Determining the Number of Latent Factors in Statistical Multi-Relational Learning

Chengchun Shi cshi4@ncsu.edu
Wenbin Lu lu@stat.ncsu.edu
Rui Song rsong@ncsu.edu
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA

Editor:

Abstract

Statistical relational learning is primarily concerned with learning and inferring relationships between entities in large-scale knowledge graphs. Nickel et al. (2011) proposed a RESCAL tensor factorization model for statistical relational learning, which achieves better or at least comparable results on common benchmark datasets when compared to other state-of-the-art methods. Given a positive integer s, RESCAL computes an s-dimensional latent vector for each entity. The latent factors can be further used for solving relational learning tasks, such as collective classification, collective entity resolution and link-based clustering.

The focus of this paper is to determine the number of latent factors in the RESCAL model. Due to the structure of the RESCAL model, its log-likelihood function is not concave. As a result, the corresponding maximum likelihood estimators (MLEs) may not be consistent. Nonetheless, we design a specific pseudometric, prove the consistency of the MLEs under this pseudometric and establish its rate of convergence. Based on these results, we propose a general class of information criteria and prove their model selection consistencies when the number of relations is either bounded or diverges at a proper rate of the number of entities. Simulations and real data examples show that our proposed information criteria have good finite sample properties.

Keywords: Information criteria; Knowledge graph; Model selection consistency; RESCAL model; Statistical relational learning; Tensor factorization.

1. Introduction

Relational data is becoming ubiquitous in artificial intelligence and social network analysis. These datasets are in the form of graphs, with nodes and edges representing entities and relationships, respectively. Recently, a number of companies have developed and released their knowledge graphs, including the Google Knowledge Graph, Microsoft Bing's Satori Knowledge Base, Yandex's Object Answer, the Linkedln Knowledge Graph, etc. These knowledge graphs are graph-structured knowledge bases that store factual information as relationships between entities. They are created via the automatic extraction of semantic relationships from semi-structured or unstructured text (see Section II.C in Nickel et al., 2016). The data may be incomplete, noisy and contain false information. It is therefore of great importance to infer the existence of a particular relationship to improve the quality of these extracted information.

Statistical relational learning is primarily concerned with learning from relational datasets, and solving tasks such as predicting whether two entities are related (link prediction), identifying equivalent entities (entity resolution), and grouping similar entities based on their relationships (link-based clustering). Statistical relational models can be roughly divided into three categories: the relational graphical models, the latent class models and the tensor factorization models. Relational graphical models include probabilistic relational models (Getoor and Mihalkova, 2011) and Markov logic networks (MLN, Richardson and Domingos, 2006). These models are constructed via Bayesian or Markov networks. In latent class models, each entity is assigned to one of the latent classes and the probability of a relationship between entities depends on their corresponding classes. Two important examples include the stochastic block model (SBM, Nowicki and Snijders, 2001) and the infinite relational model (IRM, Kemp et al., 2006). IRM can be viewed as a nonparametric extension of SBM where the total number of clusters is not prespecified. Both models have received considerable attentions in the statistics and machine learning literature for community detection in networks.

Tensors are multidimensional arrays. Tensor factorization methods such as CANDECOMP/PARAFAC (CP, Harshman and Lundy, 1994), Tucker (Tucker, 1966) and their extensions have found applications in a variety of fields. Kolda and Bader (2009) presented a thorough overview of tensor decompositions and their applications. Recently, tensor factorizations are being actively studied in the statistics literature and have becoming an emerging field of statistics. To name a few, Chi and Kolda (2012) developed a Poisson tensor factorization model for sparse count data. Yang and Dunson (2016) proposed a conditional tensor factorization model for high-dimensional classification with categorical predictors. Sun et al. (2017) proposed a sparse tensor decomposition method by incorporating a truncation step into the tensor power iteration step.

Relational datasets are typically expressed as (subject, predicate, object) triples and can be grouped as a third-order tensor. As a result, tensor factorization methods can be naturally applied to these datasets. Nickel (2013) proposed a RESCAL factorization model for statistical relational learning. Compared to other tensor factorization approaches such as CP and Tucker methods, RESCAL is more capable of detecting the correlations produced between multiple interconnected nodes. For relational data consisting of n entities, K types of relations, and a positive integer s, RESCAL computes an $n \times s$ factor matrix and an $s \times s \times K$ core tensor. The factor matrix and the core tensor can be further used for link prediction, entity resolution and link-based clustering. Nickel et al. (2011) showed that a linear RESCAL model achieved better or comparable results on common benchmark datasets when compared to other existing methods such as MLN, DEDICOM (Harshman, 1978), IRM, CP, MRC (Kok and Domingos, 2007), etc. It was shown in Nickel and Tresp (2013) that a logistic RESCAL model could further improve the link prediction results.

Central to the empirical validity of RESCAL is the correct specification of the number of latent factors. Nickel et al. (2011) proposed to select this parameter via cross-validation. As commonly known for cross-validation methods, there's no theoretical guarantee against overestimation. Besides, cross-validation can be computationally expensive, especially for large n and K. In the literature, model selection is less studied for tensor factorization methods. Allen (2012) and Sun et al. (2017) proposed to use Bayesian information criteria (BIC, Schwarz, 1978) for sparse CP decomposition. However, no theoretical results were provided for BIC. Indeed, we show in this paper that a BIC-type criterion may fail for the RESCAL model.

