenvectors.

Let $\tilde{V}_k^{(l)} = \text{Span}\{\tilde{\mathbf{v}}_1^{(l)}, \ldots, \tilde{\mathbf{v}}_k^{(l)}\}$ denote the subspace spanned by the first k leading eigenvectors. The loading instability measure at k is defined as $\sin\angle(\tilde{\mathbf{v}}_k^{(1)}, \tilde{\mathbf{v}}_k^{(2)})$. When $k = K$, this instability measure is expected to be close to zero, indicating good reproducibility of the factors. When $k > K$, the instability is expected to be close to 1, as each of $\tilde{\mathbf{v}}_k^{(1)}$ and $\tilde{\mathbf{v}}_k^{(2)}$ has at least one direction that corresponds to the noise matrix, resulting in two orthogonal directions. This phenomenon is formally stated and discussed in Section 3.3.

The measure $\sin\angle(\tilde{\mathbf{v}}_k^{(1)}, \tilde{\mathbf{v}}_k^{(2)})$ is computed using a single splitting of the data matrix X, which introduces additional randomness. To reduce this randomness, we propose to perform multiple random splittings and then take an average. Specifically, for $j = 1, \ldots, J$, let $\sin\angle(\tilde{V}_k^{(1,j)}, \tilde{V}_k^{(2,j)})$ denote the loading instabiltiy measure computed from the jth split, where J is the total number of splits. The averaged instability measure at k is defined as $\text{INS}(k) = J^{-1}\sum_{j=1}^J \sin\angle(\tilde{V}_k^{(1,j)}, \tilde{V}_k^{(2,j)})$. We use $J = 10$ for simulations and real data analysis in the rest, which seems sufficient.

3.2.3 Proposed Criteria

We propose several statistical criteria for estimating the number of factors based on the proposed instability measure. Let $\mathcal{K} = \{1, 2, ..., K_{\max}\}$. With an appropriate decreasing deterministic sequence $\{c_k\}_{k=1}^{K_{\max}}$, whose condition is given in Theorem 3.1, we can estimate K consistently by

$$\arg\min_{k\in\{1,\ldots,K_{\max}\}} \quad c_k + \text{INS}(k).$$

Here, $c_k \in [0, 1]$ is used to prevent underestimation, as $\text{INS}(k)$ is less predictable when $k < K$. In particular, the minimiser of

$$\text{SC1}(k) = \{(K_{\max} - k)/K_{\max}\} + \text{INS}(k)$$

is a consistent estimator of K.

Let $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \cdots \geq \tilde{\sigma}_p$ denote the eigenvalues of $n^{-1}X^\top X$. The following criterion estimates K by combining signal strength and stability considerations:

$$\text{SC2}(k) = l(k)/l(0) + \text{INS}(k),$$

where $l(k) = \sum_{j=k+1}^{K_{\max}} \log(\tilde{\sigma}_j + 1)$ for $k = 0, 1, \ldots, K_{\max} - 1$, and $l(K_{\max}) = 0$. Here, $l(k)/l(0)$ is analogous to the first terms of commonly used information criteria, such as the AIC and BIC proposed in Bai et al. (2018). The denominator $l(0)$, together with the addition of 1 inside the logarithm to each $l(k)$, serves to scale $l(k)/l(0)$ within the interval $[0, 1]$ to align it with the scale of the loading instability measure. This criterion aims to identify the model that balances the signal strength of the factors and their stability.

Finally, we introduce a criterion related to an information criterion in Bai and Ng (2002) $\mathrm{IC}(k) = \log(p^{-1}\sum_{j=k+1}^{p} \tilde{\sigma}_j^2) + kg(n,p)$, where $g(n,p)$ is a term depending on n and p. The second term in $IC(k)$ penalizes the number of factors. We propose a criterion in this spirit:

$$\mathrm{SC3}(k) = \log(1 + p^{-1}\sum_{j=k+1}^{p} \tilde{\sigma}_j^2)/\log(1 + p^{-1}\sum_{j=1}^{p} \tilde{\sigma}_j^2) + \mathrm{INS}(k),$$

which replaces the penalty term in the IC with the proposed instability measure. Similar to SC2, this criterion also aims to balance signal strength and stability.

Remark 3.2. *The above criteria rely on a pre-specified candidate set, where as long as $K_{\max}$ is finite, it does not affect the asymptotic theory. In practice, however, choosing $K_{\max}$ too large may make the associated penalty term overly small, which could lead to underestimation if INS(k) happens to be close to zero for some $k < K$. Since our method targets reproducible factors, the true number of factors is expected to be small; accordingly, we set $K_{\max} = 10$ in both simulations and real data analysis, which appears sufficient.*

3.3 Theoretical Results

We provide sufficient conditions under which the selection based on SC1, SC2, and SC3 is consistent. For positive sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all n. We denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Further, for two sequences of random variables A_n and B_n. We say that A_n is stochastically dominated by B_n, if for every $\epsilon > 0$ and $d > 0$ there exists $N = N(\epsilon, d)$ such that $\mathrm{pr}(A_n > n^{\epsilon}B_n) \leq n^{-d}$ for all $n \geq N$. We use the notation $A_n \prec B_n$ to denote that A_n is stochastically dominated by B_n. We use $O_{\prec}(B_n)$ to denote a term stochastically dominated by B_n. The following assumptions are made.

Assumption 3.1. *$n^{1/\kappa} \leq p \leq n^{\kappa}$ for some positive constant $\kappa > 1$.*

Assumption 3.2. *For each integer $m \geq 1$, there exists a universal constant $\kappa_m > 0$ such that*

$$\sup_{1\leq i\leq n, 1\leq j\leq p} E[|\epsilon_{ij}|^m] \leq \kappa_m \text{ and } \sup_{1\leq i\leq n, 1\leq k\leq K} E[|\gamma_{ik}|^m] \leq \kappa_m. \tag{3.3}$$

Assumption 3.3. *$(\sigma_K/\psi - 1)(n/p)^{1/2} \to +\infty$ for $n \gtrsim p$ and $(\sigma_K/\psi - 1)(n/p) \to +\infty$ for $p \gtrsim n$.*

Assumption 3.1 essentially requires that $\log(n)$ is comparable to $\log(p)$. The condition $n^{1/\kappa} \leq p$ may not be strictly necessary, since a larger n generally makes the problem easier. We retain it, however, because it facilitates symmetry in the analysis by allowing the roles of n and p to be interchanged. Assumption 3.2 is a regularity condition on the error matrix and latent factor scores, and it can be further relaxed. For example, we can require that (3.3) holds for all $m \leq M$ for a large enough constant M (Bloemendal et al., 2016). We use the current assumption for convenience. Finally, stronger signals are naturally required when n is small relative to p in order to obtain reproducible loadings, and weaker signals is sufficient when n is large compared to p. Assumption 3.3 provides the balance between the signal strength requirement from the population covariance and the ratio requirement between the number of features p and the number of samples n. Specifically, this is satisfied for bounded σ_K when n/p diverges, whereas a diverging σ_K is necessary when $n \lesssim p$. When $n \asymp p$, Assumption 3.3 holds for any diverging σ_K, regardless of the convergence rate. The following proposition characterizes the behaviour of the loading instability measure.

Proposition 3.1. *Under Assumptions 3.1 to 3.3, we have sequences a_n and b_n decaying to zero, such that*

$$\sin\angle(\tilde{V}_k^{(1)}, \tilde{V}_k^{(2)}) = 1 - O_{\prec}(a_n) \text{ for } k > K, \tag{3.4}$$

$$\sin\angle(\tilde{V}_k^{(1)}, \tilde{V}_k^{(2)}) = O_{\prec}(b_n) \text{ for } k = K. \tag{3.5}$$

Additionally, if $\sigma_1, \ldots, \sigma_K$ are distinct, then (3.5) *also holds for $k = 1, \ldots, K-1$.*

The exact expressions of a_n and b_n are provided in Section B.2.1 of the Supplementary Material. The cone concentration results for outlier eigenvectors and the delocalization results for non-outlier eigenvectors from random matrix theory (Bloemendal et al., 2016) are crucial to proving Proposition 3.1. Although the Davis-Kahan theorem (see Yu et al., 2015 and O'Rourke et al., 2018) also provide results on principal angles, it can only prove (3.5) when σ_K grows faster than the spectral norm of the error matrix, which is stronger than what is required in Theorem 3.1. The results for $k < K$ are provided for independent theoretical interest. They are not used for proving Theorem 3.1.

Theorem 3.1. *Under Assumptions 3.1 to 3.3, for any decreasing sequence $\{c_k\}_{k=1}^{K_{\max}}$ with $1 \geq c_k \geq 0$ for $k \in \mathcal{K}$, such that for some $\delta > 0$, $c_k - c_{k+1} > \delta$ for all $k \in \{1, \ldots, K-1\}$, and $1 - \delta > c_K - c_{K_{\max}}$. Define $\tilde{K} = \arg\min_{k \in \{1,\ldots,K_{\max}\}} c_k + INS(k)$. We have $\lim_{n\to\infty} pr\left(\tilde{K} = K\right) = 1$. Consequently, SC1 can consistently estimate K.*

The corollaries below give the conditions for SC2 and SC3 to be consistent.

Corollary 3.1. *Under Assumptions 3.1 to 3.3, if additionally $p \asymp n$ and $\log\sigma_1/\log\sigma_K \lesssim C$ for some constant $C > 0$, then SC2 can consistently estimate K as $n, p \to +\infty$.*

Corollary 3.2. *Under Assumptions 3.1 to 3.3, if additionally $p \asymp n$ and $\sigma_k^2 \asymp p$, $k = 1, \ldots, K$, then SC3 can consistently estimate K as $n, p \to +\infty$.*

We briefly discuss the assumptions underlying the proposed criteria and compare them with existing methods. SC1 is the least restrictive one, requiring only Assumptions 3.1 to 3.3. SC2 and SC3 require the additional condition $p \asymp n$, which is also needed for some other existing selection criteria, such as those in Bai and Ng (2002) and Bai et al. (2018). For SC2, it is further required that $\log\sigma_1$ and $\log\sigma_K$ are comparable, ensuring the gap $\{l(k-1) - l(k)\}/l(0)$ to be bounded below by a positive constant for $k = 1, \ldots, K$. Finally, SC3 imposes the strictest condition, requiring that $\sigma_k^2 \asymp p$, a crucial assumption for this type of information criterion, as in Bai and Ng (2002).

3.4 Numerical Experiments

3.4.1 Simulation Settings

We assess the finite sample performance of the proposed method via Monte Carlo simulations. In particular, we consider sample size $n = 1,500$, where p takes values from $500, 1,000, 1,500$ and $2,000$. Under each setting, 100 replications are generated. We set the true number of factors to be $K = 4$, and the candidate set be $\mathcal{K} = \{1, 2, ..., 10\}$ for model selection.

We simulate data from the model $X = \Gamma\Lambda^\top + D_\epsilon Q_\epsilon \mathcal{E}$, where the entries of Γ are independently drawn from the uniform distribution $U[-0.5, 0.5]$. The factor loadings Λ are generated as $\Lambda = QD$, with D a diagonal matrix whose diagonal elements are $\mu_1, \dots, \mu_K$. The matrix Q is orthonormal, obtained via QR decomposition of a random matrix $Z \in \mathbb{R}^{p\times K}$, where each entry of Z is independently sampled from a standard normal distribution $N(0, 1)$.

For the error term, we consider two scenarios to generate homogeneous and heterogeneous errors, respectively. In the first scenario (S1), $\mathcal{E}$ is drawn from $N(0, 1)$, with $D_\epsilon = Q_\epsilon = I_n$, ensuring homogeneous errors. In the second scenario (S2), $\mathcal{E}$ is drawn from a Student's t-distribution with 10 degrees of freedom. To introduce heteroskedasticity and test the robustness of the criteria under violation of the homoscedasticity assumption, Q_ϵ is an $n \times n$ orthonormal matrix generated similarly to Q, and $D_\epsilon = \text{diag}\{1/n, 2/n, \dots, 1\}$ is a diagonal matrix. We also consider three sets of μ_j values, the diagonal elements of D, corresponding to different signal strength settings: (i) $\{6p^{1/2}, 5p^{1/2}, 4p^{1/2}, 3p^{1/2}\}$ for strong signals, (ii) $\{6p^{1/6}, 5p^{1/6}, 3p^{1/6}, 3p^{1/6}\}$ for weak signals, (iii) $\{3p^{1/3}, 3p^{1/3}, 3p^{1/6}, 3p^{1/6}\}$ for signals of varying strengths, and (iv) $\{3p^{1/3}, 3p^{1/4}, 3p^{1/5}, 3p^{1/6}\}$ for signals with distinct growth rates. Note that under the second and third settings, the top eigenvalues of Λ are not distinct.

3.4.2 Results

We compare the stability-based criteria SC1, SC2, and SC3 proposed in this chapter with Bai and Ng (2002)'s information criterion (IC) with $g(n, p) = \{(n+p)/(np)\} \log(np/(n+p))$, Bai et al. (2018)'s AIC and BIC, and the eigenvalue ratio (ER) and growth ratio (GR) tests of Ahn and Horenstein (2013). Figure 3.1 shows the percentage of correct selections of the true number of factors K by these criteria. All methods perform well under S1(i), where the errors are homogeneous and the signal is strong. However, as expected, SC3 and IC struggle when the signal does not follow the $p^{1/2}$ order, as seen in the second to fourth columns of the figure. AIC also underperforms when faced with heterogeneous errors, as demonstrated in all S2 scenarios. While BIC shows robustness in S2, it performs poorly under S1(ii) and S1(iii). The eigenvalue ratio methods ER and GR perform well in the first two columns of the graph, where all signals grow at the same rate. However, their performance becomes poor when the signals are of varying strengths, as shown in the last two columns. Overall, SC1 and SC2 consistently select the correct number of factors across all settings, demonstrating superior performance in both homogeneous and heterogeneous error conditions.

Figure 3.2 illustrates the behaviour of the mean of the proposed instability measure across all replications for $k \in \mathcal{K}$. The instability measure is near 1 for $k > 4$ and close to 0 for $k = 4$, providing numerical support for Proposition 3.1. For $k = 1, 2, 3$, the

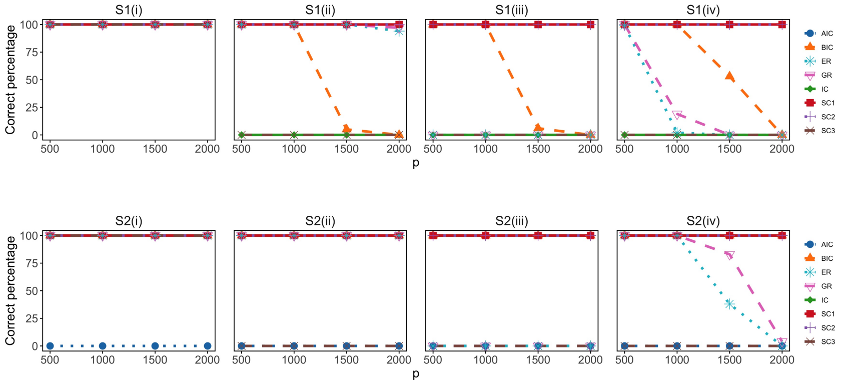

Figure 3.1: Correct selection percentages versus the number of features p across different scenarios (S1 and S2) and signal strengths (i,ii and iii). AIC (blue dotted line with circles), BIC (orange dashed line with triangles), ER (teal dotted line with stars), GR (pink dashed line with inverted triangles), IC (green solid line with diamonds), SC1 (red solid line with squares), SC2 (purple dotted line with pluses), and SC3 (brown dashed line with crosses)

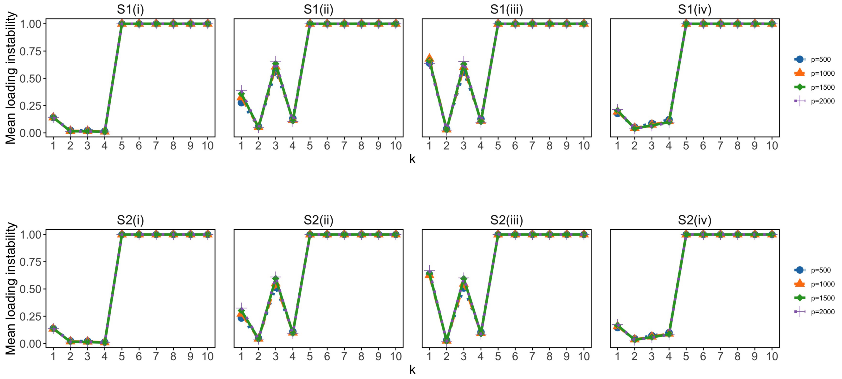

Figure 3.2: Loading instability versus k across different scenarios (S1 and S2) and signal strengths (i,ii and iii). $p = 500$ (blue dotted line with circles), $p = 1000$ (orange dashed line with triangles), $p = 1500$ (green solid line with diamonds) and $p = 2000$ (purple dotted line with pluses).

instability measures are also close to 0 in the first and fourth columns, where signals have distinct values. In the second and third columns, some measures deviate from 0 due to the presence of equal signal strengths.

3.5 Data Analysis

We consider a dataset concerning the p53 tumour suppressor protein, which plays a crucial role in cancer treatment research. This dataset includes 2D electrostatic, surface-based features, and 3D distance-based features, extracted using a method by Danziger et al. (2006), for a large collection of p53 mutations. It contains $5,208$ features for each of the $31,158$ mutations. It is important to understand the dependence between the features (see Lopes et al., 2019).

To evaluate the performance of the proposed criteria, we focus on the first $p = 1,000$ features and sample $n = 3,000$ rows without replacement from the data matrix. The row sampling is performed 100 times, resulting in 100 datasets that may be regarded as independent copies of each other. For each dataset, the features are standardized

Table 3.1: Performance of selection criteria on the p53 protein dataset

Criterion	Mode	Selection percentage(%)	Mean between-sample loading instability
SC1	2	97	0.10
SC2	2	98	0.10
SC3	2	100	0.10
IC	10	100	0.92
AIC	10	100	0.92
BIC	10	100	0.92
ER	2	100	0.10
GR	2	100	0.10

Mode: mode of the estimated number of factors for each criterion. Selection percentage(%): Percentages of instances selecting the mode. Mean between-sample loading instability: mean of the principal angles between the estimated loading spaces of all 4,950 pairs of datasets.

to have mean zero and variance one. The methods discussed in Section 3.4 are then applied to these datasets. Table 3.1 summarises the results, including the mode of the estimated number of factors, the selection percentage of the mode, and the mean between-sample loading instability. Specifically, the mean between-sample loading instability, which measures the reproducibility of the estimated factor structure, is calculated as follows. For every two datasets, we calculate the principal angle between the estimated loading spaces, which is well-defined even when they are of different dimensions. We then average the principal angle across all the 4,950 pairs of datasets.

The results show that the IC, AIC, and BIC criteria consistently estimate the number of factors to be 10, the upper bound of our candidate set. They continue to choose the maximum number in the candidate set when we increase it to $20, 30, 40$ and 50. The between-sample loading instability is 0.92, meaning the resulting ten-factor models have some unstable factors. On the other hand, all the proposed stability-based criteria, as well as the ER and GR criteria tend to select two factors, with a corresponding mean between-sample loading instability being 0.10, indicating a higher level of stability. While the true number of factors is unknown, this result shows that the proposed method is more conservative compared with information type criteria, which ensures a more reproducible factor structure.

3.6 Discussion

In this chapter, we proposed a loading stability measure for determining the number of factors in factor analysis. Using results from random matrix theory, we showed that the proposed stability-based criteria are consistent. Compared with many existing methods, the proposed one focuses directly on the stability of the estimated loading matrix and, thus, may give more reproducible results. Among the proposed criteria, SC1 is the least restrictive as it relies solely on stability, making it the most flexible. SC2 and SC3 incorporate additional signal strength conditions, with SC3 being the most conservative since it requires signals of order $p^{1/2}$. Our simulation results show that SC1 and SC2 generally provide the most stable performance, and we therefore recommend their use in practice. However, SC3 may still be suitable in applications where strong signal conditions are plausible and the validity of the estimator depends on such assumptions, for example in the type of factor model studied in Bai and Ng (2002).

Although our method is proposed under a linear factor model, the principal angle between loading spaces can be computed using two loading matrix estimates from data splitting, regardless of the specific factor model and estimator. Therefore, the same statistical criteria can be applied to determine the number of factors for other factor models, such as the generalised latent factor model (Chen and Li, 2022) that can be used to analyse binary, categorical, and count data. Specifically, we believe that a similar consistency result holds when the instability measure is constructed based on the constrained joint maximum likelihood estimator (Chen et al., 2020; Chen and Li, 2022) under the generalised latent factor model. However, establishing such results is nontrivial. The key to our consistency result is establishing the delocalization of the spurious directions of the estimated loading space. As the random matrix theory we currently use cannot be directly applied here, new technical tools are needed to establish the delocalization and further the consistency of the stability-based criteria. We leave it for future investigation.

Supplementary material

Additional simulation and real data results are presented in Appendix B.1, and the technical proofs of Proposition 3.1, Theorem 3.1, and Corollaries 1 and 2 are provided in Appendix B.2. An implementation of the proposed selection criteria for R is available at https://github.com/Arthurlee51/DNFSC. The real data in Section 3.5 is available from `https://archive.ics.uci.edu/dataset/188/p53+mutants`.

Chapter 4

Pairwise Comparisons without Stochastic Transitivity

4.1 Introduction

Pairwise comparison data have received intensive attention in statistics and machine learning, with diverse applications across domains. Such data often arise from tournaments, where each pairwise comparison outcome results from a match between two players or teams, or from crowdsourcing settings, where individuals are tasked with comparing two items, such as images, movies, or products. Specifically, the famous Thurstone (Thurstone, 1927) and Bradley-Terry (BT; Bradley and Terry, 1952) models have set a cornerstone in the field, followed by many extensions, including the parametric ordinal models proposed in Shah et al. (2016a), which broadens the class of parametric models. Oliveira et al. (2018) relax the assumption of a known link function and propose models that allow the link function to belong to a broad family of functions. Nonparametric approaches have also emerged, such as the work introduced in Shah and Wainwright (2018) based on the Borda counting algorithm, and the nonparametric Bradley-Terry models studied in Chatterjee (2015) and Chatterjee and Mukherjee (2019). Additionally, pairwise comparison models have been developed for crowdsourced settings, as discussed in Chen et al. (2013) and Chen et al. (2016), among many others. The models for pairwise comparisons have received a wide range of applications, including rank aggregation (Chen and Suh, 2015; Chen et al., 2019a; Heckel et al., 2019; Chen et al., 2022b), predicting matches/tournaments (Cattelan et al., 2013; Tsokos et al., 2019; Macrì Demartino et al., 2024), testing the efficiency of betting markets (McHale and Morton, 2011; Lyócsa and Vỳrost, 2018; Ramirez et al., 2023), and refinement of large language models based on human evaluations (Christiano et al., 2017; Ouyang et al., 2022; Zhu et al., 2023).

While the models mentioned above have made significant contributions to the field, they rely on the assumption of stochastic transitivity, which implies a strict ranking among players/teams/items. However, this assumption may be unrealistic, particularly in settings involving multiple skills or strategies, where intransitivity naturally arises. Despite its practical importance, research on models that allow intransitivity remains limited. Some notable exceptions include the work of Chen and Joachims (2016) and Spearing et al. (2023), which extend the Bradley-Terry model by introducing additional parameters to describe intransitivity alongside parameters specifying absolute strengths based on Bradley-Terry probabilities. Spearing et al. (2023) propose a Markov chain Monte Carlo algorithm for parameter estimation under a full Bayesian framework. However, their Bayesian procedure is computationally intensive and impractical for high-dimensional settings involving many players or a relatively high latent dimension. Chen and Joachims (2016) treat the parameters as fixed quantities and estimate them by optimizing a regularized objective function. However, their objective function is non-convex, and their model is highly over-parameterised. Consequently, their optimization is still computationally intensive and does not have a convergence guarantee. Moreover, no theoretical results are established in either work for their estimator.

Motivated by these challenges, we propose a general framework for modelling intransitive pairwise comparisons, assuming an approximately low-rank structure for the winning probability matrix. We propose an estimator for the probabilities, which can be efficiently solved by a convex optimization program. This estimator is shown to be optimal in the minimax sense, accommodating sparse data—a common issue when the number of players diverges. To our knowledge, this is the first framework to address intransitive comparisons with rigorous error analysis. The models presented in Chen and Joachims (2016) and Spearing et al. (2023), which assume a low-rank structure, can be seen as a special case of our framework. Furthermore, our method and computational algorithms scale efficiently to high-dimensional settings, making them suitable for applications with many players/teams/items. Empirical results on real-world datasets, including the e-sport *StarCraft II* and professional tennis, demonstrate the practical usefulness of our method, showing superior performance in intransitive settings and robust performance when transitivity largely holds.

Pairwise comparison data has been extensively studied in the statistics and machine learning literature, with numerous models and methods developed. We refer readers to Cattelan (2012) for a practical overview of the field. Theoretical properties of the BT model were first established in Simons and Yao (1999). These results were later extended to likelihood-based and spectral estimators, as well as other parametric extensions, with various losses and sparsity levels (Yan et al., 2012; Shah et al., 2016a; Negahban et al., 2017; Chen et al., 2019a; Han et al., 2020; Chen et al., 2022a). More recently, Han et al. (2023) propose a general framework covering most parametric models satisfying strong stochastic transitivity, establishing uniform consistency results under sparse and heterogeneous settings.

Our development is also closely related to the literature on generalised low-rank and approximate low-rank models (Cai and Zhou, 2013; Davenport et al., 2014; Cai and Zhou, 2016; Chen et al., 2020; Chen and Li, 2022, 2024; Lee et al., 2024). While our asymptotic results and error bounds build on techniques from these works, the parameter matrix in the current work differs in that it has a skew-symmetric structure. This structure, which arises naturally from pairwise comparison data, leads to dependent data entries and distinguishes our setting from typical low-rank models. To address this, tailored analysis is performed to establish rigorous theoretical results.

The rest of the chapter is organized as follows. Section 4.2 describes the setting, introduces the general approximate low-rank model, and proposes our estimator. Section 4.3 establishes the theoretical properties of the proposed estimator, including results on convergence and optimality. In Section 4.4, we provide an algorithm for solving the optimization problem of the proposed estimator. Section 4.5 verifies the theoretical findings and compares the proposed model with the BT model using simulations. Section 4.6 applies the proposed method to two real datasets to explore the presence of intransitivity in sports and e-sports. Finally, we conclude with discussions in Section 4.7. Detailed proofs of our main results are provided in Appendix C.

4.2 Generalised Approximate Low-rank Model for Pairwise Comparisons

4.2.1 Setting and Proposed Model

We consider a scenario with n subjects, such as players in a sports tournament. Let n_{ij} denote the total number of comparisons observed between subjects i and j, where $(n_{ij})_{n\times n}$ is a symmetric matrix. Let y_{ij} denote the observed counts where subject i beats subject j. Assuming no draws, we have $y_{ij} = n_{ji} - y_{ji}$ for $i, j \in \{1, \ldots, n\}$.

Given the total comparisons n_{ij}, we model the observed counts y_{ij} using a Binomial distribution: $y_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$, where π_{ij} denotes the probability that subject i beats subject j. A fundamental property of the probabilities is that $\pi_{ij} = 1 - \pi_{ji}$ for all $i, j \in \{1, \ldots, n\}$. This implies that the matrix $\Pi = (\pi_{ij})_{n\times n}$ is fully determined by its upper triangular part. Using the logistic link function $g(x) = (1 + \exp(-x))^{-1}$, we express the probabilities as $\pi_{ij} = g(m_{ij})$, where $M = (m_{ij})$ is a skew-symmetric matrix satisfying $M = -M^\top$. As a result, estimating the probabilities Π reduces to the problem of estimating M.

We say the model is stochastic transitive if there exists an unobserved global ranking among all the players, denoted by $i_1 \succ i_2 \succ \cdots \succ i_n$, such that the pairwise comparison probabilities for the adjacent pairs satisfy $\pi_{i_1 i_2}$, $\pi_{i_2 i_3}$, ..., $\pi_{i_{n-1} i_n} \geq 0.5$. In addition, $\pi_{ik} \geq \pi_{ij}$ whenever $j \succ k$, for all $i \neq j, k$. In other words, for two players, j and k, any player is more likely to win k than j if player j ranks higher than k. If stochastic transitivity does not hold, then we say a model is stochastic intransitive. For instance, stochastic intransitivity arises when there exists a triplet (i, j, k), such that $\pi_{ik} \geq \pi_{ij}$ and $\pi_{jk} < 0.5$.

Most traditional models for pairwise comparison assume stochastic transitivity. For example, the BT model assumes $m_{ij} = u_i - u_j$, in which case, the global ranking of the players is implied by the ordering of u_i, $i = 1, ..., n$. However, stochastic intransitivity naturally occurs in real-world competition data involving multiple strategies or skills. For example, in the professional competitions of the e-sport *StarCraft II*, players can choose from a variety of combat units with differing attributes (e.g., building cost, attack range, toughness) during the game, leading to strategic decisions that can result in intransitivity. In fact, for the best predictive model that we learned for the *StarCraft II* data, more than 70% of the (i, j, k) triplets are estimated to violate the stochastic transitivity assumption, i.e., $\pi_{ik} \geq \pi_{ij}$ and $\pi_{jk} < 0.5$; see Section 4.6 for the details.

From the modelling perspective, stochastic transitivity is achieved by imposing strong monotonicity constraints on the parameter matrix M. To allow for stochastic intransitivity, we need to relax such constraints. Given $Y = (y_{ij})_{n\times n}$, the log-likelihood is

$$\begin{aligned}\mathcal{L}(M) &= \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}\log(g(m_{ij})) \\ &= \sum_{i=1}^{n}\sum_{j>i} \left(y_{ij}\log(g(m_{ij})) + (n_{ij} - y_{ij})\log(1 - g(m_{ij}))\right).\end{aligned}$$

To prevent overfitting while accommodating stochastic intransitivity, we impose a constraint on M to reduce the size of the parameter space. Specifically, we assume that M has an approximately low-rank structure enforced through a nuclear norm constraint:

$$\|M\|_* \leq C_n n, \tag{4.1}$$

where $\|\cdot\|_*$ denotes the nuclear norm, and $C_n > 0$ is allowed to be a constant or growing with n. The maximal growth rate of C_n is restricted by the sparsity level, as will be detailed in Section 4.3. The estimator is defined as:

$$\hat{M} = \arg\max_{M} \mathcal{L}(M) \text{ subject to } \|M\|_* \leq C_n n, M = -M^\top. \tag{4.2}$$

It is easy to see that the optimization in (4.2) is convex; see Section 4.4 for its computation.

4.2.2 Comparison with Related Work

We compare the proposed model with existing parametric models in the literature. Han et al. (2023) introduce a general framework for analysing pairwise comparison data under the assumption of stochastic transitivity. In the current context, their model aligns with those proposed by Shah et al. (2016b) and Heckel et al. (2019), which are expressed as

$$\pi_{ij} = \Phi(u_i - u_j), \text{ and } \pi_{ji} = 1 - \Phi(u_i - u_j).$$

Here, $\Phi(\cdot)$ is any valid symmetric cumulative distribution function specified by the user, and $\mathbf{u} = (u_1, \ldots, u_n)^\top$ is a latent score vector representing the strengths of the teams. This framework reduces to the Bradley-Terry (BT) model when $\Phi(\cdot) = g(\cdot)$, the logistic function, and to the Thurstone model when $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Other models can be incorporated by specifying different forms of $\Phi(\cdot)$. The latent score $\mathbf{u}$ is treated as a fixed parameter to be estimated, enabling the framework to handle a large number of players effectively. This parametric form, however, enforces a rank-2 structure on the parameter matrix, given by

$$\Pi = \Phi(\mathbf{u}\mathbf{1_n}^\top - \mathbf{1_n}\mathbf{u}^\top),$$

where $\mathbf{1_n}$ is an n-dimensional vector of ones.

Several attempts have been made in the literature to generalize this parametric form, allowing the rank of the underlying parameter matrix to exceed two and accommodate stochastic intransitivity. We should note that, since M is a skew-symmetric matrix, its rank must be even (e.g., Horn and Johnson, 2013). For instance, Chen and Joachims (2016) proposed a blade-chest-inner model, which can expressed as

$$\Pi = g(AB^\top - BA^\top),$$

where A and B are $n \times K$ matrices. This model allows for a general rank-$2k$ parameter matrix, with the parameters in the frequentist sense. Similar to the parametrization in Chen and Joachims (2016), Spearing et al. (2023) propose a Bayesian model for pairwise comparison under stochastic intransitivity and further develop a Markov chain Monte Carlo algorithm for its computation. Both methods lack theoretical guarantees, such as convergence results or error bounds.

Our proposed method relaxes the requirement for an exact low-rank representation by only requiring an approximate low-rank structure specified by the nuclear norm. This offers a broad parameter space that covers the models proposed in Chen and Joachims (2016) and Spearing et al. (2023), offering greater robustness to model misspecification. This flexibility is important in real data applications, where the parameter matrix may not exhibit a clear-cut low-rank structure. In particular, if $\text{rank}(M) = 2k$ for some positive integer k, it follows that

$$\|M\|_* \leq \sqrt{2k}\|M\|_F \leq C_n n,$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and C_n depend on the magnitude of the entries of M and its rank $2k$. The subscript n in C_n indicates that both the magnitude of the entries of M and its rank are allowed to grow with n. Moreover, the proposed model imposes no distributional assumptions on the parameter matrix M, making it more scalable for handling large numbers of players. Theoretical results, including convergence and error bounds, are presented in Section 4.3. As a remark, our estimation method and theoretical framework can be easily adapted when we replace the current assumption of the logistic form of the link function $g(\cdot)$ with other functions, such as the standard normal cumulative distribution function used in the Thurstone model.

4.3 Theoretical Results

We establish convergence results and lower bounds for the estimator defined in (4.2) under settings with different data sparsity levels. For positive sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n \lesssim b_n$ if there exists a constant $\delta > 0$ that $a_n \leq \delta b_n$ for all n. Let $\mathcal{K}$ denote the parameter space, defined as

$$\mathcal{K} = \{M \in \mathbb{R}^{n\times n} : \|M\|_* \leq C_n n, \quad M = -M^\top\}. \tag{4.3}$$

We impose the following conditions:

Assumption 4.1. *The true parameter $M^* \in \mathcal{K}$.*

Assumption 4.2. *For $j = 1, \ldots, n$ and $i > j$, the variables n_{ij} are independent and follow a Binomial distribution, $n_{ij} \sim$ Binomial$(T, p_{ij,n})$, where T is a fixed integer representing the maximum possible number of comparisons between subjects, and $p_{ij,n} = p_{ji,n}$ is the success probability, which may vary across different pairs (i, j). Let $0 \leq p_n \leq q_n \leq 1$ denote the minimum and maximum comparison rates, respectively, such that $p_{ij,n} \in [p_n, q_n]$ for all $i \neq j \in \{1, \ldots, n\}$. We assume that $p_n \asymp q_n$ and $p_n \gtrsim \log(n)/n$.*

Assumption 4.1 ensures that the true parameter exhibits an approximately low-rank structure specified by our model. Assumption 4.2 deserves more explanations. The independent assumption of n_{ij} is commonly adopted in literature, as it simplifies the likelihood formulation and facilitates estimation. Moreover, under this assumption, the sparsity level of the data is characterized by the rate at which the success probabilities $p_{ij,n}$ converge to 0 as n grows. The condition $p_n \gtrsim \log(n)/n$ sets a lower bound on the sparsity level, which is the best possible threshold for pairwise comparison problems. Below this bound, the comparison graph becomes disconnected with high probability (Erdős and Rényi, 1960; Han et al., 2023). The condition $p_n \asymp q_n$ imposes homogeneity on $p_{ij,n}$, a common assumption in the literature (Simons and Yao, 1999; Chen et al., 2019a; Han et al., 2020). If q_n is larger, then more observations become available, and our method is expected to benefit from this. However, establishing theoretical guarantees under this regime requires a different proof strategy to handle the associated heterogeneous graph structure, which we leave for future investigation. The following theorem establishes the convergence rate of the proposed estimator.

Theorem 4.1. *Under Assumptions 4.1 and 4.2, let $\hat{\Pi} = (\hat{\pi}_{ij})_{n\times n}$, where $\hat{\pi}_{ij} = g(\hat{m}_{ij})$. Further let $\Pi^* = g(M^*)$. Then, with probability at least $1 - \kappa_1/n$,*

$$\frac{1}{n^2 - n}\|\hat{\Pi} - \Pi^*\|_F^2 \leq \kappa_2 C_n \sqrt{\frac{1}{p_n n}},$$

where κ_1 and κ_2 are constants that do not depend on n.

This theorem shows that convergence is guaranteed provided $C_n = o(\sqrt{p_n n})$. This means that $\|M^*\|_*$ can be as large as $o(n^{3/2}\sqrt{p_n})$. Suppose M^* has bounded entries. In the most sparse case with $p_n \asymp \log(n)/n$, this implies that $\|M^*\|_*$ can be at most of order $n\sqrt{\log(n)}$, so the rank of M^* cannot grow faster than $\sqrt{\log(n)}$. The following theorem addresses the optimality of Theorem 4.1 by establishing a lower bound.

Theorem 4.2. *Suppose* $12 \leq C_n^2 \leq \min\{1, \kappa_3^2/T\}n$, *where* κ_3 *is an absolute constant specified in* (C.6). *Consider any algorithm which, for any* $M \in \mathcal{K}$, *takes as input* Y *and returns* $\hat{M}$. *Then there exists* $M \in \mathcal{K}$ *such that with probability at least 3/8,* $\Pi = g(M)$ *and* $\hat{\Pi} = g(\hat{M})$, *satisfy*

$$\frac{1}{n^2 - n}\|\Pi - \hat{\Pi}\|_F^2 \geq \min\left\{\kappa_4, \kappa_5 C_n \sqrt{\frac{1}{np_n}}\right\} \tag{4.4}$$

for all $n > N$. *Here* $\kappa_4, \kappa_5 > 0$ *and* N *are absolute constants.*

A few technical assumptions are imposed in this theorem. The condition $C_n^2 \leq \min\{1, \kappa_3^2/T\}n$ is mild and naturally holds for sufficiently large n, provided that the rank of M does not grow at the same rate as n. We also require $C_n^2 \geq 12$ to avoid the parameter space being too small for packing set construction.

Since the rates in Theorems 4.1 and 4.2 match up to a multiplicative constant, the optimality of the proposed estimator is established.

4.4 Computation

To solve (4.2), we apply the nonmonotone spectral-projected gradient algorithm for closed convex sets proposed by Birgin et al. (2000), which guarantees convergence to a stationary point satisfying the constraints. Let Skew_n denote the space of $n \times n$ skew-symmetric matrices. Let $\mathcal{V}$ be the bijective linear mapping that vectorizes the upper-triangular part of any matrix in Skew_n into $\mathbb{R}^{0.5n(n-1)}$. For any $\mathbf{m} \in \mathbb{R}^{0.5n(n-1)}$, define $f(\mathbf{m}) = \mathcal{L}(\mathcal{V}^{-1}(\mathbf{m}))$. Then, solving (4.2) is equivalent to solving the constrained optimization problem:

$$\hat{\mathbf{m}} = \underset{\mathbf{m} \in \mathbb{R}^{0.5n(n-1)}}{\arg\max} \ f(\mathbf{m}) \quad \text{subject to } \|\mathcal{V}^{-1}(\mathbf{m})\|_* \leq \tau, \tag{4.5}$$

where $\tau = C_n n$ if C_n is known. We will later discuss an algorithm for selecting τ in practical situations where C_n is unknown.

A key step in solving (4.5) involves the orthogonal projection operator $P_\tau(\cdot)$, defined as

$$P_\tau(\mathbf{m}) = \underset{\mathbf{x} \in \mathbb{R}^{0.5n(n-1)}}{\arg\min} \ \|\mathbf{x} - \mathbf{m}\|_2 \text{ subject to } \|\mathcal{V}^{-1}(\mathbf{x})\|_* \leq \tau.$$

It is well known that the projection is equivalent to singular value soft-thresholding. Let $0_{n\times n}$ denote a $n \times n$ zero matrix, and $\max\{\cdot, \cdot\}$ be applied entry-wise for matrix inputs. The detailed procedure is presented in Algorithm 4.1.

In the last step, the projection outcome is defined as $P_\tau(\mathbf{m}) = \mathcal{V}(P_\tau(M))$, which is only valid provided that $P_\tau(M)$ is a skew-symmetric matrix. The following proposition ensures that this is always the case:

Proposition 4.1. *For any matrix* $M \in$ *Skew*$_n$, *the projection operator satisfies* $P_\tau(M) \in$ *Skew*$_n$.

Algorithm 4.1 Projection algorithm

```
Input: Parameter vector $\mathbf{m}$ and nuclear norm constraint parameter $\tau$.
  Compute $M = \mathcal{V}^{-1}(\mathbf{m})$.
  Perform singular value decomposition and obtain $M = U\Sigma V^\top$, where $U$ and $V$ are $n \times n$ orthonormal matrices and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_1, , \dots, \sigma_{n/2}, \sigma_{n/2})$ if $n$ is even and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_1, , \dots, \sigma_{\lfloor n/2 \rfloor}, \sigma_{\lfloor n/2 \rfloor}, 0)$ otherwise.
  Compute $\lambda$, the smallest value for which $\sum_{i=1}^{\lfloor n/2 \rfloor} 2\max\{\sigma_i - \lambda, 0\} \leq \tau$.
  Compute projected matrix $P_\tau(M) = U \max\{\Sigma - \lambda I_n, 0_{n\times n}\} V^\top$
Output: Projection outcome $P_\tau(\mathbf{m}) = \mathcal{V}(P_\tau(M))$.
```

Proof. We consider the case where n is even; the proof for odd n is analogous. It is well known that M can be decomposed in the Murnaghan canonical form $M = QXQ^\top$ (Murnaghan and Wintner, 1931; Benner et al., 2000), where Q is orthogonal and X is block-diagonal of the form

$$X = \begin{pmatrix} 0 & \sigma_1 & 0 & 0 & \dots & 0 & 0 \\ -\sigma_1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \sigma_2 & \dots & 0 & 0 \\ 0 & 0 & -\sigma_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & \sigma_{n/2} \\ 0 & 0 & 0 & 0 & \dots & -\sigma_{n/2} & 0 \end{pmatrix},$$

where $\sigma_1, \dots, \sigma_{n/2}$ are the singular values of M. It can be verified that the projection operator $P_\tau(M)$ preserves the Murnaghan canonical form as $\mathcal{P}_\tau(M) = QYQ^\top$, where

$$Y = \begin{pmatrix} 0 & \max\{\sigma_1 - \lambda, 0\} & 0 & \dots & 0 \\ -\max\{\sigma_1 - \lambda, 0\} & 0 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \dots & 0 \\ \vdots & \vdots & \vdots & 0 & \max\{\sigma_{n/2} - \lambda, 0\} \\ 0 & 0 & 0 & -\max\{\sigma_{n/2} - \lambda, 0\} & 0 \end{pmatrix}.$$

Hence we have $P_\tau(M) \in \mathrm{Skew}_n$. □

We now introduce the spectral projected line search method, which uses the projection operator $\mathcal{P}_\tau(\cdot)$ to ensure that each iteration's outcome remains within the feasible set defined by the nuclear norm constraint. The procedure is outlined in Algorithm 4.2.

The method employs two types of line searches. The first type performs a projection once and searches along a linear trajectory $\mathbf{m}(\alpha)$. This approach is computationally efficient since the primary computational cost lies in the projection operation. If the linear search fails to converge, the algorithm switches to a curvilinear trajectory $\mathbf{m}^{\mathrm{curve}}(\alpha)$, which requires projecting at each step. The Spectral-step length γ_{l-1} is decided using the method from Barzilai and Borwein (1988) in each iteration.

The final estimation procedure is detailed in Algorithm 4.3. The convergence criterion checks whether the optimality condition $P_\tau(\mathbf{m}^{(l)} - \nabla f(\mathbf{m}^{(l)})) = \mathbf{m}^{(l)}$ is approximately satisfied. Parts of the code are adapted from the SPGL1 package, originally implemented in Matlab (Van Den Berg and Friedlander, 2008; Davenport et al., 2014). The proposed estimator is implemented in R, and the code is available at `https://github.com/Arthurlee51/PCWST`.

```
Algorithm 4.2 Spectral projected line search
Input: Parameter vector from last iteration $\mathbf{m}^{(l-1)}$, Matrix of comparison outcomes $Y$, nuclear norm constraint parameter $\tau$ and the spectral-step length $\gamma_{l-1}$
  Compute gradient: $\mathbf{g}^{(l-1)} = \nabla f(\mathbf{m}^{(l-1)})$.
  Compute search direction: $\mathbf{d}^{(l-1)} = P_\tau(\mathbf{m}^{(l-1)} - \gamma_{l-1}\mathbf{g}^{(l-1)}) - \mathbf{m}^{(l-1)}$
  Perform line search along the linear trajectory: $\mathbf{m}(\alpha) = \mathbf{m}^{(l-1)} + \alpha\mathbf{d}^{(l-1)}$.
  if Convergence is reached then
    Set $\mathbf{m}^{(l)}$ as the result from the line search.
  else
    Perform line search along the alternative trajectory:
        $\mathbf{m}^{\text{curve}}(\alpha) = P_\tau(\mathbf{m}^{(l-1)} - \alpha\gamma_{l-1}\mathbf{g}^{(l-1)})$.
    Set $\mathbf{m}^{(l)}$ as the result from the line search.
  end if
Output: Updated parameter vector $\mathbf{m}^{(l)}$.
```

```
Algorithm 4.3 Estimation Algorithm
Input: Matrix of comparison outcomes $Y$, nuclear norm constraint parameter $\tau$.
  Initialization: Set $l = 0$, $\mathbf{m}^{(0)} = \mathbf{0}_{0.5n(n-1)}$, the zero vector and set the spectral step-length $\gamma_0 = 1$.
  while $l = 0$ or convergence criterion is not satisfied do
    Update $l \leftarrow l + 1$.
    Update $\mathbf{m}^{(l)}$ via line search using Algorithm 4.2 with inputs $\mathbf{m}^{(l-1)}$, $Y$, $\tau$ and $\gamma_{l-1}$.
    Update $\gamma_l$ as proposed by Barzilai and Borwein (1988).
  end while
Output: Estimated parameter matrix $\hat{M} = \mathcal{V}^{-1}(\mathbf{m}^{(l)})$.
```

4.5 Simulation Results

We consider three distinct scenarios characterized by varying levels of sparsity. Specifically, we define p_n as $n^{-1}\log(n)$, $n^{-1/2}$, and $1/4$, corresponding to sparse, less sparse, and dense data, respectively. The parameter q_n is given by $4p_n$. Each $p_{ij,n}$ is then generated from a uniform distribution with range $[p_n, q_n]$.