The contribution of this paper is twofold. First, we propose a general class of information criteria for the RESCAL model and prove their model selection consistency. Although we focus on the RESCAL model, our information criteria can be extended to select models for general tensor factorization methods with slight modification. The problem is nonstandard and challenging since both the factor matrix and the core tensor are not observed and need to be estimated. Besides, the model parameters are non-identifiable. Moreover, the derivation of model/tuning parameter selection consistency of information criteria usually relies on the (uniform) consistency of estimated parameters. For example, Fan and Tang (2013) derived the uniform consistency of the maximum likelihood estimators (MLEs) to prove the consistency of GIC (see Proposition 2 in that paper). Zhang et al. (2016) established the uniform consistency of the support vector machine solutions to prove the consistency of SVMIC$_H$ (see Lemma 2 in that paper). The consistency of these estimators are due to the concavity (convexity) of the likelihood (or the empirical loss) functions. In contrast, for most tensor decomposition models including RESCAL, the likelihood (or the empirical loss) function is usually non-concave (non-convex) and may have multiple local solutions. As a result, the corresponding global maximizer (minimizer) may not be consistent even with the identifiability constraints. It remains unknown how to establish the consistency of the information criterion without consistency of the estimator. A key innovation in our analysis is to design a "proper" pseudometric and show that the global optimum is consistent under this specific pseudometric. We further establish the rate of convergence of the global optimum under this pseudometric as a function of n and K. Based on these results, we establish the consistency of our information criteria when K is either bounded or diverges at a proper rate of n. No parametric assumptions are imposed on the latent factors. Second, we introduce a scalable algorithm for estimating the parameters in the logistic RESCAL model. Despite the fact that a linear RESCAL model can be conveniently solved by an alternating least square algorithm (Nickel et al., 2011), there are lack of optimization algorithms for solving general RESCAL models. The proposed algorithm is based on the alternating direction method of multipliers (ADMM, Boyd et al., 2011) and can be implemented in a parallelized fashion.

The rest of the paper is organized as follows. We formally introduce the RESCAL model and study the parameter identifiability in Section 2. Our information criteria are presented in Section 3 and their model selection properties are investigated. Numerical examples are presented in Section 4 to examine the finite sample performance of the proposed information criteria. Section 5 concludes with a summary and discussion of future extensions. All the proofs are given in the Appendix.

2. The RESCAL model

This section is structured as follows. We introduce the RESCAL model in Section 2.1. In Section 2.2, we study the identifiability of parameters in the model.

2.1 Model setup

In knowledge graphs, facts can be expressed in the form of (subject, predicate, object) triples, where subject and object are entities and predicate is the relation between entities. For example, consider the following sentence from Wikipedia:

Jon Snow is a fictional character in the A Song of Ice and Fire series of fantasy novels by American author George R. R. Martin, and its television adaptation Game of Thrones.

The information contained in this message can be summarized into the following set of (subject, predicate, object) triples:

Subject	*Predicate*	*Object*
Jon Snow	*character in*	*A Song of Ice and Fire*
Jon Snow	*character in*	*Game of Thrones*
A Song of Ice and Fire	*genre*	*novel*
Game of Thrones	*genre*	*television series*
George R.R. Martin	*author of*	*A Song of Ice and Fire*
George R.R. Martin	*profession*	*novelist*

In this example, we have a total of 7 entities, 4 types of relations and 6 triples. More generally, let $\mathcal{E} = \{e_1, \dots, e_n\}$ denote the set of all entities and $\mathcal{R} = \{r_1, \dots, r_K\}$ denote the set of all relation types. The number of relations K is either bounded or diverges with n. Assuming non-existing triples indicate false relationships, we can construct a third-order binary tensor

$$\boldsymbol{Y} = \{Y_{ijk}\}_{i,j\in\{1,\dots,n\},k\in\{1,\dots,K\}},$$

such that

$$Y_{ijk} = \begin{cases} 1, & \text{if a triple } (e_i, r_k, e_j) \text{ exists}, \\ 0, & \text{otherwise}. \end{cases}$$

The RESCAL model is defined as follows. For each entity e_i, a latent vector $\boldsymbol{a}_{i,0} \in \mathbb{R}^{s_0}$ is generated. The Y_{ijk}'s are assumed to be conditionally independent given all latent factors $\{\boldsymbol{a}_{i,0}\}_{i=1}^n$. Besides, it is assumed that

$$\Pr(Y_{ijk} = 1 | \{\boldsymbol{a}_{i,0}\}_{i=1}^n) = \mathrm{g}(\boldsymbol{a}_{i,0}^T \boldsymbol{R}_{k,0} \boldsymbol{a}_{j,0}), \tag{1}$$

for some strictly monotone link function g and $s_0 \times s_0$ matrices $\boldsymbol{R}_{1,0}, \dots, \boldsymbol{R}_{K,0}$. In Model (1), $\boldsymbol{a}_{i,0}$ corresponds to the latent representation of the ith entity and $\boldsymbol{R}_{k,0}$ specifies how these $\boldsymbol{a}_{i,0}$'s interact for the k-th relation. To account for asymmetric relations, we do not restrict $\boldsymbol{R}_{k,0}$'s to symmetric matrices. When the relations are symmetric, i.e.,

$$\Pr(Y_{ijk} = 1 | \{\boldsymbol{a}_{i,0}\}_{i=1}^n) = \Pr(Y_{jik} = 1 | \{\boldsymbol{a}_{i,0}\}_{i=1}^n), \quad \forall i, j, k,$$

one can impose the symmetry constraints and obtain a similar derivation.

For continuous Y_{ijk}, a related tensor factorization model is the TUCKER-2 decomposition, which decomposes the tensor into

$$Y_{ijk} = \boldsymbol{a}_{i,0}^T \boldsymbol{R}_{k,0} \boldsymbol{b}_{j,0} + e_{ijk}, \quad \forall i, j, k, \tag{2}$$

for some $\boldsymbol{a}_{1,0},\dots,\boldsymbol{a}_{n,0}\in\mathbb{R}^{s_1}$, $\boldsymbol{b}_{1,0},\dots,\boldsymbol{b}_{n,0}\in\mathbb{R}^{s_2}$, $\boldsymbol{R}_{1,0},\dots,\boldsymbol{R}_{K,0}\in\mathbb{R}^{s_1\times s_2}$ and some (random) errors $\{e_{ijk}\}_{ijk}$. By (1), RESCAL can be interpreted as a "nonlinear" TUCKER-2 model with the additional constraints that $s_1=s_2=s_0$ and $\boldsymbol{a}_{i,0}=\boldsymbol{b}_{i,0},\forall i$.

CP decomposition is another important tensor factorization method that decomposes a tensor into a sum of rank-1 tensors. It assumes that

$$Y_{ijk}=\sum_{s=1}^{s_0}a_{i,s}b_{j,s}r_{k,s}+e_{ijk},$$

for some $\{a_{i,s}\}_{i,s}$, $\{b_{j,s}\}_{j,s}$, $\{r_{k,s}\}_{k,s}$ and $\{e_{ijk}\}_{ijk}$. Define $\boldsymbol{a}_{i,0}=(a_{i,1},\dots,a_{i,s_0})^T$ and $\boldsymbol{b}_{i,0}=(b_{i,1},\dots,b_{i,s_0})^T$. In view of (2), CP is a special TUCKER-2 model with the constraints that $s_1=s_2=s_0$ and $\boldsymbol{R}_{k,0}=\mathrm{diag}(r_{k,1},\dots,r_{k,s_0})$ where $\mathrm{diag}(r_{k,1},\dots,r_{k,s_0})$ is a diagonal matrix with the sth diagonal elements being $r_{k,s}$.