The parameter matrix M is constructed as $\Theta J \Theta^\top$, where Θ is an $n \times 2k$ matrix, and J is a $2k \times 2k$ block diagonal matrix of the form

$$J = \begin{pmatrix} 0 & n & 0 & \dots & 0 \\ -n & 0 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \dots & 0 \\ \vdots & \vdots & \vdots & 0 & n \\ 0 & 0 & 0 & -n & 0 \end{pmatrix}.$$

The matrix Θ is orthonormal, obtained via QR decomposition of a random matrix $Z \in \mathbb{R}^{n \times 2k}$, where each entry of Z is independently sampled from a standard normal distribution $N(0,1)$. It can be verified that $\|M\|_* = 2kn$.

We conduct 50 simulations for $n = 500, 1000, 1500$, and 2000, with k ranging from 1 to 10. Recall that the rank of M is $2k$. Additionally, the maximum number of comparisons,

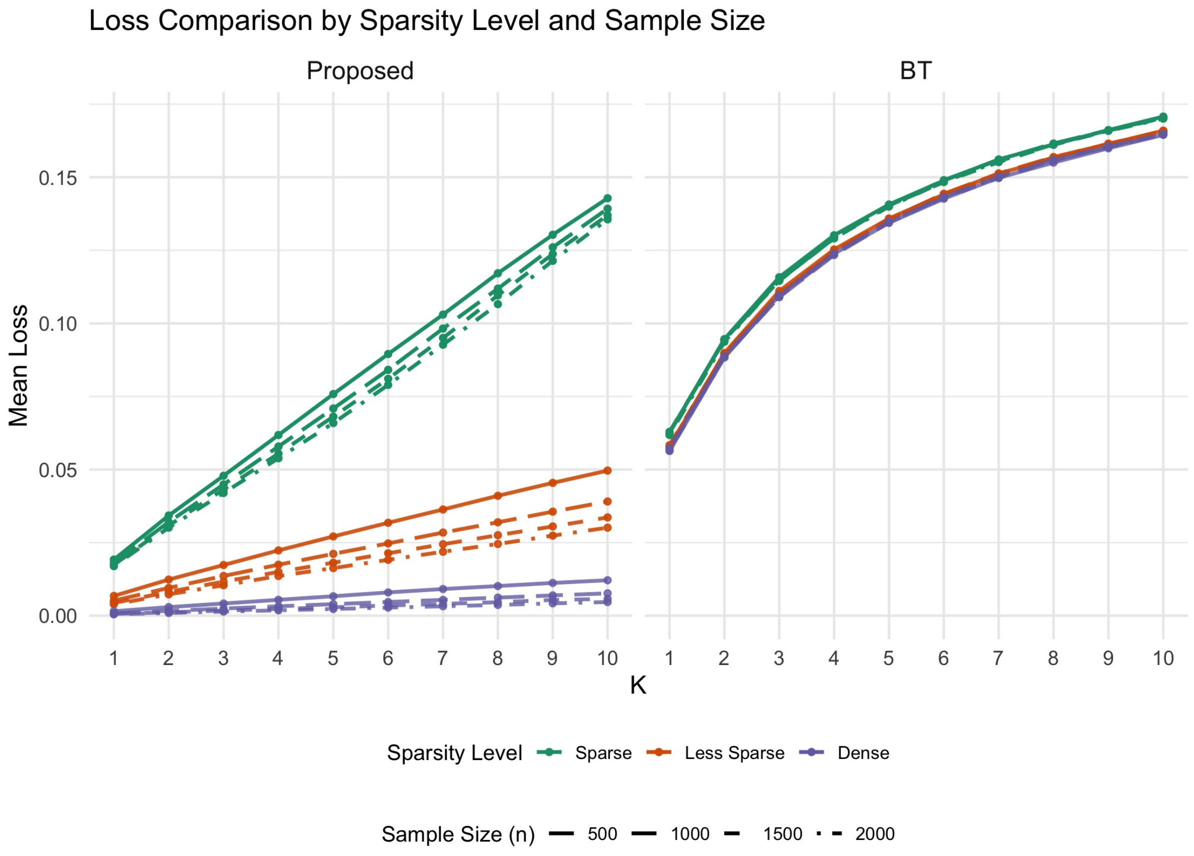

Figure 4.1: Comparison of loss between the proposed method and the Bradley-Terry (BT) model across different sparsity levels (sparse, less sparse, dense). The x-axis represents the rank parameter k, while the y-axis shows the mean loss, computed as the average of the losses defined in (4.6). Results are shown for varying sample sizes ($n = 500, 1000, 1500, 2000$)

T, is fixed at 5, across all settings. We set $C_n = 2k$. The loss is computed as

$$\text{Loss} = (n^2 - n)^{-1}\|\hat{\Pi} - \Pi^*\|_F^2, \tag{4.6}$$

and the average loss across 50 simulations is reported for each model in Figure 4.1, considering different values of n, k, and sparsity levels.

The results in Figure 4.1 show that the mean loss of the proposed estimator decreases as n increases. Moreover, the mean loss is significantly lower as the data become denser, corresponding to an increase in p_n. These observations are consistent with the results from Theorem 4.1.

Notably, the proposed and BT models incur higher losses as the rank parameter k increases, which is expected due to increasing complexity. However, the proposed model consistently outperforms the BT model across all settings. Furthermore, while the BT model's performance remains relatively unchanged as n increases, the proposed method continues to improve, showcasing its effectiveness in handling large datasets and capturing complex structures that stochastic transitivity assumptions cannot address.

4.6 Real Data Examples

In this section, we compare our model's performance with the celebrated BT model using two real datasets. Section 4.6.1 outlines the data preparation process and describes how the nuclear norm constraint parameter $\tau = C_n n$ is decided. Section 4.6.2 introduces the evaluation metrics used to compare the models. Finally, Sections 4.6.3 and 4.6.4 present detailed analyses of the results for the *StarCraft II* and tennis datasets, respectively.

4.6.1 Data Preparation and Parameter Tuning

The raw data consists of individual match records, with each comparison recorded as a separate entry. We reserve 30% of the match records for testing, while the remaining 70% is divided into 50% for training and 20% for validation.

The comparison data matrix is first constructed for the training set, with players absent from the training set removed from the validation set. The validation set is used to tune the nuclear constraint parameter C_n, as described in the sequel. After tuning, the training and validation sets are combined (including previously excluded entries), and the comparison data matrix is reconstructed from the combined dataset.

The test set is then evaluated against this combined dataset, excluding entries for players not present in the combined dataset. Although the proposed model can handle players who never lose or win any game, we still remove them in the training and combined dataset to ensure stabler results and a fair comparison with the BT model, as this is a common practice.

The nuclear norm of the parameter matrix M is unknown and is tuned on the training and validation sets using log-likelihood as the loss function. The nuclear constraint parameter $\tau = C_n n$ is determined by selecting C_n from 20 grid points, corresponding to powers of 10 evenly spaced between -1 and 1. This results in $C_n = 10^{0.47} = 2.98$ for the *StarCraft II* dataset and $C_n = 10^{-0.36} = 0.43$ for the tennis dataset.

4.6.2 Evaluation Criteria

Let $Y^{(\text{test})} = (y_{ij}^{(\text{test})})_{n\times n}$ denote the observed comparison results from the test set. Given the estimated winning probabilities $\hat{\Pi} = (\hat{\pi}_{ij})_{n\times n}$, we evaluate the performance of the estimates using two criteria. The first criterion is the log-likelihood, given by

$$L(Y^{(test)} \mid \hat{\Pi}) = \sum_{i=1}^{n} \sum_{j>i} \left(y_{ij}^{(\text{test})} \log(\hat{\pi}_{ij}) + y_{ji}^{(\text{test})} \log(1 - \hat{\pi}_{ij}) \right),$$

where a higher log-likelihood indicates a stronger agreement between the predicted probabilities and the observed results. The second criterion is the test accuracy, given by

$$A(Y^{(\text{test})} \mid \hat{\Pi}) = \frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{n} y_{ij}^{(\text{test})}} \sum_{i=1}^{n} \sum_{j>i} \left(y_{ij}^{(\text{test})} I(\hat{\pi}_{ij} \geq 0.5) + y_{ji}^{(\text{test})} I(\hat{\pi}_{ji} > 0.5) \right).$$

It measures the proportion of the comparison results correctly predicted, with higher values indicating better predictive performance. The results are presented in Table 4.1.

	StarCraft II		Tennis	
	Proposed	BT	Proposed	BT
Log-likelihood	$-1,897,946$	$-2,137,115$	$-333,076$	$-322,483$
Accuracy	0.766	0.713	0.652	0.658

Table 4.1: Comparison of model performance on StarCraft II and ATP datasets. The performance is evaluated using log-likelihood and accuracy for the proposed model and the BT model.

4.6.3 *StarCraft II* Data

StarCraft II is a military science fiction real-time strategy game developed and published by Blizzard Entertainment. The dataset comprises match results of professional *StarCraft II* players sourced from the website `aligulac.com`, covering the period from 2010 to 2016. The matches follow the most common competitive format, where two players face off against each other, and each game results in either a win or loss, with no possibility of a draw.

We specifically focus on matches played using the *StarCraft II: Heart of the Swarm* expansion, as different versions of the game are often treated as distinct games (Chen and Joachims, 2016). The training set includes $1,958$ players, with 1.9% of all player pairs competing against each other at least once. The maximum number of matches between any pair of players is 30. The dataset is available at `https://www.kaggle.com/datasets/alimbekovkz/starcraft-ii-matches-history`.

As seen in Table 4.1, the proposed model achieves a higher log-likelihood of $-1,897,946$ compared to $-2,137,115$ for the BT model. This suggests that our model provides a better fit for the observed test data. The test accuracy of the proposed model is also significantly higher at 0.766, compared to 0.711 for the BT model. Among the $1,249,168,756$ distinct triplets in the data, stochastic transitivity is violated in 70% of cases, as indicated by the matrix of estimated probabilities $\hat{\Pi}$ under the proposed model. Specifically, this occurs when there exists an ordering of the three players, denoted as i, j, and k, such that $\hat{\pi}_{ik} \geq \hat{\pi}_{ij}$ and $\hat{\pi}_{jk} < 0.5$.

These results are consistent with previous findings by Chen and Joachims (2016), who analysed a similar dataset over different time frames, suggesting that a strict ranking structure may not be appropriate in e-sports. In particular, intransitivity can naturally arise from game design, such as intransitive relationships among different unit types, which provide players with significant flexibility in choosing units and strategies. Moreover, the strong performance of our method on this dataset confirms its ability to effectively handle sparsity in real-world data, aligning with both simulation and theoretical results.

4.6.4 Tennis Data

We analyse the tennis dataset to evaluate the performance of our model in professional sports. The dataset contains the results of all men's matches organized by the Association of Tennis Professionals (ATP) from 2000 to 2018. It includes matches from major tournaments such as the Grand Slams, the ATP World Tour Masters 1000, and other professional tennis series held during this period.

The training set consists of 723 players, with 6.4% of all player pairs having competed against each other at least once. The maximum number of matches between any pair of players is 23. The data is collected from `http://www.tennis-data.co.uk`.

From Table 4.1, the BT model achieves a marginally better performance, with a log-likelihood of $-322,483$ compared to $-333,076$ for the proposed model, and a slightly higher test accuracy (0.658 vs 0.652). This advantage may come from the BT model's smaller parameter space, which is more efficient when the data aligns well with the stochastic transitivity assumption, where the level of intransitivity is minimal or absent. Nevertheless, the performance of the proposed model remains close to that of the BT model, demonstrating its robustness even in settings where transitivity holds. This flexibility is particularly useful when intransitivity is uncertain, as it maintains high accuracy without relying on strict ranking assumptions.

The lack of intransitivity in professional tennis may be due to several factors. Unlike e-sports, tennis offers limited gameplay flexibility, as adjustments to equipment like rackets and shoes have minimal impact compared to the choice of units in *StarCraft II*. Additionally, professional tennis players may be required to be well-rounded as weaknesses are quickly identified and exploited by opponents. In contrast, intransitivity may be more common at lower levels of competition, where skill imbalances are expected to be more significant. For example, a player with a strong serve but weak baseline play may be more likely to defeat one opponent while losing to another with a different style. Investigating intransitivity in lower-tier competitions remains an open question for future research.

4.7 Discussions

In this chapter, we propose a statistical framework for modelling stochastic intransitivity. The framework assumes an approximate low-rank structure in the parameter matrix, expressed through a nuclear norm constraint. Theoretical analysis demonstrates that the proposed estimator achieves optimal convergence rates under a wide range of data sparsity settings. Simulation and empirical analyses confirm that our model is superior to the Bradley-Terry model when the assumption of stochastic transitivity is violated.

Our framework stands apart from the existing literature by imposing an approximate low-rank structure. To our knowledge, all existing methods for pairwise comparison data rely on exact low-rank models, even in the limited works that allow stochastic intransitivity. By accommodating a larger parameter space, our approach offers greater flexibility and applicability to a wider range of datasets. While this may lead to slightly reduced efficiency, our analysis of the tennis dataset demonstrates that the loss of efficiency is small when stochastic transitivity largely holds. Therefore, the proposed model may predict pairwise comparison results more accurately in many real-world applications. For example, for tournament data, the proposed may better predict the champion or the number of rounds each player can play, given historical data and the current tournament schedule.

The current research may be extended in several directions. Specifically, the current theoretical analysis focuses on the convergence of the loss $\|\hat{\Pi} - \Pi^*\|^2/(n^2 - n)$, which can be seen as a notion of convergence in an average sense (across entries of the comparison probability matrix). It can be strengthened by establishing the convergence results under the matrix max-norm loss $\|\hat{\Pi} - \Pi^*\|_\infty$, which may be achieved using the refinement techniques proposed in Chen and Li (2024). This notion of convergence ensures the consistency of each $\hat{\pi}_{ij}$. Moreover, it will be useful to further establish the asymptotic normality for each $\hat{\pi}_{ij} - \pi^*_{ij}$, which can be used to quantify the uncertainty associated with the estimated winning probabilities.

From a computational perspective, the main challenge of our algorithm lies in the projection step, which requires performing singular value decomposition of an $n \times n$ matrix. This step is substantially more demanding than in the Bradley–Terry model, for which highly scalable implementations are available. While the current method is feasible for datasets with a few thousand players, it may become computationally challenging as the number of players grows larger. Optimizing the implementation will therefore be important for scaling to much larger datasets and for broadening the range of applications of the proposed method.

The proposed modelling framework also needs to be extended to accommodate more complex settings of pairwise comparisons. First, covariate information can be incorporated into the model to facilitate the prediction. For example, for many team sports tournaments (e.g., soccer and basketball), whether a team plays at their home court matters and should be included as a covariate. Second, pairwise comparison data are often collected over time, which is true for the *StarCraft II* and tennis data studied in Section 4.6. The current model ignores time information in data. To better predict future pairwise comparison results, it will be useful to model the comparison probabilities as a function of time. As a result, the estimation of these time-varying comparison probabilities will also differ substantially from the current procedure. Third, for pairwise comparison data produced by raters, which are commonly encountered in crowd-sourcing settings (e.g., Chen et al., 2013), characteristics of the raters, such as their reliability, affect the pairwise comparisons. In other words, the distribution of the comparison between two items depends not only on the pair of items but also on the rater who performs the comparison. In this regard, Chen et al. (2013) propose an extended version of the BT model that uses a rater-specific latent variable to account for raters' reliability. A similar extension can be made to the current model to simultaneously account for the raters' heterogeneity and the items' stochastic intransitivity.

Appendix A

Supplementary Materials for Chapter 2

A.1 Estimation Method

Recall that the log-likelihood function in (2.6) was defined as

$$\begin{aligned} l(\Xi) &= \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T} r_{it}\left\{y_{ijt}(\gamma_{jt} + \mathbf{a}_j^\top\boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top\mathbf{x}_i) - b_j(\gamma_{jt} + \mathbf{a}_j^\top\boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top\mathbf{x}_i)\right\} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T} r_{it}\left\{y_{ijt}(\mathbf{u}_j^\top\mathbf{e}_{it}) - b_j(\mathbf{u}_j^\top\mathbf{e}_{it})\right\}. \end{aligned}$$

We further define

$$\rho_{ijt}(t) = y_{ijt}(\mathbf{u}_j^\top\mathbf{e}_{it}) - b_j(\mathbf{u}_j^\top\mathbf{e}_{it}) \text{ and } \varrho_{ijt}(t) = r_{it}\rho_{ijt}(t), \text{ such that}$$

$$\boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i) = \sum_{t=1}^{T} r_{it}\rho_{ijt}(\mathbf{u}_j^\top\mathbf{e}_{it}) = \sum_{t=1}^{T} \varrho_{ijt}(\mathbf{u}_j^\top\mathbf{e}_{it}).$$

We now define the following objective functions:

$$l_{NJ}(\Xi) = \frac{1}{NJ}l(\Xi) = \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i),$$

$$l_{i,J}(\boldsymbol{\theta}_i, U) = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i), \quad l_{j,N}(\mathbf{u}_j, \Theta) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i).$$

Define $\nabla l_{i,J}(\boldsymbol{\theta}, U), \nabla^2 l_{i,J}(\boldsymbol{\theta}, U)$ as the gradient and Hessian matrix of $l_{i,J}(\boldsymbol{\theta}, U)$ with respect to $\boldsymbol{\theta}$, and $\nabla l_{j,N}(\mathbf{u}, \Theta), \nabla^2 l_{j,N}(\mathbf{u}, \Theta)$ as the gradient and Hessian matrix of $l_{j,N}(\mathbf{u}, \Theta)$ with respect to $\mathbf{u}$. To handle the constraint in (2.6), we introduce the projection operator

$$Prox_c(\mathbf{y}) = \underset{\mathbf{x}:\|\mathbf{x}\|\le c}{\arg\min}\,\|\mathbf{y} - \mathbf{x}\|^2 = \begin{cases}\mathbf{y} & \text{if } \|\mathbf{y}\| \le c; \\ c\|\mathbf{y}\|^{-1}\mathbf{y} & \text{otherwise.}\end{cases}$$

Recall that we defined P as the length of $\mathbf{u}_j$. We propose the following iterative algorithm to estimate the parameters of interest Ξ.

1. **Initialization:** Choose initial starting parameters $\Theta^{(0)}, U^{(0)}$.

Table A.1: Definitions of $\mathbf{u}_j$ and $\mathbf{e}_{it}$ under different model specifications

Model	$\mathbf{u}_j$	$\mathbf{e}_{it}$
Main model	$(\boldsymbol{\gamma}_j^\top, \boldsymbol{\beta}_j^\top, \mathbf{a}_j^\top)^\top$	$(\mathbf{D}_{it}^\top, \mathbf{x}_i^\top, \boldsymbol{\theta}_i^\top)^\top$
Section 2.2.3	$(\boldsymbol{\gamma}_j^\top, \boldsymbol{\beta}_j^\top, \mathbf{v}_j^\top, \mathbf{a}_j^\top)^\top$	$(\mathbf{D}_{it}^\top, \mathbf{x}_i^\top, \mathbf{z}_{it}^\top, \boldsymbol{\theta}_i^\top)^\top$
Section 2.2.4	$(\boldsymbol{\gamma}_j^\top, \boldsymbol{\beta}_{j1}^\top, \ldots, \boldsymbol{\beta}_{jT}, \mathbf{v}_j^\top, \mathbf{a}_{j1}^\top, \ldots, \mathbf{a}_{jT}^\top)^\top$	$(\mathbf{D}_{it}^\top, D_{it1}\mathbf{x}_i^\top, \ldots, D_{itT}\mathbf{x}_i^\top, \mathbf{z}_{it}^\top, D_{it1}\boldsymbol{\theta}_i^\top, \ldots, D_{itT}\boldsymbol{\theta}_i^\top)^\top$

For each model specified above, when the restriction $\gamma_{jt} = t\gamma_j$ for all $j \in \{1, \ldots, J\}$ discussed in Section 2.2.5 is imposed, we replace $\boldsymbol{\gamma}_j$ with γ in $\mathbf{u}_j$, and substitute the component $\mathbf{D}_{it}$ in $\mathbf{e}_{it}$ with t.

2. **Parameter Update:** For $l = 1, \ldots, L$, perform
   - Given $\Theta^{(l-1)}, U^{(l-1)}$, update
   $$\mathbf{u}_j^{(l)} = Prox_c\left(\mathbf{u}_j^{(l-1)} - \alpha\left(\nabla^2 l_{j,N}\left(\mathbf{u}^{(l-1)}, \Theta^{(l-1)}\right)\right)^{-1}\nabla l_{j,N}\left(\mathbf{u}_j^{(l-1)}, \Theta^{(l-1)}\right)\right)$$
   for $j = 1, \ldots, J$, where $c = 5\sqrt{P}$, $\alpha > 0$ is a step size chosen by line search.
   - Given $\Theta^{(l-1)}, U^{(l)}$, update
   $$\boldsymbol{\theta}_i^{(l)} = Prox_c\left(\boldsymbol{\theta}_i^{(l-1)} - \alpha\left(\nabla^2 l_{i,J}\left(\boldsymbol{\theta}^{(l-1)}, U^{(l)}\right)\right)^{-1}\nabla l_{i,J}\left(\boldsymbol{\theta}^{(l-1)}, U^{(l)}\right)\right),$$
   for $i = 1, \ldots, N$, where $c = 5\sqrt{K}$, $\alpha > 0$ is a step size chosen by line search.
3. **Convergence Check:** Stop iteration when $l_{NJ}(\Xi^{(L)})$ approximates $l_{NJ}(\Xi^{(L-1)})$ closely.

This algorithm guarantees that the log likelihood increases in each iteration, when the step size α is properly chosen in line search. Readers may refer to Chen et al. (2019c) and Parikh et al. (2014) for further details regarding the properties of projection operator. Moreover, when applying this estimation approach to simulation studies and data analysis, we develop a singular value decomposition (SVD) based algorithm to choose a good starting point in step 1, as presented in Section A.1.1. This algorithm can be generalised easily for mixed types of data.

Remark A.1. *The estimation procedure described above is flexible and can accommodate the extensions proposed in Sections 2.2.3 and 2.2.4, as well as the constraint on $\boldsymbol{\gamma}_j$ discussed in Section 2.2.5. For each extension, the definitions of $\mathbf{u}_j$ and $\mathbf{e}_{it}$ are modified accordingly and are summarized in Table A.1, where the definition of the parameter vector $\Xi = \left(\mathbf{u}_1{}^\top, \ldots, \mathbf{u}_J{}^\top, \boldsymbol{\theta}_1{}^\top, \ldots, \boldsymbol{\theta}_N{}^\top\right)^\top$ remains unchanged.*

A.1.1 SVD-based algorithm for obtaining initial values

The following algorithm, based on the work from Chen et al. (2019c) and Zhang et al. (2020a), works for binary variables when $N \geq J$. Modifications are needed when $J > N$. We present the algorithm in the context of the extension presented in Section 2.2.3, noting that the main model is a special case when time-dependent covariates are absent.

1. Input responses y_{ijt}, missing indicators r_{it}, dimension K of latent space, and tolerance ϵ.

2. Compute $\hat{p}_t = (\sum_{i=1}^N r_{it})/N$ as the proportion of observed responses for $t = 1, \ldots, T$.
3. Let $L_t = (l_{ijt})_{N\times J}$, where
$$l_{ijt} = \begin{cases} 2y_{ijt} - 1, & \text{if } r_{it} = 1, \\ 0, & \text{otherwise.} \end{cases}$$
4. Apply singular value decomposition to matrices $\mathcal{L}_t, t = 1, \ldots, T$ and obtain $\mathcal{L}_t = \sum_{j=1}^J \sigma_{jt}\mathbf{q}_{jt}\mathbf{h}_{jt}^\top$, where for each t, $\sigma_{1t} \geq \ldots, \geq \sigma_{Jt}$ are the singular values and $\mathbf{q}_{jt}$s and $\mathbf{h}_{jt}$s are the left and right singular vectors.
5. Let $\tilde{\mathcal{L}}_t = (\tilde{l}_{ijt})_{N\times J} = \sum_{k=1}^{\tilde{K}} \sigma_{kt}\mathbf{q}_{kt}\mathbf{h}_{kt}^\top$, where $\tilde{K} = \max\{K+1, \arg\max_k \sigma_{kt} \geq 2\sqrt{N}\hat{p}_t\}$.
6. Let $M_t = (m_{ijt})_{N\times J}$, where
$$m_{ijt} = \begin{cases} \xi^{-1}(\epsilon) & \text{if } \tilde{l}_{ijt} < -1+\epsilon, \\ \xi^{-1}(0.5(\tilde{l}_{ijt}+1)) & \text{if } -1+\epsilon \leq \tilde{l}_{ijt} \leq 1-\epsilon, \\ \xi^{-1}(1-\epsilon) & \text{if } \tilde{l}_{ijt} > 1-\epsilon. \end{cases}$$
7. Set $\Gamma_t^{(0)} = (\gamma_{1t}^{(0)}, \ldots, \gamma_{Jt}^{(0)})^\top$, where $\gamma_{jt}^{(0)} = (\sum_{i=1}^N m_{ijt})/N$.
8. Apply singular value decomposition to matrix $\tilde{M} = \left(\sum_{t=1}^T (m_{ijt} - \gamma_{jt}^{(0)})/T\right)_{N\times J}$ and obtain $\tilde{M} = \sum_{j=1}^J \tilde{\sigma}_j \tilde{\mathbf{q}}_j \tilde{\mathbf{h}}_j^\top$, where $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_J$ are the singular values and $\tilde{\mathbf{q}}_j$s and $\tilde{\mathbf{h}}_j$s are the left and right singular vectors.
9. Set $\mathbf{\Theta}_k^{(0)} = \sqrt{N}\tilde{\mathbf{q}}_k$ and $\mathbf{A}_k^{(0)} = \tilde{\sigma}_k \tilde{\mathbf{h}}_k/\sqrt{N}, k = 1, \ldots, K$.
10. Plug in $\Gamma^{(0)}$, $\Theta^{(0)} = (\mathbf{\Theta}_1^{(0)}, \ldots, \mathbf{\Theta}_K^{(0)})$ and $A^{(0)} = (\mathbf{A}_1^{(0)}, \ldots, \mathbf{A}_K^{(0)})$ in (2.5) and set $B^{(0)}$ and $V^{(0)}$ such that the log-likelihood is maximized. This can be done using the `glm` function in R.
11. Output $U^{(0)} = (\Gamma^{(0)}, B^{(0)}, V^{(0)}, A^{(0)})$ and $\Theta^{(0)}$ as starting point.

The tolerance ϵ is a positive constant that is close to 0. A default value $\epsilon = 0.01$ is used in the analysis of this appendix.

Remark A.2. *The above procedures can be easily adapted to obtain initial values for the models introduced in Sections 2.2.4 and 2.2.5. In particular, in the extension considered in Section 2.2.4, we can stack the T matrices into $L = (L_1, \ldots, L_T)$ in step 3. In step 4, we perform singular value decomposition on L, and in step 5, obtain $\tilde{L}$ by replacing $\hat{p}_t$ with the average $\sum_{t=1}^T \hat{p}_t/T$. Step 6 proceeds as before to compute $M = (M_1, \ldots, M_T)$. In step 7, each of γ_{jt} is computed from the corresponding block M_t from M, and we compute singular value decomposition in step 8 on*

$$\tilde{M} = \left((m_{ij1} - \gamma_{j1}^{(0)})_{N\times J}, \ldots, (m_{ijT} - \gamma_{jT}^{(0)})_{N\times J}\right).$$

We can then obtain $\mathbf{\Theta}^{(0)}$ and $A_1^{(0)} \ldots, A_T^{(0)}$ in step 9 and the rest of the procedure follows.

When the restriction $\gamma_{jt} = t\gamma_j$ is imposed, we compute $\Gamma^{(0)} = (\gamma_1^{(0)}, \ldots, \gamma_J^{(0)})^\top$, where $\gamma_j^{(0)}$ is the ordinary least square estimate given by $\gamma_j^{(0)} = (\sum_{t=1}^T \sum_{i=1}^N tm_{ijt})/(N \sum_{t=1}^T t^2)$ in step 7 and define $\tilde{M} = ((\sum_{t=1}^T m_{ijt} - t\gamma_j^{(0)})/T)_{N\times J}$ in Step 8. The rest of the steps follow.

Finally, when the extension Section 2.2.4 is considered and the restriction $\gamma_{jt} = t\gamma_j$ is imposed, we use a similar procefure, setting

$$\tilde{M} = \left(\left(m_{ij1} - \gamma_j^{(0)}\right)_{N\times J}, \ldots, \left(m_{ijT} - T\gamma_j^{(0)}\right)_{N\times J} \right)$$

using the ordinary least square estimate for Γ.

A.2 Normalisation criteria and algorithm

In Chapter 2, we imposed the normalisation condition as outlined in (2.4), which is fundamental for the identification of B. To ensure that the matrices $\mathit{\Gamma}$, Θ, and A are also identifiable, we can introduce additional normalisation constraints as follows:

$$\begin{aligned}
&\frac{1}{J} A^\top A = I_K, \\
&\frac{1}{N} \Theta^\top \Theta \text{ is diagonal with non-increasing diagonal elements.} \\
&\Theta^\top \mathbf{1_N} = \mathbf{0_K},
\end{aligned} \tag{A.1}$$

where $\mathbf{0_K}$ denotes a zero vector of length K. The normalisation criteria outlined above, along with (2.4), are sufficient to ensure the identifiability of the parameters in both the main model and the extension discussed in Section 2.2.3. Specifically, the parameter V associated with the time-dependent covariate is identifiable, provided there is sufficient variability in Z_t across any two time points, as detailed in Assumption A.2 in Section A.3. The following algorithm ensures the estimated parameters satisfy the normalisation criteria imposed in (2.4) and (A.1):

Suppose we have initial estimates $\mathit{\Gamma}$, A, B and Θ. For $t = 1, \ldots, T$, define $\mathit{\Gamma}_t = (\gamma_{1t}, \ldots, \gamma_{Jt})^\top$. Define $IX = (\mathbf{1_N}, X)$, $\tilde{\Theta} = \Theta - IX(IX^\top IX)^{-1} IX^\top \Theta$ and $L = (IX^\top IX)^{-1} IX^\top \Theta A^\top$. We can set $\hat{\mathit{\Gamma}} = (\hat{\mathit{\Gamma}}_1, \ldots, \hat{\mathit{\Gamma}}_T)$ and $\hat{B}$ such that

$$\begin{aligned}
\hat{\mathit{\Gamma}}_t &= \mathit{\Gamma}_t + L^\top_{[1,1:J]}, t = 1, \ldots T \\
\hat{B} &= B + L^\top_{[2:(p+1),1:J]}.
\end{aligned}$$

Now define $\mathit{\Sigma}_{JA} = A^\top A/J$ and $\mathit{\Sigma}_{NT} = \tilde{\Theta}^\top \tilde{\Theta}/N$. We apply eigendecomposition such that

$$\mathit{\Sigma}_{JA}^{1/2} \mathit{\Sigma}_{NT} \mathit{\Sigma}_{JA}^{1/2} = LDL^{-1},$$

where D is a diagonal matrix and L is the matrix whose columns are the eigenvectors of D. Then by setting

$$\hat{\Theta} = \tilde{\Theta} H^{-1} \text{ and } \hat{A} = AH^\top, \tag{A.2}$$

where $H = (\mathit{\Sigma}_{JA}^{-1/2} L)^\top$ would satisfy the required constraints.

It is simple to adjust the algorithm for the special case where only the identifiability of B is concerned. In this case we can set $L = (X^\top X)^{-1} X^\top \Theta A^\top$ and update $\hat{B} = B + L$,

$\hat{\Theta} = \Theta - X(X^\top X)^{-1}X^\top\Theta$. Similar adjustments apply to the normalisation algorithms described in Remarks A.4 and A.5, when only the identifiability of B is of interest, corresponding to the models discussed in Sections 2.2.4 and 2.2.5. Since Section A.5.1 also evaluates the recovery of latent variables, we adopt the full identifiability criterion and corresponding algorithms throughout the simulation studies and real data analysis.

Remark A.3. *We note that from this updated normalisation criteria, Assumption 2.3 is equivalent to requiring that* $N^{-1}\Theta^\top\Theta = diag(\sigma_{N1}, \ldots, \sigma_{NK})$ *with* $\sigma_{N1} \geq \sigma_{N2} \cdots \geq \sigma_{NK}$*, and* $\sigma_{Nk} \to \sigma_k$ *as* $N \to \infty$ *for all* k*. While this does not affect the identification of* $\hat{B}$*, it is well known that the latent factors remain unidentifiable without an eigengap that separates different components. Therefore, throughout this supplementary material, we additionally assume that* $\infty > \sigma_1 > \sigma_2 > \cdots > \sigma_K > 0$.

Remark A.4. *In the context of the extension discussed in Section 2.2.4, we need to adjust the normalisation constraints A.1. Specifically, the parameters in this extension are identifiable given the following normalisation criteria:*

$$\begin{aligned}
&\frac{1}{J}A_1^\top A_1 = I_K, \\
&\frac{1}{N}\Theta^\top\Theta \textit{ is diagonal with non-increasing diagonal elements.} \\
&\Theta^\top(X, \mathbf{1_N}) = 0_{K\times(p+1)},
\end{aligned} \tag{A.3}$$

where we define $A_t = (a_{jkt})_{J\times K}$ *for* $t \in \{1, \ldots, T\}$*. The normalisation algorithm described above can be adapted accordingly. Specifically, we set* $\tilde{\Theta} = \Theta - IX(IX^\top IX)^{-1}IX^\top\Theta$*. For* $t = 1, \ldots, T$*, we set* $L^{(t)} = (IX^\top IX)^{-1}IX^\top\Theta A_t^\top$*, such that* $\hat{\mathit{\Gamma}}_t$ *and* $\hat{B}_t$ *are given by*

$$\begin{aligned}
\hat{\mathit{\Gamma}}_t &= \mathit{\Gamma}_t + L^{(t)\top}_{[1,1:J]} \\
\hat{B}_t &= B_t + L^{(t)\top}_{[2:(p+1),1:J]}.
\end{aligned}$$

We can then compute $\hat{\Theta}$ *and* $\hat{A}_1$ *as in* (A.2) *and set* $\hat{A}_t = A_t H^\top$ *for* $t = 2, \ldots, T$.

Remark A.5. *When the constraint* $\gamma_{jt} = t\gamma_j$ *is imposed, the condition* $\Theta^\top\mathbf{1}_N = \mathbf{0}_K$ *is no longer required in either the main model or the extensions in Sections 2.2.3 and 2.2.4, as the remaining conditions are sufficient to ensure identifiability. The normalisation algorithm can be modified accordingly by setting* $\tilde{\Theta} = \Theta - X(X^\top X)^{-1}X\Theta$*, and defining* $L = (X^\top X)^{-1}X^\top\Theta A^\top$ *for the main model. Similarly, for the extension described in Section 2.2.4, we use* $L^{(t)} = (X^\top X)^{-1}X^\top\Theta A_t^\top$*. In both cases, there is no need to normalise* $\mathit{\Gamma} = (\gamma_1, \ldots, \gamma_J)^\top$*. The parameters* $\hat{B}$ *or* $\hat{B}_t$ *can then be updated accordingly, and the remainder of the algorithm remains unaffected by this restriction.*

A.3 Additional Assumptions and Theoretical Results

In this section, we present the additional assumptions and theoretical results corresponding to the extensions introduced in Sections 2.2.3 and 2.2.4, along with the necessary adjustments to accommodate the constraint $\gamma_{jt} = t\gamma_j$ discussed in Section 2.2.5. The definitions of $\mathbf{u}_j$ and $\mathbf{e}_{it}$ follow those provided in Table A.1, according to the relevant extension.

A.3.1 Extension in Section 2.2.3

Define $Z_t = (\mathbf{z}_{1t}, \mathbf{z}_{2t}, \ldots, \mathbf{z}_{Nt})$ and $U_t = (\mathit{\Gamma}_t, A, V, B)$ for $t = 1, \ldots T$. To establish asymptotic properties under this extension, the following additional assumptions are necessary:

Assumption A.1. *$\mathbf{z}_{it} \in \mathcal{Z}$ for all i and t, where $\mathcal{Z} \subset \mathbb{R}^{p_z}$ is compact.*

Assumption A.2. *For some $\kappa_4 > 0$, there exists $t_1, t_2 \in \{1, \ldots, T\}$ such that*

$$\liminf_{N\to\infty} \pi_{\min}\left((Z_{t_1} - Z_{t_2}, \mathbf{1_N})^\top (Z_{t_1} - Z_{t_2}, \mathbf{1_N}) \right) / N \geq \kappa_4.$$

Assumption A.3. *There exists $\kappa_5 > 0$ such that the minimum eigenvalue of the matrix*

$$\begin{pmatrix} \mathit{diag}\left(\left\{(NJ)^{-1/2}\sum_{i=1}^N \sum_{t=1}^T \mathbf{e}_{it}^* \mathbf{e}_{it}^{*\top}\right\}_{j\leq J}\right) & \left\{(NJ)^{-1/2}\sum_{t=1}^T \begin{pmatrix} \mathbf{0}_{T+p} \\ \mathbf{z}_{it} \\ \mathbf{0}_{K^*} \end{pmatrix} \mathbf{a}_j^{*\top}\right\}_{j\leq J, i\leq N} \\ (NJ)^{-1/2}\left\{\sum_{t=1}^T \mathbf{a}_j^* \left(\mathbf{0}_{T+p}^\top, \mathbf{z}_{it}^\top, \mathbf{0}_{K^*}^\top\right)\right\}_{i\leq N, j\leq J} & \mathit{diag}\left(\left\{(NJ)^{-1/2} T \sum_{j=1}^J \mathbf{a}_j^* \mathbf{a}_j^{*\top}\right\}_{i\leq N}\right) \end{pmatrix}$$

is greater that κ_5. Here $diag(\{Q_j\}_{j\leq J})$ represents a block-diagonal matrix whose jth block on the diagonal is Q_j, for any series of matrices $\{Q_j\}_{j\leq J}$. Similarly $\{Q_{ij}\}_{j\leq J, i\leq N}$ represents the block matrix where the $\{i,j\}$th block is Q_{ij}.

Assumption A.1 extends Assumption 2.1 to cater for the time-dependent covariate $\mathbf{z}_{it}$. Assumption A.2 asserts sufficient variability in Z_t across at least two time periods, ensuring the identifiability of $\mathbf{v}_j$. Assumption A.3 is a working assumption for the proof of asymptotic normality. As discussed in Section 2.2.5, when the constraint $\gamma_{jt} = t\gamma_j$ is imposed, the intercept term $\gamma_{jt}\mathbf{1_N}$ is effectively replaced by $\gamma_j(t\mathbf{1_N})$, making $t\mathbf{1_N}$ a component of the time-dependent covariate vector Z_t. In this setting, Assumptions A.1–A.3 can be naturally adapted to account for this change, and the subsequent theorems remain valid under the constraint. For instance, Assumption A.2 can be rewritten as $\liminf_{N\to\infty} \pi_{\min}\left((Z_{t_1} - Z_{t_2})^\top (Z_{t_1} - Z_{t_2})\right)/N \geq \kappa_4$.

We now present the the following theorems under the normalisation criteria discussed in (2.4) and (A.1):

Theorem A.1. *Under Assumptions 2.1 to 2.5, A.1 and A.2, we have*

$$\frac{1}{\sqrt{N}} \|\hat{\Theta} - \Theta^* \hat{S}_A\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}),$$
$$\frac{1}{\sqrt{J}} \|\hat{U}_t - U_t^* \hat{S}_U\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}).$$

Here, $\hat{S}_A$ is defined as $sgn(\hat{A}^\top A^*/J)$, where the function $sgn(X)$ yields a diagonal matrix whose diagonal elements are the signs of the diagonal elements of any square matrix X. Moreover, $\hat{S}_U$ is a $(1+p+p_z+K^*)$ by $(1+p+p_z+K^*)$ diagonal matrix whose diagonal elements are set to 1, except for the last K^* diagonal elements which are equal to $\hat{S}_A$.

Theorem A.2. *Under Assumptions 2.1 to 2.7 and A.1 to A.3, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, we have*

$$\sqrt{N}\left(\hat{\mathbf{u}}_j - \hat{S}_U \mathbf{u}_j^*\right) \xrightarrow{d} \mathcal{N}\left(0, -\mathit{\Phi}_j^{-1}\right) \text{ and } \sqrt{J}\left(\hat{\boldsymbol{\theta}}_i - \hat{S}_A \boldsymbol{\theta}_i^*\right) \xrightarrow{d} \mathcal{N}\left(0, -\mathit{\Psi}_i^{-1}\right).$$

Theorem A.3. *Suppose that Assumptions 2.1 to 2.5, A.1 and A.2 hold and $K^* \in \mathcal{K}$. If the penalty term Λ_{NJ} satisfies $\max\{N, J\} \lesssim \Lambda_{NJ} \lesssim NJ$, then*

$$\lim_{N,J\to\infty} P(\hat{K} = K^*) = 1.$$

Theorem A.1 provides the average rate of convergence of $\hat{\Theta}$ and $\hat{U}_t, t = 1, \ldots, T$. The sign matrices $\hat{S}_A$ and $\hat{S}_U$ are introduced to address the inherent sign indeterminacy in factors and loadings, that is, the factor structure remains unchanged if both the factors and loadings are multiplied by -1. Theorem A.2 establishes asymptotic normality for event-specific parameters. Similar results can also be obtained for the individual parameters $\hat{\theta}_i$, with a modified form of asymptotic variance due to the construction of the estimator. Theorem A.3 extends Theorem 2.4, proving that the number of latent factors K can be consistently estimated with time-dependent covariates. The proofs of these Theorems will be presented in Section A.4.

We point out that Theorems A.1 and A.2 hold without Assumptions A.1 to A.3 when only static covariates are considered. They can also be viewed as more general versions of Theorems 2.1 to 2.3. We will make that connection explicit in Section A.4, where we present the proofs of these theoretical results.