In this paper, the proposed information criteria are designed in particular for the RESCAL model. However, they can be extended to estimate s_0 in a more general tensor factorization framework including CP and TUCKER-2 models. We discuss this further in Section 5.

2.2 Identifiability

The parameterization in (1) is not identifiable. To see this, for any nonsingular matrix $\boldsymbol{C}\in\mathbb{R}^{s_0\times s_0}$, we define $\boldsymbol{a}_i=\boldsymbol{C}^{-1}\boldsymbol{a}_{i,0}$, $\boldsymbol{R}_k=\boldsymbol{C}^T\boldsymbol{R}_{k,0}\boldsymbol{C}$, $\forall i,k$. Observe that

$$\boldsymbol{a}_{i,0}^T\boldsymbol{R}_{k,0}\boldsymbol{a}_{j,0}=\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j,\quad\forall i,j,k,$$

and hence we have

$$\Pr(Y_{ijk}=1)=\mathrm{g}(\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j).$$

Let $\boldsymbol{A}_0=[\boldsymbol{a}_{1,0},\dots,\boldsymbol{a}_{n,0}]^T$. We impose the following condition.

(A0) (i) Assume $\boldsymbol{A}_0$ has full column rank. (ii) Assume the matrix $(\boldsymbol{R}_{1,0}^T,\dots,\boldsymbol{R}_{K,0}^T)$ has full row rank.

(A0)(i) requires the latent factors to be linearly independent. (A0)(ii) holds when at least one of the $\boldsymbol{R}_{k,0}$'s has full rank. Under Condition (A0), the following lemma states that the RESCAL model is identifiable up to a nonsingular linear transformation. In Section B of the Appendix, we show (A0) is also necessary to guarantee such identifiability when $\boldsymbol{R}_{1,0},\dots,\boldsymbol{R}_{K,0}$ are symmetric.

Lemma 1 (Identifiability)**.** *Assume (A0) holds. Assume there exist some $\{\boldsymbol{a}_i\}_i$, $\{\boldsymbol{R}_k\}_k$ such that $\boldsymbol{a}_i\in\mathbb{R}^{s_0}$, $\boldsymbol{R}_k\in\mathbb{R}^{s_0\times s_0}$ and*

$$g(\boldsymbol{a}_{i,0}^T\boldsymbol{R}_{k,0}\boldsymbol{a}_{j,0})=g(\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j),\quad\forall i,j,k.$$

Then, there exists some invertible matrix $\boldsymbol{C}\in\mathbb{R}^{s_0\times s_0}$ such that

$$\boldsymbol{a}_i=\boldsymbol{C}^{-1}\boldsymbol{a}_{i,0}\ \ \textit{and}\ \ \boldsymbol{R}_k=\boldsymbol{C}^T\boldsymbol{R}_{k,0}\boldsymbol{C}.$$

To fix the nonsingular transformation indeterminacy, we adopt a specific constrained parameterization and focus on estimating $\{\boldsymbol{a}_i^*\}_i$ and $\{\boldsymbol{R}_k^*\}_k$ where

$$\boldsymbol{a}_i^* = (\boldsymbol{A}_{s_0,0}^{-1})^T \boldsymbol{a}_{i,0} \quad \text{and} \quad \boldsymbol{R}_k^* = \boldsymbol{A}_{s_0,0} \boldsymbol{R}_{k,0} \boldsymbol{A}_{s_0,0}^T,$$

where $\boldsymbol{A}_{s_0,0} = [\boldsymbol{a}_{1,0}, \dots, \boldsymbol{a}_{s_0,0}]^T$. Observe that

$$[\boldsymbol{a}_1^*, \dots, \boldsymbol{a}_{s_0}^*] = (\boldsymbol{A}_{s_0,0}^{-1})^T [\boldsymbol{a}_{1,0}, \dots, \boldsymbol{a}_{s_0,0}] = (\boldsymbol{A}_{s_0,0}^{-1})^T \boldsymbol{A}_{s_0,0}^T = \boldsymbol{I}_{s_0},$$

where $\boldsymbol{I}_{s_0}$ stands for an $s_0 \times s_0$ identity matrix. Therefore, the first s_0 $\boldsymbol{a}_i^*$'s are fixed as long as $\boldsymbol{A}_{s_0,0}$ is nonsingular. By Lemma 1, the parameters $\{\boldsymbol{a}_i^*\}_i$ and $\{\boldsymbol{R}_k^*\}_k$ are estimable.

From now on, we only consider the logistic link function for simplicity, i.e, $\mathrm{g}(x) = 1/\{1+\exp(-x)\}$. Results for other link functions can be similarly discussed.

3. Model selection

Parameters $\{\boldsymbol{a}_i^*\}_{i=1}^n$ and $\{\boldsymbol{R}_k^*\}_{k=1}^K$ can be estimated by maximizing the (conditional) log-likelihood function. Since we use the logistic link function, the log-likelihood is equal to

$$\begin{aligned}
\ell_n(\boldsymbol{Y}; \{\boldsymbol{a}_i\}_i, \{\boldsymbol{R}_k\}_k) &= \log\left(\prod_{ijk} \mathrm{g}(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)^{Y_{ijk}} \{1 - \mathrm{g}(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)\}^{1-Y_{ijk}}\right) \\
&= \sum_{ijk} \left(Y_{ijk} \log\{\mathrm{g}(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)\} + (1 - Y_{ijk}) \log\{1 - \mathrm{g}(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)\}\right) \\
&= \sum_{ijk} \left(Y_{ijk} \boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j - \log\{1 + \exp(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)\}\right),
\end{aligned}$$

where the first equality is due to the conditional independence assumption.