A.3.2 Extension in Section 2.2.4

Define $A_t = (a_{jkt})_{J\times K}$ and $U_t = (\mathit{\Gamma}_t, A_t, V, B)$ for $t = 1, \ldots, T$ in this context, with the understanding that $\mathit{\Gamma}_t = \mathit{\Gamma} = (\gamma_1, \ldots, \gamma_J)^\top$ for all t when $\gamma_{jt} = t\gamma_j$ is imposed. To establish asymptotic properties under this extension, the following assumptions are necessary:

Assumption A.4. *For $t = 1, \ldots, T$, $J^{-1}A_t^{*\top}A_t^*$ converges to a positive definite matrix as J tends to infinity. Also, $N^{-1}\Theta^{*\top}\Theta^*$ converge to a positive definite matrix as N tends to infinity.*

Assumption A.5. *There exists $\kappa_6 > 0$ such that the minimum eigenvalue of the matrix*

$$\sum_{t=1}^{T} \begin{pmatrix} diag\left(\left\{\sum_{i=1}^{N} \mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}\right\}_{j\le J}\right) & \left\{\begin{pmatrix} \mathbf{0}_{T+Tp} \\ \mathbf{z}_{it} \\ \mathbf{0}_{K^*} \\ D_{it2}\boldsymbol{\theta}_i^* \\ \vdots \\ D_{itT}\boldsymbol{\theta}_i^* \end{pmatrix} \mathbf{a}_{jt}^{*\top}\right\}_{j\le J, i\le N} \\ \left\{\mathbf{a}_{jt}^*\left(\mathbf{0}_{T+Tp}^\top, \mathbf{z}_{it}^\top, \mathbf{0}_{K^*}^\top, D_{it2}\boldsymbol{\theta}_i^{*\top}, \ldots, D_{itT}\boldsymbol{\theta}_i^{*\top}\right)\right\}_{i\le N, j\le J} & diag\left(\left\{\sum_{j=1}^{J} \mathbf{a}_{jt}^*\mathbf{a}_{jt}^{*\top}\right\}_{i\le N}\right) \end{pmatrix}$$

is greater that $(NJ)^{1/2}\kappa_6$. When $\gamma_{jt} = t\gamma_j$, we require the minimum eigenvalue of the *matrix*

$$\sum_{t=1}^{T}\begin{pmatrix} diag\left(\left\{\sum_{i=1}^{N}\mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top}\right\}_{j\leq J}\right) & \left\{\begin{pmatrix} t \\ \mathbf{0}_{Tp} \\ \mathbf{z}_{it} \\ \mathbf{0}_{K^*} \\ D_{it2}\boldsymbol{\theta}_i^{*} \\ \vdots \\ D_{itT}\boldsymbol{\theta}_i^{*} \end{pmatrix}\mathbf{a}_{jt}^{*\top}\right\}_{j\leq J, i\leq N} \\ \left\{\mathbf{a}_{jt}^{*}\left(t, \mathbf{0}_{Tp}^{\top}, \mathbf{z}_{it}^{\top}, \mathbf{0}_{K^*}^{\top}, D_{it2}\boldsymbol{\theta}_i^{*\top}, \ldots, D_{itT}\boldsymbol{\theta}_i^{*\top}\right)\right\}_{i\leq N, j\leq J} & diag\left(\left\{\sum_{j=1}^{J}\mathbf{a}_{jt}^{*}\mathbf{a}_{jt}^{*\top}\right\}_{i\leq N}\right) \end{pmatrix}$$

to be greater than $(NJ)^{1/2}\kappa_6$*.*

Assumption A.6. *For some* $\kappa_7 > 0$*, there exists* $t_1, t_2 \in \{1, \ldots, T\}$ *such that*

$$\liminf_{N\to\infty} \pi_{\min}((\Theta^*, X, \mathbf{1_N}, Z_{t_1}, Z_{t_2})^\top(\Theta^*, X, \mathbf{1_N}, Z_{t_1}, Z_{t_2}))/N \geq \kappa_7.$$

In the case $\gamma_{jt}^* = t\gamma_j$*, we impose additional condition that*

$$\liminf_{J\to\infty} \pi_{\min}((\mathit{\Gamma}^*, A_{t_1}^*, A_{t_1}^*)^\top(\mathit{\Gamma}^*, A_1^*, A_2^*)) \geq \kappa_7.$$

Assumptions A.4 and A.5 replaces Assumption 2.3 and A.3 under this extension, respectively. Assumption A.6 replaces Assumptions 2.5 and A.2 and deserves more explanation. The condition involving $(\Theta^*, X, \mathbf{1}_N, Z_{t_1}, Z_{t_2})$ implies that the static factor Θ^* must not lie in the span of (Z_{t_1}, Z_{t_2}). We illustrate the necessity of such assumption through the following simple example: Suppose $K = 1$ and $T = 2$, and we are given $\boldsymbol{\theta}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{v}, \mathbf{z}_1$ and $\mathbf{z}_2$ such that $\boldsymbol{\theta} = \mathbf{z}_1 + \mathbf{z}_2$ and $\mathbf{a}_2 = -2\mathbf{a}_1$. We can verify that

$$\boldsymbol{\theta}\mathbf{a}_t^\top + \mathbf{z}_t\mathbf{v}^\top = \tilde{\boldsymbol{\theta}}\tilde{\mathbf{a}}_t^\top + \mathbf{z}_t\tilde{\mathbf{v}}^\top, \text{ for } t = 1, 2, \text{ where}$$

$\tilde{\boldsymbol{\theta}} = 2(\mathbf{z}_1 + 0.5\mathbf{z}_2), \tilde{\mathbf{a}}_1 = \mathbf{a}_1, \tilde{\mathbf{a}}_2 = 0.5\mathbf{a}_2$ and $\tilde{\mathbf{v}} = -\mathbf{a}_1 + \mathbf{v}$. Similarly, the additional condition when $\gamma_{jt}^* = t\gamma^*$ ensures that $\mathit{\Gamma} = (\gamma_1, \ldots, \gamma_J)$ does not lie in the span of $A_{t_1}^*$ and $A_{t_2}^*$. We note that although $\mathit{\Gamma}$ could be absorbed into $\mathbf{v}$ and $t\mathbf{1}_N$ treated as part of Z_t for estimation, the original assumption no longer ensures identification, as $t_1\mathbf{1}_N$ and $t_2\mathbf{1}_N$ become linearly dependent for all $t_1, t_2 \in \{1, \ldots, T\}$. The additional condition is therefore necessary and specifically tailored to address this scenario. Define $U_t = (\mathit{\Gamma}_t, A_t, V, B)$ for $t = 1, \ldots, T$. The following theorems, analogous to Theorems A.1, A.2 and A.3 hold for this extension.

Theorem A.4. *Under Assumptions 2.1, 2.2, A.4, 2.4, A.1 and A.6, we have*

$$\frac{1}{\sqrt{N}}\|\hat{\Theta} - \Theta^*\hat{S}_A\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}),$$
$$\frac{1}{\sqrt{J}}\|\hat{U}_t - U_t^*\hat{S}_U\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}).$$

Here, $\hat{S}_A$ *is defined as* $sgn(\hat{A}_1^\top A_1^*/J)$*, where the function* $sgn(X)$ *yields a diagonal matrix whose diagonal elements are the signs of the diagonal elements of any square matrix* X*. Moreover,* $\hat{S}_U$ *is a* $(1+p+p_z+K^*)$ *by* $(1+p+p_z+K^*)$ *diagonal matrix whose diagonal elements are set to 1, except for the last* K^* *diagonal elements which are equal to* $\hat{S}_A$*.*

Theorem A.5. *Under Assumptions 2.1, 2.2, A.4, 2.4, 2.6, 2.7, A.1,A.5 and A.6, for $i = 1, \dots, N$ and $j = 1, \dots, J$, we have*

$$\sqrt{N}\left(\hat{\mathbf{u}}_j - \hat{S}_U \mathbf{u}_j^*\right) \xrightarrow{d} \mathcal{N}\left(0, -\Phi_j^{-1}\right) \text{ and } \sqrt{J}\left(\hat{\boldsymbol{\theta}}_i - \hat{S}_A \boldsymbol{\theta}_i^*\right) \xrightarrow{d} \mathcal{N}\left(0, -\Psi_i^{-1}\right).$$

Theorem A.6. *Suppose that Assumptions 2.1, 2.2, A.4, 2.4, A.1 and A.6 hold and $K^* \in \mathcal{K}$. If the penalty term Λ_{NJ} satisfies $\max\{N, J\} \lesssim \Lambda_{NJ} \lesssim NJ$, then*

$$\lim_{N,J\to\infty} P(\hat{K} = K^*) = 1.$$

In practice, we set $\Lambda_{NJ} = \max\{N, TJ\} \times \log\left(\max\{N, TJ\}^{-1} J \sum_{i=1}^{N} \sum_{t=1}^{T} r_{it}\right)$ to reflect that under this extension, each additional latent dimension introduces TJ new parameters for each item at each time point.

The proofs of these theorems will be presented in Section A.4. We conclude this section by noting that results analogous to those stated in Theorems 2.1 to 2.3 follow from the proof of the above theorems, when the focus is on the regression coefficients B^*.

A.4 Proofs

In this section, we present the proofs of the theoretical results discussed in Chapter 2 and this appendix. Specifically, Sections A.4.1, A.4.2, A.4.3, A.4.4 and A.4.5 are dedicated to the proofs of Theorems A.1, A.2, A.3, A.4 and A.5, respectively. Moreover, Section A.4.6 establishes the consistency of the estimator introduced in Remark 2.1, for estimating the asymptotic variance of $\hat{\boldsymbol{\beta}}_j$, where $j = 1, \dots, J$. The argument is general and also applies to the estimators of asymptotic variances under the extensions discussed in Sections 2.2.3 and 2.2.4, as well as to the asymptotic variance estimators for $\hat{\mathbf{u}}_j$ and $\hat{\boldsymbol{\theta}}_i$ using the same approach.

Furthermore, it is essential to recognize that Theorem 2.1 emerges directly as a corollary of Lemma A.4, which is introduced and proved in Section A.4.2. Subsequently, Theorem 2.2 follows from the proof of Theorem A.1. Additionally, Theorem 2.3 follows from Theorem A.2. Lastly, Theorem 2.4 is a special case of Theorem A.3, and Theorem A.6 follows by an analogous argument, so the detailed derivation is omitted.

To simplify notation, the proofs of the theorems and results in Sections A.4.1 to A.4.6 are conducted under the assumption that the scale parameters ϕ_j are known and set to 1. Notably, the derivatives of $l(\Xi)$ with respect to the parameters $\mathbf{u}_j$ and $\boldsymbol{\theta}_i$ are equivalent to the derivatives of the full joint log-likelihood function. It is easy to see that Theorems A.1, A.2, A.4 and A.5 hold when ϕ_j are unknown as the estimated parameters in these theorems are unaffected by the scale parameters. Similarly, Theorem A.3 remains valid since the information criterion considered in the work does not depend on the scale parameter. Finally, it is easy to see that estimator introduced in Remark 2.1 remains consistent when ϕ_j is unknown as long as the estimate of $\hat{\phi}_j$ is consistent.

A.4.1 Proof of Theorem A.1

Throughout this section, $\delta_0, \delta_1, \delta_2, \dots$ denote positive constants that do not depend on N, J. For any random variable Y, define the Orlicz norm $\|Y\|_\Psi$ as

$$\|Y\|_\Psi = \inf\{C > 0 : E\Psi(|Y|/C) \leq 1\},$$

where Ψ is a non-decreasing, convex function with $\Psi(0)=0$. We write the norm as $\|Y\|_{\Psi_2}$ when $\Psi(x)=\exp(x^2)-1$. We use $\|\cdot\|_S$ to denote the spectral norm. Additionally, we define $C(\cdot,g,\mathcal{G})$ to denote the covering number of space $\mathcal{G}$ endowed with semimetric g. We further define $M_t=(m_{ijt})$, where $m_{ijt}=\mathbf{u}_j^\top\mathbf{e}_{it}, t=1\ldots T$. For any $\Xi^{(a)},\Xi^{(b)}\in\mathcal{H}^{K^*}$, define $d(\Xi^{(a)},\Xi^{(b)})=\max_{t:t=1,\ldots T}\|M_t^{(a)}-M_t^{(b)}\|_F/\sqrt{NJ}$. Let $\boldsymbol{\rho}_{ij}=\boldsymbol{\rho}_{ij}(\mathbf{u}_j^*,\boldsymbol{\theta}_i^*)$, $\rho_{ijt}=\rho_{ijt}(\mathbf{u}_j^{*\top}\mathbf{e}_{it}^*)$, $w_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i)=\boldsymbol{\rho}_{ij}-\boldsymbol{\rho}_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i)$ and define

$$l_{NJ}^*(\Xi)=\frac{1}{NJ}\sum_{i=1}^N\sum_{j=1}^J w_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i),\qquad \bar{l}_{NJ}^*(\Xi)=\frac{1}{NJ}\sum_{i=1}^N\sum_{j=1}^J E\left(w_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i)\right).$$

We further define

$$\mathbb{W}_{NJ}(\Xi)=l_{NJ}^*(\Xi)-\bar{l}_{NJ}^*(\Xi)=\frac{1}{NJ}\sum_{i=1}^N\sum_{j=1}^J\left(w_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i)-E\left(w_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i)\right)\right).$$

We prove the following three Lemmas. Theorem A.1 then follows from the proof of Theorem 1 in Chen et al. (2021).

Lemma A.1. *Under Assumptions 2.1,2.2, 2.4 and A.1, $d(\hat{\Xi},\Xi^*)=o_p(1)$ as $J,N\to\infty$.*

Proof: Pick any Ξ from the set $\mathcal{H}^{K^*}$. By taking $m_{ijt}=\mathbf{u}_j^\top\mathbf{e}_{it}$ and expanding around $m_{ijt}^*=\mathbf{u}_j^{*\top}\mathbf{e}_{it}^*$ for $t=1,\ldots,T$, we have

$$\begin{aligned}
&E\left(\boldsymbol{\rho}_{ij}-\boldsymbol{\rho}_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i)\right)\\
=&E\left(\boldsymbol{\rho}_{ij}-\boldsymbol{\rho}_{ij}-\sum_{t=1}^T r_{it}(m_{ijt}-m_{ijt}^*)\rho_{ijt}^{'}(m_{ijt}^*)-0.5\sum_{t=1}^T r_{it}(m_{ijt}-m_{ijt}^*)^2\rho_{ijt}^{''}(\tilde{m}_{ijt})\right)\\
=&E\left(-0.5\sum_{t=1}^T r_{it}(m_{ijt}-m_{ijt}^*)^2\rho_{ijt}^{''}(\tilde{m}_{ijt})\right)\\
=&0.5\sum_{t=1}^T\left\{(m_{ijt}-m_{ijt}^*)^2 b_j^{''}(\tilde{m}_{ijt})P(r_{it}=1)\right\},
\end{aligned}$$

where $\tilde{m}_{ijt}$ lies between m_{ijt} and $m_{ijt}^*, t=1,\ldots T$. Therefore, we have

$$\begin{gathered}E\left(\boldsymbol{\rho}_{ij}-\boldsymbol{\rho}_{ij}(\mathbf{u}_j,\boldsymbol{\theta}_i)\right)\gtrsim(m_{ijt}-m_{ijt}^*)^2, t=1,\ldots T\text{ and}\\ \bar{l}_{NJ}^*(\Xi)\gtrsim d^2(\Xi,\Xi^*).\end{gathered}\tag{A.4}$$

by Assumptions 2.2 and 2.4. Also, by the definition of $\hat{\Xi}$, we have $l_{NJ}^*(\hat{\Xi})=l_{NJ}(\Xi^*)-l_{NJ}(\hat{\Xi})\le 0$, or equivalently $\mathbb{W}_{NJ}(\hat{\Xi})+\bar{l}_{NJ}^*(\hat{\Xi})\le 0$. Combining it with (A.4), we have

$$0\le d^2(\hat{\Xi},\Xi^*)\lesssim\bar{l}_{NJ}^*(\hat{\Xi})\le\sup_{\Xi\in\mathcal{H}^{K^*}}|\mathbb{W}_{NJ}(\Xi)|.$$

So it remains to show that

$$\sup_{\Xi\in\mathcal{H}^{K^*}}|\mathbb{W}_{NJ}(\Xi)|=o_p(1).\tag{A.5}$$

By Assumption 2.1, we can choose $\delta_1>1$ large enough such that $\|\mathbf{u}_j\|,\|\boldsymbol{\theta}_i\|\le\delta_1$ for all i,j, $\mathbf{u}_j\in\mathcal{U}$ and $\boldsymbol{\theta}_i\in\boldsymbol{\Theta}$. Let $B_d(\delta_1)$ denote a Euclidean ball in $\mathbb{R}^d$ with radius δ_1 for any positive integer d. For any $\epsilon > 0$, let $\mathbf{u}_{(1)}, \ldots, \mathbf{u}_{(q_P)}$ be a maximal set of points in $B_P(\delta_1)$ such that $\|\mathbf{u}_{(h)} - \mathbf{u}_{(k)}\| > \epsilon/\delta_1$ for any $h \neq k$. Here "maximal" signifies that no point can be added without violating the validity of the inequality. Similarly, let $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(q_K)}$ be a maximal set of points in $B_K(\delta_1)$ such that $\|\boldsymbol{\theta}_{(h)} - \boldsymbol{\theta}_{(k)}\| > \epsilon/\delta_1$ for any $h \neq k$. It is well known that Q_P and Q_{K^*}, the packing numbers of $B_P(\delta_1)$ and $B_{K^*}(\delta_1)$, respectively, are bounded by $\delta_2(\delta_1/\epsilon)^P$. For any $\boldsymbol{\Xi} \in \mathcal{H}^{K^*}$, define $\bar{\boldsymbol{\Xi}} = (\bar{\mathbf{u}}_1^\top, \ldots, \bar{\mathbf{u}}_J^\top, \bar{\boldsymbol{\theta}}_1^\top, \ldots, \bar{\boldsymbol{\theta}}_N^\top)^\top$, where $\bar{\mathbf{u}}_j = \{\mathbf{u}_{(q_j)} : q_j = \min\{q : q \leq Q_P, \|\mathbf{u}_{(q)} - \mathbf{u}_j\| \leq \epsilon/\delta_1\}\}$ and $\bar{\boldsymbol{\theta}}_i = \{\boldsymbol{\theta}_{(q_i)} : q_i = \min\{q : q \leq Q_{K^*}, \|\boldsymbol{\theta}_{(q)} - \boldsymbol{\theta}_i\| \leq \epsilon/\delta_1\}\}$. This definition ensures that each $\boldsymbol{\Xi} \in \mathcal{H}^{K^*}$ is uniquely sent to a $\bar{\boldsymbol{\Xi}} \in \mathcal{H}^{K^*}$ comprised of the maximal sets of points defined previously. Thus, we can write

$$\mathbb{W}_{NJ}(\boldsymbol{\Xi}) = \mathbb{W}_{NJ}(\bar{\boldsymbol{\Xi}}) + \mathbb{W}_{NJ}(\boldsymbol{\Xi}) - \mathbb{W}_{NJ}(\bar{\boldsymbol{\Xi}}).$$

Define $\bar{\mathbf{e}}_{it} = (\mathbf{D}_{it}^\top, \mathbf{x}_i^\top, \mathbf{z}_{it}^\top, \bar{\boldsymbol{\theta}}_i^\top)^\top$. By Assumption 2.2, there exists $\delta_0 > 0$ such that

$$\begin{aligned}
&\left|\boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i) - \boldsymbol{\rho}_{ij}(\bar{\mathbf{u}}_j, \bar{\boldsymbol{\theta}}_i)\right| \\
\leq &\sum_{t=1}^{T} \left|\rho_{ijt}(\mathbf{u}_j^\top \mathbf{e}_{it}) - \rho_{ijt}(\bar{\mathbf{u}}_j^\top \bar{\mathbf{e}}_{it})\right| \\
\leq &\delta_0 \sum_{t=1}^{T} \left|\mathbf{u}_j^\top \mathbf{e}_{it} - \bar{\mathbf{u}}_j^\top \bar{\mathbf{e}_{it}}\right| \\
\leq &\delta_0 \sum_{t=1}^{T} \Big(|\gamma_{jt} - \bar{\gamma}_{jt}| + |\boldsymbol{\beta}_j^\top \mathbf{x}_i - \bar{\boldsymbol{\beta}}_j^\top \mathbf{x}_i| + |\mathbf{v}_j^\top \mathbf{z}_{it} - \bar{\mathbf{v}}_j^\top \mathbf{z}_{it}| + |\mathbf{a}_j{}^\top \boldsymbol{\theta}_i - \bar{\mathbf{a}}_j^\top \bar{\boldsymbol{\theta}}_i|\Big) \\
\leq &\delta_0 T \left(\|\boldsymbol{\gamma}_j - \bar{\boldsymbol{\gamma}}_j\| + \|\mathbf{a}_j\|\|\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}}_i\| + \|\bar{\boldsymbol{\theta}}_i\|\|\mathbf{a}_j - \bar{\mathbf{a}}_j\| + \|\mathbf{x}_i\|\|\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\| + \|\mathbf{z}_{it}\|\|\mathbf{v}_j - \bar{\mathbf{v}}_j\|\right) \\
\leq &5\delta_0 T\epsilon. \qquad\qquad (A.6)
\end{aligned}$$

Thus, we have

$$\sup_{\boldsymbol{\Xi} \in \mathcal{H}^{K^*}} \|\mathbb{W}_{NJ}(\boldsymbol{\Xi}) - \mathbb{W}_{NJ}(\bar{\boldsymbol{\Xi}})\| \leq 10\delta_0 T\epsilon. \tag{A.7}$$

Also, note that

$$\left|w_{ij}(\bar{\mathbf{u}}_j, \bar{\boldsymbol{\theta}}_i)\right| = \left|\boldsymbol{\rho}_{ij} - \boldsymbol{\rho}_{ij}(\bar{\mathbf{u}}_j, \bar{\boldsymbol{\theta}}_i)\right| \leq \delta_0 \sum_{t=1}^{T} \left|\mathbf{u}_j^{*\top} \mathbf{e}_{it}^* - \bar{\mathbf{u}}_j^\top \bar{\mathbf{e}}_{it}\right|.$$

By Cauchy-Schwarz inequality,

$$\left(\sum_{t=1}^{T} \left|\mathbf{u}_j^{*\top} \mathbf{e}_{it}^* - \bar{\mathbf{u}}_j^\top \bar{\mathbf{e}}_{it}\right|\right)^2 \leq T \sum_{t=1}^{T} \left|\mathbf{u}_j^{*\top} \mathbf{e}_{it}^* - \bar{\mathbf{u}}_j^\top \bar{\mathbf{e}}_{it}\right|^2.$$

By Hoeffding's inequality, we have

$$P\left(\left|\sqrt{NJ}\mathbb{W}_{NJ}(\bar{\boldsymbol{\Xi}})\right| > c\right) \leq 2\exp\left(-\frac{2c^2}{\delta_0^2 T^3 \cdot d^2(\bar{\boldsymbol{\Xi}}, \boldsymbol{\Xi}^*)}\right),$$

and by Lemma 2.2.1 of Van Der Vaart et al. (1996), it follows that $\|\mathbb{W}_{NJ}(\bar{\boldsymbol{\Xi}})\|_{\Psi_2} \lesssim d(\bar{\boldsymbol{\Xi}}, \boldsymbol{\Xi}^*)/\sqrt{NJ}$. Since $\bar{\boldsymbol{\Xi}}$ can take at most $Q_P^J \times Q_{K^*}^N \lesssim (\delta_1/\epsilon)^{P(N+J)}$ different values, and

$d(\bar{\Xi}, \Xi^*) \lesssim \delta_1$, it follows from Lemma 2.2.2 of Van Der Vaart et al. (1996) that

$$\begin{aligned} E\left(\sup_{\Xi \in \mathcal{H}^{K^*}} \left|\mathbb{W}_{NJ}(\bar{\Xi})\right|\right) &\leq \left\|\sup_{\Xi \in \mathcal{H}^{K^*}} \left|\mathbb{W}_{NJ}(\bar{\Xi})\right|\right\|_{\Psi_2} \\ &\lesssim \sqrt{\log(\delta_1/\epsilon)}\sqrt{P(N+J)}/\sqrt{NJ} \\ &\lesssim \sqrt{\log(\delta_1/\epsilon)} \min\left\{\sqrt{N}, \sqrt{J}\right\}^{-1}. \end{aligned} \tag{A.8}$$

Finally, by Markov's inequality and (A.7), for any $c > 0$,

$$\begin{aligned} P\left(\sup_{\Xi \in \mathcal{H}^{K^*}} |\mathbb{W}_{NJ}(\Xi)| > c\right) \leq & P\left(\sup_{\Xi \in \mathcal{H}^{K^*}} \left|\mathbb{W}_{NJ}(\bar{\Xi})\right| > \frac{c}{2}\right) \\ & + P\left(\sup_{\Xi \in \mathcal{H}^{K^*}} \left|\mathbb{W}_{NJ}(\Xi) - \mathbb{W}_{NJ}(\bar{\Xi})\right| > \frac{c}{2}\right) \\ \leq & \frac{2}{c} E\left(\sup_{\Xi \in \mathcal{H}^{K^*}} \left|\mathbb{W}_{NJ}(\bar{\Xi})\right|\right) + P\left(10\delta_0 T\epsilon > \frac{c}{2}\right). \end{aligned}$$

Thus, by choosing $c = 30\delta_0 T\epsilon$, (A.5) follows from (A.8) and the fact that ϵ is arbitrary, which concludes the proof.

Lemma A.2. *Define $\mathcal{H}^{K^*}(c) = \left\{\Xi \in \mathcal{H}^{K^*} : d(\Xi, \Xi^*) \leq c\right\}$. Under Assumptions 2.1-2.5 and A.1-A.2, for sufficiently small $c > 0$ and sufficiently large N and J, for any $\Xi \in \mathcal{H}^{K^*}(c)$, it holds that*

$$\|\Theta - \Theta^* S_A\|_F/\sqrt{N} + \|U_t - U_t^* S_U\|_F/\sqrt{J} \leq \delta_3 c, t = 1, \ldots, T,$$

where $S_A = sgn\left(A^\top A^*/J\right)$ and S_U is a $(1 + p + p_z + K^*) \times (1 + p + p_z + K^*)$ diagonal matrix whose diagonal elements are 1, except for the last K^* diagonal elements which are equal to S_A.

Proof: We first prove that $\|V - V^*\|_F/\sqrt{J} \lesssim d(\Xi, \Xi^*)$. By Assumption A.2, there exists $t_1, t_2 \in \{1, \ldots, T\}$ such that $\pi_{\min}\left((Z_{t_1} - Z_{t_2}, \mathbf{1_N})^\top (Z_{t_1} - Z_{t_2}, \mathbf{1_N})\right)/N \geq \kappa_4/2$ for sufficiently large N. Without loss of generality, assume $t_1 = 1$ and $t_2 = 2$. We have

$$\begin{aligned} & \frac{1}{\sqrt{NJ}} \left\|(Z_1 - Z_2)\left(V^\top - V^{*\top}\right) + \mathbf{1_N}\{(\mathit{\Gamma}_1 - \mathit{\Gamma}_1^*) - (\mathit{\Gamma}_2 - \mathit{\Gamma}_2^*)\}^\top\right\|_F \\ \leq & \frac{1}{\sqrt{NJ}} \sum_{t=1}^{2} \left\|\Theta A^\top - \Theta^* A^* + X(B - B^*)^\top + Z_t\left(V^\top - V^{*\top}\right) + \mathbf{1_N}(\mathit{\Gamma}_t - \mathit{\Gamma}_t^*)^\top\right\|_F \\ \leq & 2d(\Xi, \Xi^*). \end{aligned}$$

Let $Q = (V - V^{*}, (\mathit{\Gamma}_1 - \mathit{\Gamma}_1^{*}) - (\mathit{\Gamma}_2 - \mathit{\Gamma}_2^{*}))$, we have

$$\begin{aligned}
&\frac{1}{\sqrt{NJ}} \left\|(Z_1 - Z_2)\left(V^{\top} - V^{*\top}\right) + \mathbf{1_N}\left\{(\mathit{\Gamma}_1 - \mathit{\Gamma}_1^{*}) - (\mathit{\Gamma}_2 - \mathit{\Gamma}_2^{*})\right\}^{\top}\right\|_F \\
=&\frac{1}{\sqrt{NJ}} \left\|(Z_1 - Z_2, \mathbf{1_N})\, Q^{\top}\right\|_F \\
=&\frac{1}{\sqrt{NJ}} \left\|(Z_1 - Z_2, \mathbf{1_N})\, Q^{\top}Q(Q^{\top}Q)^{-1/2}\right\|_F \\
=&\frac{1}{\sqrt{NJ}} \left\|(Z_1 - Z_2, \mathbf{1_N})\left(Q^{\top}Q\right)^{1/2}\right\|_F \\
=&\frac{1}{\sqrt{NJ}}\sqrt{tr\left((Z_1 - Z_2, \mathbf{1_N})\,(Q^{\top}Q)^{1/2}\,(Q^{\top}Q)^{1/2}\,(Z_1 - Z_2, \mathbf{1_N})^{\top}\right)} \\
\geq&\frac{1}{\sqrt{J}}\sqrt{\frac{\pi_{\min}\left((Z_1 - Z_2, \mathbf{1_N})^{\top}(Z_1 - Z_2, \mathbf{1_N})\right)}{N}}\sqrt{tr(Q^{\top}Q)} \\
=&\frac{1}{\sqrt{J}}\sqrt{\frac{\pi_{\min}\left((Z_1 - Z_2, \mathbf{1_N})^{\top}(Z_1 - Z_2, \mathbf{1_N})\right)}{N}}\,\|Q\|_F
\end{aligned}$$

Therefore, we have

$$\frac{1}{\sqrt{J}}\|Q\|_F = \frac{1}{\sqrt{J}}\|V - V^{*}\|_F + \frac{1}{\sqrt{J}}\|(\mathit{\Gamma}_1 - \mathit{\Gamma}_1^{*}) - (\mathit{\Gamma}_2 - \mathit{\Gamma}_2^{*})\|_F \lesssim d(\mathbf{\Xi}, \mathbf{\Xi}^{*}). \tag{A.9}$$

By triangle inequality, we have

$$\begin{aligned}
&\frac{1}{\sqrt{NJ}}\left\|\Theta A^{\top} + XB^{\top} + Z_t V^{\top} + \mathbf{1_N}\mathit{\Gamma}_t^{\top} - \Theta^{*}A^{*\top} - XB^{*\top} - Z_t V^{*\top} - \mathbf{1_N}\mathit{\Gamma}_t^{*\top}\right\|_F \leq d(\mathbf{\Xi}, \mathbf{\Xi}^{*}), \\
&\left\|\Theta A^{\top} + XB^{\top} + \mathbf{1_N}\mathit{\Gamma}_t^{\top} - \Theta^{*}A^{*\top} - XB^{*\top} - \mathbf{1_N}\mathit{\Gamma}_t^{*\top}\right\|_F \leq \sqrt{NJ}d(\mathbf{\Xi}, \mathbf{\Xi}^{*}) + \left\|Z_t\,(V - V^{*})^{\top}\right\|_F .
\end{aligned} \tag{A.10}$$

Let $\pi_{\max}(\cdot)$ refers to the largest eigenvalue of a matrix. Since $\mathcal{Z}$ is bounded, we have

$$\begin{aligned}
\frac{1}{\sqrt{NJ}}\left\|Z_t\,(V - V^{*})^{\top}\right\|_F =&\frac{1}{\sqrt{NJ}}\sqrt{tr(Z_t\,(V - V^{*})^{\top}\,(V - V^{*})\,Z_t^{\top})} \\
\leq&\frac{1}{\sqrt{J}}\sqrt{tr(Z_t^{\top}Z_t/N \cdot (V - V^{*})^{\top}\,(V - V^{*}))} \\
\leq&\frac{1}{\sqrt{J}}\pi_{\max}\left(\sqrt{tr(Z_t^{\top}Z_t/N)}\right)\left\|(V - V^{*})^{\top}\right\|_F \\
\lesssim& d(\mathbf{\Xi}, \mathbf{\Xi}^{*}).
\end{aligned} \tag{A.11}$$

Thus by (A.10) and (A.11), we have

$$\frac{1}{\sqrt{NJ}}\left\|\Theta A^{\top} + XB^{\top} + \mathbf{1_N}\mathit{\Gamma}_t^{\top} - \Theta^{*}A^{*\top} - XB^{*\top} - \mathbf{1_N}\mathit{\Gamma}_t^{*\top}\right\|_F \lesssim d(\mathbf{\Xi}, \mathbf{\Xi}^{*}).$$

Define $H_1 = \Theta A^{\top} - \Theta^{*}A^{*\top}$ and $H_2 = X(B - B^{*})^{\top} + \mathbf{1_N}(\mathit{\Gamma}_t - \mathit{\Gamma}_t^{*})^{\top}$,

$$\begin{aligned}
&\left\|\Theta A^{\top} + XB^{\top} + \mathbf{1_N}\mathit{\Gamma}_t{}^{\top} - \Theta^{*}A^{*\top} - XB^{*\top} - \mathbf{1_N}\mathit{\Gamma}_t^{*\top}\right\|_F^2 \\
=&\|H_1\|_F^2 + \|H_2\|_F^2 + 2tr(H_1^{\top}H_2) \\
=&\|H_1\|_F^2 + \|H_2\|_F^2
\end{aligned}$$

because $\Theta^{*\top}(\mathbf{1_N}, X) = \Theta^\top(\mathbf{1_N}, X) = 0_{K^*\times(1+p)}$. Therefore, we have

$$\frac{1}{\sqrt{NJ}} \left\| \Theta A^\top - \Theta^* A^{*\top} \right\|_F \lesssim d(\Xi, \Xi^*) \tag{A.12}$$

$$\text{and } \frac{1}{\sqrt{NJ}} \left\| X(B - B^*)^\top + \mathbf{1_N}(\mathit{\Gamma}_t - \mathit{\Gamma}_t^*)^\top \right\|_F \lesssim d(\Xi, \Xi^*). \tag{A.13}$$

From (A.12) and the arguments in Lemma 2 of Chen et al. (2021), we can prove that

$$\left\| \Theta - \Theta^* S_A \right\|_F / \sqrt{N} + \left\| A - A^* S_A \right\|_F / \sqrt{J} \lesssim d(\Xi, \Xi^*). \tag{A.14}$$

Finally, from (A.13),

$$\frac{1}{\sqrt{N}} \left\| X(B - B^*)^\top + \mathbf{1_N}(\mathit{\Gamma}_t - \mathit{\Gamma}_t^*)^\top \right\|_F \geq \sqrt{\frac{\pi_{\min}\left(X, \mathbf{1_N}\right)^\top \left(X, \mathbf{1_N}\right)}{N}} \left\| (B, \mathit{\Gamma}_t) - (B^*, \mathit{\Gamma}_t^*) \right\|_F.$$

Hence

$$\left\| (B, \mathit{\Gamma}_t) - (B^*, \mathit{\Gamma}_t^*) \right\|_F / \sqrt{J} \lesssim d(\Xi, \Xi^*) \tag{A.15}$$

for sufficiently large N and J by Assumption 2.5. The proof of Lemma A.2 is thus complete by (A.9), (A.14) and (A.15).

Lemma A.3. *Under Assumptions 2.1-2.5 and A.1-A.2, for sufficiently small c and sufficiently large N and J, it holds that*

$$E\left(\sup_{\Xi \in \mathcal{H}^{K^*}(c)} |\mathbb{W}_{NJ}(\Xi)|\right) \lesssim \frac{c}{\min\left\{\sqrt{N}, \sqrt{J}\right\}}.$$

Proof: In the proof of Lemma A.1 we have shown that, for $\Xi^{(a)}, \Xi^{(b)} \in \mathcal{H}^{K^*}$,

$$\left\| \sqrt{NJ} \left| \mathbb{W}_{NJ}\left(\Xi^{(a)}\right) - \mathbb{W}_{NJ}\left(\Xi^{(b)}\right) \right| \right\|_{\Psi_2} \lesssim d\left(\Xi^{(a)}, \Xi^{(b)}\right).$$

Since the process $\mathbb{W}_{NJ}(\Xi)$ is separable, it follows from Theorem 2.2.4 of Van Der Vaart et al. (1996) that

$$\sqrt{NJ} E\left(\sup_{\Xi \in \mathcal{H}^{K^*}(c)} |\mathbb{W}_{NJ}(\Xi)|\right) \lesssim \sqrt{NJ} \left\| \sup_{\Xi \in \mathcal{H}^{K^*}(c)} |\mathbb{W}_{NJ}(\Xi)| \right\|_{\Psi_2} \lesssim \int_0^c \sqrt{\log D(\epsilon, d, \mathcal{H}^{K^*}(c))} d\epsilon.$$

Thus, it remains to be shown that

$$\int_0^c \sqrt{\log D(\epsilon, d, \mathcal{H}^{K^*}(c))} d\epsilon = O\left(\sqrt{N + J} c\right). \tag{A.16}$$

To prove (A.16), first note that Lemma A.2 implies that

$$\mathcal{H}^{K^*}(c) \subset \bigcup_{E \in \mathcal{S}} \mathcal{H}^{K^*}(c; E),$$

where

$$\mathcal{S} = \Big\{ E \in \mathbb{R}^{(1+p+p_z+K^*)\times(1+p+p_z+K^*)} : E = \operatorname{diag}(1, \dots, 1, u_{(2+p+p_z)}, \dots, u_{(1+p+p_z+K^*)}), \\ u_k \in \{-1, 1\} \text{ for } k = 2+p+p_z, \dots, 1+p+p_z+K^* \Big\}$$

is the set of diagonal matrices E that has ones on the diagonal in all positions except for the $(2+p+p_z)$th to $(1+p+p_z+K^*)$th positions, which can take either $+1$ or -1, and

$$\mathcal{H}^{K^*}(c; E) = \Big\{ \Xi \in \mathcal{H}^{K^*} : \|\Theta - \Theta^* E_{[(2+p+p_z):(1+p+p_z+K^*),(2+p+p_z):(1+p+p_z+K^*)]}\|_F / \sqrt{N} \\ + \max_{t=1,\dots,T} \|U_t - U_t^* E\|_F / \sqrt{J} \le \delta_3 c \Big\}.$$

Since there are 2^{K^*} elements in $\mathcal{S}$ and K^* is fixed, it suffices to show that

$$\int_0^c \sqrt{\log D(\epsilon, d, \mathcal{H}^{K^*}(c; E))} d\epsilon = O\left(\sqrt{N+J} c\right)$$

for each $E \in \mathcal{S}$. Without loss of generality, we focus on the case $E = I_{1+K^*+p+p_z}$. Second, for any $\Xi^{(a)}, \Xi^{(b)} \in \mathcal{H}^{K^*}$,

$$\begin{aligned} d(\Xi^{(a)}, \Xi^{(b)}) =& \frac{1}{\sqrt{NJ}} \max_{t=1,\dots T} \left\| (\mathbf{1_N}, X, Z_t, \Theta^{(a)}) {U_t^{(a)}}^\top - (\mathbf{1_N}, X, Z_t, \Theta^{(b)}) {U_t^{(b)}}^\top \right\|_F \\ \le& \frac{1}{\sqrt{NJ}} \max_{t=1,\dots T} \left\| (\mathbf{1_N}, X, Z_t, \Theta^{(a)}) {U_t^{(a)}}^\top - (\mathbf{1_N}, X, Z_t, \Theta^{(b)}) {U_t^{(a)}}^\top \right\|_F \\ &+ \frac{1}{\sqrt{NJ}} \max_{t=1,\dots T} \left\| (\mathbf{1_N}, X, Z_t, \Theta^{(b)}) {U_t^{(a)}}^\top - (\mathbf{1_N}, X, Z_t, \Theta^{(b)}) {U_t^{(b)}}^\top \right\|_F \\ \le& \frac{1}{\sqrt{NJ}} \left\| \left(\Theta^{(a)} - \Theta^{(b)}\right) {A^{(a)}}^\top \right\|_F + \frac{\|(\mathbf{1_N}, X, Z_t, \Theta^{(b)})\|_F}{\sqrt{N}} \cdot \max_{t=1,\dots,T} \frac{\left\| U_t^{(a)} - U_t^{(b)} \right\|_F}{\sqrt{J}} \\ \le& \delta_4 \left(\frac{\left\|\Theta^{(a)} - \Theta^{(b)}\right\|_F}{\sqrt{N}} + \max_{t=1,\dots,T} \frac{\left\| U_t^{(a)} - U_t^{(b)} \right\|_F}{\sqrt{J}} \right) \\ \le& \delta_4 \left(\frac{\left\|\Theta^{(a)} - \Theta^{(b)}\right\|_F}{\sqrt{N}} + \frac{\left\| \left(\mathit{\Gamma}^{(a)}, B^{(a)}, V^{(a)}, A^{(a)}\right) - \left(\mathit{\Gamma}^{(b)}, B^{(b)}, V^{(b)}, A^{(b)}\right) \right\|_F}{\sqrt{J}} \right). \end{aligned}$$

Now define

$$d^*(\Xi^{(a)}, \Xi^{(b)}) = 2\delta_4 \sqrt{ \frac{\|\Theta^{(a)} - \Theta^{(b)}\|_F^2}{N} + \frac{\|(\mathit{\Gamma}^{(a)}, B^{(a)}, V^{(a)}, A^{(a)}) - (\mathit{\Gamma}^{(b)}, B^{(b)}, V^{(b)}, A^{(b)})\|_F^2}{J} }.$$

It follows from $\sqrt{x} + \sqrt{y} \le 2\sqrt{x+y}$ that $d(\Xi^{(a)}, \Xi^{(b)}) \le d^*(\Xi^{(a)}, \Xi^{(b)})$. Moreover, for any $\Xi \in \mathcal{H}^{K^*}(c; I_{1+p+p_z+K^*})$, we have

$$\left(\frac{\|\Theta - \Theta^*\|_F}{\sqrt{N}} + \frac{\|(\mathit{\Gamma}, A, B, V) - (\mathit{\Gamma}^*, A^*, B^*, V^*)\|_F}{\sqrt{J}} \right) \le T\delta_3 c.$$

Thus it follows from $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ that

$$\mathcal{H}^{K^*}(c; I_{1+K^*+p+p_z}) \subset \mathcal{H}^{K*}(c) = \left\{ \Xi \in \mathcal{H}^{K^*} : d^*(\Xi, \Xi^*) \le \delta_5 c \right\}$$

where $\delta_5 = 2T\delta_3\delta_4$. The remainder of the proof follows from the argument presented in lemma 3 of Chen et al. (2021).

A.4.2 Proof of Theorem A.2

It suffices to prove the result for $\mathbf{u}_j$, as the argument for $\boldsymbol{\theta}_i$ is symmetric. Without loss of generality, we assume that $\hat{S}_A = I_{K^*}$ to simplify the notation. Define

$$l^*_{j,N}(\mathbf{u}_j, \Theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{\rho}_{ij}(\mathbf{u}^*_j, \boldsymbol{\theta}^*_i) - \boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i)\right),$$

$$\bar{l}^*_{j,N}(\mathbf{u}_j, \Theta) = \frac{1}{N}\sum_{i=1}^{N} E\left(\boldsymbol{\rho}_{ij}(\mathbf{u}^*_j, \boldsymbol{\theta}^*_i) - \boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i)\right),$$

$$\breve{\varrho}_{ijt}(\cdot) = \varrho_{ijt}(\cdot) - E(\varrho_{ijt}(\cdot)), \quad \breve{\varrho}_{ijt} = \breve{\varrho}_{ijt}(\mathbf{u}^{*\top}_j \mathbf{e}^*_{it}).$$

We first prove the following Lemmas:

Lemma A.4. *Under Assumptions 2.1-2.5 and A.1-A.2, we have*

$$\left\|\hat{\mathbf{u}}_j - \mathbf{u}^*_j\right\|_F = o_P(1) \textit{ for each } j.$$

Proof: Note that $\hat{\mathbf{u}}_j = \arg\min_{\mathbf{u}_j \in \mathcal{U}} l^*_{j,N}(\mathbf{u}_j, \hat{\Theta})$. First, we show that

$$\sup_{\mathbf{u}_j \in \mathcal{U}}\left|l^*_{j,N}(\mathbf{u}_j, \hat{\Theta}) - \bar{l}^*_{j,N}(\mathbf{u}_j, \Theta^*)\right| = o_P(1). \tag{A.17}$$

Note that

$$\sup_{\mathbf{u}_j \in \mathcal{U}}\left|l^*_{j,N}(\mathbf{u}_j, \hat{\Theta}) - \bar{l}^*_{j,N}(\mathbf{u}_j, \Theta^*)\right| \leq \sup_{\mathbf{u}_j \in \mathcal{U}}\left|l^*_{j,N}(\mathbf{u}_j, \hat{\Theta}) - l^*_{j,N}(\mathbf{u}_j, \Theta^*)\right| + \sup_{\mathbf{u}_j \in \mathcal{U}}\left|l^*_{j,N}(\mathbf{u}_j, \Theta^*) - \bar{l}^*_{j,N}(\mathbf{u}_j, \Theta^*)\right|.$$

It is easy to show that

$$\begin{aligned}\sup_{\mathbf{u}_j \in \mathcal{U}}\left|l^*_{j,N}(\mathbf{u}_j, \hat{\Theta}) - l^*_{j,N}(\mathbf{u}_j, \Theta^*)\right| &\lesssim \sup_{\mathbf{u}_j \in \mathcal{U}}\|\mathbf{a}_j\| \cdot \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*_i\right\| \\ &\lesssim \left\|\hat{\Theta} - \Theta^*\right\|_F / \sqrt{N} = O_P\left(\min\{\sqrt{N}, \sqrt{J}\}^{-1}\right), \\ \sup_{\mathbf{u}_j \in \mathcal{U}}\left|l^*_{j,N}(\mathbf{u}_j, \Theta^*) - \bar{l}^*_{j,N}(\mathbf{u}_j, \Theta^*)\right| &= o_P(1),\end{aligned}$$

thus showing that (A.17) holds.
Second, we can show that for any $\epsilon > 0$, and $B_j(\epsilon) = \left\{\mathbf{u}_j \in \mathcal{U} : \left\|\mathbf{u}_j - \mathbf{u}^*_j\right\| \leq \epsilon\right\}$,

$$\inf_{\mathbf{u}_j \in B^C_j(\epsilon)} \bar{l}^*_{j,N}(\mathbf{u}_j, \Theta^*) > \bar{l}^*_{j,N}(\mathbf{u}^*_j, \Theta^*) = 0, \tag{A.18}$$

where $B^C_j(\epsilon)$ denotes the complement of $B_j(\epsilon)$. It follows from standard argument by noting that $\bar{l}^*_{j,N}(\mathbf{u}_j, \Theta^*)$ is convex and differentiable, holding Θ^* fixed, for example, the proof of Proposition 3.1 of Galvao and Kato (2016b).
Finally, given (A.17) and (A.18), the proof is complete by standard consistency proof of M-estimator(see Theorem 2.1 of Newey and McFadden (1994).)

Lemma A.5. *Under Assumptions 2.1-2.7, A.1 and A.2, we have*

$$\left\|\hat{\mathbf{u}}_j - \mathbf{u}^*_j\right\| = O_p(N^{-1/2}) \textit{ for each } j.$$

Proof: For any fixed $\mathbf{u}_j \in \mathcal{U}$ and $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$, expanding $\rho^{'}_{ijt}(\mathbf{u}_j^\top \mathbf{e}_{it})\mathbf{e}_{it}$ gives

$$
\begin{aligned}
&\rho^{'}_{ijt}(\mathbf{u}_j^\top \mathbf{e}_{it})\mathbf{e}_{it}\\
=&\rho^{'}_{ijt}(\mathbf{u}_j^{*\top} \mathbf{e}_{it})\mathbf{e}_{it} + \rho^{''}_{ijt}(\mathbf{u}_j^{*\top} \mathbf{e}_{it})\mathbf{e}_{it}\left\{\mathbf{e}_{it}^\top(\mathbf{u}_j - \mathbf{u}_j^*)\right\} + 0.5\rho^{'''}_{ijt}(\tilde{\mathbf{u}}_j^\top \mathbf{e}_{it})\mathbf{e}_{it}\left\{\mathbf{e}_{it}^\top(\mathbf{u}_j - \mathbf{u}_j^*)\right\}^2\\
=&\left\{\rho^{'}_{ijt} + \rho^{''}_{ijt}(\mathbf{u}_j^{*\top}\tilde{\mathbf{e}}_{it})\mathbf{u}_j^{*\top}\left(\mathbf{e}_{it} - \mathbf{e}_{it}^*\right)\right\}\mathbf{e}_{it}\\
&+\left\{\rho^{''}_{ijt} + \rho^{'''}_{ijt}(\mathbf{u}_j^{*\top}\tilde{\mathbf{e}}_{it})\mathbf{u}_j^{*\top}\left(\mathbf{e}_{it} - \mathbf{e}_{it}^*\right)\right\}\mathbf{e}_{it}\left\{\mathbf{e}_{it}^\top(\mathbf{u}_j - \mathbf{u}_j^*)\right\}\\
&+0.5\rho^{'''}_{ijt}(\tilde{\mathbf{u}}_j^\top \mathbf{e}_{it})\mathbf{e}_{it}\left\{\mathbf{e}_{it}^\top(\mathbf{u}_j - \mathbf{u}_j^*)\right\}^2\\
=&\rho^{'}_{ijt}\mathbf{e}_{it}^* + \rho^{'}_{ijt}\left(\mathbf{e}_{it} - \mathbf{e}_{it}^*\right) + \rho^{''}_{ijt}(\mathbf{u}_j^{*\top}\tilde{\mathbf{e}}_{it})\mathbf{e}_{it}\mathbf{u}_j^{*\top}\left(\mathbf{e}_{it} - \mathbf{e}_{it}^*\right) + \rho^{''}_{ijt}\mathbf{e}_{it}\left\{\mathbf{e}_{it}^\top(\mathbf{u}_j - \mathbf{u}_j^*)\right\}\\
&+\rho^{'''}_{ijt}(\mathbf{u}_j^{*\top}\tilde{\mathbf{e}}_{it})\mathbf{e}_{it}\mathbf{u}_j^{*\top}\left(\mathbf{e}_{it} - \mathbf{e}_{it}^*\right)\left\{\mathbf{e}_{it}^\top(\mathbf{u}_j - \mathbf{u}_j^*)\right\} + 0.5\rho^{'''}_{ijt}(\tilde{\mathbf{u}}_j^\top \mathbf{e}_{it})\mathbf{e}_{it}\left\{\mathbf{e}_{it}^\top(\mathbf{u}_j - \mathbf{u}_j^*)\right\}^2,
\end{aligned}
$$

where $\tilde{\mathbf{u}}_j$ lies between $\mathbf{u}_j$ and $\mathbf{u}_j^*$ and $\tilde{\mathbf{e}}_{it}$ lies between $\mathbf{e}_{it}$ and $\mathbf{e}_{it}^*$. Taking expectations on both sides of the above equation, and setting $\mathbf{u}_j = \hat{\mathbf{u}}_j, \mathbf{e}_i = \hat{\mathbf{e}}_i$, it follows that

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} E\left(\sum_{t=1}^{T} r_{it}\rho^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it}\right)\\
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it})\right)\hat{\mathbf{e}}_{it}\\
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{'}_{ijt}\right)\mathbf{e}_{it}^* + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(E\left(\varrho^{''}_{ijt}\right)\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^\top\right)(\hat{\mathbf{u}}_j - \mathbf{u}_j^*)\\
&+O_P\left(N^{-1/2}\left\|\hat{\Theta} - \Theta^*\right\|_F\right) + O_P\left(\left\|\hat{\mathbf{u}}_j - \mathbf{u}_j^*\right\|\right)\cdot O_P\left(N^{-1/2}\left\|\hat{\Theta} - \Theta^*\right\|_F\right) + O_P(\left\|\hat{\mathbf{u}}_j - \mathbf{u}_j^*\right\|^2).
\end{aligned}
$$

Note that we have $N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{'}_{ijt}\right)\mathbf{e}_{it}^* = 0$. Also, by Theorem A.1,

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{''}_{ijt}\right)\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^\top &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{''}_{ijt}\right)\mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top} + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{''}_{ijt}\right)\left(\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^\top - \mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}\right)\\
&= \mathit{\Phi}_j + o_p(1).
\end{aligned}
$$

Then, by Theorem A.1 and Lemma A.4, we have

$$
\mathit{\Phi}_j(\hat{\mathbf{u}}_j - \mathbf{u}_j^*) + o_p\left(\left\|\hat{\mathbf{u}}_j - \mathbf{u}_j^*\right\|\right) = O_P\left(\min\left\{\sqrt{N}, \sqrt{J}\right\}^{-1}\right) + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it})\right)\hat{\mathbf{e}}_{it}. \tag{A.19}
$$

Note that we can write

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it})\right)\hat{\mathbf{e}}_{it}\\
=&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}^{'}_{ijt}\mathbf{e}_{it}^* - \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\breve{\varrho}^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it} - \breve{\varrho}^{'}_{ijt}\mathbf{e}_{it}^*\right)\\
=&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}^{'}_{ijt}\mathbf{e}_{it}^* - \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\breve{\varrho}^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it} - \breve{\varrho}^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \mathbf{e}_{it}^*)\mathbf{e}_{it}^*\right)\\
&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\breve{\varrho}^{'}_{ijt}(\hat{\mathbf{u}}_j^\top \mathbf{e}_{it}^*) - \breve{\varrho}^{'}_{ijt}\right)\mathbf{e}_{it}^*.
\end{aligned}
$$

The first term of the RHS of the above equation is clearly $O_P(N^{-1/2})$ by central limit theorem. For the second term on the RHS of the equation, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\breve{\varrho}'_{ijt}(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it}-\breve{\varrho}'_{ijt}(\hat{\mathbf{u}}_j^\top\mathbf{e}_{it}^*)\mathbf{e}_{it}^*\right)\\
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}(\hat{\mathbf{u}}_j^\top\mathbf{e}_{it}^*)\left(\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*\right)+\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}(\hat{\mathbf{u}}_j^\top\tilde{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it}\hat{\mathbf{u}}_j^\top\left(\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*\right), \qquad \text{(A.20)}
\end{aligned}$$

where $\tilde{\mathbf{e}}_{it}$ lies between $\hat{\mathbf{e}}_{it}$ and $\mathbf{e}_{it}^*$. The first term of the RHS of (A.20) is $O_P\left(1/\min\left\{\sqrt{N},\sqrt{J}\right\}\right)$ due to the boundedness of $\breve{\varrho}'_{ijt}$, and by Theorem A.1, we have $N^{-1}\sum_{i=1}^{N}\|\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*\| = O_P\left(1/\min\left\{\sqrt{N},\sqrt{J}\right\}\right)$. Similarly, the second term on the RHS of (A.20) is also $O_P\left(1/\min\left\{\sqrt{N},\sqrt{J}\right\}\right)$. Finally, we can show that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\breve{\varrho}'_{ijt}(\hat{\mathbf{u}}_j^\top\mathbf{e}_{it}^*)-\breve{\varrho}'_{ijt}\right)\mathbf{e}_{it}^*\\
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\breve{\varrho}'_{ijt}+\breve{\varrho}''_{ijt}(\tilde{\mathbf{u}}_j^\top\mathbf{e}_{it}^*)\mathbf{e}_{it}^{*\top}\left(\hat{\mathbf{u}}_j-\mathbf{u}_j^*\right)-\breve{\varrho}'_{ijt}\right\}\mathbf{e}_{it}^*\\
=&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\breve{\varrho}''_{ijt}(\tilde{\mathbf{u}}_j^\top\mathbf{e}_{it}^*)\mathbf{e}_{it}^{*\top}\left(\hat{\mathbf{u}}_j-\mathbf{u}_j^*\right)\right\}\mathbf{e}_{it}^*\\
=&O_P(1/\sqrt{N})\cdot o_p(\|\hat{\mathbf{u}}_j-\mathbf{u}_j^*\|)
\end{aligned}$$

by central limit theorem, where $\tilde{\mathbf{u}}_j$ lies between $\hat{\mathbf{u}}_j$ and $\mathbf{u}_j^*$. Combining the above results yields

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varrho'_{ijt}(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it})\right)\hat{\mathbf{e}}_{it}=O_P\left(1/\min\left\{\sqrt{N},\sqrt{J}\right\}\right)+o_p(\|\hat{\mathbf{u}}_j-\mathbf{u}_j^*\|). \qquad \text{(A.21)}$$

Thus the desired result follows from (A.19), (A.21) and Assumption 2.6. To derive the asymptotic distribution of $\hat{\mathbf{u}}_j$, it is essential to obtain the stochastic expansion of $\hat{\boldsymbol{\theta}}_i$. Define

$$\begin{aligned}
&\mathit{\Phi}_{N,j}=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varrho''_{ijt}\right)\mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}, \quad \mathit{\Psi}_{J,i}=\frac{1}{J}\sum_{j=1}^{J}\sum_{t=1}^{T}E\left(\varrho''_{ijt}\right)\mathbf{a}_j^*\mathbf{a}_j^{*\top},\\
&\mathbb{P}_{NJ}(\boldsymbol{\Xi})=T\delta\Bigg\{\frac{1}{2J}\sum_{l=1}^{K^*}\sum_{q>l}^{K^*}\Bigg(\sum_{j=1}^{J}a_{jl}a_{jq}\Bigg)^2+\frac{1}{2N}\sum_{l=1}^{K^*}\sum_{q>l}^{K^*}\Bigg(\sum_{i=1}^{N}\theta_{il}\theta_{iq}\Bigg)^2+\frac{1}{8J}\sum_{k=1}^{K^*}\Bigg(\sum_{j=1}^{J}a_{jk}^2-J\Bigg)^2\\
&\qquad\qquad+\frac{1}{2N}\sum_{l=1}^{K^*}\Bigg(\sum_{i=1}^{N}\theta_{il}\Bigg)^2+\frac{1}{2N}\sum_{k=1}^{K^*}\sum_{l=1}^{p}\Bigg(\sum_{i=1}^{N}\theta_{ik}x_{il}\Bigg)\Bigg\}
\end{aligned}$$

for some $\delta > 0$. We further define

$$S^*(\Xi)$$

$$= \left(\underbrace{\ldots, \frac{1}{\sqrt{NJ}} \sum_{i=1}^{N} \sum_{t=1}^{T} E\left(\varrho_{ijt}^{'}(\mathbf{u}_j^\top \mathbf{e}_{it}) \right) \mathbf{e}_{it}^\top, \ldots,}_{1 \times JP} \underbrace{\ldots, \frac{1}{\sqrt{NJ}} \sum_{j=1}^{J} \sum_{t=1}^{T} E\left(\varrho_{ijt}^{'}(\mathbf{u}_j^\top \mathbf{e}_{it}) \right) \mathbf{a}_j^\top, \ldots}_{1 \times NK^*} \right)^\top,$$

$$S(\Xi) = S^*(\Xi) + \partial \mathbb{P}_{NJ}(\Xi)/\partial \Xi, \qquad \mathbb{H}(\Xi) = \partial S^*(\Xi)/\partial \Xi^\top + \partial \mathbb{P}_{NJ}(\Xi)/\partial \Xi \partial \Xi^\top$$

and let $\mathbb{H} = \mathbb{H}(\Xi^*)$. Expanding $S(\hat{\Xi})$ around $S(\Xi^*)$ gives

$$S(\hat{\Xi}) = S(\Xi^*) + \mathbb{H} \cdot (\hat{\Xi} - \Xi^*) + 0.5\mathcal{R}(\hat{\Xi}), \tag{A.22}$$

where

$$\mathcal{R}(\hat{\Xi}) = \left\{ \sum_{m=1}^{JP+NK^*} \partial \mathbb{H}(\tilde{\Xi})/\partial \Xi_m \cdot (\hat{\Xi}_m - \Xi_m^*) \right\} (\hat{\Xi} - \Xi^*),$$

$\tilde{\Xi}$ lies between $\hat{\Xi}$ and Ξ^*. Further, define

$$\mathbb{H}_d = \begin{pmatrix} \mathbb{H}_d^{\mathcal{U}} & 0 \\ 0 & \mathbb{H}_d^{\Theta} \end{pmatrix}, \mathbb{H}_d^{\mathcal{U}} = \frac{\sqrt{N}}{\sqrt{J}} \text{diag}\left(\varPhi_{N,1}, \ldots, \varPhi_{N,J} \right), \mathbb{H}_d^{\Theta} = \frac{\sqrt{J}}{\sqrt{N}} \text{diag}\left(\varPsi_{J,1}, \ldots, \varPsi_{J,N} \right),$$

we have the following lemma.

Lemma A.6. *Under Assumptions 2.1-2.7, A.1-A.3 and the condition that* $\sum_{i=1}^N \mathbf{x}_i = \mathbf{0}_p$ *and* $\sum_{i=1}^N x_{ik} x_{il} = 0$ *for* $l, k \in \{1, \ldots, p\}, l \neq k$, *the matrix* $\mathbb{H}$ *is invertible and* $\left\| \mathbb{H}^{-1} - \mathbb{H}_d^{-1} \right\|_{\max} = O(1/N)$.

Proof: We assume $p = T = K^* = 2$ for simplicity, which can be generalised easily. We consider

$$\mathbb{P}_{NJ}(\Xi) = 2\delta \Bigg\{ \frac{1}{2N} \left(\sum_{i=1}^N \theta_{i1}\theta_{i2} \right)^2 + \frac{1}{2J} \left(\sum_{j=1}^J a_{j1} a_{j2} \right)^2 + \frac{1}{8J} \left(\sum_{j=1}^J a_{j1}^2 - J \right)^2 + \frac{1}{8J} \left(\sum_{j=1}^J a_{j2}^2 - J \right)^2$$
$$+ \frac{1}{2N} \left(\sum_{i=1}^N \theta_{i1} \right)^2 + \frac{1}{2N} \left(\sum_{i=1}^N \theta_{i2} \right)^2 + \frac{1}{2N} \sum_{k=1}^{2} \sum_{p=1}^{2} \left(\sum_{i=1}^N \theta_{ik} x_{ip} \right)^2 \Bigg\}.$$

Then we can define

$$\boldsymbol{\mu}_1 = ((\mathbf{0}_{4+p_z}^\top, a_{11}^*, 0), \ldots, (\mathbf{0}_{4+p_z}^\top, a_{J1}^*, 0), \mathbf{0}_{2N}^\top)^\top / \sqrt{J},$$
$$\boldsymbol{\mu}_2 = ((\mathbf{0}_{4+p_z}^\top, 0, a_{12}^*), \ldots, (\mathbf{0}_{4+p_z}^\top, 0, a_{J2}^*), \mathbf{0}_{2N}^\top)^\top / \sqrt{J},$$
$$\boldsymbol{\mu}_3 = ((\mathbf{0}_{4+p_z}^\top, a_{12}^*, a_{11}^*), \ldots, (\mathbf{0}_{4+p_z}^\top, a_{J2}^*, a_{J1}^*), \mathbf{0}_{2N}^\top)^\top / \sqrt{J},$$
$$\boldsymbol{\mu}_4 = (\mathbf{0}_{PJ}^\top, (\theta_{12}^*, \theta_{11}^*), \ldots, (\theta_{N2}^*, \theta_{N1}^*))^\top / \sqrt{N}$$
$$\boldsymbol{\mu}_5 = (\mathbf{0}_{PJ}^\top, (1, 0), \ldots, (1, 0))^\top / \sqrt{N},$$
$$\boldsymbol{\mu}_6 = (\mathbf{0}_{PJ}^\top, (0, 1), \ldots, (0, 1))^\top / \sqrt{N},$$
$$\boldsymbol{\mu}_7 = (\mathbf{0}_{PJ}^\top, (x_{11}, 0), \ldots, (x_{N1}, 0))^\top / \sqrt{N},$$
$$\boldsymbol{\mu}_8 = (\mathbf{0}_{PJ}^\top, (0, x_{11}), \ldots, (0, x_{N1}))^\top / \sqrt{N},$$
$$\boldsymbol{\mu}_9 = (\mathbf{0}_{PJ}^\top, (x_{12}, 0), \ldots, (x_{N2}, 0))^\top / \sqrt{N},$$
$$\boldsymbol{\mu}_{10} = (\mathbf{0}_{PJ}^\top, (0, x_{12}), \ldots, (0, x_{N2}))^\top / \sqrt{N},$$

such that $\partial\mathbb{P}_{NJ}(\boldsymbol{\Xi}^*)/\partial\boldsymbol{\Xi}\partial\boldsymbol{\Xi}^\top = 2\delta\left(\sum_{m=1}^{10}\boldsymbol{\mu}_m\boldsymbol{\mu}_m^\top\right)$. We further define

$$\boldsymbol{\omega}_{1,1} = \left(\underbrace{(\mathbf{0}_{4+p_z}^\top, a_{11}^*/\sqrt{J}, 0), \dots, (\mathbf{0}_{4+p_z}^\top, a_{J1}^*/\sqrt{J}, 0)}_{\boldsymbol{\omega}_{1\mathcal{U},1}^\top}, \underbrace{(-\theta_{11}^*/\sqrt{N}, 0), \dots, (-\theta_{N1}^*/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{1\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{2,1} = \left(\underbrace{(\mathbf{0}_{4+p_z}^\top, 0, a_{12}^*/\sqrt{J}), \dots, (\mathbf{0}_{4+p_z}^\top, 0, a_{J2}^*/\sqrt{J})}_{\boldsymbol{\omega}_{2\mathcal{U},1}^\top}, \underbrace{(0, -\theta_{12}^*/\sqrt{N}), \dots, (0, -\theta_{N2}^*/\sqrt{N})}_{\boldsymbol{\omega}_{2\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{3,1} = \left(\underbrace{(\mathbf{0}_{4+p_z}^\top, a_{12}^*/\sqrt{J}, 0), \dots, (\mathbf{0}_{4+p_z}^\top, a_{J2}^*/\sqrt{J}, 0)}_{\boldsymbol{\omega}_{3\mathcal{U},1}^\top}, \underbrace{(0, -\theta_{11}^*/\sqrt{N}), \dots, (0, -\theta_{N1}^*/\sqrt{N})}_{\boldsymbol{\omega}_{3\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{4,1} = \left(\underbrace{(\mathbf{0}_{4+p_z}^\top, 0, a_{11}^*/\sqrt{J}), \dots, (\mathbf{0}_{4+p_z}^\top, 0, a_{J1}^*/\sqrt{J})}_{\boldsymbol{\omega}_{4\mathcal{U},1}^\top}, \underbrace{(-\theta_{12}^*/\sqrt{N}, 0), \dots, (-\theta_{N2}^*/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{4\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{5,1} = \left(\underbrace{(a_{11}^*/\sqrt{J}, \mathbf{0}_{5+p_z}^\top), \dots, (a_{J1}^*/\sqrt{J}, \mathbf{0}_{5+p_z}^\top)}_{\boldsymbol{\omega}_{5\mathcal{U},1}^\top}, \underbrace{(-1/\sqrt{N}, 0), \dots, (-1/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{5\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{6,1} = \left(\underbrace{(a_{12}^*/\sqrt{J}, \mathbf{0}_{5+p_z}^\top), \dots, (a_{J2}^*/\sqrt{J}, \mathbf{0}_{5+p_z}^\top)}_{\boldsymbol{\omega}_{6\mathcal{U},1}^\top}, \underbrace{(0, -1/\sqrt{N}), \dots, (0, -1/\sqrt{N})}_{\boldsymbol{\omega}_{6\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{7,1} = \left(\underbrace{(0, 0, a_{11}^*/\sqrt{J}, \mathbf{0}_{3+p_z}^\top), \dots, (0, 0, a_{J1}^*/\sqrt{J}, \mathbf{0}_{3+p_z}^\top)}_{\boldsymbol{\omega}_{7\mathcal{U},1}^\top}, \underbrace{(-x_{11}/\sqrt{N}, 0), \dots, (-x_{N1}/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{7\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{8,1} = \left(\underbrace{(0, 0, a_{12}^*/\sqrt{J}, \mathbf{0}_{3+p_z}^\top), \dots, (0, 0, a_{J2}^*/\sqrt{J}, \mathbf{0}_{3+p_z}^\top)}_{\boldsymbol{\omega}_{8\mathcal{U},1}^\top}, \underbrace{(0, -x_{11}/\sqrt{N}), \dots, (0, -x_{N1}/\sqrt{N})}_{\boldsymbol{\omega}_{8\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{9,1} = \left(\underbrace{(0, 0, 0, a_{11}^*/\sqrt{J}, \mathbf{0}_{2+p_z}^\top), \dots, (0, 0, 0, a_{J1}^*/\sqrt{J}, \mathbf{0}_{2+p_z}^\top)}_{\boldsymbol{\omega}_{9\mathcal{U},1}^\top}, \underbrace{(-x_{12}/\sqrt{N}, 0), \dots, (-x_{N2}/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{9\Theta,1}^\top}\right)^\top,$$

$$\boldsymbol{\omega}_{10,1} = \left(\underbrace{(0, 0, 0, a_{12}^*/\sqrt{J}, \mathbf{0}_{2+p_z}^\top), \dots, (0, 0, 0, a_{J2}^*/\sqrt{J}, \mathbf{0}_{2+p_z}^\top)}_{\boldsymbol{\omega}_{10\mathcal{U},1}^\top}, \underbrace{(0, -x_{12}/\sqrt{N}), \dots, (0, -x_{N2}/\sqrt{N})}_{\boldsymbol{\omega}_{10\Theta,1}^\top}\right)^\top,$$

and $W_1 = (\boldsymbol{\omega}_{1,1}, \boldsymbol{\omega}_{2,1}, \dots, \boldsymbol{\omega}_{10,1})$. It is easy to check that $\boldsymbol{\omega}_{p,1}^\top\boldsymbol{\omega}_{q,1} = 0$ for $p \neq q$. Moreover, we have

$$
\begin{aligned}
W_1 W_1^\top &= \sum_{k=1}^{10} \boldsymbol{\omega}_{k,1} \boldsymbol{\omega}_{k,1}^\top \\
&= \begin{pmatrix} \sum_{k=1}^{10} \boldsymbol{\omega}_{k\mathcal{U},1} \boldsymbol{\omega}_{k\mathcal{U},1}^\top & -(NJ)^{-1/2} \left\{ \begin{pmatrix} \mathbf{D}_{i1} \\ \mathbf{x}_i \\ \mathbf{0}_{p_z} \\ \theta_i^* \end{pmatrix} \mathbf{a}_j^{*\top} \right\}_{j \le J, i \le N} \\ -(NJ)^{-1/2} \left\{ \mathbf{a}_j^* \left(\mathbf{D}_{i1}^\top, \mathbf{x}_i^\top, \mathbf{0}_{p_z}^\top, \theta_i^{*\top} \right) \right\}_{i \le N, j \le J} & \sum_{k=1}^{10} \boldsymbol{\omega}_{k\Theta,1} \boldsymbol{\omega}_{k\Theta,1}^\top \end{pmatrix}.
\end{aligned} \tag{A.23}
$$

Further, it is easy to see that under our normalisation,

$$
\begin{aligned}
W_1^\top W_1 = \operatorname{diag}\Big(& \sigma_{N1} + 1, \sigma_{N2} + 1, \sigma_{N1} + 1, \sigma_{N2} + 1, 2, 2, \\
& 1 + N^{-1} \sum_{i=1}^N x_{i1}^2, 1 + N^{-1} \sum_{i=1}^N x_{i1}^2, 1 + N^{-1} \sum_{i=1}^N x_{i2}^2, 1 + N^{-1} \sum_{i=1}^N x_{i2}^2 \Big).
\end{aligned}
$$

Next, we project $\boldsymbol{\mu}_k$ onto W_1, and write $\boldsymbol{\mu}_k = W_1 \mathbf{s}_{k,1} + \boldsymbol{\zeta}_{k,1}$ for $k = 1, \ldots 10$, where $\mathbf{s}_{k,1} = (W_1^\top W_1)^{-1} W_1^\top \boldsymbol{\mu}_k$. For example, we have

$$
\mathbf{s}_{1,1} = \begin{pmatrix} \frac{1}{\sigma_{N1}+1} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{s}_{2,1} = \begin{pmatrix} 0 \\ \frac{1}{\sigma_{N2}+1} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \mathbf{s}_{3,1} = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{\sigma_{N1}+1} \\ \frac{1}{\sigma_{N2}+1} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{s}_{4,1} = \begin{pmatrix} 0 \\ 0 \\ -\frac{\sigma_{N1}}{\sigma_{N1}+1} \\ -\frac{\sigma_{N2}}{\sigma_{N2}+1} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{s}_{5,1} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -0.5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.
$$

Define $\mathcal{S}_{N,1} = \sum_{k=1}^{10} \mathbf{s}_{k,1} \mathbf{s}_{k,1}^\top$. We have

$$
\mathcal{S}_{N,1} = \left(\begin{array}{cc} \begin{matrix} \frac{1}{(1+\sigma_{N1})^2} & 0 & 0 & 0 \\ 0 & \frac{1}{(1+\sigma_{N2})^2} & 0 & 0 \\ 0 & 0 & \frac{1+\sigma_{N1}^2}{(1+\sigma_{N1})^2} & \frac{1+\sigma_{N1}\sigma_{N2}}{(\sigma_{N1}+1)(\sigma_{N2}+1)} \\ 0 & 0 & \frac{1+\sigma_{N1}\sigma_{N2}}{(\sigma_{N1}+1)(\sigma_{N2}+1)} & \frac{1+\sigma_{N2}^2}{(1+\sigma_{N2})^2} \end{matrix} & 0_{4\times 6} \\ 0_{6\times 4} & \operatorname{diag} \begin{pmatrix} 0.25 \\ 0.25 \\ \frac{\left(N^{-1} \sum_{i=1}^N x_{i1}^2\right)^2}{\left(1+N^{-1} \sum_{i=1}^N x_{i1}^2\right)^2} \\ \frac{\left(N^{-1} \sum_{i=1}^N x_{i1}^2\right)^2}{\left(1+N^{-1} \sum_{i=1}^N x_{i1}^2\right)^2} \\ \frac{\left(N^{-1} \sum_{i=1}^N x_{i2}^2\right)^2}{\left(1+N^{-1} \sum_{i=1}^N x_{i2}^2\right)^2} \\ \frac{\left(N^{-1} \sum_{i=1}^N x_{i2}^2\right)^2}{\left(1+N^{-1} \sum_{i=1}^N x_{i2}^2\right)^2}) \end{pmatrix} \end{array} \right).
$$

It is easy to show that there exists $\underline{\pi} > 0$ such that $\pi_{\min}(\mathcal{S}_{N,1}) > \underline{\pi}$ for all large N as long as $\sigma_{N1} - \sigma_{N2}$ is bounded below by a positive constant for all large N, which is true under our assumption that $\sigma_{N1} \to \sigma_1$, $\sigma_{N2} \to \sigma_2$, and $\sigma_1 > \sigma_2$ as well as Assumption 2.5. Likewise, we can define $\boldsymbol{\omega}_{1,2}, \ldots, \boldsymbol{\omega}_{10,2}$ and $W_2 = (\boldsymbol{\omega}_{1,2}, \boldsymbol{\omega}_{2,2}, \ldots, \boldsymbol{\omega}_{10,2})$ such that

$$\boldsymbol{\omega}_{5,2} = \left(\underbrace{(0, a^*_{11}/\sqrt{J}, \mathbf{0}^\top_{4+p_z}), \ldots, (0, a^*_{J1}/\sqrt{J}, \mathbf{0}^\top_{4+p_z})}_{\boldsymbol{\omega}^\top_{5\mathcal{U},2}}, \underbrace{(-1/\sqrt{N}, 0), \ldots, (-1/\sqrt{N}, 0)}_{\boldsymbol{\omega}^\top_{5\Theta,2}} \right)^\top,$$

$$\boldsymbol{\omega}_{6,2} = \left(\underbrace{(0, a^*_{12}/\sqrt{J}, \mathbf{0}^\top_{4+p_z}), \ldots, (0, a^*_{J2}/\sqrt{J}), \mathbf{0}^\top_{4+p_z})}_{\boldsymbol{\omega}^\top_{6\mathcal{U},2}}, \underbrace{(0, -1/\sqrt{N}), \ldots, (0, -1/\sqrt{N})}_{\boldsymbol{\omega}^\top_{6\Theta,2}} \right)^\top,$$

and $\boldsymbol{\omega}_{k,2} = \boldsymbol{\omega}_{k,1}$ for $k = 1, \ldots, 4$ and $k = 7, \ldots, 10$. We can easily verify that $\boldsymbol{\omega}^\top_{p,2}\boldsymbol{\omega}_{q,2} = 0$ for $p \neq q$,

$$\begin{aligned} W_2 W_2^\top &= \sum_{k=1}^{10} \boldsymbol{\omega}_{k,2}\boldsymbol{\omega}^\top_{k,2} \\ &= \begin{pmatrix} \sum_{k=1}^{10} \boldsymbol{\omega}_{k\mathcal{U},2}\boldsymbol{\omega}^\top_{k\mathcal{U},2} & -(NJ)^{-1/2} \left\{ \begin{pmatrix} \mathbf{D}_{i2} \\ \mathbf{x}_i \\ \mathbf{0}_{p_z} \\ \theta^*_i \end{pmatrix} \mathbf{a}^{*\top}_j \right\}_{j \le J, i \le N} \\ -(NJ)^{-1/2} \left\{ \mathbf{a}^*_j \left(\mathbf{D}^\top_{i2}, \mathbf{x}^\top_i, \mathbf{0}^\top_{p_z}, \theta^{*\top}_i \right) \right\}_{i \le N, j \le J} & \sum_{k=1}^{10} \boldsymbol{\omega}_{k\Theta,2}\boldsymbol{\omega}^\top_{k\Theta,2} \end{pmatrix}, \end{aligned} \tag{A.24}$$

and $W_2^\top W_2 = W_1^\top W_1$. Similarly, we can write $\boldsymbol{\mu}_k = W_2 \mathbf{s}_{k,2} + \boldsymbol{\zeta}_{k,2}$ for $k = 1, \ldots 10$, where $\mathbf{s}_{k,2} = (W_2^\top W_2)^{-1} W_2^\top \boldsymbol{\mu}_k$. We can easily verify that $\mathbf{s}_{k,2} = \mathbf{s}_{k,1}$ for all k and thus $\mathcal{S}_{N,2} = \sum_{k=1}^{10} \mathbf{s}_{k,2}\mathbf{s}^\top_{k,2} = \mathcal{S}_{N,1}$. Therefore, we write $\mathbf{s}_k = \mathbf{s}_{k,1} = \mathbf{s}_{k,2}$ and $\mathcal{S}_N = \mathcal{S}_{N,1} = \mathcal{S}_{N,2}$. It then follows that

$$\begin{aligned} \partial \mathbb{P}_{NJ}(\Xi^*)/\partial \Xi \partial \Xi^\top &= 2\delta \left(\sum_{k=1}^{10} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top \right) \\ &= \delta \sum_{l=1}^{2} W_l \left(\sum_{k=1}^{10} \mathbf{s}_k \mathbf{s}_k^\top \right) W_l^\top + \delta \sum_{l=1}^{2} \left(\sum_{k=1}^{10} \boldsymbol{\zeta}_{k,l} \boldsymbol{\zeta}^\top_{k,l} \right) \\ &= \delta \underline{\pi} \sum_{l=1}^{2} W_l W_l^\top + \delta \sum_{l=1}^{2} W_l \left(\mathcal{S}_N - \underline{\pi} I_{10} \right) W_l^\top + \delta \sum_{l=1}^{2} \left(\sum_{k=1}^{10} \boldsymbol{\zeta}_{k,l} \boldsymbol{\zeta}^\top_{k,l} \right). \end{aligned} \tag{A.25}$$

Note that there exists $\kappa_8 > 0$ such that $E(\varrho''_{ijt}(\mathbf{u}_j^\top \mathbf{e}_{it})) > \kappa_8$ by Assumptions 2.1, 2.2 and

2.4. Now let $\underline{\delta} = \min\{\kappa_8, \delta\underline{\pi}\}$. Then it follows from (A.25) that

$$\begin{aligned}
\mathbb{H} =& \partial S^*(\boldsymbol{\Xi}^*)/\partial\boldsymbol{\Xi}^\top + \partial\mathbb{P}_{NJ}(\boldsymbol{\Xi}^*)/\partial\boldsymbol{\Xi}\partial\boldsymbol{\Xi}^\top \\
=& \partial S^*(\boldsymbol{\Xi}^*)/\partial\boldsymbol{\Xi}^\top + \underline{\delta}\sum_{l=1}^{2} W_l W_l^\top + (\delta\underline{\pi} - \underline{\delta})\sum_{l=1}^{2} W_l W_l^\top + \delta\sum_{l=1}^{2} W_l\left(\mathcal{S}_N - \underline{\pi} I_{10}\right)W_l^\top \\
&+ \delta\sum_{l=1}^{2}\left(\sum_{k=1}^{10}\boldsymbol{\zeta}_{k,l}\boldsymbol{\zeta}_{k,l}^\top\right) \\
\geq & \partial S^*(\boldsymbol{\Xi}^*)/\partial\boldsymbol{\Xi}^\top + \underline{\delta}\sum_{l=1}^{2} W_l W_l^\top .
\end{aligned}$$

Moreover, we can write

$$\begin{aligned}
&(NJ)^{1/2}\partial S^*(\boldsymbol{\Xi}^*)/\partial\boldsymbol{\Xi}^\top \\
=&\begin{pmatrix} \operatorname{diag}\left(\left\{\sum_{i=1}^N\sum_{t=1}^T E\left(\varrho''_{ijt}\right)\mathbf{e}^*_{it}\mathbf{e}^{*\top}_{it}\right\}_{j\le J}\right) & \left\{\sum_{t=1}^T E\left(\varrho''_{ijt}\right)\mathbf{e}^*_{it}\mathbf{a}^{*\top}_{j}\right\}_{j\le J, i\le N} \\ \left\{\sum_{t=1}^T E\left(\varrho''_{ijt}\right)\mathbf{a}^*_{j}\mathbf{e}^{*\top}_{it}\right\}_{i\le N, j\le J} & \operatorname{diag}\left(\left\{\sum_{j=1}^J\sum_{t=1}^T E\left(\varrho''_{ijt}\right)\mathbf{a}^*_{j}\mathbf{a}^{*\top}_{j}\right\}_{i\le N}\right)\end{pmatrix} \\
=&\underbrace{\underline{\delta}\begin{pmatrix} \operatorname{diag}\left(\left\{\sum_{i=1}^N\sum_{t=1}^T \mathbf{e}^*_{it}\mathbf{e}^{*\top}_{it}\right\}_{j\le J}\right) & \left\{\sum_{t=1}^T\begin{pmatrix}\mathbf{0}_4\\ \mathbf{z}_{it}\\ \mathbf{0}_2\end{pmatrix}\mathbf{a}^{*\top}_j\right\}_{j\le J, i\le N} \\ \left\{\sum_{t=1}^T \mathbf{a}^*_j\left(\mathbf{0}_4^\top, \mathbf{z}_{it}^\top, \mathbf{0}_2^\top\right)\right\}_{i\le N, j\le J} & \operatorname{diag}\left(\left\{T\sum_{j=1}^J \mathbf{a}^*_j\mathbf{a}^{*\top}_j\right\}_{i\le N}\right)\end{pmatrix}}_{(NJ)^{1/2}I} \\
&+\underbrace{\underline{\delta}\begin{pmatrix} 0_{(6+p_z)J\times(6+p_z)J} & \left\{\sum_{t=1}^T\begin{pmatrix}\mathbf{D}_{it}\\ \mathbf{x}_i\\ \mathbf{0}_{p_z}\\ \boldsymbol{\theta}^*_i\end{pmatrix}\mathbf{a}^{*\top}_j\right\}_{j\le J, i\le N} \\ \left\{\mathbf{a}^*_j\left(\mathbf{D}_{it}^\top, \mathbf{x}_i^\top, \mathbf{0}_{p_z}^\top, \boldsymbol{\theta}^{*\top}_i\right)\right\}_{i\le N, j\le J} & 0_{2N\times 2N}\end{pmatrix}}_{(NJ)^{1/2}II} \\
&+\underbrace{\sum_{t=1}^T\begin{pmatrix} \operatorname{diag}\left(\left\{\sum_{i=1}^N\left(E\left(\varrho''_{ijt}\right)-\underline{\delta}\right)\mathbf{e}^*_{it}\mathbf{e}^{*\top}_{it}\right\}_{j\le J}\right) & \left\{\left(E\left(\varrho''_{ijt}\right)-\underline{\delta}\right)\mathbf{e}^*_{it}\mathbf{a}^{*\top}_j\right\}_{j\le J, i\le N} \\ \left\{\left(E\left(\varrho''_{ijt}\right)-\underline{\delta}\right)\mathbf{a}^*_j\mathbf{e}^{*\top}_{it}\right\}_{i\le N, j\le J} & \operatorname{diag}\left(\left\{\sum_{j=1}^J\left(E\left(\varrho''_{ijt}\right)-\underline{\delta}\right)\mathbf{a}^*_j\mathbf{a}^{*\top}_j\right\}_{i\le N}\right)\end{pmatrix}}_{(NJ)^{1/2}III}.
\end{aligned}$$

For I, we have

$$I \geq \kappa_5 \cdot I_{(6+p_z)J+2N} \tag{A.26}$$

by Assumption A.3. From (A.23) and (A.24), we have

$$II + \underline{\delta}\sum_{l=1}^{2} W_l W_l^\top = \underline{\delta}\cdot\begin{pmatrix}\sum_{l=1}^2\sum_{k=1}^{10}\boldsymbol{\omega}_{k\mathcal{U},l}\boldsymbol{\omega}_{k\mathcal{U},l}^\top & 0_{(6+p_z)J\times 2N} \\ 0_{2N\times(6+p_z)J} & \sum_{l=1}^2\sum_{k=1}^{10}\boldsymbol{\omega}_{k\Theta,l}\boldsymbol{\omega}_{k\Theta,l}^\top\end{pmatrix} \geq 0_{PJ+2N,PJ+2N}. \tag{A.27}$$

For the last term, we have, for J, N large enough,

$$III = \frac{1}{NJ}\sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(E\left(\varrho_{ijt}^{''}\right)-\underline{\delta}\right)\varsigma_{ijt}\varsigma_{ijt}^{\top} \geq 0_{PJ+2N,PJ+2N}, \tag{A.28}$$

where $\varsigma_{ijt} = \left(\underbrace{0,\ldots,\mathbf{e}_{it}^{*\top},\ldots,0}_{(6+p_z)J},\underbrace{0,\ldots,\mathbf{a}_j^{*\top},\ldots,0}_{2N}\right)^{\top}$, by the definition of $\underline{\delta}$. It then follows from (A.26), (A.27) and (A.28) that

$$\mathbb{H} \geq \partial S^*(\Xi^*)/\partial\Xi^\top + \underline{\delta}\cdot\boldsymbol{\omega}\boldsymbol{\omega}^\top = I + II + III + \underline{\delta}\cdot\boldsymbol{\omega}\boldsymbol{\omega}^\top \geq \kappa_5\cdot I_{(6+p_z)J+2N},$$

and thus

$$\mathbb{H}^{-1} \leq \kappa_5^{-1}\cdot I_{(6+p_z)J+2N}. \tag{A.29}$$

Finally, write $\mathbb{H} = \mathbb{H}_d + \mathcal{C}$, where

$$\mathcal{C} = \begin{pmatrix} 0_{(6+p_z)J\times(6+p_z)J} & (NJ)^{-1/2}\left\{\sum_{t=1}^{T}E\left(\varrho_{ijt}^{''}\right)\mathbf{e}_{it}^{*}\mathbf{a}_j^{*\top}\right\}_{j\leq J, i\leq N} \\ (NJ)^{-1/2}\left\{\sum_{t=1}^{T}E\left(\varrho_{ijt}^{''}\right)\mathbf{a}_j^{*}\mathbf{e}_{it}^{*\top}\right\}_{i\leq N, j\leq J} & 0_{2N\times 2N} \end{pmatrix}$$
$$+ 2\delta\left(\sum_{k=1}^{10}\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\top}\right).$$

Note that

$$\mathbb{H}^{-1} - \mathbb{H}_d^{-1} = -\mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}_d^{-1} + \mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}^{-1}\mathcal{C}\mathbb{H}_d^{-1},$$

and thus $\left\|\mathbb{H}^{-1} - \mathbb{H}_d^{-1}\right\|_{\max} \leq \left\|\mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}_d^{-1}\right\|_{\max} + \left\|\mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}^{-1}\mathcal{C}\mathbb{H}_d^{-1}\right\|_{\max}$. Inequality (A.29) implies that $\mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}^{-1}\mathcal{C}\mathbb{H}_d^{-1} \leq \kappa_5^{-1}\mathbb{H}_d^{-1}\mathcal{C}^2\mathbb{H}_d^{-1}$, and thus the l-th diagonal element of $\mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}^{-1}\mathcal{C}\mathbb{H}_d^{-1}$ is smaller than the l-th diagonal element of $\kappa_5^{-1}\mathbb{H}_d^{-1}\mathcal{C}^2\mathbb{H}_d^{-1}$. It then follows that $\left\|\mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}^{-1}\mathcal{C}\mathbb{H}_d^{-1}\right\|_{\max} \leq \kappa_5^{-1}\left\|\mathbb{H}_d^{-1}\mathcal{C}^2\mathbb{H}_d^{-1}\right\|_{\max}$ and therefore

$$\left\|\mathbb{H}^{-1} - \mathbb{H}_d^{-1}\right\|_{\max} \leq \left\|\mathbb{H}_d^{-1}\mathcal{C}\mathbb{H}_d^{-1}\right\|_{\max} + \kappa_5^{-1}\left\|\mathbb{H}_d^{-1}\mathcal{C}^2\mathbb{H}_d^{-1}\right\|_{\max}$$

because the entry with the largest absolute value of a positive semidefinite matrix is always on the diagonal. Since $\mathbb{H}_d^{-1}$ is a block diagonal matrix whose elements are all $O(1)$ from Assumptions 2.6 and $\|\mathcal{C}\|_{\max}$ and $\|\mathcal{C}^2\|_{\max}$ are $O(N^{-1})$, the proof is complete.