For any $s \in \{1, \dots, s_{\max}\}$ where $s_{\max}$ is a bounded integer such that $s_0 \le s_{\max}$, we define the following constrained maximum likelihood estimator

$$(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) = \operatorname*{arg\,max}_{\substack{\boldsymbol{a}_1^{(s)}, \dots, \boldsymbol{a}_n^{(s)} \in \Theta_a^{(s)} \\ \mathrm{vec}(\boldsymbol{R}_1^{(s)}), \dots, \mathrm{vec}(\boldsymbol{R}_K^{(s)}) \in \Theta_r^{(s)}}} \ell_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k), \tag{3}$$

$$\text{subject to} \qquad [\boldsymbol{a}_1^{(s)}, \dots, \boldsymbol{a}_s^{(s)}] = \boldsymbol{I}_s, \tag{4}$$

for some $\Theta_a^{(s)} \subseteq \mathbb{R}^s$, $\Theta_r^{(s)} \subseteq \mathbb{R}^{s^2}$, where the vec$(\cdot)$ operator stacks the entries of a matrix into a column vector. To estimate the number of latent factors, we define the following likelihood-based information criteria

$$\mathrm{IC}(s) = 2\ell_n(\boldsymbol{Y}; \{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - s\kappa(n, K),$$

for some penalty functions $\kappa(\cdot,\cdot)$. The estimated number of latent factors is given by

$$\hat{s} = \operatorname*{arg\,max}_{s \in \{1, \dots, s_{\max}\}} \mathrm{IC}(s). \tag{5}$$

A major technical difficulty in establishing the consistency of IC is due to the nonconcavity of the objective function (3). For any $\{\boldsymbol{a}_j\}_{j\in\{1,\dots,n\}}$, $\{\boldsymbol{R}_k\}_{k\in\{1,\dots,K\}}$, let

$$\boldsymbol{\beta} = (\boldsymbol{a}_1^T, \dots, \boldsymbol{a}_n^T, \text{vec}(\boldsymbol{R}_1)^T, \dots, \text{vec}(\boldsymbol{R}_K)^T)^T,$$

be the set of parameters.

For any $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n \in \mathbb{R}^s$, $\boldsymbol{T}_1, \dots, \boldsymbol{T}_K \in \mathbb{R}^{s\times s}$, we define

$$\boldsymbol{\zeta} = (\boldsymbol{b}_1^T, \dots, \boldsymbol{b}_n^T, \text{vec}(\boldsymbol{T}_1)^T, \dots, \text{vec}(\boldsymbol{T}_K)^T)^T.$$

With some calculations, we can show that

$$\begin{aligned} -\boldsymbol{\zeta}^T \frac{\partial^2 \ell_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \boldsymbol{\zeta} &= \underbrace{\sum_{ijk} \pi_{ijk}(1-\pi_{ijk})(\boldsymbol{b}_i^T \boldsymbol{R}_k \boldsymbol{a}_j + \boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{b}_j + \boldsymbol{a}_i^T \boldsymbol{T}_k \boldsymbol{a}_j)^2}_{I_1} \\ &+ \underbrace{\sum_{ijk} (\pi_{ijk} - Y_{ijk})(2\boldsymbol{b}_i^T \boldsymbol{R}_k \boldsymbol{b}_j + \boldsymbol{b}_i^T \boldsymbol{T}_k \boldsymbol{a}_j + \boldsymbol{a}_i^T \boldsymbol{T}_k \boldsymbol{b}_j),}_{I_2} \end{aligned}$$

where $\pi_{ijk} = \exp(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)/\{1 + \exp(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)\}$. Here, I_1 is nonnegative. However, I_2 can be negative for some $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$. Therefore, the negative Hessian matrix is not positive semidefinite and the likelihood function is not concave. As a result, $\widehat{\boldsymbol{a}}_i^{s_0}$ and $\widehat{\boldsymbol{R}}_k^{s_0}$ may not be consistent to $\boldsymbol{a}_i^*$ and $\boldsymbol{R}_k^*$, even with the identifiability constraints in (4). Here, the presence of I_2 is due to the bilinear formulation of the RESCAL model.

Let $\theta_{ijk} = \boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j$. Note that ℓ_n is concave in θ_{ijk}, $\forall i, j, k$. This motivates us to consider the following pseudometric:

$$\begin{aligned} & d\left(\{\boldsymbol{a}_{i,1}^{(s_1)}\}_i, \{\boldsymbol{R}_{k,1}^{(s_1)}\}_{k_1}; \{\boldsymbol{a}_{i,2}^{(s_2)}\}_i, \{\boldsymbol{R}_{k,2}^{(s_2)}\}_{k_2}\right) \\ = & \left\{ \frac{1}{n^2 K} \sum_{ijk} \left((\boldsymbol{a}_{i,1}^{(s_1)})^T (\boldsymbol{R}_{k,1}^{(s_1)})^T \boldsymbol{a}_{j,1}^{(s_1)} - (\boldsymbol{a}_{i,2}^{(s_2)})^T (\boldsymbol{R}_{k,2}^{(s_2)})^T \boldsymbol{a}_{j,2}^{(s_2)} \right)^2 \right\}^{1/2}, \end{aligned}$$

for any integers s_1, s_2 and $\boldsymbol{a}_{i,1}^{(s_1)} \in \mathbb{R}^{s_1}$, $\boldsymbol{R}_{k,1}^{(s_1)} \in \mathbb{R}^{s_1 \times s_1}$, $\boldsymbol{a}_{i,2}^{(s_2)} \in \mathbb{R}^{s_2}$, $\boldsymbol{R}_{k,2}^{(s_2)} \in \mathbb{R}^{s_2 \times s_2}$. Apparently, $d(\cdot, \cdot)$ is nonnegative, symmetric and satisfies the triangle inequality. Below, we establish the convergence rate of

$$d\left(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k; \{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k\right).$$