By this lemma, (A.22), and the fact that $\partial\mathbb{P}_{NJ}(\hat{\Xi})/\partial\Xi = \partial\mathbb{P}_{NJ}(\Xi^*)/\partial\Xi = 0$, we can write

$$\hat{\Xi} - \Xi^* = \mathbb{H}^{-1}S^*(\hat{\Xi}) - \mathbb{H}^{-1}S^*(\Xi^*) - 0.5\mathbb{H}^{-1}\mathcal{R}(\hat{\Xi}). \tag{A.30}$$

Define

$$S_{NJ}^*(\Xi) = \left(\underbrace{\ldots,\frac{1}{\sqrt{NJ}}\sum_{i=1}^{N}\sum_{t=1}^{T}\varrho_{ijt}^{'}(\mathbf{u}_j^\top\mathbf{e}_{it})\mathbf{e}_{it}^\top,\ldots,}_{JK}\underbrace{\ldots,\frac{1}{\sqrt{NJ}}\sum_{j=1}^{J}\sum_{t=1}^{T}\varrho_{ijt}^{'}(\mathbf{u}_j^\top\mathbf{e}_{it})\mathbf{a}_j^\top,\ldots}_{NK^*}\right)^{\top},$$

$\breve{S}^*(\Xi) = S^*_{NJ}(\Xi) - S^*(\Xi)$ and $\mathcal{D} = \mathbb{H}^{-1} - \mathbb{H}_d^{-1}$. Note that by the first-order conditions, $S^*_{NJ}(\hat{\Xi}) = 0$. As a result, we can write

$$\begin{aligned}
\mathbb{H}^{-1}S^*(\hat{\Xi}) =& \mathbb{H}_d^{-1}S^*(\hat{\Xi}) + \mathcal{D}S^*(\hat{\Xi}) \\
=& -\mathbb{H}_d^{-1}\breve{S}^*(\hat{\Xi}) + \mathcal{D}S^*(\hat{\Xi}) \\
=& -\mathbb{H}_d^{-1}\breve{S}^*(\Xi^*) - \mathbb{H}_d^{-1}\left(\breve{S}^*(\hat{\Xi}) - \breve{S}^*(\Xi^*)\right) + \mathcal{D}S^*(\hat{\Xi}) \qquad \text{(A.31)} \\
=& -\mathbb{H}_d^{-1}\breve{S}^*(\Xi^*) - \mathbb{H}_d^{-1}\left(\breve{S}^*(\hat{\Xi}) - \breve{S}^*(\Xi^*)\right) \\
& - \mathcal{D}\breve{S}^*(\Xi^*) - \mathcal{D}\left(\breve{S}^*(\hat{\Xi}) - \breve{S}^*(\Xi^*)\right). \qquad \text{(A.32)}
\end{aligned}$$

Next, let $\mathcal{R}(\hat{\boldsymbol{\zeta}})_m$ denote the vector containing the $\{(m-1)(P)+1\}$th to the mPth elements of $\mathcal{R}(\hat{\boldsymbol{\zeta}})$ for $m = 1, \dots J$, and $\{JP + (m-J-1)(K^*)+1\}$th to the $\{JP+(m-J)(K^*)\}$th elements of $\mathcal{R}(\hat{\boldsymbol{\zeta}})$ for $m = J+1 \dots J+N$. We further let $\bar{O}_P(\cdot)$ denote a stochastic order that is uniform in i and j. For example, $Q_{ij} = \bar{O}_P(1)$ means that $\max_{i\le N, j\le J}\|Q_{ij}\| = O_P(1)$. Then, by the result of Theorem A.1, it can be shown that

$$\mathcal{R}(\hat{\Xi})_m = \bar{O}_P(1)\|\hat{\mathbf{u}}_j - \mathbf{u}_j^*\|^2 + \bar{O}_P(1/\sqrt{N})\|\hat{\mathbf{u}}_j - \mathbf{u}_j^*\| + \bar{O}_P(1/N) \qquad \text{(A.33)}$$

for $m = 1, \dots, J$ and

$$\mathcal{R}(\hat{\Xi})_{J+m} = \bar{O}_P(1)\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*\|^2 + \bar{O}_P(1/\sqrt{N})\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*\| + \bar{O}_P(1/N) \qquad \text{(A.34)}$$

for $i = 1, \dots, N$. We define $\mathcal{D}_{m,s}$ such that

$$\mathcal{D}_{m,s} = \begin{cases} \mathcal{D}_{[(m-1)P:mP,(s-1)P:sP]}, \text{if } m, s \in \{1, \dots, J\} \\ \mathcal{D}_{[(m-1)P:mP,JP+(s-J-1)K^*:JP+(s-J)K^*]}, \text{if } m \in \{1, \dots, J\}, s \in \{J+1, \dots N\} \\ \mathcal{D}_{[JP+(m-J-1)K^*:JP+(m-J)K^*,(s-1)P:sP]}, \text{if } m \in \{J+1, \dots N\}, s \in \{1, \dots, J\} \\ \mathcal{D}_{[JP+(m-J-1)K^*:JP+(m-J)K^*,JP+(s-J-1)K^*:JP+(s-J)K^*]}, \text{if } m, s \in \{J+1, \dots N\} \end{cases}.$$

Note that $S^*(\Xi^*) = 0$. Then, from equations (A.30) to (A.34), recall that we defined

$\breve{\varrho}'_{ijt}(\mathbf{u}_j^\top \mathbf{e}_{it}) = \varrho'_{ijt}(\mathbf{u}_j{}^\top \mathbf{e}_{it}) - E(\varrho'_{ijt}(\mathbf{u}_j{}^\top \mathbf{e}_{it}))$ and $\breve{\varrho}'_{ijt} = \breve{\varrho}'_{ijt}(\mathbf{u}_j^{*\top} \mathbf{e}^*_{it})$, we can write

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^* = &- (\mathit{\Psi}_{J,i})^{-1} \frac{1}{J} \sum_{j=1}^{J} \sum_{t=1}^{T} \breve{\varrho}'_{ijt} \mathbf{a}_j^* - \frac{1}{\sqrt{NJ}} \sum_{s=1}^{J} \sum_{m=1}^{N} \mathcal{D}_{J+i,s} \cdot \sum_{t=1}^{T} \breve{\varrho}'_{mst} \mathbf{e}^*_{mt} \\
&- \frac{1}{\sqrt{NJ}} \sum_{s=1}^{J} \sum_{m=1}^{N} \mathcal{D}_{J+i,J+m} \cdot \sum_{t=1}^{T} \breve{\varrho}'_{mst} \mathbf{a}_s^* \\
&- (\mathit{\Psi}_{J,i})^{-1} \frac{1}{J} \sum_{j=1}^{J} \sum_{t=1}^{T} \left(\breve{\varrho}'_{ijt} (\hat{\mathbf{u}}_j^\top \hat{\mathbf{e}}_{it}) \hat{\mathbf{a}}_j - \breve{\varrho}'_{ijt} \mathbf{a}_j^* \right) \\
&- \frac{1}{\sqrt{NJ}} \sum_{s=1}^{J} \sum_{m=1}^{N} \mathcal{D}_{J+i,s} \sum_{t=1}^{T} \left(\breve{\varrho}'_{mst} (\hat{\mathbf{u}}_s^\top \hat{\mathbf{e}}_{mt}) \hat{\mathbf{e}}_{mt} - \breve{\varrho}'_{mst} \mathbf{e}^*_{mt} \right) \\
&- \frac{1}{\sqrt{NJ}} \sum_{s=1}^{J} \sum_{m=1}^{N} \mathcal{D}_{J+i,J+m} \sum_{t=1}^{T} \left(\breve{\varrho}'_{mst} (\hat{\mathbf{u}}_s^\top \hat{\mathbf{e}}_{mt}) \hat{\mathbf{a}}_s - \breve{\varrho}'_{mst} \mathbf{a}_s^* \right) \\
&- 0.5 (\mathit{\Psi}_{J,i})^{-1} \mathcal{R}(\hat{\boldsymbol{\Xi}})_{J+i} - 0.5 \sum_{s=1}^{J} \mathcal{D}_{J+i,s} \mathcal{R}(\hat{\boldsymbol{\Xi}})_s \\
&- 0.5 \sum_{m=1}^{N} \mathcal{D}_{J+i,J+m} \mathcal{R}(\hat{\boldsymbol{\Xi}})_{J+m}.
\end{aligned}
\tag{A.35}
$$

Lemma A.7. *Let $c_1, \ldots, c_N$ be a sequence of uniformly bounded constants. Then, under the conditions in Lemma A.6, we have*

$$
\frac{1}{N} \sum_{i=1}^{N} c_i (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*) = O_P\left(\frac{1}{N}\right).
$$

Proof: Define $d_s = \sqrt{NJ} \cdot N^{-1} \sum_{i=1}^{N} c_i \mathcal{D}_{J+i,s}$, for $s = 1, \ldots, N+J$. Lemma A.6 implies that $\max_{1 \leq s \leq N+J} \|d_s\|$ is bounded. From (A.35), we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} c_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*) \\
=& -\frac{1}{NJ}\sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{t=1}^{T} c_i(\mathit{\Psi}_{J,i})^{-1}\breve{\varrho}'_{ijt}\mathbf{a}_j^* \\
&-\frac{1}{NJ}\sum_{m=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T} d_s\breve{\varrho}'_{mst}\cdot\mathbf{e}_{mt}^* - \frac{1}{NJ}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T} d_{J+m}\breve{\varrho}'_{mst}\cdot\mathbf{a}_s^* \\
&-\frac{1}{NJ}\sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{t=1}^{T} c_i(\mathit{\Psi}_{J,i})^{-1}\left(\breve{\varrho}'_{ijt}(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it})\hat{\mathbf{a}}_j - \breve{\varrho}'_{ijt}\mathbf{a}_j^*\right) \\
&-\frac{1}{NJ}\sum_{m=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T} d_s\left(\breve{\varrho}'_{mst}(\hat{\mathbf{u}}_s^\top\hat{\mathbf{e}}_{mt})\hat{\mathbf{e}}_{mt} - \breve{\varrho}'_{mst}\mathbf{e}_{mt}^*\right) \\
&-\frac{1}{NJ}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T} d_{J+m}\left(\breve{\varrho}'_{mst}(\hat{\mathbf{u}}_s^\top\hat{\mathbf{e}}_{mt})\hat{\mathbf{a}}_s - \breve{\varrho}'_{mst}\mathbf{a}_s^*\right) - 0.5\frac{1}{N}\sum_{i=1}^{N} c_i(\mathit{\Psi}_{J,i})^{-1}\mathcal{R}(\hat{\boldsymbol{\zeta}})_{J+i} \\
&-0.5\frac{1}{\sqrt{NJ}}\sum_{s=1}^{J} d_s\mathcal{R}(\hat{\boldsymbol{\Xi}})_s - 0.5\frac{1}{\sqrt{NJ}}\sum_{m=1}^{N} d_{J+m}\mathcal{R}(\hat{\boldsymbol{\Xi}})_{J+m}. \qquad \text{(A.36)}
\end{aligned}
$$

First, by Lyapunov's CLT, it is easy to see that the first three terms on the RHS of (A.36) are all $O_P(1/\sqrt{NJ})$. Next, it follows from Theorem A.1, (A.33) and (A.34) that the last three terms on the RHS of (A.36) are all $O_P(1/\min\{N,J\}^2)$. Finally, we will show that the remaining three terms on the RHS of (A.36) are all $O_P(1/N)$, from which the desired result follows. Define

$$
\mathbb{V}_{NJ}(\boldsymbol{\Xi}) = \frac{1}{NJ}\sum_{m=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T} d_s\left(\breve{\varrho}'_{mst}(\mathbf{u}_s^\top\mathbf{e}_{mt})\mathbf{e}_{mt} - \breve{\varrho}'_{mst}\mathbf{e}_{mt}^*\right),
$$

and $\Delta_{NJ}(\boldsymbol{\Xi}^{(a)}, \boldsymbol{\Xi}^{(b)}) = \sqrt{NJ}\left(\mathbb{V}_{NJ}(\boldsymbol{\Xi}^{(a)}) - \mathbb{V}_{NJ}(\boldsymbol{\Xi}^{(b)})\right)$. Note that

$$
\begin{aligned}
\Delta_{NJ}(\boldsymbol{\Xi}^{(a)}, \boldsymbol{\Xi}^{(b)}) =& \underbrace{\frac{1}{\sqrt{NJ}}\sum_{m=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T} d_s\breve{\varrho}'_{mst}(\mathbf{u}_s^{(a)\top}\mathbf{e}_{mt}^{(a)})(\mathbf{e}_{mt}^{(a)} - \mathbf{e}_{mt}^{(b)})}_{\Delta_{1,NJ}(\boldsymbol{\Xi}^{(a)},\boldsymbol{\Xi}^{(b)})} \\
&+ \underbrace{\frac{1}{\sqrt{NJ}}\sum_{m=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T} d_s\left(\breve{\varrho}'_{mst}(\mathbf{u}_s^{(a)\top}\mathbf{e}_{mt}^{(a)}) - \breve{\varrho}'_{mst}(\mathbf{u}_s^{(b)\top}\mathbf{e}_{mt}^{(b)})\right)\mathbf{e}_{mt}^{(b)}}_{\Delta_{2,NJ}(\boldsymbol{\zeta}^{(a)},\boldsymbol{\zeta}^{(b)})}.
\end{aligned}
$$

Note that we have

$$
\begin{aligned}
\left\|\sum_{t=1}^{T} d_s\breve{\varrho}'_{mst}(\mathbf{u}_s^{(a)\top}\mathbf{e}_{mt}^{(a)})(\mathbf{e}_{mt}^{(a)} - \mathbf{e}_{mt}^{(b)})\right\| &\lesssim \left\|\boldsymbol{\theta}_m^{(a)} - \boldsymbol{\theta}_m^{(b)}\right\| \text{ and} \\
\left|\sum_{t=1}^{T} d_s\left(\breve{\varrho}'_{mst}(\mathbf{u}_s^{(a)\top}\mathbf{e}_{mt}^{(a)}) - \breve{\varrho}'_{mst}(\mathbf{u}_s^{(b)\top}\mathbf{e}_{mt}^{(b)})\right)\right| &\lesssim \left|\mathbf{u}_s^{(a)\top}\mathbf{e}_m^{(a)} - \mathbf{u}_s^{(b)\top}\mathbf{e}_m^{(b)}\right|.
\end{aligned}
$$

By Hoeffding's inequality, Lemma 2.2.1 of van der Vaart and Wellner (1996), and arguments similar to the proof of Lemma A.2, we can show that for $d(\Xi^{(a)}, \Xi^{(b)})$ sufficiently small,

$$\begin{aligned} &\left\|\Delta_{1,NJ}(\Xi^{(a)}, \Xi^{(b)})\right\|_{\Psi_2} \lesssim \left\|\Theta^{(a)} - \Theta^{(b)}\right\|_F \lesssim d(\Xi^{(a)}, \Xi^{(b)}), \\ &\left\|\Delta_{2,NJ}(\Xi^{(a)}, \Xi^{(b)})\right\|_{\Psi_2} \lesssim d(\Xi^{(a)}, \Xi^{(b)}). \end{aligned}$$

Thus,

$$\left\|\Delta_{NJ}(\Xi^{(a)}, \Xi^{(b)})\right\|_{\Psi_2} \leq \left\|\Delta_{1,NJ}(\Xi^{(a)}, \Xi^{(b)})\right\|_{\Psi_2} + \left\|\Delta_{2,NJ}(\Xi^{(a)}, \Xi^{(b)})\right\|_{\Psi_2} \lesssim d(\Xi^{(a)}, \Xi^{(b)}).$$

Therefore, similar to the proof of Lemma A.3, we can show that for sufficiently small $c > 0$,

$$E\left(\sup_{\Xi \in \mathcal{H}^{K^*}(c)} |\mathbb{V}_{NJ}(\Xi)|\right) \lesssim \frac{c}{\min\{\sqrt{N}, \sqrt{J}\}}. \tag{A.37}$$

It then follows from (A.37) and Theorem A.1 that $\mathbb{V}_{NJ}(\hat{\Xi}) = O_P(1/N)$. Thus the fifth term of the right of (A.36) is $O_P(1/N)$. Similar results can be obtained for the fourth term and sixth term on the right of (A.36), and the desired result follows.

Lemma A.8. *Under the conditions in lemma A.6, for each j we have*

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \breve{\varrho}'_{ijt}(\hat{\mathbf{e}}_{it} - \mathbf{e}^*_{it}) = O_P\left(\frac{1}{N}\right) \text{ and } \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \breve{\varrho}''_{ijt}\mathbf{e}^*_{it}(\hat{\mathbf{e}}_{it} - \mathbf{e}^*_{it})^\top = O_P\left(\frac{1}{N}\right).$$

Proof: It suffices to prove that for each $l \in \{1, \ldots, T\}$,

$$\frac{1}{N}\sum_{i=1}^{N} \breve{\varrho}'_{ijl}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*_i) = O_P\left(\frac{1}{N}\right) \text{ and } \frac{1}{N}\sum_{i=1}^{N} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*_i)^\top = O_P\left(\frac{1}{N}\right).$$

The proof of the two results are similar, thus we only prove the second one to save space. By (A.31) and (A.35), we have

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*_i)^\top \\ = &- \frac{1}{NJ}\sum_{i=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T} \breve{\varrho}''_{ijl}\breve{\varrho}'_{ist}\mathbf{e}^*_{il}\mathbf{a}^{*\top}_s(\mathit{\Psi}_{J,i})^{-1} \\ &- \frac{1}{NJ}\sum_{i=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}\left(\breve{\varrho}'_{ist}(\hat{\mathbf{u}}_s^\top\hat{\mathbf{e}}_{it})\hat{\mathbf{a}}_s - \breve{\varrho}'_{ist}\mathbf{a}^*_s\right)(\mathit{\Psi}_{J,i})^{-1} \\ &- \frac{1}{\sqrt{NJ}}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}\hat{\mathbf{e}}_{mt}^\top\mathcal{D}_{J+i,s}^\top\right)E\left(\varrho'_{mst}(\hat{\mathbf{u}}_s^\top\hat{\mathbf{e}}_{mt})\right) \\ &- \frac{1}{\sqrt{NJ}}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}\hat{\mathbf{a}}_s^\top\mathcal{D}_{J+i,J+m}^\top\right)E\left(\varrho'_{mst}(\hat{\mathbf{u}}_s^\top\hat{\mathbf{e}}_{mt})\right) \\ &- \frac{1}{2N}\sum_{i=1}^{N} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}\mathcal{R}(\hat{\boldsymbol{\Xi}})_{J+i}^\top(\mathit{\Psi}_{J,i})^{-1} - \frac{1}{2N}\sum_{i=1}^{N}\sum_{s=1}^{J} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}\mathcal{R}(\hat{\boldsymbol{\Xi}})_s^\top\mathcal{D}_{J+i,s}^\top \\ &- \frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{N} \breve{\varrho}''_{ijl}\mathbf{e}^*_{il}\mathcal{R}(\hat{\boldsymbol{\Xi}})_{J+m}^\top\mathcal{D}_{J+i,J+m}^\top. \end{aligned} \tag{A.38}$$

First, we can write

$$
\begin{aligned}
&-\frac{1}{NJ}\sum_{i=1}^{N}\sum_{s=1}^{J}\sum_{t=1}^{T}\breve{\varrho}^{''}_{ijl}\breve{\varrho}^{'}_{ist}\mathbf{e}^{*}_{il}\mathbf{a}^{*\top}_{s}(\mathit{\Psi}_{J,i})^{-1}\\
=&-\frac{1}{NJ}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}^{''}_{ijl}\breve{\varrho}^{'}_{ijt}\mathbf{e}^{*}_{il}\mathbf{a}^{*\top}_{j}(\mathit{\Psi}_{J,i})^{-1}-\frac{1}{NJ}\sum_{i=1}^{N}\sum_{s=1,s\neq j}^{J}\sum_{t=1}^{T}\breve{\varrho}^{''}_{ijl}\breve{\varrho}^{'}_{ist}\mathbf{e}^{*}_{il}\mathbf{a}^{*\top}_{s}(\mathit{\Psi}_{J,i})^{-1}.
\end{aligned}
$$

Since $\breve{\varrho}^{''}_{ijl}, \breve{\varrho}^{'}_{ist}$ are bounded and $\max_{i\leq N}\|(\mathit{\Psi}_{J,i})^{-1}\| = O(1)$ for large J by Assumption 2.6, the first term of the RHS of the above equation is $O_P(J^{-1})$. By Lyapunov's CLT, the second term on the RHS of the above equation can be shown to be $O_P((NJ)^{-1/2})$. Thus, the first term on the RHS of (A.38) is $O_P(N^{-1})$.
Second, for the second term on the RHS of (A.38), it can be written as

$$
O_P(J^{-1})-\frac{1}{NJ}\sum_{i=1}^{N}\sum_{s=1,s\neq j}^{J}\sum_{t=1}^{T}\breve{\varrho}^{''}_{ijl}\mathbf{e}^{*}_{il}\left(\breve{\varrho}^{'}_{ist}(\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{it})\hat{\mathbf{a}}_s-\breve{\varrho}^{'}_{ist}\mathbf{a}^{*}_{s}\right)(\mathit{\Psi}_{J,i})^{-1}.
$$

Similar to the proof of Lemma A.7, the second term of the above expression can be shown to be $O_P(N^{-1})$. So the second term of the RHS of (A.38) is $O_P(N^{-1})$. For the third term on the RHS of (A.38), its p, qth element is given by

$$
-\frac{1}{NJ}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T}\chi_{j,s}E(\varrho^{'}_{mst}(\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt}))\hat{\mathbf{e}}_{mt},
$$

where $\chi_{j,s} = N^{-1}\sum_{i=1}^{N}\left(\sqrt{NJ}\mathcal{D}_{J+i,s,q}\right)\breve{\varrho}^{''}_{ijl}\mathbf{e}^{*}_{il,p}$, and $\mathcal{D}_{J+i,s,q}$ is the qth row of $\mathcal{D}_{J+i,s}$. Therefore,

$$
\begin{aligned}
&\left|-\frac{1}{NJ}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T}\chi_{j,s}E(\varrho^{'}_{mst}(\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt}))\hat{\mathbf{e}}_{mt}\right|\\
\lesssim&\sqrt{\frac{1}{J}\sum_{s=1}^{J}\|\chi_{j,s}\|^2}\cdot\sqrt{\frac{1}{NJ}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T}E(\varrho^{'}_{mst}(\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt}))^2\|\hat{\mathbf{e}}_{mt}\|^2}.
\end{aligned}
$$

Since $|\sqrt{NJ}\mathcal{D}_{J+i,s,q}|$ is uniformly bounded by Lemma A.6, it can be shown that $E\|\chi_{j,s}\|^2$ is $O(N^{-1})$. Moreover, for some $\tilde{c}_{mst}$ between $\mathbf{u}^{*\top}_{s}\mathbf{e}^{*}_{mt}$ and $\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt}$,

$$
\begin{aligned}
E\left(\varrho^{'}_{mst}(\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt})\right)^2&=\left\{E(\varrho^{'}_{mst})+E(\varrho^{''}_{mst}(\tilde{c}_{mst}))(\mathbf{u}^{*\top}_{s}\mathbf{e}^{*}_{mt}-\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt})\right\}^2\\
&\lesssim(\mathbf{u}^{*\top}_{s}\mathbf{e}^{*}_{mt}-\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt})^2.
\end{aligned}
$$

Thus,

$$
\sqrt{\frac{1}{NJ}\sum_{s=1}^{J}\sum_{m=1}^{N}\sum_{t=1}^{T}E(\varrho^{'}_{mst}(\hat{\mathbf{u}}_s^{\top}\hat{\mathbf{e}}_{mt}))^2\|\hat{\mathbf{e}}_{mt}\|^2}\lesssim d(\hat{\mathbf{\Xi}},\mathbf{\Xi}^{*})=O_P(1/\min\{\sqrt{N},\sqrt{J}\})
$$

by Theorem A.1. So, the third term on the RHS of (A.38) is $O_P(N^{-1})$, and the fourth term of the RHS of (A.38) can be shown to be $O_P(N^{-1})$ in the same way. Finally, it follows from Theorem A.1 and (A.34) that the fifth term on the RHS of (A.38) is $O_P((\min\{\sqrt{N},\sqrt{J}\}^{-1})^2) = O_P(N^{-1})$. For the sixth term, the absolute value of the p, qth element can be written as

$$\frac{1}{2\sqrt{NJ}}\left|\sum_{s=1}^{J}\chi_{j,s}\mathcal{R}(\hat{\Xi})_s\right| \leq \frac{\sqrt{J}}{2\sqrt{N}}\sqrt{\frac{1}{J}\sum_{s=1}^{J}\|\chi_{j,s}\|^2}\sqrt{\frac{1}{J}\sum_{s=1}^{J}\|\mathcal{R}(\hat{\Xi})_s\|^2} = O_P(N^{-3/2}).$$

The same bound for the seventh term on the RHS of (A.38) can be obtained using the same argument. Therefore, we get

$$\frac{1}{N}\sum_{i=1}^{N}\breve{\varrho}''_{ijl}\mathbf{e}^*_{il}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*_i)^\top = O_P(N^{-1})$$

as desired by combining the results.

Proof of Theorem A.2 We first consider the case where the conditions in Lemma A.6 hold. From the expansion in the proof of Lemma A.5, we can derive that

$$\begin{aligned}\mathit{\Phi}_{N,j}(\hat{\mathbf{u}}_j - \mathbf{u}^*_j) =& \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varrho'_{ijt}(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it})\right)\hat{\mathbf{e}}_{it} - \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varrho'_{ijt}\right)(\hat{\mathbf{e}}_{it} - \mathbf{e}^*_{it})\\ &- \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varrho''_{ijt}(\mathbf{u}_j^{*\top}\tilde{\mathbf{e}}_{it})\right)\mathbf{e}^*_{it}\mathbf{u}_j^{*\top}(\hat{\mathbf{e}}_{it} - \mathbf{e}^*_{it})\\ &+ O_P\left(N^{-1}\left\|\hat{\Theta} - \Theta^*\right\|_F^2\right) + o_P(\left\|\hat{\mathbf{u}}_j - \mathbf{u}^*_j\right\|).\end{aligned}$$

Then, it follows from Theorem A.1 and Lemma A.7 that

$$\mathit{\Phi}_{N,j}(\hat{\mathbf{u}}_j - \mathbf{u}^*_j) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varrho'_{ijt}(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it})\right)\hat{\mathbf{e}}_{it} + O_P(N^{-1}) + o_P(\left\|\hat{\mathbf{u}}_j - \mathbf{u}^*_j\right\|).$$

Note that

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varrho'_{ijt}(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it})\right)\hat{\mathbf{e}}_{it}\\ =&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it}\\ =&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}\hat{\mathbf{e}}_{it} - \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}\cdot(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it} - \mathbf{u}_j^{*\top}\mathbf{e}^*_{it})\hat{\mathbf{e}}_{it}\\ &- 0.5\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'''_{ijt}(\tilde{m}_{ijt})(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it} - \mathbf{u}_j^{*\top}\mathbf{e}^*_{it})^2\hat{\mathbf{e}}_{it},\end{aligned}$$

where $\tilde{m}_{ijt}$ is between $\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it}$ and $\mathbf{u}_j^{*\top}\mathbf{e}^*_{it}$.
By Lemma A.8, we have

$$\begin{aligned}-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}\hat{\mathbf{e}}_{it} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}\mathbf{e}^*_{it} - \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}(\hat{\mathbf{e}}_{it} - \mathbf{e}^*_{it})\\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}\mathbf{e}^*_{it} + O_P(N^{-1}).\end{aligned}$$

Second,

$$\begin{aligned}
&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}\cdot(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it}-\mathbf{u}_j^{*\top}\mathbf{e}_{it}^*)\hat{\mathbf{e}}_{it}\\
=&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}\hat{\mathbf{e}}_{it}(\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*)^\top\hat{\mathbf{u}}_j-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}\hat{\mathbf{e}}_{it}\mathbf{e}_{it}^{*\top}(\hat{\mathbf{u}}_j-\mathbf{u}_j^*)\\
=&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}\mathbf{e}_{it}^*(\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*)^\top\hat{\mathbf{u}}_j-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}(\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*)(\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*)^\top\hat{\mathbf{u}}_j\\
&-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}\mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}(\hat{\mathbf{u}}_j-\mathbf{u}_j^*)-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}(\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*)\mathbf{e}_{it}^{*\top}(\hat{\mathbf{u}}_j-\mathbf{u}_j^*). \qquad (A.39)
\end{aligned}$$

It then follows from Lemma A.8 and Theorem A.1 that the first two terms on the RHS of (A.39) are $O_P(N^{-1})$, respectively. It is easy to show that the last two terms on the right of (A.39) are both $o_P(\|\hat{\mathbf{u}}_j-\mathbf{u}_j^*\|)$. Thus, we have

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}''_{ijt}\cdot(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it}-\mathbf{u}_j^{*\top}\mathbf{e}_{it}^*)\hat{\mathbf{e}}_{it}=O_P(N^{-1})+o_P(\|\hat{\mathbf{u}}_j-\mathbf{u}_j^*\|).$$

Next, it is easy to show that

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'''_{ijt}(\tilde{m}_{ijt})(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it}-\mathbf{u}_j^{*\top}\mathbf{e}_{it}^*)^2\hat{\mathbf{e}}_{it}\right\|\\
\lesssim&\|\hat{\mathbf{u}}_j-\mathbf{u}_j^*\|^2\frac{1}{N}\sum_{t=1}^{T}|\breve{\varrho}'''_{ijt}(\tilde{m}_{ijt})|+\frac{1}{N}\sum_{t=1}^{T}|\breve{\varrho}'''_{ijt}(\tilde{m}_{ijt})|\cdot\|\hat{\mathbf{e}}_{it}-\mathbf{e}_{it}^*\|^2.
\end{aligned}$$

Therefore, from Theorem A.1 and Lemma A.5, we have

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'''_{ijt}(\tilde{m}_{ijt})(\hat{\mathbf{u}}_j^\top\hat{\mathbf{e}}_{it}-\mathbf{u}_j^{*\top}\mathbf{e}_{it}^*)^2\hat{\mathbf{e}}_{it}=O_P(N^{-1}). \qquad (A.40)$$

Finally, combining all the results, we get

$$\mathit{\Phi}_{N,j}(\hat{\mathbf{u}}_j-\mathbf{u}_j^*)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}\mathbf{e}_{it}^*+o_P(\|\hat{\mathbf{u}}_j-\mathbf{u}_j^*\|)+O_P(N^{-1}).$$

By Assumption 2.6, we can show that

$$\mathit{\Phi}_{N,j}\to\mathit{\Phi}_j>0 \text{ and } -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}\mathbf{e}_{it}^*\xrightarrow{d}\mathcal{N}\left(0,\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left((\varrho'_{ijt})^2\right)\mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}\right). \qquad (A.41)$$

Thus, the desired result follows from (A.40) and (A.41) and the fact that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left((\varrho'_{ijt})^2\right)\mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}=-\mathit{\Phi}_j.$$

For the general case where the additional assumptions for X in Lemma A.6 are not satisfied, we note that by similar arguments in the normalisation algorithm mentioned in Section A.2, there exists a $(1+p+p_z+K^*)\times(1+p+p_z+K^*)$ matrix Q such that $(\Gamma_1, X, Z_t, \Theta)Q^\top$ satisfies the conditions in Lemma A.6. The corresponding estimates and values of U_t will be $\hat{U}_tQ^{-1}$ and $U_t^*Q^{-1}$ accordingly. Define

$$\varPhi^Q_{N,j} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} E\left(\varrho''_{ijt}\right) Q\mathbf{e}^*_{it}\mathbf{e}^{*\top}_{it}Q^\top = Q\varPhi_{N,j}Q^\top.$$

By the previous result, we have

$$\begin{aligned}\varPhi^Q_{N,j}Q^{-1^\top}(\hat{\mathbf{u}}_j-\mathbf{u}^*_j) &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}Q\mathbf{e}^*_{it} + o_P(\|\hat{\mathbf{u}}_j-\mathbf{u}^*_j\|) + O_P(N^{-1}), \text{ and thus}\\ \varPhi_{N,j}(\hat{\mathbf{u}}_j-\mathbf{u}^*_j) &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\varrho}'_{ijt}\mathbf{e}^*_{it} + o_P(\|\hat{\mathbf{u}}_j-\mathbf{u}^*_j\|) + O_P(N^{-1})\end{aligned}$$

by straightforward calculations. This completes the proof of Theorem A.2.

A.4.3 Proof of Theorem A.3

Recall that we defined $M_t = (m_{ijt})$, where $m_{ijt} = \mathbf{u}_j^\top\mathbf{e}_{it}$, $t = 1,\ldots,T$. With a little abuse of notation, for $\Xi_{K_1}\in\mathcal{H}^{K_1}$ and $\Xi_{K_2}\in\mathcal{H}^{K_2}$, let

$$d(\Xi_{K_1},\Xi_{K_2}) = \max_{t:t=1,\ldots,T}\left\|M_{K_1,t}-M_{K_2,t}\right\|_F.$$

We further define the following functions for $\Xi_K\in\mathcal{H}^K$ and the corresponding $\mathbf{u}_{K,j}$s and $\boldsymbol{\theta}_{K,i}$s:

$$\begin{aligned} w_{ij}(\mathbf{u}_{K,j},\boldsymbol{\theta}_{K,i}) &= \boldsymbol{\rho}_{ij}(\mathbf{u}^*_j,\boldsymbol{\theta}^*_i) - \boldsymbol{\rho}_{ij}(\mathbf{u}_{K,j},\boldsymbol{\theta}_{K,i}),\\ l^*_{NJ}(\Xi_K) &= \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J} w_{ij}(\mathbf{u}_{K,j},\boldsymbol{\theta}_{K,i}),\\ \bar{l}^*_{NJ}(\Xi_K) &= \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J} E\left(w_{ij}(\mathbf{u}_{K,j},\boldsymbol{\theta}_{K,i})\right),\\ \mathbb{W}_{NJ}(\Xi_K) &= l^*_{NJ}(\Xi_K) - \bar{l}^*_{NJ}(\Xi_K) = \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\left(w_{ij}(\mathbf{u}_{K,j},\boldsymbol{\theta}_{K,i}) - E\left(w_{ij}(\mathbf{u}_{K,j},\boldsymbol{\theta}_{K,i})\right)\right).\end{aligned}$$

Following the proof of Bai and Ng (2002) and Chen et al. (2021), it suffices to show that for some $\delta>0$,

$$-l_{NJ}(\hat{\Xi}_K) + l_{NJ}(\hat{\Xi}_{K^*}) > \delta + o_P(1) \text{ for } K<K^*, \tag{A.42}$$

$$-l_{NJ}(\hat{\Xi}_K) + l_{NJ}(\hat{\Xi}_{K^*}) = O_P(\min\{N,J\}^{-1}) \text{ for } K>K^*. \tag{A.43}$$

Proof of (A.42) Suppose $K<K^*$. We can write

$$-l_{NJ}(\hat{\Xi}_K) + l_{NJ}(\hat{\Xi}_{K^*}) = l^*_{NJ}(\hat{\Xi}_K) - l^*_{NJ}(\hat{\Xi}_{K^*}) = \bar{l}^*_{NJ}(\hat{\Xi}_K) + \mathbb{W}_{NJ}(\hat{\Xi}_K) - l^*_{NJ}(\hat{\Xi}_{K^*}).$$

By the arguments in the proof of Lemma A.1, it can be shown that $|\mathbb{W}_{NJ}(\hat{\Xi}_K)| = o_P(1)$. Further, since $|\boldsymbol{\rho}_{ij}(\mathbf{u}_j^*, \boldsymbol{\theta}_i^*) - \boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i)| \lesssim \sum_{t=1}^T |\mathbf{u}_j^{*\top}\mathbf{e}_{it}^* - \mathbf{u}_j^\top \mathbf{e}_{it}|$, it follows from Lemma A.1 that

$$|l_{NJ}^*(\hat{\Xi}_{K^*})| \lesssim \frac{1}{NJ}\sum_{i=1}^N\sum_{j=1}^J\sum_{t=1}^T |(\mathbf{u}_j^*)^\top \mathbf{e}_{it}^* - (\hat{\mathbf{u}}_{K^*,j})^\top \hat{\mathbf{e}}_{K^*,it}| \leq T \cdot d(\Xi^*, \hat{\Xi}_{K^*}) = o_P(1).$$

Therefore, it remains to show that $\bar{l}_{NJ}^*(\hat{\Xi}_K) \geq \delta$. By the arguments in Lemma A.1, we can show that $\bar{l}_{NJ}^*(\hat{\Xi}_K) \gtrsim d^2(\hat{\Xi}_K, \Xi^*)$. Next, similar to the derivation of (A.12), we have

$$(NJ)^{-1/2}\|\hat{\Theta}_K(\hat{A}_K)^\top - \Theta^*(A^*)^\top\|_F \lesssim d(\hat{\Xi}_K, \Xi^*).$$

Then, by the arguments in Lemma 2 of Chen et al. (2021), we have

$$\left\|\left[I_J - \hat{A}_K\left\{(\hat{A}_K)^\top \hat{A}_K\right\}^{-1}(\hat{A}_K)^\top\right]A^*\right\|_F / \sqrt{J} \lesssim d(\hat{\Xi}_K, \Xi^*).$$

It then follows from the arguments of (A.34) to (A.37) in Chen et al. (2021) that (A.42) holds.

Proof of (A.43) Suppose $K > K^*$. we can write

$$-l_{NJ}(\hat{\Xi}_K) + l_{NJ}(\hat{\Xi}_{K^*}) = \mathbb{W}_{NJ}(\hat{\Xi}_K) - \mathbb{W}_{NJ}(\hat{\Xi}_{K^*}) + \bar{l}_{NJ}^*(\hat{\Xi}_K) - \bar{l}_{NJ}^*(\hat{\Xi}_{K^*}).$$

Similar to the proof of Lemma A.3 and Theorem A.1, we can show that for sufficiently small c and sufficiently large N and J, it holds that

$$E\left(\sup_{d(\Xi_K, \Xi^*)\leq c} |\mathbb{W}_{NJ}(\Xi_K)|\right) \lesssim \frac{c}{\min\left\{\sqrt{N}, \sqrt{J}\right\}}.$$

and $d(\hat{\Xi}_K, \Xi^*) = O_P(\min\{\sqrt{N}, \sqrt{J}\}^{-1})$. It then follows that

$$\mathbb{W}_{NJ}(\hat{\Xi}_K) = O_P(\min\{\sqrt{N}, \sqrt{J}\}^{-2}).$$

Similarly we can show that

$$\mathbb{W}_{NJ}(\hat{\Xi}_{K^*}) = O_P(\min\{\sqrt{N}, \sqrt{J}\}^{-2}).$$

Finally, consider $\bar{l}_{NJ}^*(\hat{\Xi}_K) - \bar{l}_{NJ}^*(\hat{\Xi}_{K^*})$. By Taylor expansion and the fact that $\exp(c)/(1+\exp(c))^2$ is bounded above for $c \in \mathbb{R}$, it is easy to show that

$$|\bar{l}_{NJ}^*(\hat{\Xi}_K)| \lesssim d^2(\hat{\Xi}_K, \Xi^*) \text{ and } |\bar{l}_{NJ}^*(\hat{\Xi}_{K^*})| \lesssim d^2(\hat{\Xi}_{K^*}, \Xi^*).$$

The rest of the proof follows from the remaining arguments in the proof of Theorem 3 in Chen et al. (2021).

A.4.4 Proof of Theorem A.4

Following the proof of Theorem A.1, it suffices to show the following three lemmas:

Lemma A.9. *Under Assumptions 2.1,2.2, 2.4 and A.1, $d(\hat{\Xi}, \Xi^*) = o_p(1)$ as $J, N \to \infty$.*

Proof: The proof largely follows the arguments in the proof of Lemma A.1. Therefore, we omit the full details, and instead highlight the key step that in this context, we can derive an upper bound similar to that in (A.6), given by

$$
\begin{aligned}
&\left|\boldsymbol{\rho}_{ij}(\mathbf{u}_j, \boldsymbol{\theta}_i) - \boldsymbol{\rho}_{ij}(\bar{\mathbf{u}}_j, \bar{\boldsymbol{\theta}}_i)\right| \\
\leq &\delta_6 \sum_{t=1}^{T} \Big(|\gamma_{jt} - \bar{\gamma}_{jt}| + |\boldsymbol{\beta}_{jt}^\top \mathbf{x}_i - \bar{\boldsymbol{\beta}}_{jt}^\top \mathbf{x}_i| + |\mathbf{v}_j^\top \mathbf{z}_{it} - \bar{\mathbf{v}}_j^\top \mathbf{z}_{it}| + |\mathbf{a}_{jt}{}^\top \boldsymbol{\theta}_i - \bar{\mathbf{a}}_{jt}^\top \bar{\boldsymbol{\theta}}_i| \Big) \\
\leq &\delta_6 T \bigg(\|\boldsymbol{\gamma}_j - \bar{\boldsymbol{\gamma}}_j\| + (\max_t \|\mathbf{a}_{jt}\|) \|\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}}_i\| + \|\bar{\boldsymbol{\theta}}_i\| (\max_t \|\mathbf{a}_{jt} - \bar{\mathbf{a}}_{jt}\|) + \|\mathbf{x}_i\| \max_t \|\boldsymbol{\beta}_{jt} - \bar{\boldsymbol{\beta}}_{jt}\| \\
&+ \|\mathbf{z}_{it}\| \|\mathbf{v}_j - \bar{\mathbf{v}}_j\| \bigg) \\
\leq &5\delta_6 T \epsilon.
\end{aligned}
$$

Lemma A.10. *Define $\mathcal{H}^{K^*}(c) = \left\{\Xi \in \mathcal{H}^{K^*} : d(\Xi, \Xi^*) \leq c\right\}$. Under Assumptions 2.1, 2.2, A.4, 2.4, A.1 and A.6, for sufficiently small $c > 0$ and sufficiently large N and J, for any $\Xi \in \mathcal{H}^{K^*}(c)$, it holds that*

$$
\|\Theta - \Theta^* S_A\|_F / \sqrt{N} + \|U_t - U_t^* S_U\|_F / \sqrt{J} \leq \delta_7 c, t = 1, \ldots, T,
$$

where $S_A = sgn\left(A_1^\top A_1^* / J\right)$ and S_U is a $(1 + p + p_z + K^*) \times (1 + p + p_z + K^*)$ diagonal matrix whose diagonal elements are 1, except for the last K^* diagonal elements which are equal to S_A.

Proof: Without loss of generality, assume that Assumption A.6 holds with $t_1 = 1$ and $t_2 = 2$. Recall that for $t = 1, \ldots, T$, we have

$$
\frac{1}{NJ} \|\Theta A_t^\top - \Theta^* A_t^{*\top} + X(B_t - B_t^*)^\top + \mathbf{1_N}(\mathit{\Gamma}_t - \mathit{\Gamma}_t^*)^\top + Z_t(V - V^*)^\top\|_F^2 \lesssim d^2(\Xi, \Xi^*).
$$

For $t = 1, 2$, project Θ on $(\Theta^*, X, \mathbf{1_N}, Z_t)$ and express it as $\Theta = \Theta^* M_{1t} + X M_{2t} + \mathbf{1_N} M_{3t} + Z_t M_{4t} + \mathbb{R}_{1t}$, where M_{1t}, M_{2t}, M_{3t} and M_{4t} are projection matrices of dimensions $K^* \times K^*$, $p \times K^*$, $1 \times K^*$ and $p_z \times K^*$, respectively. The term $\mathbb{R}_{1t}$ is an $N \times K$ matrix that is orthogonal to Θ^*, X, Z_t and $\mathbf{1_N}$. We have

$$
\begin{aligned}
&\frac{1}{NJ} \|(\Theta - \mathbb{R}_{1t}) A_t^\top - \Theta^* A_t^{*\top} + X(B_t - B_t^*)^\top + \mathbf{1_N}(\mathit{\Gamma}_t - \mathit{\Gamma}_t^*)^\top + Z_t(V - V^*)^\top\|_F^2 \\
+ &\frac{1}{NJ} \|\mathbb{R}_{1t} A_t^\top\|_F^2 \lesssim d^2(\Xi, \Xi^*). \qquad \text{(A.44)}
\end{aligned}
$$

Since $(J^{-1/2} \|A_t\|_F) N^{-1/2} \|\mathbb{R}_{1t}\|_F \leq (NJ)^{-1/2} \|\mathbb{R}_{1t} A_t^\top\|_F$, we can derive that

$$
\frac{1}{\sqrt{N}} \|\mathbb{R}_{1t}\|_F \lesssim d(\Xi, \Xi^*). \qquad \text{(A.45)}
$$

Combining (A.44) and (A.45), we obtain

$$\begin{aligned}&\frac{1}{\sqrt{NJ}}\|\Theta^*(M_{1t}A_t^\top - A_t^{*\top}) + X\{M_{2t}A_t^\top + (B_t - B_t^*)^\top\}\\ &\qquad + \mathbf{1_N}\{M_{3t}A_t^\top + (\mathit{\Gamma}_t - \mathit{\Gamma}_t^*)^\top\} + Z_t\{M_{4t}A_t^\top + (V - V^*)^\top\}\|_F \lesssim d(\Xi, \Xi^*).\end{aligned}$$

By Assumption A.6, we have $\pi_{\min}((\Theta^*, X, \mathbf{1_N}, Z_t)^\top(\Theta^*, X, \mathbf{1_N}, Z_t))/N \geq \kappa_7$ for $t = 1, 2$. Hence we have

$$\frac{1}{\sqrt{J}}\|M_{1t}A_t^\top - A_t^{*\top}\|_F \lesssim d(\Xi, \Xi^*) \text{ and } \frac{1}{\sqrt{J}}\|M_{4t}A_t^\top + (V - V^*)^\top\|_F \lesssim d(\Xi, \Xi^*) \quad \text{(A.46)}$$

for $t = 1, 2$. On the other hand, we have

$$\begin{aligned}&N^{-1/2}\|\Theta^*(M_{11} - M_{12}) + X(M_{21} - M_{22}) + \mathbf{1_N}(M_{31} - M_{32}) + Z_1M_{41} - Z_2M_{42}\|\\ =&N^{-1/2}\|\mathbb{R}_{11} - \mathbb{R}_{12}\| \lesssim d(\Xi, \Xi^*).\end{aligned}$$

Under Assumption A.6, we have

$$\begin{aligned}\|M_{4t}\| \lesssim d(\Xi, \Xi^*) &\text{ for } t = 1, 2 \text{ and}\\ \|M_{j1} - M_{j2}\| \lesssim d(\Xi, \Xi^*) &\text{ for } j = 1, 2, 3.\end{aligned} \quad \text{(A.47)}$$

Combining (A.46) and (A.47), we can show that

$$\frac{1}{\sqrt{J}}\|V - V^*\| \lesssim d(\Xi, \Xi^*). \quad \text{(A.48)}$$

Therefore, by (A.48), we have

$$\frac{1}{\sqrt{NJ}}\|\Theta A_t^\top - \Theta^* A_t^{*\top} + \mathbf{1_N}(\mathit{\Gamma}_t - \mathit{\Gamma}_t^*)^\top + X(B_t - B_t^*)^\top\|_F \lesssim d(\Xi, \Xi^*) \text{ for } t = 1, \ldots, T. \quad \text{(A.49)}$$

Since $\Theta^\top(\mathbf{1_N}, X) = 0_{K^*,1+p}$ by the normalisation condition, the rest of the proof follows from similar argument in the proof of Lemma A.2.

Remark A.6. *The above proof works for general $\mathit{\Gamma}_t^*$. When $\gamma_{jt} = \gamma_j$ such that $\mathit{\Gamma}_t^* = \mathit{\Gamma}^* = (\gamma_1^*, \ldots, \gamma_J^*)^\top$, we can still arrive at* (A.49)*. However, we no longer impose the normalisation constraint $\Theta^\top\mathbf{1_N} = \mathbf{0}_{K^*}$. To proceed, we express Θ as $\Theta = \Theta^*M_1 + XM_2 + \mathbf{1_N}M_3 + \mathbb{R}_1$ using* (A.47)*, where $N^{-1/2}\|\mathbb{R}_1\| \lesssim d(\Xi, \Xi^*)$. Similar to the derivation of* (A.46)*, we can show that*

$$\frac{1}{\sqrt{J}}\|M_1A_t^\top - A_t^{*\top}\| \lesssim d(\Xi, \Xi^*) \textit{ for } t = 1, 2.$$

Since A_t^* has rank K^*, M_1 must be invertible and thus we can show that

$$\begin{aligned}(NJ)^{-1/2}\|\mathbf{1_N}\{M_3M_1^{-1}A_t^{*\top} + t(\mathit{\Gamma} - \mathit{\Gamma}^*)^\top\}\|_F &\lesssim d(\Xi, \Xi^*)\\ (J)^{-1/2}\|M_3M_1^{-1}A_t^{*\top} + t(\mathit{\Gamma} - \mathit{\Gamma}^*)^\top\|_F &\lesssim d(\Xi, \Xi^*),\end{aligned}$$

where the second line follows from the additional condition in Assumption A.6. We write $\mathit{\Gamma} = m_{5t}\mathit{\Gamma}^* + A_t^*M_{6t} + r_{3t}$ for $t = 1, 2$, such that r_{3t} is orthogonal to $(\mathit{\Gamma}^*, A_t^*)$. We have

$$\begin{aligned}(J)^{-1/2}\|M_3M_1^{-1}A_t^{*\top} + t(m_{5t}\mathit{\Gamma}^* + A_t^*M_{6t} + r_{3t} - \mathit{\Gamma}^*)^\top\|_F &\lesssim d(\Xi, \Xi^*),\\ (J)^{-1/2}\|(M_3M_1^{-1} + M_{6t}^\top)A_t^{*\top} + t(m_{5t} - 1)\mathit{\Gamma}^{*\top}\|_F &\lesssim d(\Xi, \Xi^*).\end{aligned}$$

Hence we have $|m_{51} - 1| \lesssim d(\boldsymbol{\Xi}, \boldsymbol{\Xi}^*)$ *by Assumption A.6. On the other hand, we have*

$$J^{-1/2}\|(m_{52} - m_{51})\mathit{\Gamma}^* + A_1^* M_{61} + A_2^* M_{62}\|_F \lesssim d(\boldsymbol{\Xi}, \boldsymbol{\Xi}^*).$$

Hence we have $\|M_{61}\| \lesssim d(\boldsymbol{\Xi}, \boldsymbol{\Xi}^*)$ *and thus* $J^{-1/2}\|\mathit{\Gamma} - \mathit{\Gamma}^*\| \lesssim d(\boldsymbol{\Xi}, \boldsymbol{\Xi}^*)$*. We can then obtain*

$$\frac{1}{\sqrt{NJ}}\|\Theta A_t^\top - \Theta^* A_t^{*\top} + X(B_t - B_t^*)^\top\|_F \lesssim d(\boldsymbol{\Xi}, \boldsymbol{\Xi}^*) \text{ for } t = 1, \ldots, T,$$

and the rest follows from the argument in the proof of Lemma 2 in Chen et al. (2021).

Lemma A.11. *Under Assumptions 2.1, 2.2, A.4, 2.4, A.1 and A.6, for sufficiently small c and sufficiently large N and J, it holds that*

$$E\left(\sup_{\boldsymbol{\Xi} \in \mathcal{H}^{K^*}(c)} |\mathbb{W}_{NJ}(\boldsymbol{\Xi})|\right) \lesssim \frac{c}{\min\left\{\sqrt{N}, \sqrt{J}\right\}}.$$

The proof of Lemma A.11 follows directly from the argument in the proof of Lemma A.3. We omit the detailed proof here.

A.4.5 Proof of Theorem A.5

The proof of this theorem largely follows from the arguments in the proof of Theorem A.2. Therefore, we only present the proof of a key lemma essential to the derivation of the stochastic expansion of $\hat{\boldsymbol{\theta}}_i$. In the context of this extension, define

$$\mathit{\Phi}_{N,j} = \frac{1}{N}\sum_{i=1}^N \sum_{t=1}^T E\left(\varrho_{ijt}^{''}\right)\mathbf{e}_{it}^*\mathbf{e}_{it}^{*\top}, \quad \mathit{\Psi}_{J,i} = \frac{1}{J}\sum_{j=1}^J \sum_{t=1}^T E\left(\varrho_{ijt}^{''}\right)\mathbf{a}_{jt}^*\mathbf{a}_{jt}^{*\top} \text{ and}$$

$$\begin{aligned}\mathbb{P}_{NJ}(\boldsymbol{\Xi}) = T\delta\Bigg\{&\frac{1}{2J}\sum_{l=1}^{K^*}\sum_{q>l}^{K^*}\Bigg(\sum_{j=1}^J a_{jl1}a_{jq1}\Bigg)^2 + \frac{1}{2N}\sum_{l=1}^{K^*}\sum_{q>l}^{K^*}\Bigg(\sum_{i=1}^N \theta_{il}\theta_{iq}\Bigg)^2 + \frac{1}{8J}\sum_{t=1}^T\sum_{k=1}^{K^*}\Bigg(\sum_{j=1}^J a_{jk1}^2 - J\Bigg)^2 \\ &+ \frac{1}{2N}\sum_{l=1}^{K^*}\Bigg(\sum_{i=1}^N \theta_{il}\Bigg)^2 + \frac{1}{2N}\sum_{k=1}^{K^*}\sum_{l=1}^{p}\Bigg(\sum_{i=1}^N \theta_{ik}x_{il}\Bigg)\Bigg\}\end{aligned}$$

for some $\delta > 0$. We further define

$$\begin{aligned}&S^*(\boldsymbol{\Xi}) \\ &= \left(\underbrace{\ldots, \frac{1}{\sqrt{NJ}}\sum_{i=1}^N\sum_{t=1}^T E\left(\varrho_{ijt}^{'}(\mathbf{u}_j^\top \mathbf{e}_{it})\right)\mathbf{e}_{it}^\top, \ldots}_{1\times JP}, \underbrace{\ldots, \frac{1}{\sqrt{NJ}}\sum_{j=1}^J\sum_{t=1}^T E\left(\varrho_{ijt}^{'}(\mathbf{u}_j^\top \mathbf{e}_{it})\right)\mathbf{a}_{jt}^\top, \ldots}_{1\times NK^*}\right)^\top,\end{aligned}$$

$$S(\boldsymbol{\Xi}) = S^*(\boldsymbol{\Xi}) + \partial\mathbb{P}_{NJ}(\boldsymbol{\Xi})/\partial\boldsymbol{\Xi}, \qquad \mathbb{H}(\boldsymbol{\Xi}) = \partial S^*(\boldsymbol{\Xi})/\partial\boldsymbol{\Xi}^\top + \partial\mathbb{P}_{NJ}(\boldsymbol{\Xi})/\partial\boldsymbol{\Xi}\partial\boldsymbol{\Xi}^\top$$

and let $\mathbb{H} = \mathbb{H}(\boldsymbol{\Xi}^*)$. Further, define

$$\mathbb{H}_d = \begin{pmatrix}\mathbb{H}_d^{\mathcal{U}} & 0 \\ 0 & \mathbb{H}_d^{\Theta}\end{pmatrix}, \mathbb{H}_d^{\mathcal{U}} = \frac{\sqrt{N}}{\sqrt{J}}\text{diag}\left(\mathit{\Phi}_{N,1}, \ldots, \mathit{\Phi}_{N,J}\right), \mathbb{H}_d^{\Theta} = \frac{\sqrt{J}}{\sqrt{N}}\text{diag}\left(\mathit{\Psi}_{J,1}, \ldots, \mathit{\Psi}_{J,N}\right).$$

We have the following lemma:

Lemma A.12. *Under Assumptions 2.1, 2.2, A.4, 2.4, 2.6, 2.7, A.1, A.5, A.6 and the additional condition that* $\sum_{i=1}^N \mathbf{x}_i = \mathbf{0}_p$ *and* $\sum_{i=1}^N x_{ik}x_{il} = 0$ *for* $l, k \in \{1, \dots, p\}, l \neq k$, *the matrix* $\mathbb{H}$ *is invertible and* $\left\|\mathbb{H}^{-1} - \mathbb{H}_d^{-1}\right\|_{\max} = O(1/N)$.

Proof: We assume $p = T = K^* = 2$ for simplicity, which can be generalised easily. Consider

$$\mathbb{P}_{NJ}(\boldsymbol{\Xi}) = 2\delta \left\{ \frac{1}{2N}\left(\sum_{i=1}^N \theta_{i1}\theta_{i2}\right)^2 + \frac{1}{2J}\left(\sum_{j=1}^J a_{j11}a_{j21}\right)^2 + \frac{1}{8J}\left(\sum_{j=1}^J a_{j11}^2 - J\right)^2 + \frac{1}{8J}\left(\sum_{j=1}^J a_{j21}^2 - J\right)^2 \right.$$
$$\left. + \frac{1}{2N}\left(\sum_{i=1}^N \theta_{i1}\right)^2 + \frac{1}{2N}\left(\sum_{i=1}^N \theta_{i2}\right)^2 + \frac{1}{2N}\sum_{k=1}^2\sum_{p=1}^2\left(\sum_{i=1}^N \theta_{ik}x_{ip}\right)^2 \right\}.$$

We further define

$$\begin{aligned}
\boldsymbol{\mu}_1 &= ((\mathbf{0}_{6+p_z}^\top, a_{111}^*, 0, \mathbf{0}_2^\top), \dots, (\mathbf{0}_{6+p_z}^\top, a_{J11}^*, 0, \mathbf{0}_2^\top), \mathbf{0}_{2N}^\top)^\top/\sqrt{J},\\
\boldsymbol{\mu}_2 &= ((\mathbf{0}_{6+p_z}^\top, 0, a_{121}^*, \mathbf{0}_2^\top), \dots, (\mathbf{0}_{6+p_z}^\top, 0, a_{J21}^*, \mathbf{0}_2^\top), \mathbf{0}_{2N}^\top)^\top/\sqrt{J},\\
\boldsymbol{\mu}_3 &= ((\mathbf{0}_{6+p_z}^\top, a_{121}^*, a_{111}^*, \mathbf{0}_2^\top), \dots, (\mathbf{0}_{6+p_z}^\top, a_{J21}^*, a_{J11}^*, \mathbf{0}_2^\top), \mathbf{0}_{2N}^\top)^\top/\sqrt{J},\\
\boldsymbol{\mu}_4 &= (\mathbf{0}_{PJ}^\top, (\theta_{12}^*, \theta_{11}^*), \dots, (\theta_{N2}^*, \theta_{N1}^*))^\top/\sqrt{N}\\
\boldsymbol{\mu}_5 &= (\mathbf{0}_{PJ}^\top, (1, 0), \dots, (1, 0))^\top/\sqrt{N},\\
\boldsymbol{\mu}_6 &= (\mathbf{0}_{PJ}^\top, (0, 1), \dots, (0, 1))^\top/\sqrt{N},\\
\boldsymbol{\mu}_7 &= (\mathbf{0}_{PJ}^\top, (x_{11}, 0), \dots, (x_{N1}, 0))^\top/\sqrt{N},\\
\boldsymbol{\mu}_8 &= (\mathbf{0}_{PJ}^\top, (0, x_{11}), \dots, (0, x_{N1}))^\top/\sqrt{N},\\
\boldsymbol{\mu}_9 &= (\mathbf{0}_{PJ}^\top, (x_{12}, 0), \dots, (x_{N2}, 0))^\top/\sqrt{N},\\
\boldsymbol{\mu}_{10} &= (\mathbf{0}_{PJ}^\top, (0, x_{12}), \dots, (0, x_{N2}))^\top/\sqrt{N},
\end{aligned}$$

such that $\partial\mathbb{P}_{NJ}(\boldsymbol{\Xi}^*)/\partial\boldsymbol{\Xi}\partial\boldsymbol{\Xi}^\top = 2\delta\left(\sum_{m=1}^{10}\boldsymbol{\mu}_m\boldsymbol{\mu}_m^\top\right)$. At $t = 2$, there exists a $K^* \times K^*$ matrix $Q_2 = (q_{ij2})$ such that $\Theta A_2^\top = \Theta Q_2^{-1}Q_2A_2^\top$, where

$$(\Theta Q_2^{-1})^\top\Theta Q_2^{-1}/N \text{ is diagonal and } (A_2Q_2^\top)^\top(A_2Q_2^\top)/J = I_{K^*}.$$

Define $\breve{\Theta}_2 = \Theta Q_2^{-1}$ and $\breve{A}_2 = A_2 Q_2^\top$. At $t = 1$, define

$$\boldsymbol{\omega}_{1,1} = \left(\underbrace{(\mathbf{0}_{6+p_z}^\top, a_{111}^*/\sqrt{J}, 0, \mathbf{0}_2^\top), \ldots, (\mathbf{0}_{6+p_z}^\top, a_{J11}^*/\sqrt{J}, 0, \mathbf{0}_2^\top)}_{\boldsymbol{\omega}_{1\mathcal{U},1}^\top}, \underbrace{(-\theta_{11}^*/\sqrt{N}, 0), \ldots, (-\theta_{N1}^*/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{1\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{2,1} = \left(\underbrace{(\mathbf{0}_{6+p_z}^\top, 0, a_{121}^*/\sqrt{J}, \mathbf{0}_2^\top), \ldots, (\mathbf{0}_{6+p_z}^\top, 0, a_{J21}^*/\sqrt{J}, \mathbf{0}_2^\top)}_{\boldsymbol{\omega}_{2\mathcal{U},1}^\top}, \underbrace{(0, -\theta_{12}^*/\sqrt{N}), \ldots, (0, -\theta_{N2}^*/\sqrt{N})}_{\boldsymbol{\omega}_{2\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{3,1} = \left(\underbrace{(\mathbf{0}_{6+p_z}^\top, a_{121}^*/\sqrt{J}, 0, \mathbf{0}_2^\top), \ldots, (\mathbf{0}_{6+p_z}^\top, a_{J21}^*/\sqrt{J}, 0, \mathbf{0}_2^\top)}_{\boldsymbol{\omega}_{3\mathcal{U},1}^\top}, \underbrace{(0, -\theta_{11}^*/\sqrt{N}), \ldots, (0, -\theta_{N1}^*/\sqrt{N})}_{\boldsymbol{\omega}_{3\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{4,1} = \left(\underbrace{(\mathbf{0}_{6+p_z}^\top, 0, a_{111}^*/\sqrt{J}, \mathbf{0}_2^\top), \ldots, (\mathbf{0}_{6+p_z}^\top, 0, a_{J11}^*/\sqrt{J}, \mathbf{0}_2^\top)}_{\boldsymbol{\omega}_{4\mathcal{U},1}^\top}, \underbrace{(-\theta_{12}^*/\sqrt{N}, 0), \ldots, (-\theta_{N2}^*/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{4\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{5,1} = \left(\underbrace{(a_{111}^*/\sqrt{J}, \mathbf{0}_{9+p_z}^\top), \ldots, (a_{J11}^*/\sqrt{J}, \mathbf{0}_{9+p_z}^\top)}_{\boldsymbol{\omega}_{5\mathcal{U},1}^\top}, \underbrace{(-1/\sqrt{N}, 0), \ldots, (-1/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{5\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{6,1} = \left(\underbrace{(a_{121}^*/\sqrt{J}, \mathbf{0}_{9+p_z}^\top), \ldots, (a_{J21}^*/\sqrt{J}, \mathbf{0}_{9+p_z}^\top)}_{\boldsymbol{\omega}_{6\mathcal{U},1}^\top}, \underbrace{(0, -1/\sqrt{N}), \ldots, (0, -1/\sqrt{N})}_{\boldsymbol{\omega}_{6\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{7,1} = \left(\underbrace{(0, 0, a_{111}^*/\sqrt{J}, \mathbf{0}_{7+p_z}^\top), \ldots, (0, 0, a_{J11}^*/\sqrt{J}, \mathbf{0}_{7+p_z}^\top)}_{\boldsymbol{\omega}_{7\mathcal{U},1}^\top}, \underbrace{(-x_{11}/\sqrt{N}, 0), \ldots, (-x_{N1}/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{7\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{8,1} = \left(\underbrace{(0, 0, a_{121}^*/\sqrt{J}, \mathbf{0}_{7+p_z}^\top), \ldots, (0, 0, a_{J21}^*/\sqrt{J}, \mathbf{0}_{7+p_z}^\top)}_{\boldsymbol{\omega}_{8\mathcal{U},1}^\top}, \underbrace{(0, -x_{11}/\sqrt{N}), \ldots, (0, -x_{N1}/\sqrt{N})}_{\boldsymbol{\omega}_{8\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{9,1} = \left(\underbrace{(0, 0, 0, a_{111}^*/\sqrt{J}, \mathbf{0}_{6+p_z}^\top), \ldots, (0, 0, 0, a_{J11}^*/\sqrt{J}, \mathbf{0}_{6+p_z}^\top)}_{\boldsymbol{\omega}_{9\mathcal{U},1}^\top}, \underbrace{(-x_{12}/\sqrt{N}, 0), \ldots, (-x_{N2}/\sqrt{N}, 0)}_{\boldsymbol{\omega}_{9\Theta,1}^\top} \right)^\top,$$

$$\boldsymbol{\omega}_{10,1} = \left(\underbrace{(0, 0, 0, a_{121}^*/\sqrt{J}, \mathbf{0}_{6+p_z}^\top), \ldots, (0, 0, 0, a_{J21}^*/\sqrt{J}, \mathbf{0}_{6+p_z}^\top)}_{\boldsymbol{\omega}_{10\mathcal{U},1}^\top}, \underbrace{(0, -x_{12}/\sqrt{N}), \ldots, (0, -x_{N2}/\sqrt{N})}_{\boldsymbol{\omega}_{10\Theta,1}^\top} \right)^\top.$$

On the other hand, at $t = 2$, we define

$$\boldsymbol{\omega}_{5,2} = \left(\underbrace{(0, \breve{a}^*_{112}/\sqrt{J}, \mathbf{0}^\top_{8+p_z}), \ldots, (0, \breve{a}^*_{J12}/\sqrt{J}, \mathbf{0}^\top_{8+p_z})}_{\boldsymbol{\omega}^\top_{5\mathcal{U},2}}, \underbrace{(-1/\sqrt{N}, 0), \ldots, (-1/\sqrt{N}, 0)}_{\boldsymbol{\omega}^\top_{5\Theta,2}} \right)^\top,$$

$$\boldsymbol{\omega}_{6,2} = \left(\underbrace{(0, \breve{a}^*_{122}/\sqrt{J}, \mathbf{0}^\top_{8+p_z}), \ldots, (0, \breve{a}^*_{J22}/\sqrt{J}, \mathbf{0}^\top_{8+p_z})}_{\boldsymbol{\omega}^\top_{6\mathcal{U},2}}, \underbrace{(0, -1/\sqrt{N}), \ldots, (0, -1/\sqrt{N})}_{\boldsymbol{\omega}^\top_{6\Theta,2}} \right)^\top,$$

$$\boldsymbol{\omega}_{7,2} = \left(\underbrace{(\mathbf{0}^\top_4, \breve{a}^*_{112}/\sqrt{J}, \mathbf{0}^\top_{5+p_z}), \ldots, (\mathbf{0}^\top_4, \breve{a}^*_{J12}/\sqrt{J}, \mathbf{0}^\top_{5+p_z})}_{\boldsymbol{\omega}^\top_{7\mathcal{U},2}}, \underbrace{(-x_{11}/\sqrt{N}, 0), \ldots, (-x_{N1}/\sqrt{N}, 0)}_{\boldsymbol{\omega}^\top_{7\Theta,2}} \right)^\top,$$

$$\boldsymbol{\omega}_{8,2} = \left(\underbrace{(\mathbf{0}^\top_4, \breve{a}^*_{122}/\sqrt{J}, \mathbf{0}^\top_{5+p_z}), \ldots, (\mathbf{0}^\top_4, \breve{a}^*_{J22}/\sqrt{J}, \mathbf{0}^\top_{5+p_z})}_{\boldsymbol{\omega}^\top_{8\mathcal{U},2}}, \underbrace{(0, -x_{11}/\sqrt{N}), \ldots, (0, -x_{N1}/\sqrt{N})}_{\boldsymbol{\omega}^\top_{8\Theta,2}} \right)^\top,$$

$$\boldsymbol{\omega}_{9,2} = \left(\underbrace{(\mathbf{0}^\top_5, \breve{a}^*_{112}/\sqrt{J}, \mathbf{0}^\top_{4+p_z}), \ldots, (\mathbf{0}^\top_5, \breve{a}^*_{J12}/\sqrt{J}, \mathbf{0}^\top_{4+p_z})}_{\boldsymbol{\omega}^\top_{9\mathcal{U},2}}, \underbrace{(-x_{12}/\sqrt{N}, 0), \ldots, (-x_{N2}/\sqrt{N}, 0)}_{\boldsymbol{\omega}^\top_{9\Theta,2}} \right)^\top,$$

$$\boldsymbol{\omega}_{10,2} = \left(\underbrace{(\mathbf{0}^\top_5, \breve{a}^*_{122}/\sqrt{J}, \mathbf{0}^\top_{4+p_z}), \ldots, (\mathbf{0}^\top_5, \breve{a}^*_{J22}/\sqrt{J}, \mathbf{0}^\top_{4+p_z})}_{\boldsymbol{\omega}^\top_{10\mathcal{U},2}}, \underbrace{(0, -x_{12}/\sqrt{N}), \ldots, (0, -x_{N2}/\sqrt{N})}_{\boldsymbol{\omega}^\top_{10\Theta,2}} \right)^\top.$$

We define $W_1 = (\boldsymbol{\omega}_{1,1}, \boldsymbol{\omega}_{2,1}, \ldots, \boldsymbol{\omega}_{10,1})$ and $W_2 = (\boldsymbol{\omega}_{5,2}, \boldsymbol{\omega}_{6,2}, \ldots, \boldsymbol{\omega}_{10,2})$. It is easy to check that $\boldsymbol{\omega}^\top_{p,t}\boldsymbol{\omega}_{q,t} = 0$ for $p \neq q$, $t = 1, 2$. We can verify that

$$\begin{aligned} W_1 W_1^\top &= \sum_{k=1}^{10} \boldsymbol{\omega}_{k,1}\boldsymbol{\omega}^\top_{k,1} \\ &= \begin{pmatrix} \sum_{k=1}^{10} \boldsymbol{\omega}_{k\mathcal{U},1}\boldsymbol{\omega}^\top_{k\mathcal{U},1} & -(NJ)^{-1/2}\left\{ \begin{pmatrix} \mathbf{D}_{i1} \\ \mathbf{x}_i \\ \mathbf{0}_{2+p_z} \\ \boldsymbol{\theta}^*_i \\ \mathbf{0}_2 \end{pmatrix} \mathbf{a}^{*}_{j1}{}^\top \right\}_{j\leq J, i\leq N} \\ -(NJ)^{-1/2}\left\{ \mathbf{a}^*_{j1}\left(\mathbf{D}^\top_{i1}, \mathbf{x}^\top_i, \mathbf{0}^\top_{2+p_z}, \boldsymbol{\theta}^{*}_i{}^\top, \mathbf{0}^\top_2 \right) \right\}_{i\leq N, j\leq J} & \sum_{k=1}^{10} \boldsymbol{\omega}_{k\Theta,1}\boldsymbol{\omega}^\top_{k\Theta,1} \end{pmatrix}, \end{aligned} \tag{A.50}$$

$$
\begin{aligned}
W_2W_2^\top &= \sum_{k=5}^{10} \boldsymbol{\omega}_{k,2}\boldsymbol{\omega}_{k,2}^\top \\
&= \begin{pmatrix} \sum_{k=5}^{10} \boldsymbol{\omega}_{k\mathcal{U},2}\boldsymbol{\omega}_{k\mathcal{U},2}^\top & -(NJ)^{-1/2}\left\{\begin{pmatrix}\mathbf{D}_{i2}\\ \mathbf{0}_2\\ \mathbf{x}_i\\ \mathbf{0}_{p_z+4}\end{pmatrix}\breve{\mathbf{a}}_{j2}^{*\top}\right\}_{j\le J,i\le N} \\ -(NJ)^{-1/2}\left\{\breve{\mathbf{a}}_{j2}^{*}\left(\mathbf{D}_{i2}^\top, \mathbf{0}_2^\top, \mathbf{x}_i^\top, \mathbf{0}_{p_z+4}^\top\right)\right\}_{i\le N,j\le J} & \sum_{k=5}^{10} \boldsymbol{\omega}_{k\Theta,2}\boldsymbol{\omega}_{k\Theta,2}^\top \end{pmatrix} \\
&= \begin{pmatrix} \sum_{k=5}^{10} \boldsymbol{\omega}_{k\mathcal{U},2}\boldsymbol{\omega}_{k\mathcal{U},2}^\top & -(NJ)^{-1/2}\left\{\begin{pmatrix}\mathbf{D}_{i2}\\ \mathbf{0}_2\\ \mathbf{x}_i\\ \mathbf{0}_{p_z+4}\end{pmatrix}\mathbf{a}_{j2}^{*}{}^\top Q_2^\top\right\}_{j\le J,i\le N} \\ -(NJ)^{-1/2}\left\{Q_2\mathbf{a}_{j2}^{*}\left(\mathbf{D}_{i2}^\top, \mathbf{0}_2^\top, \mathbf{x}_i^\top, \mathbf{0}_{p_z+4}^\top\right)\right\}_{i\le N,j\le J} & \sum_{k=5}^{10} \boldsymbol{\omega}_{k\Theta,2}\boldsymbol{\omega}_{k\Theta,2}^\top \end{pmatrix}.
\end{aligned}
\tag{A.51}
$$

By writing

$$
\breve{W}_2 = \begin{pmatrix} I_{PJ} & 0 \\ 0 & \operatorname{diag}(\{Q_2^{-1}\}_{i\le N}) \end{pmatrix} W_2, \text{ and}
$$
$$
\breve{\boldsymbol{\omega}}_{k\Theta,2} = \operatorname{diag}(\{Q_2^{-1}\}_{i\le N})\boldsymbol{\omega}_{k\Theta,2} \text{ for } k = 5, 6, \dots, 10,
$$

we can express $\breve{W}_2\breve{W}_2^\top$ as

$$
\begin{aligned}
&\breve{W}_2\breve{W}_2^\top \\
&= \begin{pmatrix} \sum_{k=5}^{10} \boldsymbol{\omega}_{k\mathcal{U},2}\boldsymbol{\omega}_{k\mathcal{U},2}^\top & -(NJ)^{-1/2}\left\{\begin{pmatrix}\mathbf{D}_{i2}\\ \mathbf{0}_2\\ \mathbf{x}_i\\ \mathbf{0}_{p_z+4}\end{pmatrix}\mathbf{a}_{j2}^{*}{}^\top\right\}_{j\le J,i\le N} \\ -(NJ)^{-1/2}\left\{\mathbf{a}_{j2}^{*}\left(\mathbf{D}_{i2}^\top, \mathbf{0}_2^\top, \mathbf{x}_i^\top, \mathbf{0}_{p_z+4}^\top\right)\right\}_{i\le N,j\le J} & \sum_{k=5}^{10} \breve{\boldsymbol{\omega}}_{k\Theta,2}\breve{\boldsymbol{\omega}}_{k\Theta,2}^\top \end{pmatrix}.
\end{aligned}
$$

Further, it is easy to see that under our normalisation criteria,

$$
\begin{aligned}
W_1^\top W_1 = \operatorname{diag}\Big(&\sigma_{N1}+1, \sigma_{N2}+1, \sigma_{N1}+1, \sigma_{N2}+1, 2, 2 \\
&1 + N^{-1}\sum_{i=1}^N x_{i1}^2, 1 + N^{-1}\sum_{i=1}^N x_{i1}^2, 1 + N^{-1}\sum_{i=1}^N x_{i2}^2, 1 + N^{-1}\sum_{i=1}^N x_{i2}^2\Big).
\end{aligned}
$$

Moreover, we have

$$
W_2^\top W_2 = \operatorname{diag}\Big(2, 2, 1 + N^{-1}\sum_{i=1}^N x_{i1}^2, 1 + N^{-1}\sum_{i=1}^N x_{i1}^2, 1 + N^{-1}\sum_{i=1}^N x_{i2}^2, 1 + N^{-1}\sum_{i=1}^N x_{i2}^2\Big).
$$

At $t = 1$, for $k = 1, \dots, 10$, we project $\boldsymbol{\mu}_k$ onto W_1 and write $\boldsymbol{\mu}_k = W_1\mathbf{s}_{k,1} + \boldsymbol{\zeta}_{k,1}$ for $k = 1, \dots 10$, where $\mathbf{s}_{k,1} = (W_1^\top W_1)^{-1}W_1^\top\boldsymbol{\mu}_k$. At $t = 2$, we write $\boldsymbol{\mu}_k = W_2\mathbf{s}_{k,2} + \boldsymbol{\zeta}_{k,2}$ for

$k=5,6,\dots,10$, where $\mathbf{s}_{k,2}=(W_2^\top W_2)^{-1}W_2^\top\boldsymbol{\mu}_k$. For example, at $t=2$, we have

$$\mathbf{s}_{5,2}=\begin{pmatrix}-0.5\\0\\0\\0\\0\\0\end{pmatrix},\mathbf{s}_{6,2}=\begin{pmatrix}0\\-0.5\\0\\0\\0\\0\end{pmatrix}\mathbf{s}_{7,2}=\begin{pmatrix}0\\0\\\frac{-N^{-1}\sum_{i=1}^N x_{i1}^2}{1+N^{-1}\sum_{i=1}^N x_{i1}^2}\\0\\0\\0\end{pmatrix},$$

$$\mathbf{s}_{8,2}=\begin{pmatrix}0\\0\\0\\\frac{-N^{-1}\sum_{i=1}^N x_{i1}^2}{1+N^{-1}\sum_{i=1}^N x_{i1}^2}\\0\\0\end{pmatrix},\mathbf{s}_{9,2}=\begin{pmatrix}0\\0\\0\\0\\\frac{-N^{-1}\sum_{i=1}^N x_{i2}^2}{1+N^{-1}\sum_{i=1}^N x_{i2}^2}\\0\end{pmatrix},\mathbf{s}_{10,2}=\begin{pmatrix}0\\0\\0\\0\\0\\\frac{-N^{-1}\sum_{i=1}^N x_{i2}^2}{1+N^{-1}\sum_{i=1}^N x_{i2}^2}\end{pmatrix}.$$

We now define $\mathcal{S}_{N,1}=\sum_{k=1}^{10}\mathbf{s}_{k,1}\mathbf{s}_{k,1}^\top$ and $\mathcal{S}_{N,2}=\sum_{k=5}^{10}\mathbf{s}_{k,2}\mathbf{s}_{k,2}^\top$. We set $\boldsymbol{\zeta}_{k,2}=\boldsymbol{\mu}_k$ for $k=1,\dots,4$. Recall that we have shown in the proof of Lemma A.6 that there exists $\underline{\pi}$ such that $\pi_{\min}(\mathcal{S}_{N,1})>\underline{\pi}$ for all large N. It is easy to see that $\pi_{\min}(\mathcal{S}_{N,2})>\underline{\pi}$. We further Let

$$\bar{\pi}=\pi_{\max}\left(\begin{pmatrix}I_{PJ}&0\\0&\operatorname{diag}(\{Q_2^{-1}\}_{i\le N})\end{pmatrix}^\top\begin{pmatrix}I_{PJ}&0\\0&\operatorname{diag}(\{Q_2^{-1}\}_{i\le N})\end{pmatrix}\right)=\max(1,\pi_{\max}(Q_2^{-1\top}Q_2^{-1})).$$

We can verify that $\bar{\pi}W_2W_2^\top-\breve{W}_2\breve{W}_2^\top\ge 0_{PJ+2N,PJ+2N}$ and thus we have

$$\begin{aligned}\delta\left\{W_2\mathcal{S}_{N,2}W_2^\top-\bar{\pi}^{-1}\underline{\pi}\breve{W}_2\breve{W}_2^\top\right\}=&\delta\{W_2\mathcal{S}_{N,2}W_2^\top-\underline{\pi}W_2W_2^\top\}+\delta\underline{\pi}\{W_2W_2^\top-\bar{\pi}^{-1}\breve{W}_2\breve{W}_2^\top\}\\\ge&0_{PJ+2N,PJ+2N}.\end{aligned}$$

It then follows that

$$\begin{aligned}\partial\mathbb{P}_{NJ}(\Xi^*)/\partial\Xi\partial\Xi^\top&=2\delta\left(\sum_{k=1}^{10}\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top\right)\\&=\delta\left\{W_1\left(\sum_{k=1}^{10}\mathbf{s}_k\mathbf{s}_k^\top\right)W_1^\top+W_2\left(\sum_{k=5}^{10}\mathbf{s}_k\mathbf{s}_k^\top\right)W_2^\top\right\}+\delta\sum_{l=1}^{2}\left(\sum_{k=1}^{10}\boldsymbol{\zeta}_{k,l}\boldsymbol{\zeta}_{k,l}^\top\right)\\&=\delta\underline{\pi}W_1W_1^\top+\delta\bar{\pi}^{-1}\underline{\pi}\breve{W}_2\breve{W}_2^\top+\delta\left\{W_1\left(\mathcal{S}_{N,1}-\underline{\pi}I_{10}\right)W_1^\top\right\}\\&\quad+\delta\left\{W_2\mathcal{S}_{N,2}W_2^\top-\bar{\pi}^{-1}\underline{\pi}\breve{W}_2\breve{W}_2^\top\right\}+\delta\sum_{l=1}^{2}\left(\sum_{k=1}^{10}\boldsymbol{\zeta}_{k,l}\boldsymbol{\zeta}_{k,l}^\top\right).\qquad\text{(A.52)}\end{aligned}$$

Recall that there exists $\kappa_9>0$ such that $E(\varrho_{ijt}^{''}(\mathbf{u}_j^\top\mathbf{e}_{it}))>\kappa_9$ by Assumptions 2.1, 2.2 and 2.4. Now let $\underline{\delta}=\min\{\kappa_9,\delta\bar{\pi}^{-1}\underline{\pi}\}$. Then it follows from (A.52) that

$$\begin{aligned}\mathbb{H}&=\partial S^*(\Xi^*)/\partial\Xi^\top+\partial\mathbb{P}_{NJ}(\Xi^*)/\partial\Xi\partial\Xi^\top\\&\ge\partial S^*(\Xi^*)/\partial\Xi^\top+\underline{\delta}\sum_{l=1}^{2}W_lW_l^\top.\qquad\text{(A.53)}\end{aligned}$$

Moreover, we can write

$$
\begin{aligned}
&(NJ)^{1/2}\partial S^*(\Xi^*)/\partial \Xi^\top \\
&= \begin{pmatrix} \operatorname{diag}\left(\left\{\sum_{i=1}^N \sum_{t=1}^T E\left(\varrho''_{ijt}\right) \mathbf{e}^*_{it}\mathbf{e}^{*\top}_{it}\right\}_{j\le J}\right) & \left\{\sum_{t=1}^T E\left(\varrho''_{ijt}\right) \mathbf{e}^*_{it}\mathbf{a}^{*\top}_{jt}\right\}_{j\le J, i\le N} \\ \left\{\sum_{t=1}^T E\left(\varrho''_{ijt}\right) \mathbf{a}^*_{jt}\mathbf{e}^{*\top}_{it}\right\}_{i\le N, j\le J} & \operatorname{diag}\left(\left\{\sum_{j=1}^J \sum_{t=1}^T E\left(\varrho''_{ijt}\right) \mathbf{a}^*_{jt}\mathbf{a}^{*\top}_{jt}\right\}_{i\le N}\right) \end{pmatrix} \\
&= \underbrace{\underline{\delta} \begin{pmatrix} \operatorname{diag}\left(\left\{\sum_{i=1}^N \sum_{t=1}^T \mathbf{e}^*_{it}\mathbf{e}^{*\top}_{it}\right\}_{j\le J}\right) & \left\{\sum_{t=1}^T \begin{pmatrix} \mathbf{0}_6 \\ \mathbf{z}_{it} \\ \mathbf{0}_2 \\ D_{it2}\boldsymbol{\theta}^*_i \end{pmatrix} \mathbf{a}^{*\top}_{jt}\right\}_{j\le J, i\le N} \\ \left\{\sum_{t=1}^T \mathbf{a}^*_{jt}\left(\mathbf{0}_6^\top, \mathbf{z}_{it}^\top, \mathbf{0}_2^\top, D_{it2}\boldsymbol{\theta}^{*\top}_i\right)\right\}_{i\le N, j\le J} & \operatorname{diag}\left(\left\{\sum_{j=1}^J \sum_{t=1}^T \mathbf{a}^*_{jt}\mathbf{a}^{*\top}_{jt}\right\}_{i\le N}\right) \end{pmatrix}}_{(NJ)^{1/2} I} \\
&\quad + \underbrace{\sum_{t=1}^T \underline{\delta} \begin{pmatrix} 0_{PJ\times PJ} & \left\{\begin{pmatrix} \mathbf{D}_{it} \\ D_{it1}\mathbf{x}_i \\ D_{it2}\mathbf{x}_i \\ \mathbf{0}_{p_z} \\ D_{it1}\boldsymbol{\theta}^*_i \\ \mathbf{0}_2 \end{pmatrix} \mathbf{a}^{*\top}_{jt}\right\}_{j\le J, i\le N} \\ \left\{\mathbf{a}^*_{jt}\left(\mathbf{D}_{it}^\top, D_{it1}\mathbf{x}_i{}^\top, D_{it2}\mathbf{x}_i{}^\top, \mathbf{0}_{p_z}^\top, D_{it1}\boldsymbol{\theta}^{*\top}_i, \mathbf{0}_2\right)\right\}_{i\le N, j\le J} & 0_{2N\times 2N} \end{pmatrix}}_{(NJ)^{1/2} II} \\
&\quad + \underbrace{\sum_{t=1}^T \begin{pmatrix} \operatorname{diag}\left(\left\{\sum_{i=1}^N \left(E\left(\varrho''_{ijt}\right) - \underline{\delta}\right) \mathbf{e}^*_{it}\mathbf{e}^{*\top}_{it}\right\}_{j\le J}\right) & \left\{\left(E\left(\varrho''_{ijt}\right) - \underline{\delta}\right) \mathbf{e}^*_{it}\mathbf{a}^{*\top}_{jt}\right\}_{j\le J, i\le N} \\ \left\{\left(E\left(\varrho''_{ijt}\right) - \underline{\delta}\right) \mathbf{a}^*_{jt}\mathbf{e}^{*\top}_{it}\right\}_{i\le N, j\le J} & \operatorname{diag}\left(\left\{\sum_{j=1}^J \left(E\left(\varrho''_{ijt}\right) - \underline{\delta}\right) \mathbf{a}^*_{jt}\mathbf{a}^{*\top}_{jt}\right\}_{i\le N}\right) \end{pmatrix}}_{(NJ)^{1/2} III}.
\end{aligned}
$$

The rest of the proof the follows from Assumption A.5, (A.50), (A.51), (A.53) and the arguments from the proof of Lemma A.6.

Remark A.7. *When the restriction $\gamma_{jt} = t\gamma_j$ is imposed, the above derivations remain valid with minor modifications. Specifically, the normalisation constraint $\Theta^\top \mathbf{1}_N = \mathbf{0}_{K^*}$ no longer applies. Therefore, in the case where $p = T = K^* = 2$, we define*

$$
\begin{aligned}
\mathbb{P}_{NJ}(\Xi) = 2\delta \Bigg\{ &\frac{1}{2N}\left(\sum_{i=1}^N \theta_{i1}\theta_{i2}\right)^2 + \frac{1}{2J}\left(\sum_{j=1}^J a_{j11}a_{j21}\right)^2 + \frac{1}{8J}\left(\sum_{j=1}^J a_{j11}^2 - J\right)^2 \\
&+ \frac{1}{8J}\left(\sum_{j=1}^J a_{j21}^2 - J\right)^2 + \frac{1}{2N}\sum_{k=1}^2 \sum_{p=1}^2 \left(\sum_{i=1}^N \theta_{ik}x_{ip}\right)^2 \Bigg\}.
\end{aligned}
$$

The definitions of $\boldsymbol{\mu}_k$ and $\boldsymbol{\omega}_{k,t}$ remain the same, except that $\boldsymbol{\mu}_5$, $\boldsymbol{\mu}_6$, $\boldsymbol{\omega}_{5,t}$, and $\boldsymbol{\omega}_{6,t}$ are omitted to adjust for the normalisation constraints. The rest of the proof then proceeds accordingly.

A.4.6 Proof of consistency results for asymptotic variance

It suffices to show that for $j = 1, \ldots, J$, $\|\hat{\mathit{\Phi}}_j - \mathit{\Phi}_j\|_F = o_P(1)$. Assume that Assumptions 2.1 to 2.7 and A.1 to A.3 hold. Note that

$$\begin{aligned}
\hat{\mathit{\Phi}}_j =& \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\rho_{ijt}^{''}(\hat{\mathbf{u}}_j^{\top}\hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^{\top} \\
=& \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\rho_{ijt}^{''}(\hat{\mathbf{u}}_j^{\top}\hat{\mathbf{e}}_{it})(\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^{\top} - \mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top}) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\left(\rho_{ijt}^{''}(\hat{\mathbf{u}}_j^{\top}\hat{\mathbf{e}}_{it}) - \rho_{ijt}^{''}\right)\mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top} \\
&+ \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\rho_{ijt}^{''}\mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top}.
\end{aligned}$$

For the first term,

$$\begin{aligned}
\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\rho_{ijt}^{''}(\hat{\mathbf{u}}_j^{\top}\hat{\mathbf{e}}_{it})(\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^{\top} - \mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top})\right\|_F &\lesssim \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^{\top} - \mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top}\|_F \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N}\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^{*}\| \\
&= O_P(N^{-1/2})
\end{aligned}$$

by Theorem A.2. For the second term, we can show that

$$\begin{aligned}
\frac{1}{N}\left\|\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\left(\rho_{ijt}^{''}(\hat{\mathbf{u}}_j^{\top}\hat{\mathbf{e}}_{it}) - \rho_{ijt}^{''}\right)\mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top}\right\|_F &\lesssim \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}|\hat{\mathbf{u}}_j^{\top}\hat{\mathbf{e}}_{it} - \mathbf{u}_j^{*\top}\mathbf{e}_{it}^{*}| \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(\|\hat{\mathbf{e}}_{it} - \mathbf{e}_{it}^{*}\| + \|\mathbf{u}_j^{*} - \hat{\mathbf{u}}_j\|) \\
&= O_P(N^{-1/2})
\end{aligned}$$

by Theorem A.2. Therefore, we have

$$\hat{\mathit{\Phi}}_j - \mathit{\Phi}_{N,j} = O_P(N^{-1/2}) + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{it} - E(r_{it}))\rho_{ijt}^{''}\mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top} = O_P(N^{-1/2})$$

and consequently $\|\hat{\mathit{\Phi}}_j - \mathit{\Phi}_j\|_F = o_P(1)$ since $\mathit{\Phi}_j = \lim_{N\to\infty}\mathit{\Phi}_{N,j}$.

A.5 Additional Simulation and Real Data Analysis result

This section provides additional numerical results. Section A.5.1 reports additional simulation results for the main model. Section A.5.2 presents results for the model extensions introduced in Sections 2.2.4 and 2.2.5. Finally, Section A.5.3 reports the number of estimated factors in the real data analysis under both the main model and its extensions.

Table A.2: Summary statistics for the constraint parameter sensitivity analysis. The results for the proposed method under different choices of the constraint parameter c across various combinations of N and J are reported, for $K^* = 3$ and $K^* = 8$, respectively. Definitions of Loss, Bloss, and MMSE are provided in Table 2.1.

		$N = 5J$				$N = 10J$			
K^*	Metric (c)	$J = 100$	$J = 200$	$J = 300$	$J = 400$	$J = 100$	$J = 200$	$J = 300$	$J = 400$
3	Loss ($c = 3$)	0.54	0.36	0.28	0.24	0.48	0.32	0.26	0.22
	Loss ($c = 4$)	0.55	0.36	0.28	0.25	0.48	0.32	0.26	0.22
	Loss ($c = 5$)	0.55	0.36	0.29	0.24	0.48	0.32	0.26	0.22
	Loss ($c = 6$)	0.55	0.36	0.29	0.24	0.48	0.32	0.26	0.22
	Loss ($c = 7$)	0.55	0.36	0.28	0.24	0.48	0.32	0.26	0.22
	Bloss ($c = 3$)	0.48	0.32	0.25	0.22	0.33	0.22	0.18	0.16
	Bloss ($c = 4$)	0.48	0.32	0.25	0.22	0.33	0.22	0.18	0.16
	Bloss ($c = 5$)	0.49	0.32	0.25	0.22	0.33	0.22	0.18	0.16
	Bloss ($c = 6$)	0.48	0.32	0.25	0.22	0.33	0.22	0.18	0.16
	Bloss ($c = 7$)	0.49	0.32	0.25	0.22	0.33	0.22	0.18	0.16
	MMSE ($c = 3$)	0.11	0.05	0.03	0.03	0.06	0.02	0.02	0.01
	MMSE ($c = 4$)	0.15	0.05	0.04	0.03	0.06	0.03	0.02	0.01
	MMSE ($c = 5$)	0.12	0.06	0.03	0.03	0.06	0.03	0.02	0.02
	MMSE ($c = 6$)	0.13	0.06	0.03	0.03	0.06	0.03	0.02	0.01
	MMSE ($c = 7$)	0.13	0.05	0.03	0.03	0.06	0.02	0.02	0.01
8	Loss ($c = 3$)	1.28	0.69	0.52	0.44	1.06	0.62	0.48	0.41
	Loss ($c = 4$)	1.31	0.69	0.52	0.44	1.08	0.63	0.48	0.41
	Loss ($c = 5$)	1.33	0.68	0.52	0.44	1.09	0.62	0.48	0.41
	Loss ($c = 6$)	1.38	0.68	0.52	0.44	1.10	0.62	0.48	0.41
	Loss ($c = 7$)	1.35	0.69	0.52	0.44	1.10	0.63	0.48	0.41
	Bloss ($c = 3$)	0.66	0.39	0.31	0.26	0.43	0.27	0.22	0.19
	Bloss ($c = 4$)	0.67	0.39	0.31	0.26	0.44	0.27	0.21	0.19
	Bloss ($c = 5$)	0.67	0.39	0.31	0.26	0.44	0.27	0.22	0.19
	Bloss ($c = 6$)	0.69	0.39	0.31	0.26	0.44	0.27	0.22	0.19
	Bloss ($c = 7$)	0.68	0.39	0.31	0.26	0.44	0.27	0.21	0.19
	MMSE ($c = 3$)	1.21	0.09	0.06	0.04	0.15	0.04	0.03	0.02
	MMSE ($c = 4$)	1.33	0.08	0.05	0.04	0.13	0.05	0.02	0.02
	MMSE ($c = 5$)	1.10	0.08	0.05	0.04	0.13	0.04	0.03	0.02
	MMSE ($c = 6$)	2.21	0.09	0.07	0.04	0.13	0.05	0.02	0.02
	MMSE ($c = 7$)	1.03	0.09	0.06	0.03	0.15	0.05	0.03	0.02

A.5.1 Additional Simulation results for the Main Model

We present three additional sets of simulation results for the main model. Table A.2 reports the sensitivity analysis results for the constraint parameter c, as discussed in Section 2.4.

We also compare the computation time of the proposed method using the candidate sets $\mathcal{K} = \{K^*\}$ and $\mathcal{K} = \{1, \ldots, 10\}$ with that of standard logistic regression (LR) method and logistic regression with a random intercept (LRRI). Specifically, Table A.3 reports the total time required to complete five replications across the settings described in Section 2.4. When the candidate set $\mathcal{K}$ equals the true latent dimension K^*, the proposed method is significantly faster than LRRI, one of the simplest traditional latent variable models that does not account for dependence between items. This demonstrates the superior computational efficiency of our method over existing latent variable approaches. Even with a broader candidate set $\mathcal{K} = \{1, \ldots, 10\}$, the proposed method remains comparable in speed to LRRI and is faster in some cases, highlighting its strong scalability for larger datasets even when the number of latent factors needs to be estimated.