We first introduce some notation. For any $s > s_0$, we define

$$\boldsymbol{a}_{i,0}^{(s)} = \begin{cases} ((\boldsymbol{a}_{i,0})^T, \boldsymbol{0}_{s-s_0}^T)^T, & i \notin \{s_0+1, \dots, s\}, \\ ((\boldsymbol{a}_{i,0})^T, \underbrace{0, \dots, 0}_{i-s_0-1}, 1, \underbrace{0, \dots, 0}_{s-i})^T, & i \in \{s_0+1, \dots, s\}, \end{cases}$$

and

$$\boldsymbol{R}_{k,0}^{(s)} = \begin{pmatrix} \boldsymbol{R}_{k,0} & \boldsymbol{O}_{r,s-r} \\ \boldsymbol{O}_{s-r,r} & \boldsymbol{O}_{s-r,s-r} \end{pmatrix},$$

where $\mathbf{0}_q$ denotes a q-dimensional zero vector and $\boldsymbol{O}_{p,q}$ is an $p \times q$ zero matrix. With a slight abuse of notation, we write $\boldsymbol{a}_{i,0}^{(s_0)} = \boldsymbol{a}_{i,0}$ and $\boldsymbol{R}_{k,0}^{(s_0)} = \boldsymbol{R}_{k,0}$. Clearly, for any $s \geq s_0$, we have

$$(\boldsymbol{a}_{i,0}^{(s)})^T \boldsymbol{R}_{k,0}^{(s)} \boldsymbol{a}_{j,0}^{(s)} = \boldsymbol{a}_{i,0}^T \boldsymbol{R}_{k,0} \boldsymbol{a}_{j,0}, \quad \forall i, j, k,$$

and hence

$$(\{\boldsymbol{a}_{i,0}^{(s)}\}_i, \{\boldsymbol{R}_{k,0}^{(s)}\}_k) = \operatorname*{arg\,max}_{\{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k} \mathrm{E}\ell_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k).$$

Let

$$\boldsymbol{a}_i^{(s)*} = (\boldsymbol{A}_{s,0}^{-1})^T \boldsymbol{a}_{i,0}^{(s)} \quad \text{and} \quad \boldsymbol{R}_k^{(s)*} = \boldsymbol{A}_{s,0} \boldsymbol{R}_{k,0} \boldsymbol{A}_{s,0}^T,$$

where $\boldsymbol{A}_{s,0} = [\boldsymbol{a}_{1,0}^{(s)}, \ldots, \boldsymbol{a}_{s,0}^{(s)}]^T$. When $\boldsymbol{A}_{s_0,0}$ is invertible, $\boldsymbol{A}_{s,0}$'s are invertible for all $s > s_0$. The defined $\{\boldsymbol{a}_i^{(s)*}\}$'s satisfy the identifiability constraints in (4) for all $s \geq s_0$. We make the following assumption.

(A1) Assume $\boldsymbol{a}_i^{(s)*} \in \Theta_a^{(s)}$ and $\mathrm{vec}(\boldsymbol{R}_k^{(s)*}) \in \Theta_w^{(s)}$, $\forall i = 1, \ldots, n$, $k = 1, \ldots, K$ and $s_0 \leq s \leq s_{\max}$. Besides, assume $\Theta_a^{(s)}$ and $\Theta_r^{(s)}$ are bounded subsets in $\mathbb{R}^s$ and $\mathbb{R}^{s^2}$, $\forall s \in \{1, \ldots, s_{\max}\}$.

Lemma 2. *Assume (A1) holds. Then for any $s \in \{s_0, \ldots, s_{\max}\}$, we have*

$$n^2 K d^2 \left(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k; \{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k \right) = O_p\left((n+K)(\log n + \log K)\right).$$

In Condition (A1), we require the subsets $\Theta_a^{(s)}$ and $\Theta_r^{(s)}$ to be bounded for $s = 1, \ldots, s_{\max}$. The boundedness assumption can be dropped if we show

$$\max_{\substack{s \in \{1, \ldots, s_{\max}\} \\ i \in \{1, \ldots, n\}}} \|\widehat{\boldsymbol{a}}_i^{(s)}\|_2 = O_p(1) \quad \text{and} \quad \max_{\substack{s \in \{1, \ldots, s_{\max}\} \\ k \in \{1, \ldots, K\}}} \|\widehat{\boldsymbol{R}}_k^{(s)}\|_F = O_p(1).$$

By Lemma 2, when $\log K = o(n^2)$, we have that

$$\frac{(n+K)(\log n + \log K)}{n^2 K} \leq \frac{2n \log n + 2K \log K}{n^2 K} = o(1).$$

Hence, $\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i$ and $\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k$ are consistent under the pseudometric d for all overfitted models. On the contrary, for underfitted models, we require the following conditions.

(A2) Assume there exists some constant $\bar{c} > 0$ such that $\lambda_{\min}(\boldsymbol{A}_0^T \boldsymbol{A}_0) \geq n\bar{c}$.

(A3) Let $\bar{K} = \lambda_{\min}(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T \boldsymbol{R}_{k,0})$. Assume $\liminf_n \bar{K} > 0$.

Lemma 3. *Assume (A2) and (A3) hold. The for any $s \in \{1, 2, \ldots, s_0 - 1\}$, we have*

$$n^2 K d^2 \left(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k; \{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k \right) \geq \bar{c}^2 n^2 \bar{K},$$

where $\bar{c}$ and $\bar{K}$ are defined in (A2) and (A3), respectively.

Assume $\boldsymbol{a}_{i,0}$'s are i.i.d according to some distribution function. By the law of large numbers, we have

$$\frac{1}{n}\boldsymbol{A}_0^T\boldsymbol{A}_0 \xrightarrow{P} \mathrm{E}\boldsymbol{a}_{1,0}\boldsymbol{a}_{1,0}^T = \boldsymbol{\Sigma}_0,$$

for some $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{s_0\times s_0}$. Assume $\boldsymbol{\Sigma}_0$ is positive definite. Then, Assumption (A2) holds for sufficiently large n.

Assumption (A3) holds if there exists some $k_0 \in \{1,\dots,K\}$ such that

$$\liminf_n \lambda_{\min}(\boldsymbol{R}_{k_0,0}\boldsymbol{R}_{k_0,0}^T) > 0.$$

When $\bar{K} \geq c'K$ for some constant $c' > 0$, it follows from Lemma 3 that

$$\liminf_n d\left(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k; \{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k\right) > 0.$$

Based on these results, we establish the consistency of $\hat{s}$ defined in (5) below. For any sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \sim b_n$ if there exist some universal constants $c_1, c_2 > 0$ such that $c_1 a_n \leq b_n \leq c_2 a_n$.

Theorem 3.1. *Assume (A1)-(A3) hold. Assume $K \sim n^{l_0}$ for some $0 \leq l_0 \leq 1$ and $\liminf_n \bar{K} \geq \sqrt{\log n}$ if $l_0 = 1$. Assume $\kappa(n,K)$ satisfies*

$$(n+K)(\log n + \log K) \ll \kappa(n,K) \ll n^2\bar{K}. \tag{6}$$

Then, we have $Pr(\hat{s} = s_0) \to 1$ where $\hat{s}$ is defined in (5).