Finally, we evaluate the performance of the proposed estimator in estimating the latent variables. As noted in Remark A.3, having distinct eigenvalues in $(\Theta^*)^\top \Theta^*/N$ is

Table A.3: Computation time (in minutes) over 5 simulations under various settings.

N	J	Proposed($\mathcal{K} = \{K^*\}$)	Proposed($\mathcal{K} = \{1, \dots, 10\}$)	LR	LRRI
$K^* = 3$					
5J	100	0.09	22.72	0.01	9.54
5J	200	0.28	78.91	0.05	37.40
5J	300	0.63	163.12	0.10	82.80
5J	400	1.15	386.44	0.18	152.36
10J	100	0.15	45.14	0.02	17.46
10J	200	0.55	202.89	0.09	77.52
10J	300	1.23	584.78	0.19	180.16
10J	400	2.25	1132.89	0.35	332.14
$K^* = 8$					
5J	100	0.24	11.46	0.01	12.12
5J	200	0.70	40.65	0.04	44.56
5J	300	1.54	74.08	0.10	100.88
5J	400	2.51	119.48	0.17	185.35
10J	100	0.42	21.09	0.02	21.92
10J	200	1.38	73.46	0.09	91.71
10J	300	2.57	174.58	0.19	213.08
10J	400	4.59	229.82	0.37	389.39

essential for identifying the latent factors. To ensure this, we generate each θ_{ik} from a truncated standard normal distribution on $[-1, 1]$, multiplied by a constant factor $k/2$. All other parameters are generated as in the procedures described in Section 2.4.

Following the evaluation criteria outlined in Section 2.4.2, we compute the Frobenius losses for $\hat{A}$ and $\hat{\Theta}$, denoted by "Aloss" and "Tloss", respectively:

$$\text{ALoss} = \frac{1}{\sqrt{J}} \left\| \hat{A} - A^* \hat{S}_A \right\|_F, \quad \text{TLoss} = \frac{1}{\sqrt{N}} \left\| \hat{\Theta} - \Theta^* \hat{S}_A \right\|_F,$$

where $\hat{S}_A$ is a diagonal matrix correcting for sign indeterminacy, as defined in Theorem A.1.

To further assess the estimator's performance at the individual parameter level, we compute the mean squared errors for each a_{jk} and θ_{ik}, where $j = 1, \dots, J$, $i = 1, \dots, N$, and $k = 1, \dots, K^*$. The maximum MSE across all simulation trials for the entries in A and Θ are reported as "MAMSE" and "MTMSE", repsectively.

In addition, we construct 95% confidence intervals for a_{jk} and θ_{ik} using the asymptotic variance estimation method described in Remark 2.1. The empirical coverage probabilities (AECP for A and TECP for Θ) are computed by aggregating the coverage rates across all parameters and simulation repetitions. The results are summarized in Table A.4.

The results validate Theorem A.1, indicated by decreasing trends in "ALoss", "TLoss", "MAMSE", and "MTMSE" with larger N and J. Empirical coverage probabilities (AECP and TECP) approach the nominal 95% level as N increases, supporting Theorem A.2. We note that coverage rates worsen as K^* increases from 3 to 8, especially when $J = 100$. Therefore, while the theoretical properties are supported in large samples, caution is required when conducting inference for latent variables in practical settings, especially when the estimated number of factors is large relative to the sample size.

Table A.4: Summary statistics on latent variables estimations for the proposed method across different values of N, K^* and J.

N	J	Aloss	Tloss	MAMSE	MTMSE	AECP	TECP
$K^* = 3$							
5J	100	0.39	0.36	0.28	0.11	0.88	0.94
5J	200	0.23	0.24	0.06	0.06	0.92	0.94
5J	300	0.20	0.19	0.05	0.03	0.93	0.95
5J	400	0.17	0.17	0.04	0.02	0.94	0.95
10J	100	0.25	0.35	0.08	0.12	0.89	0.94
10J	200	0.18	0.23	0.07	0.05	0.91	0.95
10J	300	0.14	0.19	0.03	0.04	0.93	0.95
10J	400	0.12	0.17	0.02	0.03	0.94	0.95
$K^* = 8$							
5J	100	1.86	3.69	3.74	31.94	0.36	0.65
5J	200	0.46	0.89	0.29	1.38	0.81	0.91
5J	300	0.32	0.61	0.14	0.49	0.87	0.93
5J	400	0.27	0.50	0.09	0.36	0.89	0.93
10J	100	0.82	2.16	0.81	14.85	0.55	0.85
10J	200	0.32	0.82	0.11	1.67	0.81	0.93
10J	300	0.23	0.58	0.06	0.56	0.87	0.94
10J	400	0.19	0.48	0.04	0.42	0.90	0.94

Aloss / Tloss: Frobenius loss measuring convergence of $\hat{A}/\hat{\Theta}$.
MAMSE / MTMSE: Maximum mean squared error across all estimated a_{jk}/θ_{ik}s.
AECP / TECP: Empirical coverage probability of the confidence intervals across all estimated a_{jk}/θ_{ik}s.

Table A.5: Simulation results for the model variants introduced in Sections 2.2.4 and 2.2.5 across different combinations of N, J, and K^*.

		Section 2.2.4				**Section 2.2.5**				**Sections 2.2.4 and 2.2.5**			
N	J	Loss	$P(\hat{K} = K^*)$	Bloss	MMSE	Loss	$P(\hat{K} = K^*)$	Bloss	MMSE	Loss	$P(\hat{K} = K^*)$	Bloss	MMSE
$K^* = 3$													
5J	100	0.88	1	1.06	0.85	0.51	1	0.42	0.11	0.83	1	0.94	1.02
5J	200	0.51	1	0.67	0.29	0.34	1	0.29	0.04	0.50	1	0.60	0.23
5J	300	0.40	1	0.55	0.16	0.28	1	0.23	0.03	0.38	1	0.48	0.12
5J	400	0.34	1	0.45	0.11	0.24	1	0.20	0.02	0.33	1	0.41	0.08
10J	100	0.66	1	0.71	0.35	0.46	1	0.29	0.05	0.63	1	0.63	0.33
10J	200	0.40	1	0.47	0.11	0.31	1	0.20	0.02	0.40	1	0.42	0.11
10J	300	0.32	1	0.38	0.08	0.25	1	0.16	0.02	0.31	1	0.34	0.06
10J	400	0.27	1	0.32	0.06	0.22	1	0.14	0.01	0.27	1	0.29	0.05
$K^* = 8$													
5J	100	7.63	1	3.55	61.55	1.24	1	0.58	0.31	7.68	1	3.03	51.17
5J	200	1.01	1	0.86	0.53	0.67	1	0.35	0.09	0.99	1	0.75	0.35
5J	300	0.73	1	0.64	0.25	0.51	1	0.28	0.04	0.72	1	0.58	0.17
5J	400	0.59	1	0.55	0.17	0.43	1	0.24	0.03	0.58	1	0.49	0.13
10J	100	2.96	1	1.24	24.35	1.09	1	0.40	0.15	3.21	1	1.23	15.53
10J	200	0.79	1	0.57	0.21	0.62	1	0.25	0.04	0.78	1	0.50	0.19
10J	300	0.59	1	0.45	0.11	0.48	1	0.20	0.02	0.59	1	0.41	0.10
10J	400	0.49	1	0.38	0.08	0.40	1	0.17	0.01	0.49	1	0.34	0.07

Loss: Frobenius loss measuring the convergence of $\hat{\Xi}$.
P($\hat{K} = K^*$): Proportion of instances where the correct number of factors is identified.
Bloss: Frobenius loss measuring convergence of $\hat{B}$.
MMSE: Maximum mean squared error across all elements in $\hat{B}$.

Table A.6: Number of factors $\hat{K}$ selected by each method for varying values of J in real data analysis.

Method	$J = 100$	$J = 200$	$J = 300$	$J = 400$
Prop	8	8	11	11
Prop (2.2.4)	4	2	2	1
Prop (2.2.5)	10	11	11	14
Prop (2.2.4 & 2.2.5)	5	3	3	3

A.5.2 Additional Simulation Results for Models in Section 2.2.4 and 2.2.5

To assess the performance of the model variants introduced in Sections 2.2.4 and 2.2.5, we conduct additional simulations using the same combinations of N, J, and K^* as in Section 2.4. For each setting, we consider the candidate set $\mathcal{K} = \{1, 2, \ldots, 10\}$ and generate 100 independent replications.

For the extended model in Section 2.2.4, data are generated from the logistic model:

$$P(y_{ijt} = 1 \mid \gamma_{jt}, \mathbf{a}_{jt}, \boldsymbol{\theta}_i, \boldsymbol{\beta}_{jt}, \mathbf{x}_i) = \frac{\exp(\gamma_{jt} + \sum_{k=1}^{K^*} a_{jkt}\theta_{ik} + \sum_{l=1}^{5} \beta_{jlt}x_{il})}{1 + \exp(\gamma_{jt} + \sum_{k=1}^{K^*} a_{jkt}\theta_{ik} + \sum_{l=1}^{5} \beta_{jlt}x_{il})}. \tag{A.54}$$

The variables are generated following a similar procedure to that described in Section 2.4 for the main model. As before, we slightly abuse notation by using the same symbols before and after normalisation. Specifically, the covariates $\mathbf{x}_i$, intercepts γ_{jt}, preliminary latent factors θ_{ik}, and missingness indicators $\mathbf{r}_i$ are generated as in the main setting. The preliminary time-varying regression coefficients β_{jlt} are independently sampled from a uniform distribution $U[0.5, 1]$, and the time-varying factor loadings a_{jkt} are sampled from truncated standard normal distributions on $[-3, 3]$. We then apply the normalisation procedure described in Section A.2 to ensure identifiability conditions in this model are satisfied. Following normalisation, we set half of the coefficient pairs $(\beta_{j1t}, \beta_{j2t})$ to zero. The same procedure is applied independently to the pairs $(\beta_{j3t}, \beta_{j4t})$, and half of the individual coefficients β_{j5t} are also set to zero.

For simulations under the constraint $\gamma_{jt} = t\gamma_j$ (Section 2.2.5), we substitute γ_{jt} accordingly in both (2.10) and (A.54), where γ_j is drawn from $U[-0.25, 0.25]$.

The models described above are evaluated using the performance metrics analogous to those introduced in Section 2.4.2, focusing on the convergence of parameters and accuracy in determining the number of factors. In particular, for the extension in Section 2.2.4, the convergence of the estimated time-varying regression coefficients $\hat{B} = (\hat{B}_1, \ldots, \hat{B}_T)$ is assessed by

$$\text{Bloss} = \frac{1}{\sqrt{J}} \max_{t=1,\ldots,T} \left\| \hat{B}_t - B_t^* \right\|_F,$$

and the convergence of $\hat{\boldsymbol{\Xi}}$ is evaluated by

$$\text{Loss} = \max_{t=1,\ldots,T} \frac{\left\| \hat{\Theta}\hat{A}_t^\top - \Theta^* A_t^{*\top} + X(\hat{B}_t - B_t^*)^\top + \mathbf{1_N}(\hat{\mathit{\Gamma}}_t - \mathit{\Gamma}_t^*)^\top \right\|_F}{\sqrt{NJ}}.$$

We replace $\hat{\varGamma}_t$ and $\varGamma_t^*$ by $t\hat{\varGamma}$ and $t\varGamma^*$, respectively, in the "Loss" metric when the model in Section 2.2.5 is evaluated. Table A.5 presents the simulation results for the model variants described in Sections 2.2.4 and 2.2.5, as well as the combined model incorporating both specifications. The results are consistent with the theory of these methods. In particular, the performance metrics "Loss", "Bloss", and "MMSE" decrease as N and J increase, regardless of the number of factors or the ratio between N and J. Furthermore, the proposed information criterion is effective, with $P(\hat{K} = K^*)$ consistently equals 1 across all settings and model variants.

A.5.3 Additional Results in Real Data Analysis

Table A.6 reports the number of latent factors $\hat{K}$ selected by each of the proposed methods described in Section 2.5, for different numbers of items $J = 100, 200, 300, 400$.

Appendix B

Supplementary Materials for Chapter 3

B.1 Additional Simulation and Real Data Results

We present additional simulation results for the proposed criteria SC1, SC2 and SC3, where the total number of splits J is set to 1. Under this setup, only one random split is performed to compute the instability measure $\mathrm{INS}(k)$. Figure B.1 shows the percentage of correct selections of the true number of factors K by these criteria. The result demonstrates that the correct selection percentages of the proposed criteria with a single split are nearly identical to those obtained at $J = 10$, as shown in figure 3.1.

This observation is further supported by Table B.1, which presents the modes of the selected number of factors and the corresponding selection percentages for each criteria in the analysis of the p53 protein dataset using a single split. As shown in the table, all criteria with a single split consistently select 2 as the mode, which matches the result in Table 3.1 at $J = 10$.

These findings suggest that the randomness introduced by splitting does not have severe impact on the performance of the proposed selection criteria. Therefore, using $J = 10$ or any other reasonable number of splits is adequate in practice.

B.2 Proof

With slight abuse of notation, we assume $n_1 = n_2 = n$ in this section for convenience. To simplify notation, we assume $\psi = 1$ without loss of generality. Recall that the population covariance matrix is

$$\Sigma = \Lambda^\top \Lambda + I_p = \sum_{j=1}^{p} \sigma_j \mathbf{v}_j \mathbf{v}_j^\top .$$

We express the eigenvalues σ_j of Σ as

$$\sigma_j = 1 + \phi_n^{1/2} d_j, j = 1, \dots, p,$$

where $\phi_n = p/n$ is the dimensional ratio. For notational convenience, we suppress the dependence on n and write ϕ. It is easy to see that $\sigma_j = 1$ and $d_j = 0$ for $j > K$. For any vector $\mathbf{w} \in \mathbb{R}^T$, we use $w_j = \langle \mathbf{v}_j, \mathbf{w} \rangle, j = 1, \dots, p$ to denote the components of $\mathbf{w}$ in the eigenbasis of Σ.

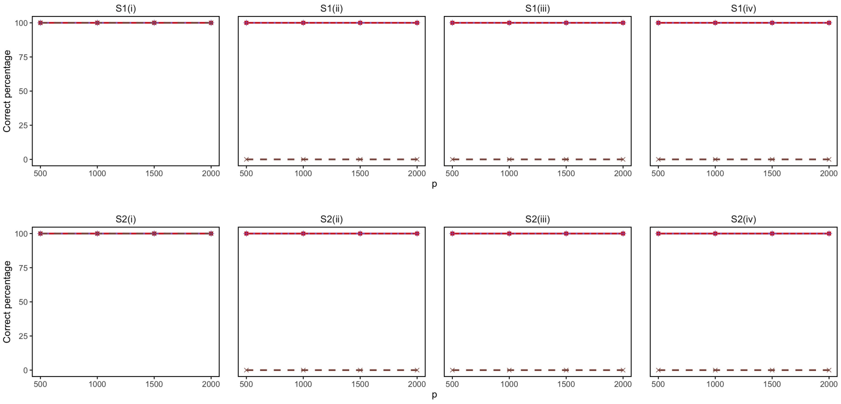

Figure B.1: Correct selection percentages versus the number of features p across different scenarios (S1 and S2) and signal strengths (i,ii and iii). All criteria are evaluated using $J = 1$ for the calculation of INS(k). SC1 (red solid line with squares), SC2 (purple dotted line with pluses), and SC3 (brown dashed line with crosses)

Table B.1: Performance of the Proposed Criteria ($J = 1$) on the p53 Protein Dataset

Criterion	Mode	Selection percentage(%)
SC1	2	95
SC2	2	95
SC3	2	98

Mode: mode of the estimated number of factors for each criterion. Selection percentage(%): Percentages of instances selecting the mode.

B.2.1 Proof of Proposition 3.1

The following lemma establishes a delocalization bound for the $(K+1)$-th to $K_{\max}$-th eigenvectors of the sample covariance matrix of $X^{(l_1)}$ with respect to all eigenvectors of the sample covariance matrix of $X^{(l_2)}$ for the candidate set $\mathcal{K}$, where $l_1 \neq l_2$.

Lemma B.1. *Under Assumption 3.1 to 3.3, for any $k \in \{K+1, \dots, K_{\max}\}$ and $s \in \mathcal{K}$, we have*

$$\langle \tilde{\mathbf{v}}_k^{(1)}, \tilde{\mathbf{v}}_s^{(2)} \rangle^2 \prec (p \min\{n,p\})^{-1/2} \text{ and } \langle \tilde{\mathbf{v}}_k^{(2)}, \tilde{\mathbf{v}}_s^{(1)} \rangle^2 \prec (p \min\{n,p\})^{-1/2}.$$

Proof: We prove the first statement; the proof for the second is identical. For any deterministic unit vector $\mathbf{w}$, by Theorem 2.17 of Bloemendal et al. (2016), we have

$$\langle \tilde{\mathbf{v}}_k^{(1)}, \mathbf{w} \rangle^2 \prec \frac{\|\mathbf{w}\|^2}{p} + \sum_{j=1}^{p} \frac{\sigma_j w_j^2}{p(d_j - 1)^2 + \kappa_j}, \tag{B.1}$$

where $\kappa_j = \min\{n,p\}^{-2/3} \min\{j, \min\{n,p\}+1-j\}^{2/3}$. Since $d_j = 0$ and $\sigma_j = 1$ for $j > K$, we have

$$\sum_{j=K+1}^{p} \frac{\sigma_j w_j^2}{p\{(d_j-1)^2 + \kappa_j\}} \leq \sum_{j=K+1}^{p} \frac{w_j^2}{p} \leq \frac{1}{p}. \tag{B.2}$$

On the other hand, for $1 \leq j \leq K$, by Assumption 3.3, we have $d_j \to +\infty$. Therefore, we have

$$\frac{\sigma_j w_j^2}{p\{(d_j-1)^2+\kappa_j\}} = \frac{(1+\phi^{1/2}d_j)w_j^2}{p\{(d_j-1)^2+\kappa_j\}} \lesssim \frac{1}{(np)^{-1/2}d_j}. \tag{B.3}$$

Combining equations (B.1), (B.2), and (B.3), we have

$$\langle \tilde{\mathbf{v}}_k^{(1)}, \mathbf{w}\rangle^2 \prec (p\min\{n,p\})^{-1/2} \tag{B.4}$$

for any deterministic unit vector $\mathbf{w}$. Note that $\tilde{\mathbf{v}}_k^{(1)}$ and $\tilde{\mathbf{v}}_s^{(2)}$ are independent. Therefore, by conditioning on $\tilde{\mathbf{v}}_s^{(2)}$, we can apply (B.4) and hence the lemma is proved.

We now turn to the proof of (3.4) in Proposition 3.1. It is sufficient to show that for $k > K$,

$$\min_{\boldsymbol{\xi}^{(1)}\in\tilde{V}_k^{(1)};\|\boldsymbol{\xi}^{(1)}\|=1}\ \max_{\boldsymbol{\xi}^{(2)}\in\tilde{V}_k^{(2)};\|\boldsymbol{\xi}^{(2)}\|=1} \langle \boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}\rangle \prec (p\min\{n,p\})^{-1/2}. \tag{B.5}$$

Note that Lemma B.1 implies

$$\max_{\boldsymbol{\xi}^{(2)}\in\tilde{V}_k^{(2)};\|\boldsymbol{\xi}^{(2)}\|=1} \langle \tilde{\mathbf{v}}_{K+1}^{(1)}, \boldsymbol{\xi}^{(2)}\rangle \prec (p\min\{n,p\})^{-1/2}.$$

Since $\tilde{\mathbf{v}}_{K+1}^{(1)} \in \tilde{V}_k^{(1)}$ and $\|\tilde{\mathbf{v}}_{K+1}^{(1)}\| = 1$, the proof is complete by the definition of (B.5). We now proceed to prove (3.5) in Proposition 3.1. First, we introduce some definitions and supporting lemmas. Let $A \subseteq \{1,\dots,p\}$ be a subset of integers from 1 to p. For $l = 1,2$, define the random spectral projection

$$P_A^{(l)} = \sum_{k\in A} \tilde{\mathbf{v}}_k^{(l)}(\tilde{\mathbf{v}}_k^{(l)})^\top.$$

For $k = 1,\dots,p$, we also define

$$\nu_k(A) = \begin{cases} \min_{j\notin A}|d_k - d_j| \text{ if } k \in A \\ \min_{j\in A}|d_k - d_j| \text{ if } k \notin A \end{cases}.$$

Here $\nu_k(A)$ is the distance from d_k to either $\{d_j\}_{j\in A}$ or $\{d_j\}_{j\notin A}$, whichever it does not belong to. We further define the deterministic positive quadratic form

$$\langle \mathbf{w}, Z_A\mathbf{w}\rangle = \sum_{j\in A}\mu(d_j)w_j^2, \text{ where}$$
$$\mu(d_j) = \frac{\sigma_j}{\phi^{1/2}\theta(d_j)}(1-d_j^{-2}),\ \theta(d_j) = \phi^{1/2} + \phi^{-1/2} + d_j + d_j^{-1}.$$

The following lemma describes the behaviour of eigenvectors associated with a subset $A \subseteq \{1,\dots,K\}$, in comparison with the corresponding deterministic positive quadratic form.

Lemma B.2. *Under Assumptions 3.1 to 3.3, let $A \subseteq \{1,\dots,K\}$ such that $d_j = d_A$ for all $j \in A$. Then for any deterministic unit vector $\mathbf{w} \in \mathbb{R}^T$, we have*

$$\langle \mathbf{w}, P_A^{(l)}\mathbf{w}\rangle = \langle \mathbf{w}, Z_A\mathbf{w}\rangle + O_\prec(n^{-1/2}), l = 1,2.$$

Proof: From Theorem 2.16 of Bloemendal et al. (2016), we have

$$\langle \mathbf{w}, P_A^{(l)} \mathbf{w} \rangle = \langle \mathbf{w}, Z_A \mathbf{w} \rangle + O_\prec \Bigg(\frac{1}{p^{1/2}(\phi^{1/2}+d_A)} \sum_{j \in A} \sigma_j w_j^2 + \left(1 + \frac{\phi^{1/2} d_A^2}{\phi^{1/2}+d_A}\right) \sum_{j=1}^{p} \frac{\sigma_j w_j^2}{p \nu_j(A)^2} + \frac{d_A}{\phi^{1/2}+d_A} \left(\sum_{j \in A} \sigma_j w_j^2 \right)^{1/2} \left(\sum_{j \notin A} \frac{\sigma_j w_j^2}{p \nu_j(A)^2} \right)^{1/2} \Bigg). \tag{B.6}$$

Recall that $\sigma_j = 1 + \phi^{1/2} d_A$ for $j \in A$. Thus, we have

$$\frac{1}{p^{1/2}(\phi^{1/2}+d_A)} \sum_{j \in A} \sigma_j w_j^2 \lesssim \frac{\phi^{1/2} d_A}{p^{1/2} d_A} \lesssim \frac{1}{n^{1/2}}. \tag{B.7}$$

Moreover, we can verify that

$\nu_j(A) \asymp d_A$ for $j \in A$,
$\nu_j(A) \asymp \max\{d_j, d_A\}$ for $j \in \{1, \ldots, K\} \setminus A$ and
$\nu_j(A) \asymp d_A$ for $j \in \{K+1, \ldots, p\}$.

Therefore, we have

$$\begin{aligned} &\left(1 + \frac{\phi^{1/2} d_A^2}{\phi^{1/2}+d_A}\right) \sum_{j=1}^{p} \frac{\sigma_j w_j^2}{p \nu_j(A)^2} \\ =&\left(1 + \frac{\phi^{1/2} d_A^2}{\phi^{1/2}+d_A}\right) \left(\sum_{j=1}^{K} \frac{\sigma_j w_j^2}{p \nu_j(A)^2} + \sum_{j=K+1}^{p} \frac{\sigma_j w_j^2}{p \nu_j(A)^2} \right) \\ \lesssim& \frac{\phi d_A}{p(\phi^{1/2}+d_A)} + \frac{\phi^{1/2}}{p(\phi^{1/2}+d_A)} \\ \lesssim& \frac{1}{n} + \frac{1}{p + (np)^{1/2} d_A}. \end{aligned} \tag{B.8}$$

Finally,

$$\begin{aligned} &\frac{d_A}{\phi^{1/2}+d_A} \left(\sum_{j \in A} \sigma_j w_j^2 \right)^{1/2} \left(\sum_{j \notin A} \frac{\sigma_j w_j^2}{p \nu_j(A)^2} \right)^{1/2} \\ \lesssim& (\phi^{1/2} d_A)^{1/2} \left(\frac{\phi^{1/2}}{p d_A} \right)^{1/2} \\ \lesssim& n^{-1/2}. \end{aligned} \tag{B.9}$$

Therefore, the lemma is proved by (B.6), (B.7), (B.8) and (B.9).

The following lemma states some useful results derived from Lemma B.2:

Lemma B.3. *Under the conditions of Lemma B.2, we have*

$$\sum_{k\in A}\langle \mathbf{v}, \tilde{\mathbf{v}}_k^{(l)}\rangle^2 = 1 + O_{\prec}(n^{-1/2} + \phi^{1/2}d_A^{-1} + d_A^{-2}) \text{ and} \tag{B.10}$$

$$\sum_{k\in A}\langle \boldsymbol{\xi}^{(l)}, \mathbf{v}_k\rangle^2 = 1 + O_{\prec}(n^{-1/2} + \phi^{1/2}d_A^{-1} + d_A^{-2}) \tag{B.11}$$

for any $\mathbf{v} \in$ *Span*$\{\mathbf{v}_j : j \in A\}$ *with* $\|\mathbf{v}\| = 1$ *and* $\boldsymbol{\xi}^{(l)} \in$ *Span*$\{\tilde{\mathbf{v}}_j^{(l)} : j \in A\}$ *with* $\|\boldsymbol{\xi}^{(l)}\| = 1$. *Moreover, for* $j \notin A$, *we have*

$$\sum_{k\in A}\langle \mathbf{v}_j, \tilde{\mathbf{v}}_k^{(l)}\rangle^2 = O_{\prec}(n^{-1/2}) \text{ and}$$

$$\sum_{k\in A}\langle \tilde{\mathbf{v}}_j^{(l)}, \mathbf{v}_k\rangle^2 = O_{\prec}(n^{-1/2}).$$

Proof: We only prove (B.10) and (B.11) to save space as the derivations are similar. We first show (B.10). Let $\mathbf{v} \in \text{Span}\{\mathbf{v}_i : i \in A\}$ with length 1. By Lemma B.2, for $l = 1, 2$, we have

$$\langle \mathbf{v}, P_A^{(l)}\mathbf{v}\rangle = \langle \mathbf{v}, Z_A^{(l)}\mathbf{v}\rangle + O_{\prec}(n^{-1/2}).$$

Note that

$$\begin{aligned}\langle \mathbf{v}, P_A^{(l)}\mathbf{v}\rangle &= \langle \mathbf{v}, \sum_{k\in A}\tilde{\mathbf{v}}_k^{(l)}(\tilde{\mathbf{v}}_k^{(l)})^\top \mathbf{v}\rangle \\ &= \sum_{k\in A}\langle \mathbf{v}, \tilde{\mathbf{v}}_k^{(l)}(\tilde{\mathbf{v}}_k^{(l)})^\top \mathbf{v}\rangle \\ &= \sum_{k\in A}\langle \mathbf{v}, \tilde{\mathbf{v}}_k^{(l)}\rangle^2. \end{aligned} \tag{B.12}$$

On the other hand,

$$\langle \mathbf{v}, Z_A^{(l)}\mathbf{v}\rangle = \sum_{j\in A}\mu(d_j)\langle \mathbf{v}_j, \mathbf{v}\rangle^2. \tag{B.13}$$

Note that

$$\begin{aligned}\mu(d_j) &= \frac{\sigma_j}{\phi^{1/2}(\phi^{1/2} + \phi^{-1/2} + d_j + d_j^{-1})}(1 - d_j^{-2}) \\ &= \left(1 - \frac{\phi + \phi^{1/2}d_j^{-1}}{\phi + 1 + \phi^{1/2}d_j + \phi^{1/2}d_j^{-1}}\right)(1 - d_j^{-2}) \\ &= 1 - O_{\prec}(\phi^{1/2}d_j^{-1} + d_j^{-2}).\end{aligned}$$

Hence (B.10) is proved by (B.12), (B.13) and the fact that $\sum_{j\in A}\langle \mathbf{v}_j, \mathbf{v}\rangle^2 = 1$. Following similar argument in Example 2.15 of Bloemendal et al. (2016), we can interchange the role of $\{\mathbf{v}_i\}_{i\in A}$ and $\{\tilde{\mathbf{v}}_i^{(l)}\}_{i\in A}$ to get (B.11), and the proof is complete.

We proceed to generalise the result of Lemma B.3 to $\{1, \ldots, K\}$. Define a partition of $\{1, \ldots, K\}$ by $\cup_{s=1}^S A_s$, where S is the number of distinct values in $\{\sigma_1, \ldots, \sigma_K\}$. Specifically, define

$$\begin{aligned}&A_1 = \{j : \sigma_j = \sigma_1\} \text{ and} \\ &A_s = \{j : \sigma_j = \sigma_{\sum_{j=1}^{s-1}|A_j|+1}\} \text{ for } s = 2, \ldots, S.\end{aligned}$$

For $l = 1, 2$, pick any $\boldsymbol{\xi}^{(l)} \in \text{Span}\{\tilde{\mathbf{v}}_1^{(l)}, \ldots, \tilde{\mathbf{v}}_K^{(l)}\}$. We can write

$$\boldsymbol{\xi}^{(l)} = \sum_{s=1}^{S} \boldsymbol{\xi}_s^{(l)},$$

where $\boldsymbol{\xi}_s^{(l)} \in \text{Span}\{\tilde{\mathbf{v}}_j^{(l)} : j \in A_s\}$. We have the following Lemma:

Lemma B.4. *Under Assumption 3.1 to 3.3, for $l = 1, 2$, for any $\boldsymbol{\xi}^{(l)} \in Span\{\tilde{\mathbf{v}}_1^{(l)}, \ldots, \tilde{\mathbf{v}}_K^{(l)}\}$, we have*

$$\max_{\mathbf{v} \in Span\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}, \|\mathbf{v}\|=1} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v} \rangle = 1 + O_{\prec}(n^{-1/2} + \phi^{1/2} d_K^{-1} + d_K^{-2}).$$

Proof: From Lemma B.3, for $j \notin A_s$, we have

$$\sum_{k \in A_s} \langle \tilde{\mathbf{v}}_j^{(l)}, \mathbf{v}_k \rangle^2 = O_{\prec}(n^{-1/2}).$$

Therefore, we have

$$\sum_{k \in A_s} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v}_k \rangle^2 = \sum_{k \in A_s} \langle \boldsymbol{\xi}_s^{(l)}, \mathbf{v}_k \rangle^2 + O_{\prec}(n^{-1/2}).$$

Hence we have

$$\begin{aligned}
\sum_{k=1}^{K} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v}_k \rangle^2 &= \sum_{s=1}^{S} \sum_{k \in A_s} \langle \boldsymbol{\xi}_s^{(l)}, \mathbf{v}_k \rangle^2 + O_{\prec}(n^{-1/2}) \\
&= \sum_{s=1}^{S} \|\boldsymbol{\xi}_s^{(l)}\|^2 + O_{\prec}(n^{-1/2} + \phi^{1/2} d_K^{-1} + d_K^{-2}) \\
&= 1 + O_{\prec}(n^{-1/2} + \phi^{1/2} d_K^{-1} + d_K^{-2})
\end{aligned}$$

by Lemma B.3. Finally, since $\langle \boldsymbol{\xi}^{(l)}, \mathbf{v} \rangle$ is maximised by taking

$$\mathbf{v} = \frac{\sum_{k=1}^{K} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v}_k \rangle \mathbf{v}_k}{(\sum_{k=1}^{K} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v}_k \rangle^2)^{1/2}},$$

we have

$$\begin{aligned}
\max_{\mathbf{v} \in \text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{K^*}\}, \|\mathbf{v}\|=1} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v} \rangle &= \frac{\sum_{k=1}^{K} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v}_k \rangle^2}{(\sum_{k=1}^{K} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v}_k \rangle^2)^{1/2}} \\
&= 1 + O_{\prec}(n^{-1/2} + \phi^{1/2} d_K^{-1} + d_K^{-2})
\end{aligned}$$

by taylor's expansion. Thus the proof is complete.
We are now ready to prove (3.5) and complete the proof of Proposition 3.1:
From Lemma B.4, for $l = 1, 2$, we have

$$\min_{\boldsymbol{\xi}^{(l)} \in \tilde{V}_K^{(l)}; \|\boldsymbol{\xi}^{(l)}\|=1} \max_{\mathbf{v} \in V_K; \|\mathbf{v}\|=1} \langle \boldsymbol{\xi}^{(l)}, \mathbf{v} \rangle = 1 + O_{\prec}(n^{-1/2} + \phi^{1/2} d_K^{-1} + d_K^{-2}).$$

This implies

$$\max_{\boldsymbol{\xi}^{(l)}\in\tilde{V}_K^{(l)};\boldsymbol{\xi}^{(l)}\neq 0}\ \min_{\mathbf{v}\in V_K;\mathbf{v}\neq 0}\sin\angle(\boldsymbol{\xi}^{(l)},\mathbf{v}) = O_{\prec}(n^{-1/2}+\phi^{1/2}d_K^{-1}+d_K^{-2}). \tag{B.14}$$

Note that the roles of $\boldsymbol{\xi}^{(l)}$ and $\mathbf{v}$ in Lemma B.4 are interchangeable, as demonstrated in the proof. Hence, by the swapped version of (B.14), we can show that

$$\max_{\mathbf{v}\in V_K;\mathbf{v}\neq 0}\ \min_{\boldsymbol{\xi}^{(l)}\in\tilde{V}_K^{(l)};\boldsymbol{\xi}^{(l)}\neq 0}\sin\angle(\mathbf{v},\boldsymbol{\xi}^{(l)}) = O_{\prec}(n^{-1/2}+\phi^{1/2}d_K^{-1}+d_K^{-2}) \text{ for } l=1,2.$$

Therefore, the proof is complete.

Remark B.1. *When the eigenvalues are distinct, for $k=1,\dots,K-1$, it is easy to derive result analogous to Lemma B.4 for $\boldsymbol{\xi}^{(l)}\in$ Span$\{\mathbf{v}_1,\dots,\mathbf{v}_k\}$. Hence we can verify that* (3.5) *holds.*

B.2.2 Proof of Theorem 3.1

Proof: Without loss of generality, we prove under the assumption that $J=1$. It suffices to show that

$$c_k+\sin\angle(\tilde{V}_k^{(1)},\tilde{V}_k^{(2)}) \gtrsim c_K+\sin\angle(\tilde{V}_K^{(1)},\tilde{V}_K^{(2)})+\delta/2 \text{ for } k\in\mathcal{K}, k\neq K$$

as $n\to\infty$.

For $k\le K-1$, since $c_k-c_{k+1}>\delta$, it follows that $c_k-c_K>(K-k)\delta$. Hence, we have

$$\begin{aligned}&c_k+\sin\angle(\tilde{V}_k^{(1)},\tilde{V}_k^{(2)})-c_K-\sin\angle(\tilde{V}_K^{(1)},\tilde{V}_K^{(2)})\\ >&(K-k)\delta-O_{\prec}(n^{-1/2}+\phi^{1/2}d_K^{-1}+d_K^{-2})\\ \gtrsim&\delta/2\end{aligned} \tag{B.15}$$

as $n\to\infty$ by Proposition 3.1.

On the other hand, for $k>K$, by Proposition 3.1, we have

$$\begin{aligned}&c_k+\sin\angle(\tilde{V}_k^{(1)},\tilde{V}_k^{(2)})-c_K-\sin\angle(\tilde{V}_K^{(1)},\tilde{V}_K^{(2)})\\ =&c_k-c_K+1-O_{\prec}((p\min\{n,p\})^{-1/2})\\ >&\delta-1+1-O_{\prec}((p\min\{n,p\})^{-1/2})\\ \gtrsim&\delta/2\end{aligned} \tag{B.16}$$

as $n\to\infty$. Combining (B.15) and (B.16), the proof is complete.

B.2.3 Proof of Corollary 3.1

Recall that

$$\mathrm{SC2}(k) = \frac{l(k)}{l(0)} + \mathrm{INS}(k),$$

where $l(k) = \sum_{j=k+1}^{K_{\max}} \log(\tilde{\sigma}_j + 1)$ for $k = 0, 1, \ldots, K_{\max} - 1$ and $l(K_{\max}) = 0$. It is obvious that $1 > l(1)/l(0) > \cdots > l(K_{\max})/l(1) = 0$. Moreover, from the assumption in the theorem, we have

$$\log(\sigma_K)/\log(\sigma_1) \gtrsim 1/C. \tag{B.17}$$

From Theorem 2.3 of Bloemendal et al. (2016), for $j \in \{1, \ldots, K\}$, we have

$$|\tilde{\sigma}_j - \phi^{1/2} - \phi^{-1/2} - d_j - d_j^{-1}| \prec \left(1 + \frac{d_j}{1 + \phi^{-1/2}}\right) \min\{p, n\}^{-1/2}.$$

Since $\sigma_j \asymp \phi^{1/2} d_j$, by (B.17) and the assumption that $p \asymp n$, we have

$$\log(\tilde{\sigma}_K + 1)/\log(\tilde{\sigma}_1 + 1) \gtrsim 1/C_2$$

for some $C_2 > 1$. This implies that for $k \in \{1, \ldots, K - 1\}$,

$$\frac{l(k)}{l(0)} - \frac{l(k+1)}{l(0)} = \frac{\log(\tilde{\sigma}_{k+1} + 1)}{l(0)} \gtrsim \frac{1}{K_{\max} C_2}.$$

On the other hand,

$$\frac{l(K)}{l(0)} - \frac{l(K_{\max})}{l(0)} \leq \frac{l(1)}{l(0)} - \frac{l(K_{\max})}{l(0)} = 1 - \frac{\log(\tilde{\sigma}_1 + 1)}{\sum_{j=1}^{K_{\max}} \log(\tilde{\sigma}_j + 1)} < 1 - \frac{1}{K_{\max}}.$$

Hence the proof is complete by Theorem 3.1, taking $\delta = K_{\max} C_2$.

B.2.4 Proof of Corollary 3.2

Recall that

$$\mathrm{SC3}(k) = \frac{\log(1 + p^{-1} \sum_{j=k+1}^{p} \tilde{\sigma}_j^2)}{\log(1 + p^{-1} \sum_{j=1}^{p} \tilde{\sigma}_j^2)} + \mathrm{INS}(k).$$

It is obvious that the first term of SC3 lies between 0 and 1 for $k \in \mathcal{K}$. Note that $\sigma_k^2 \asymp p$ for $k = 1, \ldots, K$. Therefore, by Theorem 2.3 of Bloemendal et al. (2016) and the assumption that $p \asymp n$, we can show that $\tilde{\sigma}_k^2 \asymp p$ for $k = 1, \ldots, K$. Let $L(0) = \log(1 + p^{-1} \sum_{j=1}^{p} \tilde{\sigma}_j^2)$. For $k \in \{1, \ldots, K - 1\}$, we have

$$\begin{aligned}\frac{\log(1 + p^{-1} \sum_{j=k+1}^{p} \tilde{\sigma}_j^2)}{L(0)} - \frac{\log(1 + p^{-1} \sum_{j=k+2}^{p} \tilde{\sigma}_j^2)}{L(0)} &= \frac{1}{L(0)} \log\left(1 + \frac{p^{-1} \tilde{\sigma}_{k+1}^2}{1 + p^{-1} \sum_{j=k+2}^{p} \tilde{\sigma}_j^2}\right) \\ &\geq \frac{1}{L(0)} \log\left(1 + \frac{p^{-1} \tilde{\sigma}_K^2}{1 + p^{-1} \sum_{j=1}^{p} \tilde{\sigma}_j^2}\right) > 0.\end{aligned} \tag{B.18}$$

On the other hand,

$$\begin{aligned}
&\frac{\log(1+p^{-1}\sum_{j=K+1}^{p}\tilde{\sigma}_j^2)}{L(0)}-\frac{\log(1+p^{-1}\sum_{j=K_{\max}+1}^{p}\tilde{\sigma}_j^2)}{L(0)}\\
\leq&\frac{\log(1+p^{-1}\sum_{j=K+1}^{p}\tilde{\sigma}_j^2)}{L(0)}\\
\leq&1-\frac{L(0)-\log(1+p^{-1}\sum_{j=K+1}^{p}\tilde{\sigma}_j^2)}{L(0)}\\
\leq&1-\frac{1}{L(0)}\log\left(1+\frac{p^{-1}\tilde{\sigma}_1^2}{1+p^{-1}\sum_{j=K+1}^{p}\tilde{\sigma}_j^2}\right)<1. \qquad \text{(B.19)}
\end{aligned}$$

Hence by (B.18) and (B.19), the conditions of Theorem 3.1 are satisfied and proof is complete.

Appendix C

Supplementary Materials for Chapter 4

C.1 Proofs

The appendix presents the proofs of the main results. Section C.1.1 provides the proof of Theorem 4.1, while Section C.1.2 provides the proof of Theorem 4.2. Throughout this section, $\delta_0, \delta_1, \dots$ denote positive constants that do not depend on n. For two probability distributions $\mathcal{P}$ and Q on a finite set A, $D(\mathcal{P}\|Q)$ will denote the Kullback-Leibler (KL) divergence,

$$D(\mathcal{P}\|Q) = \sum_{x\in A} \mathcal{P}(x) \log\left(\frac{\mathcal{P}(x)}{Q(x)}\right).$$

C.1.1 Proof of Theorem 4.1

For two scalars $x, z \in [0,1]$, define the Hellinger distance as

$$d_H^2(x,z) = (\sqrt{x} - \sqrt{z})^2 + (\sqrt{1-x} - \sqrt{1-z})^2.$$

For $n \times n$ matrices $X = (x_{ij})_{n\times n}$ and $Z = (z_{ij})_{n\times n}$ where $X, Z \in [0,1]^{n\times n}$, define

$$d_H^2(X,Z) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_H^2(x_{ij}, z_{ij}).$$

It is straightforward to show that $d_H^2(X,Z) \gtrsim \|X-Z\|_F^2/(n^2-n)$. Moreover, let $\|X\|_\infty = \max_{i,j}|x_{ij}|$ denotes the entry-wise infinity norm of X. We will first prove the theorem under an additional constraint that $\|M^*\|_\infty \leq \gamma$ and $\|\hat{M}\|_\infty \leq \gamma$ for some $\gamma > 0$, then send $\gamma \to \infty$ to recover Theorem 4.1. Formally, we prove the following theorem:

Theorem C.1. *Under the conditions in Theorem 4.1, suppose in addition that $\|M^*\|_\infty \leq \gamma$. Let $\hat{M}$ be a solution to* (4.2) *under the additional constraint that $\|\hat{M}\|_\infty \leq \gamma$. Then with probability at least $1 - \delta_1/n$,*

$$d_H^2(\hat{\Pi}, \Pi^*) \leq \delta_2 C_n \sqrt{\frac{1}{p_n n}},$$

where δ_1 and δ_2 are absolute constants.

Proof. Define $\bar{\mathcal{L}}(M) = \mathcal{L}(M) - \mathcal{L}(0_{n\times n})$. The following lemma is essential to proving Theorem C.1:

Lemma C.1. *Under the conditions in Theorem C.1, we have*

$$P\left(\frac{1}{n^2}\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M)-E(\bar{\mathcal{L}}(M))|\geq \delta_0 C_n\sqrt{\frac{Tq_n}{n}}\right)\leq \frac{\delta_1}{n},$$

where δ_0 is an absolute constant, and $\mathcal{G}\subset\mathbb{R}^{n\times n}$ is defined as

$$\mathcal{G}=\{M\in\mathbb{R}^{n\times n}:\|M\|_*\leq C_n n, \|M\|_\infty\leq\gamma, M=-M^\top\}.$$

Before proving the lemma, we first show how Lemma C.1 implies Theorem C.1. For two scalars $x,z\in[0,1]$, we abuse the notation of $D(\cdot\|\cdot)$ and define the divergence measure as

$$D(x\|z)=x\log\left(\frac{x}{z}\right)+(1-x)\log\left(\frac{1-x}{1-z}\right).$$

Similarly, for two matrices $X,Z\in[0,1]^{n\times n}$, define

$$D(X\|Z)=\sum_{i=1}^n\sum_{j=1}^n D(x_{ij}\|z_{ij}).$$

For any choice of $M\in\mathcal{G}$, we have

$$\begin{aligned}
&E(\bar{\mathcal{L}}(M)-\bar{\mathcal{L}}(M^*))\\
=&E(\mathcal{L}(M)-\mathcal{L}(M^*))\\
=&\sum_{i=1}^n\sum_{j>i}E\left(y_{ij}\log\left(\frac{g(m_{ij})}{g(m_{ij}^*)}\right)+(n_{ij}-y_{ij})\log\left(\frac{1-g(m_{ij})}{1-g(m_{ij}^*)}\right)\right)\\
=&\sum_{i=1}^n\sum_{j>i}E\left(n_{ij}g(m_{ij}^*)\log\left(\frac{g(m_{ij})}{g(m_{ij}^*)}\right)+n_{ij}(1-g(m_{ij}^*))\log\left(\frac{1-g(m_{ij})}{1-g(m_{ij}^*)}\right)\right)\\
=&-T\sum_{i=1}^n\sum_{j>i}p_{ij,n}D(g(m_{ij}^*)\|g(m_{ij}))\\
\leq&-0.5Tp_nD(\Pi^*\|\Pi).
\end{aligned}$$

Note that $M^*\in\mathcal{G}$ by assumption. Therefore, for any $M\in\mathcal{G}$, we have

$$\begin{aligned}
\bar{\mathcal{L}}(M)-\bar{\mathcal{L}}(M^*)&=E(\bar{\mathcal{L}}(M)-\bar{\mathcal{L}}(M^*))+(\bar{\mathcal{L}}(M)-E(\bar{\mathcal{L}}(M)))-(\bar{\mathcal{L}}(M^*)-E(\bar{\mathcal{L}}(M^*)))\\
&\leq E(\bar{\mathcal{L}}(M)-\bar{\mathcal{L}}(M^*))+2\sup_{X\in\mathcal{G}}|\bar{\mathcal{L}}(X)-E(\bar{\mathcal{L}}(X))|\\
&\leq -0.5Tp_nD(\Pi^*\|\Pi)+2\sup_{X\in\mathcal{G}}|\bar{\mathcal{L}}(X)-E(\bar{\mathcal{L}}(X))|.
\end{aligned}$$

Moreover, from the definition of $\hat{M}$, we have $\hat{M}\in\mathcal{G}$ and $\mathcal{L}(\hat{M})\geq\mathcal{L}(M^*)$. Therefore, we obtain

$$0\leq -0.5Tp_nD(\Pi^*\|\hat{\Pi})+2\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M)-E(\bar{\mathcal{L}}(M))|.$$

Applying Lemma C.1, then with probability at least $1-\delta_1/n$, we have

$$0\leq\frac{-0.5Tp_nD(\Pi^*\|\hat{\Pi})}{n^2}+2\delta_0C_n\sqrt{\frac{Tq_n}{n}}.$$

This implies that

$$\frac{D(\Pi^*\|\hat{\Pi})}{n^2} \leq \frac{4\delta_0 C_n}{Tp_n}\sqrt{\frac{Tq_n}{n}} \lesssim \frac{4\delta_0 C_n}{\sqrt{Tp_n}}\sqrt{\frac{1}{n}}$$

by Assumption 4.2. Note that $d_H^2(\hat{\Pi}, \Pi^*) \leq n^{-2}D(\Pi^*\|\hat{\Pi})$ by Jensen's inequality combined with the fact that $(1-x) \leq \log(x)$. Hence Theorem C.1 is proved. Theorem 4.1 then follows by the fact that $d_H^2(\hat{\Pi}, \Pi^*) \gtrsim \|\hat{\Pi} - \Pi^*\|_F^2/(n^2 - n)$ and taking the limit as $\gamma \to \infty$. □

We now begin to prove Lemma C.1.

Proof. For any $h > 0$, using Markov's inequality, we have

$$\begin{aligned}
&P\left(\frac{1}{n^2}\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M) - E(\bar{\mathcal{L}}(M))| \geq \delta_0 C_n\sqrt{Tq_n/n}\right) \\
=&P\left(\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M) - E(\bar{\mathcal{L}}(M))|^h \geq \left(\delta_0 C_n n^{1.5}\sqrt{Tq_n}\right)^h\right) \\
\leq&\frac{E\left(\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M) - E(\bar{\mathcal{L}}(M))|^h\right)}{\left(\delta_0 C_n n^{1.5}\sqrt{Tq_n}\right)^h}. \qquad \text{(C.1)}
\end{aligned}$$

The bound in Lemma C.1 will be established by combining (C.1), deriving an upper bound on $E\left(\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M) - E(\bar{\mathcal{L}}(M))|^h\right)$ and setting $h = \log(n)$. Note that we can write $\bar{\mathcal{L}}(M)$ as

$$\bar{\mathcal{L}}(M) = \sum_{i=1}^{n}\sum_{j>i} y_{ij}\log\left(\frac{g(m_{ij})}{g(0)}\right) + (n_{ij} - y_{ij})\log\left(\frac{1-g(m_{ij})}{1-g(0)}\right).$$

By a symmetrization argument (Lemma 6.3 in Ledoux and Talagrand (1991)), we have

$$\begin{aligned}
&E\left(\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M) - E(\bar{\mathcal{L}}(M))|^h\right) \\
\leq&2^h E\left(\sup_{M\in\mathcal{G}}\left|\sum_{i=1}^{n}\sum_{j>i}\epsilon_{ij}\left\{y_{ij}\log\left(\frac{g(m_{ij})}{g(0)}\right) + (n_{ij} - y_{ij})\log\left(\frac{1-g(m_{ij})}{1-g(0)}\right)\right\}\right|^h\right),
\end{aligned}$$

where $\epsilon_{i,j}$ are i.i.d. Rademacher random variables for $i, j = 1, \ldots, n$. To bound the latter term, we apply a contraction principle (Theorem 4.12 in Ledoux and Talagrand (1991)). From the assumption that $\|M\|_\infty \leq \gamma$, conditional on n_{ij}, for $n_{ij} \geq 1$,

$$n_{ij}^{-1}\left(y_{ij}\log\left(\frac{g(m_{ij})}{g(0)}\right) + (n_{ij} - y_{ij})\log\left(\frac{1-g(m_{ij})}{1-g(0)}\right)\right)$$

is a contraction that vanish at 0. Thus, we have

$$\begin{aligned}
E\left(\sup_{M\in\mathcal{G}}|\bar{\mathcal{L}}(M) - E(\bar{\mathcal{L}}(M))|^h\right) \leq&(2^h)(2^h)E\left(\sup_{M\in\mathcal{G}}\left|\sum_{i=1}^{n}\sum_{j>i} n_{ij}\epsilon_{ij}m_{ij}\right|^h\right) \\
=&4^h E\left(\sup_{M\in\mathcal{G}}\left|\sum_{i=1}^{n}\sum_{j>i} n_{ij}\epsilon_{ij}m_{ij}\right|^h\right). \qquad \text{(C.2)}
\end{aligned}$$

To bound $E\left(\sup_{M\in\mathcal{G}}\left|\sum_{i=1}^n\sum_{j>i}n_{ij}\epsilon_{ij}m_{ij}\right|^h\right)$, we apply the skew-symmetric property of M and the fact that $n_{ij}=n_{ji}$ for $i,j\in\{1,\ldots,n\}$. For any $M\in\mathcal{G}$, we have

$$\sum_{i=1}^n\sum_{j=1}^n n_{ij}\epsilon_{ij}m_{ij}=\sum_{i=1}^n\sum_{j>i}n_{ij}(\epsilon_{ij}-\epsilon_{ji})m_{ij}.$$

On the other hand, for $h>1$, by the convexity of $|\cdot|^h$, we have

$$\begin{aligned}\left|\sum_{i=1}^n\sum_{j>i}n_{ij}\epsilon_{ij}m_{ij}\right|^h&=\left|0.5\left\{\sum_{i=1}^n\sum_{j>i}n_{ij}(\epsilon_{ij}-\epsilon_{ji})m_{ij}+\sum_{i=1}^n\sum_{j>i}n_{ij}(\epsilon_{ij}+\epsilon_{ji})m_{ij}\right\}\right|^h\\&\le 0.5\left(\left|\sum_{i=1}^n\sum_{j>i}n_{ij}(\epsilon_{ij}-\epsilon_{ji})m_{ij}\right|^h+\left|\sum_{i=1}^n\sum_{j>i}n_{ij}(\epsilon_{ij}+\epsilon_{ji})m_{ij}\right|^h\right).\end{aligned}$$

Since ϵ_{ji} and $-\epsilon_{ji}$ have identical distribution, after taking expectation, we have

$$\begin{aligned}E\left(\sup_{M\in\mathcal{G}}\left|\sum_{i=1}^n\sum_{j>i}n_{ij}\epsilon_{ij}m_{ij}\right|^h\right)&\le E\left(\sup_{M\in\mathcal{G}}\left|\sum_{i=1}^n\sum_{j>i}n_{ij}(\epsilon_{ij}-\epsilon_{ji})m_{ij}\right|^h\right)\\&=E\left(\sup_{M\in\mathcal{G}}\left|\sum_{i=1}^n\sum_{j=1}^n n_{ij}\epsilon_{ij}m_{ij}\right|^h\right)\\&=E\left(\sup_{M\in\mathcal{G}}|\langle\mathcal{E}\circ\mathcal{N},M\rangle|^h\right).\qquad\text{(C.3)}\end{aligned}$$

Here, $\mathcal{E}=(\epsilon_{ij})_{n\times n}$, $\mathcal{N}=(n_{ij})_{n\times n}$ and $\mathcal{E}\circ\mathcal{N}$ represents the hadamard product between $\mathcal{E}$ and $\mathcal{N}$, and $\langle X,Z\rangle=\sum_{i=1}^n\sum_{j=1}^n x_{ij}z_{ij}$ for any $n\times n$ matrices X and Z. Note that $|\langle X,Z\rangle|\le\|X\|_{op}\|Z\|_*$, where $\|\cdot\|_{op}$ is the Euclidean operator norm. Hence we have

$$\begin{aligned}E\left(\sup_{M\in\mathcal{G}}|\langle\mathcal{E}\circ\mathcal{N},M\rangle|^h\right)&\le E\left(\sup_{M\in\mathcal{G}}\|\mathcal{E}\circ\mathcal{N}\|_{op}^h\|M\|_*^h\right)\\&\le(C_nn)^hE\left(\|\mathcal{E}\circ\mathcal{N}\|_{op}^h\right).\qquad\text{(C.4)}\end{aligned}$$

We can write $\mathcal{E}\circ\mathcal{N}=\sum_{i=1}^n\sum_{j=1}^n\epsilon_{ij}n_{ij}E_{ij}$, where E_{ij} is a $n\times n$ matrix with 1 at the (i,j)th entry and 0 otherwise. Following arguments similar to Section 4.3 of Tropp et al. (2015), and applying Theorem 4.1.1 in Tropp et al. (2015), for $t>0$, we set $s=-t^{2/h}/(2\max_j\{\sum_{i=1}^n n_{ij}^2\})$ such that

$$\begin{aligned}E(\|\mathcal{E}\circ\mathcal{N}\|_{op}^h\mid\mathcal{N})&=\left(\int_0^\infty P(\|\mathcal{E}\circ\mathcal{N}\|_{op}^h\ge t)dt\right)\\&\le\left(\int_0^\infty 2n\exp\left(\frac{-t^{2/h}}{2\max_j\{\sum_{i=1}^n n_{ij}^2\}}\right)dt\right)\\&=\left(\int_0^\infty 2n\left(\frac{h}{2}(2\max_j\{\sum_{i=1}^n n_{ij}^2\})^{h/2}s^{h/2-1}\right)\exp(-s)\,ds\right)\\&=\left((nh)(2\max_j\{\sum_{i=1}^n n_{ij}^2\})^{h/2}\int_0^\infty s^{h/2-1}\exp(-s)\,ds\right)\\&=nh\Gamma(h/2)(2\max_j\{\sum_{i=1}^n n_{ij}^2\})^{h/2},\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. Taking expectation, we have

$$E(\|\mathcal{E}\circ\mathcal{N}\|_{op}^h) \le nh\Gamma(h/2)2^{h/2}E(\max_j\{\sum_{i=1}^n n_{ij}^2\}^{h/2}). \tag{C.5}$$

We aim to find a bound for $E(\max_j\{\sum_{i=1}^n n_{ij}^2\}^{h/2})$. Using Bernstein's inequality, for each j and all $t>0$, we have

$$\begin{aligned}P\left(\left|\sum_{i=1}^n\left(n_{ij}^2-E(n_{ij}^2)\right)\right|>t\right) &\le 2\exp\left(\frac{-t^2/2}{\sum_{i=1}^n\{E(n_{ij}^4)-(E(n_{ij}^2))^2\}+T^2t/3}\right)\\ &\le 2\exp\left(\frac{-t^2/2}{nT^4q_n+T^2t/3}\right).\end{aligned}$$

In particular, for $t\ge 6nT^2q_n$, we have

$$P\left(\left|\sum_{i=1}^n\left(n_{ij}^2-E(n_{ij}^2)\right)\right|>t\right)\le 2\exp\left(-t/T^2\right)=2P(U_j>t/T^2),$$

where $U_1,\ldots,U_n$ are independent and identically distributed exponential random variables. Hence, we have

$$\begin{aligned}&\left(E\left(\max_j\left\{\sum_{i=1}^n n_{ij}^2\right\}^{h/2}\right)\right)^{1/h}\\
=&\left(E\left(\max_j\left|\sum_{i=1}^n n_{ij}^2-E(n_{ij}^2)+E(n_{ij}^2)\right|^{h/2}\right)\right)^{1/h}\\
\le&2\left(E\left(\max_j\left|\sum_{i=1}^n n_{ij}^2-E(n_{ij}^2)\right|^{h/2}\right)\right)^{1/h}+2\left(E\left(\max_j\left|\sum_{i=1}^n E(n_{ij}^2)\right|^{h/2}\right)\right)^{1/h}\\
\le&2\sqrt{nT^2q_n}+2\left(E\left(\max_j\left|\sum_{i=1}^n n_{ij}^2-E(n_{ij}^2)\right|^{h}\right)\right)^{1/2h}\\
=&2\sqrt{nT^2q_n}+2\left(\int_0^\infty P\left(\max_j\left|\sum_{i=1}^n n_{ij}^2-E(n_{ij}^2)\right|^{h}\ge t\right)dt\right)^{1/2h}\\
\le&2\sqrt{nT^2q_n}+2\left\{(6nT^2q_n)^h+\int_{(6nT^2q_n)^h}^\infty P\left(\max_j\left|\sum_{i=1}^n n_{ij}^2-E(n_{ij}^2)\right|^{h}\ge t\right)dt\right\}^{1/2h}\\
\le&2\sqrt{nT^2q_n}+2\left\{(6nT^2q_n)^h+2\int_{(6nT^2q_n)^h}^\infty P\left(\max_j\{U_j\}^h\ge t/T^{2h}\right)dt\right\}^{1/2h}\\
\le&2\sqrt{nT^2q_n}+2\left\{(6nT^2q_n)^h+2E\left(\max_j\{T^2U_j\}^h\right)\right\}^{1/2h}\\
=&2\sqrt{nT^2q_n}+2\left\{(6nT^2q_n)^h+2T^{2h}E\left((\max_j\{U_j\})^h\right)\right\}^{1/2h}.\end{aligned}$$

By standard computations for exponential random variables, we can obtain the inequality $E\left((\max_j\{U_j\})^h\right) \leq 2h! + \log^h(n)$. Thus, we have

$$\begin{aligned}\left(E\left(\max_j\left\{\sum_{i=1}^n n_{ij}^2\right\}^{h/2}\right)\right)^{1/h} &\leq 2\sqrt{nT^2q_n} + 2\left\{(6nT^2q_n)^h + 2T^{2h}(2h! + \log^h(n))\right\}^{1/2h} \\ &\leq 2T(1+\sqrt{6})\sqrt{nq_n} + 2T(2)^{1/2h}(\sqrt{\log(n)} + 2^{1/2h}\sqrt{h}) \\ &\leq 2T(1+\sqrt{6})\sqrt{nq_n} + 2T(2+\sqrt{2})\sqrt{\log(n)}\end{aligned}$$

using the choice $h = \log(n)$ in the final line. Combining this result with (C.5), we have

$$\begin{aligned}E(\|\mathcal{E}\circ\mathcal{N}\|_{op}^h)^{1/h} &\leq (nh\Gamma(h/2))^{1/h}\sqrt{2}\{2T(1+\sqrt{6})\sqrt{nq_n} + 2T(2+\sqrt{2})\sqrt{\log(n)}\} \\ &\leq \delta_3 T\sqrt{nq_n}\end{aligned}$$

for some constant $\delta_3 > 0$ by Assumption 4.2. Combining this with (C.2), (C.3) and (C.4), we obtain

$$E\left(\sup_{M\in\mathcal{G}} |\bar{\mathcal{L}}(M) - E(\bar{\mathcal{L}}(M))|^h\right)^{1/h} \leq (4T)\,(C_n n)(\delta_3)\sqrt{nq_n}.$$

Plugging this into (C.1), the probability in (C.1) is upper bounded by

$$\left\{\frac{(4T)\,(C_n n)(\delta_3)\sqrt{nq_n}}{\delta_0 C_n n^{1.5}\sqrt{Tq_n}}\right\}^h \leq \left(\frac{4\sqrt{T}\delta_3}{\delta_0}\right)^{\log(n)} \leq \frac{\delta_1}{n},$$

provided that $\delta_0 \geq 4\sqrt{T}\delta_3/e$, which establishes the lemma. □

C.1.2 Proof of Theorem 4.2

We first quote the following lemma from Davenport et al. (2014):

Lemma C.2. *Suppose $x, z \in (0,1)$. Then*

$$D(x\|z) \leq \frac{(x-z)^2}{z(1-z)}.$$

The following lemma constructs a packing set $\mathcal{X} \subset \mathcal{K}$ such that, for any distinct $X^{(a)}, X^{(b)} \in \mathcal{X}$, $\|X^{(a)} - X^{(b)}\|_F^2$ is large:

Lemma C.3. *Let $\mathcal{K}$ be defined as in (4.3), and k a positive integer. Let $\gamma \leq 1$ be such that k/γ^2 is an integer, and suppose $k/\gamma^2 \leq n$. Then, there exists a set $\mathcal{X} \subset \mathcal{K}$ satisfying*

$$|\mathcal{X}| \geq \exp\left(\frac{kn}{25600\gamma^2}\right)$$

with the following properties:

1. *For all $X = (x_{ij})_{n\times n} \in \mathcal{X}$, each entry of X satisfies $|x_{ij}| \leq C_n\gamma/\sqrt{2k}$.*
2. *For all $X^{(a)} \neq X^{(b)} \in \mathcal{X}$,*

$$\|X^{(a)} - X^{(b)}\|_F^2 > \frac{C_n^2\gamma^2 n^2}{16k}.$$

Proof. We use a probabilistic argument. The set will be constructed by drawing

$$|\mathcal{X}| = \left\lceil \exp\left(\frac{kn}{25600\gamma^2}\right)\right\rceil$$

matrices independently from the following distribution. Set $B = k/\gamma^2$. Each matrix in $\mathcal{X}$ is constructed of the form $S - S^\top$, where $S = (s_{ij})_{n\times n}$ consists of blocks of dimension $B \times n$, stacked vertically. The entries of the first block are independent and identically distributed symmetric random variables taking values $\pm C_n\gamma/(2\sqrt{2k})$. Then S is filled out by copying this block as many times as it fits. That is,

$$s_{ij} = s_{i^{'}j}, \text{ where } i^{'} = i \pmod B + 1.$$

Now we argue that with nonzero probability, this set will have all the desired properties. For $X \in \mathcal{X}$, it is easy to verify that $X = -X^\top$. Moreover, we have

$$\|X\|_\infty \leq 2\{C_n\gamma/(2\sqrt{2k})\} \leq C_n/\sqrt{2k}.$$

Further, since $\text{rank}(X) \leq 2\text{rank}(S) \leq 2B$,

$$\|X\|_* \leq \sqrt{2B}\|X\|_F \leq \sqrt{2k/\gamma^2}n(C_n\gamma/\sqrt{2k}) = C_n n.$$

Thus $\mathcal{X} \subset \mathcal{K}$, and it remains to show that $\mathcal{X}$ satisfies property 2 in Lemma C.3. Let $p = \lfloor n/B \rfloor$. Consider the submatrix of S containing the first B rows, denoted by $S_{[1:B,:]}$. This can be written as

$$S_{[1:B,:]} = (S_1, S_2, \ldots, S_p, S_{p+1}),$$

where $S_1, \ldots, S_p$ are matrices of dimension $B \times B$, and S_{p+1} accounts for the remaining part of $S_{[1:B,:]}$. If n is divisible by B, then S_{p+1} is an empty matrix. For $X^{(a)} = S^{(a)} - (S^{(a)})^\top$ and $X^{(b)} = S^{(b)} - (S^{(b)})^\top$, drawn from the above distribution, define

$$\Theta_i = \frac{\sqrt{2k}}{C_n\gamma}\left(S_i^{(a)} - S_i^{(b)}\right), \quad \text{for } i = 1, \ldots, p.$$

Each Θ_i is a $B \times B$ matrix, and we write $\Theta_i = (\theta_{i,sl})_{B\times B}$, where each $\theta_{i,sl}$ is independent and identically distributed random variables such that for each $s, l \in \{1, \ldots, B\}$, we have

$$P(\theta_{i,sl} = 1) = P(\theta_{i,sl} = -1) = 0.25 \text{ and } P(\theta_{i,sl} = 0) = 0.5.$$

Hence we can write

$$\begin{aligned}
\|X^{(a)} - X^{(b)}\|_F^2 &\geq \sum_{i=1}^p \sum_{j=1}^p \|S_i^{(a)} - (S_j^{(a)})^\top - S_i^{(b)} + (S_j^{(b)})^\top\|_F^2 \\
&= \frac{C_n^2\gamma^2}{2k} \sum_{i=1}^p \sum_{j=1}^p \|\Theta_i - \Theta_j^\top\|_F^2 \\
&= \frac{C_n^2\gamma^2}{2k} \sum_{i=1}^p \sum_{j=1}^p (\|\Theta_i\|_F^2 + \|\Theta_j^\top\|_F^2 - 2tr(\Theta_i\Theta_j)) \\
&= \frac{C_n^2\gamma^2}{2k} \left\{ 2p\sum_{i=1}^p (\|\Theta_i\|_F^2) - 2tr\left((\sum_{i=1}^p \Theta_i)(\sum_{i=1}^p \Theta_i) \right) \right\}.
\end{aligned}$$

The trace can be expanded as:

$$\begin{aligned}tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) &= \sum_{s=1}^{B}\sum_{k=1}^{B}(\sum_{i=1}^{p}\theta_{i,sl})(\sum_{i=1}^{p}\theta_{i,ls})\\ &= 2\sum_{s=1}^{B}\sum_{k>s}(\sum_{i=1}^{p}\theta_{i,sl})(\sum_{i=1}^{p}\theta_{i,ls}) + \sum_{s=1}^{B}(\sum_{i=1}^{p}\theta_{i,ss})^2.\end{aligned}$$

Hence we can write

$$\begin{aligned}&2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) - 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right)\\ =&2p\sum_{i=1}^{p}\sum_{s=1}^{B}\sum_{k=1}^{B}\theta_{i,sl}^2 - 4\sum_{s=1}^{B}\sum_{k>s}(\sum_{i=1}^{p}\theta_{i,sl})(\sum_{i=1}^{p}\theta_{i,ls}) - 2\sum_{s=1}^{B}(\sum_{i=1}^{p}\theta_{i,ss})^2\\ =&2\sum_{s=1}^{B}\{p\sum_{i=1}^{p}(\theta_{i,ss}^2) - (\sum_{i=1}^{p}\theta_{i,ss})^2\} + 2\sum_{s=1}^{B}\sum_{k>s}\{p\sum_{i=1}^{p}(\theta_{i,sl}^2+\theta_{i,ls}^2) - 2(\sum_{i=1}^{p}\theta_{i,sl})(\sum_{i=1}^{p}\theta_{i,ls})\}.\end{aligned}$$

Taking expectation, we have

$$\begin{aligned}E\left(2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) - 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right)\right) =&2B\{p(0.5p) - 0.5p\} + \frac{2B(B-1)p^2}{2}\\ =&Bp^2 - Bp + B^2p^2 - Bp^2\\ =&B^2p^2 - Bp.\end{aligned}$$

Using the fact that $p = \lfloor n/B \rfloor \geq n/2B$ and $p \leq n/B$, we have

$$\begin{aligned}&P\left(2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) - 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) \leq n^2/8\right)\\ =&P\left(-2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) + 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) + B^2p^2 - Bp \geq -n^2/8 + B^2p^2 - Bp\right)\\ \leq&P\left(-2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) + 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) + B^2p^2 - Bp \geq -n^2/8 + n^2/4 - n\right)\\ \leq&P\left(-2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) + 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) + B^2p^2 - Bp \geq n^2/16\right),\end{aligned}$$

where the last inequality holds as long as $n \geq 16$. Using McDiarmid's inequality, we can obtain the bound

$$\begin{aligned}P\left(2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) - 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) \leq n^2/8\right) &\leq \exp\left(-\frac{2(n^2/16)^2}{\sum_{s=1}^{B}\sum_{k=1}^{B}\sum_{i=1}^{p}(10p)^2}\right)\\ &= \exp\left(-\frac{n^4}{12800B^2p^3}\right)\\ &\leq \exp\left(-\frac{nB}{12800}\right).\end{aligned}$$

Using Union bound, we have that

$$P\left(\min_{X^{(a)}\neq X^{(b)}\in\mathcal{X}} 2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) - 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) \leq n^2/8\right) \leq \binom{|\mathcal{X}|}{2}\exp\left(-\frac{nB}{12800}\right),$$

which is less than 1 given the size of $\mathcal{X}$. Thus the event that

$$2p\sum_{i=1}^{p}(\|\Theta_i\|_F^2) - 2tr\left((\sum_{i=1}^{p}\Theta_i)(\sum_{i=1}^{p}\Theta_i)\right) > n^2/8$$

for all $X^{(a)} \neq X^{(b)} \in \mathcal{X}$ has non-zero probability. In this event,

$$\|X^{(a)} - X^{(b)}\|_F^2 > \frac{C_n^2\gamma^2}{2k}(n^2/8) = \frac{C_n^2\gamma^2 n^2}{16k}.$$

The proof of the lemma is thus complete. □

We now proceed to prove the following theorem, which concerns the lower bound treating n_{ij} as given.

Theorem C.2. *Suppose* $12 \leq C_n^2 \leq \min\{1, \kappa_3^2/T\}n$. *For any given* n_{ij}, $i, j \in \{1, \dots, n\}, j > i$, *consider any algorithm which, for any* $M \in \mathcal{K}$, *takes as input* Y *and returns* $\hat{M}$. *Then there exists* $M \in \mathcal{K}$ *such that with probability at least 3/4,* $\Pi = g(M)$ *and* $\hat{\Pi} = g(\hat{M})$ *satisfy*

$$\frac{1}{n^2-n}\|\Pi - \hat{\Pi}\|_F^2 \geq \min\left\{\kappa_4, \kappa_3 C_n\sqrt{\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}}}\right\}. \tag{C.6}$$

for all $n > N$. *Here* $\kappa_3, \kappa_4 > 0$ *and* N *are absolute constants.*

Proof. Let $c = g^{'}(-1) = g(-1)(1-g(-1))$, and let $c^{'} = g(-1)$. Note that for all $x \in [-1,1]$, we have $g^{'}(x) \geq c$ and $c^{'} \leq g(x) \leq 1 - c^{'}$. We begin by choosing ϵ so that

$$\epsilon^2 = \min\left\{\frac{c}{64}, \kappa_3 C_n\sqrt{\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}}}\right\}, \tag{C.7}$$

where κ_3 is an absolute constant to be determined. Let $k = 6$ and choose γ so that k/γ^2 is an integer and

$$4\sqrt{2}\frac{\epsilon\sqrt{2k}}{C_n c} \leq \gamma \leq \frac{8\epsilon\sqrt{2k}}{C_n c}.$$

This is possible since by assumption $C_n \geq \sqrt{12}$, $\epsilon \leq c/8$ and $c = 0.197$. One can check that γ satisfies the assumptions of Lemma C.3. Note that for $X^{(i)} \neq X^{(j)} \in \mathcal{X}$,

$$\|g(X^{(i)}) - g(X^{(j)})\|_F^2 \geq c^2\|X^{(i)} - X^{(j)}\|_F^2 > c^2 C_n^2\gamma^2 n^2/16k \geq 4\epsilon^2 n^2. \tag{C.8}$$

Now suppose for the sake of a contradiction that there exists an algorithm such that for any $X \in \mathcal{K}$, it returns an $\hat{X}$ such that

$$\frac{1}{n^2}\|g(X) - g(\hat{X})\|_F^2 \leq \epsilon^2. \tag{C.9}$$

with probability at least 1/4. Define

$$X^* = \arg\min_{X^{(a)} \in \mathcal{X}} \frac{1}{n^2} \|g(X^{(a)}) - g(\hat{X})\|_F^2.$$

If (C.9) holds, then (C.8) implies that $X^* = X$. Thus, if (C.9) holds with probability at least 1/4 then

$$P(X \neq X^*) \leq 3/4.$$

However, by a variant of Fano's inequality, we have

$$P(X \neq X^*) \geq 1 - \frac{n^2 \max_{X^{(a)} \neq X^{(b)}} D(Y \mid X^{(a)} \| Y \mid X^{(b)}) + 1}{\log |\mathcal{X}|}. \tag{C.10}$$

Since $y_{ij} + y_{ji} = n_{ij}$ (with n_{ij} given), the value of y_{ji} is determined by y_{ij}. Moreover, y_{ij} are independent for $i = 1, \ldots, n, j > i$. Therefore,

$$D(Y \mid X^{(a)} \| Y \mid X^{(b)}) = \sum_{i=1}^{n} \sum_{j>i} D(y_{ij} \mid x_{ij}^{(a)} \| y_{ij} \mid x_{ij}^{(b)}).$$

Using Lemma C.2, we have

$$\begin{aligned} D(y_{ij} \mid x_{ij}^{(a)} \| y_{ij} \mid x_{ij}^{(b)}) &\leq \frac{(g(C_n\gamma/\sqrt{2k}) - g(-C_n\gamma/\sqrt{2k}))^2}{g(C_n\gamma/\sqrt{2k})(1 - g(C_n\gamma/\sqrt{2k}))} \\ &\leq \frac{4(g^{'}(\xi))^2 C_n^2 \gamma^2/(2k)}{g(C_n\gamma/\sqrt{2k})(1 - g(C_n\gamma/\sqrt{2k}))} \\ &= \frac{4\{g(\xi)(1 - g(\xi))\}^2 C_n^2 \gamma^2/(2k)}{g(C_n\gamma/\sqrt{2k})(1 - g(C_n\gamma/\sqrt{2k}))} \end{aligned}$$

for some $|\xi| \leq C_n\gamma/\sqrt{2k}$. Since $c^{'} < g(x) < 1 - c^{'}$ for $|x| < 1$, $g(\xi) \leq g(C_n\gamma/\sqrt{2k})$, and that

$$C_n\gamma/\sqrt{2k} \leq C_n \frac{8\epsilon\sqrt{2k}}{C_n c\sqrt{2k}} = \frac{8\epsilon}{c} \leq 1,$$

we have

$$D(y_{ij} \mid x_{ij}^{(a)} \| y_{ij} \mid x_{ij}^{(b)}) \leq \frac{4(1 - c^{'})}{c^{'}} \frac{64\epsilon^2}{c^2} = \delta_4 \epsilon^2,$$

where $\delta_4 = 256(1 - c^{'})/(c^{'}c^2)$. Thus, from (C.10), we have

$$\begin{aligned} \frac{1}{4} \leq \frac{\delta_4(\sum_{i=1}^n \sum_{j=1}^n y_{ij})\epsilon^2 + 1}{\log(|\mathcal{X}|)} &\leq \frac{25600\gamma^2}{kn} \{\delta_4(\sum_{i=1}^n \sum_{j=1}^n y_{ij})\epsilon^2 + 1\} \\ &\leq \frac{3276800}{c^2} \epsilon^2 \left(\frac{\delta_4(\sum_{i=1}^n \sum_{j=1}^n y_{ij})\epsilon^2 + 1}{nC_n^2} \right). \end{aligned}$$

We now argue that this leads to a contradiction. Specifically, if $\delta_4(\sum_{i=1}^n \sum_{j=1}^n y_{ij})\epsilon^2 \leq 1$, then together with (C.7) implies that $nC_n^2 \leq 409600/c$. Since $C_n^2 \geq 2k$ by assumption, if we set $N > 204800/(kc)$, this would lead to a contradiction. Thus, suppose now that $\delta_4(\sum_{i=1}^n \sum_{j=1}^n y_{ij})\epsilon^2 > 1$, in which case we have

$$\epsilon^2 \geq \frac{cC_n\sqrt{n}}{5120\sqrt{\delta_4(\sum_{i=1}^n \sum_{j=1}^n y_{ij})}}.$$

Thus setting $\kappa_3 \leq c/(5120\sqrt{\delta_4})$ in (C.7) leads to a contradiction, and hence (C.9) must fail to hold with probability at least 3/4, which completes the proof. □

We now apply Theorem C.2 to prove Theorem 4.2. For any $\epsilon > 0$, Hoeffding's inequality allows us to derive that

$$\begin{aligned}
&P\left(\sqrt{\frac{n}{\sum_{i=1}^n \sum_{j=1}^n y_{ij}}} \geq \epsilon\sqrt{\frac{1}{np_n}}\right) \\
=&P\left(\sum_{i=1}^n \sum_{j=1}^n y_{ij} \leq \frac{n^2p_n}{\epsilon^2}\right) \\
=&1 - P\left(\sum_{i=1}^n \sum_{j>i}(n_{ij} - Tp_{ij,n}) \geq \frac{n^2p_n}{\epsilon^2} - T\sum_{i=1}^n \sum_{j>i} p_{ij,n}\right) \\
\geq&1 - P\left(\sum_{i=1}^n \sum_{j>i}(n_{ij} - Tp_{ij,n}) \geq \frac{n^2p_n}{\epsilon^2} - \frac{Tn(n-1)q_n}{2}\right) \\
\geq&1 - \exp\left(\frac{-2[(n^2p_n/\epsilon^2) - \{Tn(n-1)q_n\}/2]^2}{T^2n(n-1)/2}\right) \\
=&1 - \exp\left(\frac{-\{(2n^2p_n/\epsilon^2) - Tn(n-1)q_n\}^2}{T^2n(n-1)}\right).
\end{aligned}$$

To apply Theorem C.2, it suffices to find ϵ such that

$$1 - \exp\left(\frac{-\{(2n^2p_n/\epsilon^2) - Tn(n-1)q_n\}^2}{T^2n(n-1)}\right) \geq 0.5$$

for sufficiently large n. From Assumption 4.2, we have $p_n \asymp q_n$ and $q_n \gtrsim \log(n)/n$. Consequently, there exists $\delta_5, \delta_6 > 0$ such that $p_n \geq \delta_5 q_n$ and $q_n \geq \delta_6 \log(n)/n$. Taking $\epsilon = \sqrt{\delta_5/T}$, we have

$$\begin{aligned}
P\left(\sqrt{\frac{n}{\sum_{i=1}^n \sum_{j=1}^n y_{ij}}} \geq \sqrt{\frac{\delta_5}{T}}\sqrt{\frac{1}{np_n}}\right) \geq&P\left(\sqrt{\frac{n}{\sum_{i=1}^n \sum_{j=1}^n y_{ij}}} \geq \sqrt{\frac{p_n}{Tq_n}}\sqrt{\frac{1}{np_n}}\right) \\
\geq&1 - \exp\left(-\frac{(2Tn^2q_n - Tn(n-1)q_n)^2}{T^2n(n-1)}\right) \\
=&1 - \exp\left(-\frac{T^2n^2q_n^2(n+1)^2}{T^2n(n-1)}\right) \\
\geq&1 - \exp\left(-q_n^2(n+1)^2\right) \\
\geq&1 - \exp(-\delta_6^2(\log(n))^2) \\
\geq&1/2
\end{aligned}$$

for sufficiently large n. Therefore, the proof of Theorem 4.2 is complete by setting $\kappa_5 = \kappa_3\sqrt{\delta_5/T}$.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Anderson, T. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1956*, pages 111–150. University of California Press.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

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.

Bai, Z., Choi, K. P., and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*, 46(3):1050–1076.

Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148.

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

Benner, P., Byers, R., Fassbender, H., Mehrmann, V., and Watkins, D. (2000). Cholesky-like factorizations of skew-symmetric matrices. *Electronic Transactions on Numerical Analysis*, 11:85–93.

Bertsekas, D. (1999). *Nonlinear Programming.* Athena Scientific.

Birgin, E. G., Martínez, J. M., and Raydan, M. (2000). Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211.

Bloemendal, A., Knowles, A., Yau, H.-T., and Yin, J. (2016). On the principal components of sample covariance matrices. *Probability Theory and Related Fields*, 164(1):459–552.

Bolger, N. and Laurenceau, J.-P. (2013). *Intensive longitudinal methods : An introduction to diary and experience sampling research.* Guilford Press, New York, NY.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Cai, T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14(114):3619–3647.

Cai, T. and Zhou, W.-X. (2016). Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10:1493–1525.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3):412–433.

Cattelan, M., Varin, C., and Firth, D. (2013). Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 62(1):135–150.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Chatterjee, S. and Mukherjee, S. (2019). Estimation in tournaments and graphs under monotonicity constraints. *IEEE Transactions on Information Theory*, 65(6):3525–3539.

Chen, L., Dolado, J. J., and Gonzalo, J. (2021). Quantile factor models. *Econometrica*, 89(2):875–910.

Chen, P., Gao, C., and Zhang, A. Y. (2022a). Partial recovery for top-k ranking: optimality of mle and suboptimality of the spectral method. *The Annals of Statistics*, 50(3):1618–1652.

Chen, S. and Joachims, T. (2016). Modeling intransitivity in matchup and comparison data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 227–236.

Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202.

Chen, X., Chen, Y., and Li, X. (2022b). Asymptotically optimal sequential design for rank aggregation. *Mathematics of Operations Research*, 47(3):2310–2332.

Chen, X., Jiao, K., and Lin, Q. (2016). Bayesian decision process for cost-efficient dynamic ranking via crowdsourcing. *Journal of Machine Learning Research*, 17(216):1–40.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika*, 85(4):1052–1075.

Chen, Y., Fan, J., Ma, C., and Wang, K. (2019a). Spectral method and regularized mle are both optimal for top-k ranking. *The Annals of Statistics*, 47(4):2204.

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. and Li, X. (2024). A note on entrywise consistency for mixed-data matrix completion. *Journal of Machine Learning Research*, 25(343):1–66.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2019b). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, 10:486.

Chen, Y., Li, X., and Zhang, S. (2019c). 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.

Chen, Y. and Suh, C. (2015). Spectral mle: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pages 371–380. PMLR.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Cook, R. J. and Lawless, J. F. (2007). *The statistical analysis of recurrent events*. Springer, New York, NY.

Danziger, S. A., Swamidass, S. J., Zeng, J., Dearth, L. R., Lu, Q., Chen, J. H., Cheng, J., Hoang, V. P., Saigo, H., Luo, R., et al. (2006). Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):114–125.

Davenport, M. A., Plan, Y., Van Den Berg, E., and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223.

Dobriban, E. and Owen, A. B. (2019). Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1):163–183.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 5:17–61.

Fan, J., Lv, J., and Qi, L. (2011). Sparse high-dimensional models in economics. *Annu. Rev. Econ.*, 3(1):291–317.

Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56(3):468–477.

French, S. A., Tangney, C. C., Crane, M. M., Wang, Y., and Appelhans, B. M. (2019). Nutrition quality of food purchases varies by household income: the shopper study. *BMC Public Health*, 19:1–7.

Galvao, A. F. and Kato, K. (2016a). Smoothed quantile regression for panel data. *Journal of Econometrics*, 193(1):92–112.

Galvao, A. F. and Kato, K. (2016b). Smoothed quantile regression for panel data. *Journal of econometrics*, 193(1):92–112.

Gray, S. M. and Brookmeyer, R. (2000). Multidimensional longitudinal data: estimating a treatment effect from continuous, discrete, or time-to-event response variables. *Journal of the American Statistical Association*, 95(450):396–406.

Han, R., Xu, Y., and Chen, K. (2023). A general pairwise comparison model for extremely sparse networks. *Journal of the American Statistical Association*, 118(544):2422–2432.

Han, R., Ye, R., Tan, C., and Chen, K. (2020). Asymptotic theory of sparse bradley–terry model. *The Annals of Applied Probability*, 30(5):2491–2515.

Heckel, R., Shah, N. B., Ramchandran, K., and Wainwright, M. J. (2019). Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126.

Horn, R. A. and Johnson, C. R. (2013). *Matrix analysis.* Cambridge university press.

Hsieh, C.-A., von Eye, A. A., and Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with delinquent peers in the national youth survey. *Multivariate Behavioral Research*, 45(3):508–552.

Ke, Z. T., Ma, Y., and Lin, X. (2023). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. *Journal of the American Statistical Association*, 118(541):374–392.

Lambert, P. and Vandenhende, F. (2002). A copula-based model for multivariate nonnormal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, 21(21):3197–3217.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media.

Lee, S. M., Chen, Y., and Sit, T. (2024). A latent variable approach to learning high-dimensional multivariate longitudinal data. *arXiv preprint arXiv:2405.15053*.

Lian, W., Henao, R., Rao, V., Lucas, J., and Carin, L. (2015). A multitask point process predictive model. In *International Conference on Machine Learning*, pages 2030–2038.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liang, K.-Y. and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association*, 84(406):447–451.

Lim, C. and Yu, B. (2016). Estimation stability with cross-validation (ESCV). *Journal of Computational and Graphical Statistics*, 25(2):464–492.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems*, 23.

Liu, L. C. and Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, 62(1):261–268.

Liu, W., Lin, H., Zheng, S., and Liu, J. (2023a). Generalized factor model for ultra-high dimensional correlated variables with mixed types. *Journal of the American Statistical Association*, 118(542):1385–1401.

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

Lopes, M. E., Blandino, A., and Aue, A. (2019). Bootstrapping spectral statistics in high dimensions. *Biometrika*, 106(4):781–801.

Lyócsa, Š. and Vỳrost, T. (2018). To bet or not to bet: a reality check for tennis betting market efficiency. *Applied Economics*, 50(20):2251–2272.

Macrì Demartino, R., Egidi, L., and Torelli, N. (2024). Alternative ranking measures to predict international football results. *Computational Statistics*, pages 1–19.

McHale, I. and Morton, A. (2011). A bradley-terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619–630.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data.* Springer, New York, NY.

Murnaghan, F. and Wintner, A. (1931). A canonical form for real matrices under orthogonal transformations. *Proceedings of the National Academy of Sciences*, 17(7):417–420.

Negahban, S., Oh, S., and Shah, D. (2017). Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Oliveira, I. F., Ailon, N., and Davidov, O. (2018). A new and flexible approach to the analysis of paired comparison data. *Journal of Machine Learning Research*, 19(60):1–29.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479.

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 54(1):49–78.

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.

Ounajim, A., Slaoui, Y., Louis, P.-Y., Billot, M., Frasca, D., and Rigoard, P. (2023). Mixture of longitudinal factor analyzers and their application to the assessment of chronic pain. *Statistics in Medicine*, 42(18):3259–3282.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239.

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P. (2021). Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048.

Proust-Lima, C., Amieva, H., and Jacqmin-Gadda, H. (2013). Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *British Journal of Mathematical and Statistical Psychology*, 66(3):470–487.

Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A.-L., Migliavacca, E., Binder, H., Michiels, S., Sauerbrei, W., et al. (2023). Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. *BMC medicine*, 21(1):182.

Ramirez, P., Reade, J. J., and Singleton, C. (2023). Betting on a buzz: Mispricing and inefficiency in online sportsbooks. *International Journal of Forecasting*, 39(3):1413–1423.

Rohe, K. and Zeng, M. (2023). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85:1037–1060.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramch, K., and Wainwright, M. J. (2016a). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. J. (2016b). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 63(2):934–959.

Shah, N. B. and Wainwright, M. J. (2018). Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38.

Simons, G. and Yao, Y.-C. (1999). Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27:1041–1060.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Smith, M., Min, A., Almeida, C., and Czado, C. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association*, 105(492):1467–1479.

Sørensen, Ø., Fjell, A. M., and Walhovd, K. B. (2023). Longitudinal modeling of age-dependent latent traits with generalized additive latent and mixed models. *Psychometrika*, 88(2):456–486.

Spearing, H., Tawn, J., Irons, D., and Paulden, T. (2023). Modeling intransitivity in pairwise comparisons with application to baseball data. *Journal of Computational and Graphical Statistics*, 32(4):1383–1392.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Sun, W., Wang, J., and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research*, 14(1):3419–3440.

Sun, W. W., Qiao, X., and Cheng, G. (2016). Stabilized nearest neighbor classifier and its statistical properties. *Journal of the American Statistical Association*, 111(515):1254–1265.

Ten Have, T. R. and Morabia, A. (1999). Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics*, 55(1):85–93.

Thurstone, L. L. (1927). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21:384 – 400.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G., and Király, F. (2019). Modeling outcomes of soccer matches. *Machine Learning*, 108:77–95.

Van Den Berg, E. and Friedlander, M. P. (2008). Probing the pareto frontier for basis pursuit solutions. *Siam Journal on Scientific Computing*, 31(2):890–912.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes : with applications to statistics*. Springer series in statistics. Springer, New York.

Van Der Vaart, A. W., Wellner, J. A., van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Wan, M., Wang, D., Goldman, M., Taddy, M., Rao, J., Liu, J., Lymberopoulos, D., and McAuley, J. (2017). Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1103–1112.

Wan, M., Wang, D., Liu, J., Bennett, P., and McAuley, J. (2018). Representing and recommending shopping baskets with complementarity, compatibility and loyalty. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1133–1142.

Wang, C., Kohli, N., and Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3):455–465.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4):893–904.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. (2014). Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14.

Yan, T., Yang, Y., and Xu, J. (2012). Sparse paired comparisons in the bradley-terry model. *Statistica Sinica*, pages 1305–1318.

Yu, B. (2013). Stability. *Bernoulli*, 19(4):1484–1500.

Yu, B. and Kumbier, K. (2020). Inaugural article by a recently elected academy member: Veridical data science. *Proceedings of the National Academy of Sciences of the United States of America*, 117(8):3920.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zeng, L. and Cook, R. J. (2007). Transition models for multivariate longitudinal binary data. *Journal of the American Statistical Association*, 102(477):211–223.

Zhang, H., Chen, Y., and Li, X. (2020a). A note on exploratory item factor analysis by singular value decomposition. *psychometrika*, 85:358–372.

Zhang, W., Kuang, Z., Peissig, P., and Page, D. (2020b). Adverse drug reaction discovery from electronic health records with deep neural networks. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 30–39.

Zhu, B., Jordan, M., and Jiao, J. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR.

Zhu, Y., Shen, X., and Ye, C. (2016). Personalized prediction and sparsity pursuit in latent factor models. *Journal of the American Statistical Association*, 111(513):241–252.