Let $c(n,K) = \kappa(n,K)(n+K)^{-1}(\log n + \log K)^{-1}$. It follows from Theorem 3.1 that IC is consistent provided that $c(n,K) \to \infty$ and $c(n,K) = o(n\bar{K}/\log n)$. Define

$$\tau_\alpha(n,K) = \frac{(n+K)^\alpha}{\max^\alpha(n,K)},$$

for some $\alpha \geq 0$. Note that

$$1 \leq \tau_\alpha(n,K) \leq 2^\alpha. \tag{7}$$

Consider the following criteria:

$$\mathrm{IC}_\alpha(s) = 2\ell_n(\boldsymbol{Y}; \{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - s\tau_\alpha(n,K)(n+K)(\log n + \log K)\log\{\log(nK)\}. \tag{8}$$

Note that the term $\log\{\log(nK)\}$ satisfies that $\log\{\log(nK)\} \to \infty$ and $\log\{\log(nK)\} = o(n/\log n)$. It follows from (7) and Theorem 3.1 that IC_α is consistent for all $\alpha \geq 0$. When $\alpha > 0$, the term $\tau_\alpha(n,K)$ adjust the model complexity penalty upwards. We notice that Bai and Ng (2002) used a similar finite sample correction term in their proposed information criteria for approximate factor models. Our simulation studies show that such adjustment is essential to achieve selection consistency for large K.

Observe that we have a total of $n \times n \times K = n^2K$ observations. Consider the following BIC-type criterion:

$$\mathrm{BIC}(s) = 2\ell_n(\boldsymbol{Y}; \{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - s\log(n^2K). \tag{9}$$

The model complexity penalty in BIC satisfies

$$\log(n^2K) = 2\log n + \log K \ll (n+K)(\log n + \log K).$$

Hence, it does not meet Condition (6) in Theorem 3.1. As a result, BIC may fail to identity the true model. As shown in our simulation studies, BIC will choose overfitted models and is not selection consistent.

4. Numerical experiments

This section is organized as follows. In Section 4.1, we introduce our algorithm for computing the maximum likelihood estimators of a logistic RESCAL model. Simulation studies are presented in Section 4.2. In Section 4.3, we apply the proposed information criteria to a real dataset.

4.1 Implementation

In this section, we propose an algorithm for computing $\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i$ and $\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k$. The algorithm is based upon a 3-block alternating direction method of multipliers (ADMM). Set $\Theta_a^{(s)} = \mathbb{R}^s$, $\Theta_r^{(s)} = \mathbb{R}^{s\times s}$, $[\boldsymbol{a}_1^{(s)}, \dots, \boldsymbol{a}_s^{(s)}] = \boldsymbol{I}_s$, $\widehat{\boldsymbol{a}}_i^{(s)}$'s and $\widehat{\boldsymbol{R}}_k^{(s)}$'s are defined by

$$(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_{i=(s+1)}^n, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) = \operatorname*{arg\,max}_{\{\boldsymbol{a}_i^{(s)}\}_{i=s+1}^n, \{\boldsymbol{R}_k^{(s)}\}_k} \ell_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k), \tag{10}$$

where

$$\ell_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k) = \sum_{ijk} \left(Y_{ijk}(\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)} - \log[1 + \exp\{(\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)}\}] \right).$$

For any $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n \in \mathbb{R}^s$, define

$$\bar{\ell}_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k, \{\boldsymbol{b}_i^{(s)}\}_i) = \sum_{ijk} \left(Y_{ijk}(\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{b}_j^{(s)} - \log[1 + \exp\{(\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{b}_j^{(s)}\}] \right).$$

Fix $[\boldsymbol{b}_1^{(s)}, \dots, \boldsymbol{b}_s^{(s)}] = \boldsymbol{I}_s$, the optimization problem in (10) is equivalent to

$$\begin{aligned}
(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_{i=s+1}^n, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k, \{\widehat{\boldsymbol{b}}_i^{(s)}\}_{i=s+1}^n) = & \operatorname*{arg\,max}_{\substack{\{\boldsymbol{a}_i^{(s)}\}_{i=s+1}^n, \{\boldsymbol{R}_k^{(s)}\}_k \\ \{\boldsymbol{b}_i^{(s)}\}_{i=s+1}^n}} \bar{\ell}_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k, \{\boldsymbol{b}_i^{(s)}\}_i), \\
& \text{subject to} \qquad \boldsymbol{a}_i^{(s)} = \boldsymbol{b}_i^{(s)}, \forall i = s+1, \dots, n.
\end{aligned}$$

We then derive its augmented Lagrangian, which gives us

$$\begin{aligned}
& L_\rho(\{\boldsymbol{a}_i^{(s)}\}_{i=s+1}^n, \{\boldsymbol{R}_k^{(s)}\}_k, \{\boldsymbol{b}_i^{(s)}\}_{i=s+1}^n, \{\boldsymbol{v}_i^{(s)}\}_{i=s+1}^n) \\
= \quad & -\bar{\ell}_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k, \{\boldsymbol{b}_i^{(s)}\}_i) + \sum_{i=s+1}^n \rho(\boldsymbol{a}_i^{(s)} - \boldsymbol{b}_i^{(s)})^T \boldsymbol{v}_i^{(s)} + \sum_{i=s+1}^n \frac{\rho}{2}\|\boldsymbol{a}_i^{(s)} - \boldsymbol{b}_i^{(s)}\|_2^2,
\end{aligned}$$

where $\rho > 0$ is a penalty parameter and $\boldsymbol{v}_{s+1}^{(s)}, \dots, \boldsymbol{v}_n^{(s)} \in \mathbb{R}^s$.

Applying dual descent method yields the following steps, with l denotes the iteration number:

$$\{\boldsymbol{a}_{i,l+1}^{(s)}\}_{i=s+1}^n = \operatorname*{arg\,min}_{\{\boldsymbol{a}_i^{(s)}\}_{i=s+1}^n} L_\rho(\{\boldsymbol{a}_i^{(s)}\}_{i=s+1}^n, \{\boldsymbol{R}_{k,l}^{(s)}\}_k, \{\boldsymbol{b}_{i,l}^{(s)}\}_{i=s+1}^n, \{\boldsymbol{v}_{i,l}^{(s)}\}_{i=s+1}^n), \quad (11)$$

$$\{\boldsymbol{R}_{k,l+1}^{(s)}\}_{k=1}^K = \operatorname*{arg\,min}_{\{\boldsymbol{R}_k^{(s)}\}_{k=1}^K} L_\rho(\{\boldsymbol{a}_{i,l+1}^{(s)}\}_{i=s+1}^n, \{\boldsymbol{R}_k^{(s)}\}_k, \{\boldsymbol{b}_{i,l}^{(s)}\}_{i=s+1}^n, \{\boldsymbol{v}_{i,l}^{(s)}\}_{i=s+1}^n), \quad (12)$$

$$\{\boldsymbol{b}_{i,l+1}^{(s)}\}_{i=s+1}^n = \operatorname*{arg\,min}_{\{\boldsymbol{b}_i^{(s)}\}_{i=s+1}^n} L_\rho(\{\boldsymbol{a}_{i,l+1}^{(s)}\}_{i=s+1}^n, \{\boldsymbol{R}_{k,l+1}^{(s)}\}_k, \{\boldsymbol{b}_i^{(s)}\}_{i=s+1}^n, \{\boldsymbol{v}_{i,l}^{(s)}\}_{i=s+1}^n), (13)$$

$$\boldsymbol{v}_{i,l+1}^{(s)} = \boldsymbol{v}_{i,l}^{(s)} + \boldsymbol{a}_{i,l}^{(s)} - \boldsymbol{b}_{i,l}^{(s)}, \qquad \forall i = s+1, \dots, n.$$

Let us examine (11)-(13) in more details. In (11), we rewrite the objective function as

$$\begin{aligned} & L_\rho(\{\boldsymbol{a}_i^{(s)}\}_{i=s+1}^n, \{\boldsymbol{R}_{k,l}^{(s)}\}_k, \{\boldsymbol{b}_{i,l}^{(s)}\}_{i=s+1}^n, \{\boldsymbol{v}_{i,l}^{(s)}\}_{i=s+1}^n) \\ = & \sum_{i=s+1}^n \left\{ \sum_{j,k} \left(\log[1 + \exp\{(\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_{k,l}^{(s)} \boldsymbol{b}_{j,l}^{(s)}\}] - Y_{ijk} (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_{k,l}^{(s)} \boldsymbol{b}_{j,l}^{(s)} \right) \right. \\ + & \left. \rho(\boldsymbol{a}_i^{(s)} - \boldsymbol{b}_{i,l}^{(s)})^T \boldsymbol{v}_{i,l}^{(s)} + \frac{\rho}{2} \|\boldsymbol{a}_i^{(s)} - \boldsymbol{b}_{i,l}^{(s)}\|_2^2 \right\}. \end{aligned}$$

Note that L_ρ can be represented as a separable sum of functions. As a result, $\boldsymbol{a}_{i,l+1}^{(s)}$'s can be solved in parallel. More specifically, we have

$$\begin{aligned} \boldsymbol{a}_{i,l+1}^{(s)} = & \operatorname*{arg\,min}_{\boldsymbol{a}^{(s)}} \left\{ \sum_{j,k} \left(\log[1 + \exp\{(\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_{k,l}^{(s)} \boldsymbol{b}_{j,l}^{(s)}\}] - Y_{ijk} (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_{k,l}^{(s)} \boldsymbol{b}_{j,l}^{(s)} \right) \right. \\ + & \left. \rho(\boldsymbol{a}_i^{(s)} - \boldsymbol{b}_{i,l}^{(s)})^T \boldsymbol{v}_{i,l}^{(s)} + \frac{\rho}{2} \|\boldsymbol{a}_i^{(s)} - \boldsymbol{b}_{i,l}^{(s)}\|_2^2 \right\}. \end{aligned}$$

Hence, each $\boldsymbol{a}_{i,l+1}^{(s)}$ can be computed by solving a ridge type logistic regression with responses $\{Y_{ijk}\}_{j,k}$ and covariates $\{\boldsymbol{R}_{k,l}^{(s)} \boldsymbol{b}_{j,l}^{(s)}\}_{j,k}$.

In (12), each $\boldsymbol{R}_{k,l+1}^{(s)}$ can be independently updated by solving a logistic regression with responses $\{Y_{ijk}\}_{i,j}$ and covariates $\boldsymbol{b}_{j,l}^{(s)} \otimes \boldsymbol{a}_{i,l+1}^{(s)}$, i.e,

$$\operatorname{vec}(\boldsymbol{R}_{k,l+1}^{(s)}) = \operatorname*{arg\,min}_{\boldsymbol{r}_k^{(s)} \in \mathbb{R}^{s^2}} \sum_{ij} \left\{ \log\left(1 + \exp[\{(\boldsymbol{b}_{j,l}^{(s)})^T \otimes (\boldsymbol{a}_{i,l+1}^{(s)})^T\} \boldsymbol{r}_k^{(s)}]\right) - Y_{ijk} \{(\boldsymbol{b}_{j,l}^{(s)})^T \otimes (\boldsymbol{a}_{i,l+1}^{(s)})^T\} \boldsymbol{r}_k^{(s)} \right\},$$

where $\otimes$ denotes the Kronecker product.

Similar to (11), each $\boldsymbol{b}_{i,l+1}^{(s)}$ in (13) can be independently computed by solving a ridge type regression with responses $\{Y_{ijk}\}_{j,k}$ and covariates $\{(\boldsymbol{R}_{k,l+1}^{(s)})^T \boldsymbol{a}_{j,l+1}^{(s)}\}_{j,k}$.

Using similar arguments in Theorem 2 in Wang et al. (2017), we can show that the proposed 3-block ADMM algorithm converges for any sufficiently large ρ. In our implementation, we set $\rho = nK/2$. To guarantee global convergence, we randomly generate multiple initial estimators and solve the optimization problem multiple times based on these initial values.

Table 1: Simulation results for Setting I, II and III (standard errors in parenthesis)

	$s_0 = 2$		$s_0 = 4$		$s_0 = 8$	
$n = 100, K = 3$	TP	$\hat{s}$	TP	$\hat{s}$	TP	$\hat{s}$
IC_0	0.97 (0.02)	2.03 (0.02)	0.97 (0.02)	4.03 (0.02)	0.90(0.03)	7.90 (0.03)
$\text{IC}_{0.5}$	0.97 (0.02)	2.03 (0.02)	0.98 (0.01)	4.02 (0.01)	0.90(0.03)	7.90 (0.03)
IC_1	0.97 (0.02)	2.03 (0.02)	0.98 (0.01)	4.02 (0.01)	0.89(0.03)	7.89 (0.03)
BIC	0.00 (0.00)	11.99 (0.01)	0.00 (0.00)	12.00 (0.00)	0.00 (0.00)	11.99 (0.01)
$n = 150, K = 3$	TP	$\hat{s}$	TP	$\hat{s}$	TP	$\hat{s}$
IC_0	0.99 (0.01)	2.01 (0.01)	0.97 (0.02)	4.03 (0.02)	0.96(0.02)	8.04 (0.02)
$\text{IC}_{0.5}$	0.99 (0.01)	2.01 (0.01)	0.97 (0.02)	4.03 (0.02)	0.96(0.02)	8.04 (0.02)
IC_1	0.99 (0.01)	2.01 (0.01)	0.97 (0.02)	4.03 (0.02)	0.96(0.02)	8.04 (0.02)
BIC	0.00 (0.00)	12.00 (0.00)	0.00 (0.00)	12.00 (0.00)	0.00 (0.00)	11.98 (0.01)
$n = 200, K = 3$	TP	$\hat{s}$	TP	$\hat{s}$	TP	$\hat{s}$
IC_0	0.99 (0.01)	2.01 (0.01)	0.95 (0.02)	4.05 (0.02)	0.95(0.02)	8.05 (0.02)
$\text{IC}_{0.5}$	0.99 (0.01)	2.01 (0.01)	0.95 (0.02)	4.05 (0.02)	0.95(0.02)	8.05 (0.02)
IC_1	0.99 (0.01)	2.01 (0.01)	0.95 (0.02)	4.05 (0.02)	0.95(0.02)	8.05 (0.02)
BIC	0.00 (0.00)	12.00 (0.00)	0.00 (0.00)	11.99 (0.01)	0.00 (0.00)	11.98 (0.01)

4.2 Simulations

We simulate $\{Y_{ijk}\}_{ijk}$ from the following model:

$$\begin{aligned}
&\Pr(Y_{ijk} = 1 | \{\boldsymbol{a}_i\}_i, \{\boldsymbol{R}_k\}_k) = \frac{\exp(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)}{1 + \exp(\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j)}, \\
&\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \overset{iid}{\sim} N(0, 1), \\
&\boldsymbol{R}_1 = \boldsymbol{R}_2 = \cdots = \boldsymbol{R}_K = \text{diag}(\underbrace{1, -1, 1, -1, \ldots, 1, -1}_{r}),
\end{aligned}$$

where $N(0, 1)$ stands for a standard normal random variable and $\text{diag}(v_1, \ldots, v_q)$ denotes a $q \times q$ diagonal matrix with the jth element equal to v_j.

We consider six simulation settings. In the first three settings, we fix $K = 3$ and set $n = 100, 150$ and 200, respectively. In the last three settings, we increase K to 10, 20, 50, and set $n = 50$. In each setting, we further consider three scenarios, by setting $s_0 = 2, 4$ and 8. Let $s_{\max} = 12$. The ADMM algorithm proposed in Section 4.1 is implemented in R. Some subroutines of the algorithm are written in C with the GNU Scientific Library (GSL, Galassi et al., 2015) to facilitate the computation. We compare the proposed IC_α (8) with the BIC-type criterion (9). In IC_α, we set $\alpha = 0, 0.5$ and 1. Note that when $\alpha = 0$, we have

$$\tau_\alpha(n, K) = \frac{(n + K)^\alpha}{\max^\alpha(n, K)} = 1.$$

Reported in Table 1 and 2 are the percentage of selecting the true models (TP) and the average of $\hat{s}$ selected by IC_0, $\text{IC}_{0.5}$, IC_1 and BIC over 100 replications.

It can be seen from Table 1 and 2 that BIC fails in all settings. It always selects overfitted models. On the contrary, the proposed information criteria are consistent for most of the

Table 2: Simulation results for Setting IV, V and VI (standard errors in parenthesis)

	$s_0 = 2$		$s_0 = 4$		$s_0 = 8$	
$n = 50, K = 10$	TP	$\hat{s}$	TP	$\hat{s}$	TP	$\hat{s}$
IC_0	1.00 (0.00)	2.00 (0.00)	0.97 (0.02)	4.03 (0.02)	0.69(0.05)	7.91 (0.06)
$\text{IC}_{0.5}$	1.00 (0.00)	2.00 (0.00)	0.97 (0.02)	4.03 (0.02)	0.66(0.05)	7.75 (0.06)
IC_1	1.00 (0.00)	2.00 (0.00)	0.98 (0.01)	4.02 (0.01)	0.60(0.05)	7.62 (0.06)
BIC	0.00 (0.00)	11.81 (0.06)	0.00 (0.00)	11.60 (0.06)	0.01 (0.01)	11.67 (0.07)
$n = 50, K = 20$	TP	$\hat{s}$	TP	$\hat{s}$	TP	$\hat{s}$
IC_0	0.97 (0.02)	2.03 (0.02)	0.95 (0.02)	4.05 (0.02)	0.73(0.04)	8.46 (0.10)
$\text{IC}_{0.5}$	0.97 (0.02)	2.03 (0.02)	0.98 (0.01)	4.02 (0.01)	0.87(0.03)	8.09 (0.03)
IC_1	0.98 (0.01)	2.02 (0.02)	1.00 (0.00)	4.00 (0.00)	0.79(0.04)	7.99 (0.05)
BIC	0.00 (0.00)	12.00 (0.00)	0.00 (0.00)	11.92 (0.03)	0.00 (0.00)	11.99 (0.01)
$n = 50, K = 50$	TP	$\hat{s}$	TP	$\hat{s}$	TP	$\hat{s}$
IC_0	0.98 (0.01)	2.02 (0.01)	0.93 (0.03)	4.07 (0.03)	0.17(0.04)	11.24 (0.15)
$\text{IC}_{0.5}$	0.99 (0.01)	2.01 (0.01)	0.97 (0.02)	4.03 (0.02)	0.76(0.04)	8.24 (0.05)
IC_1	1.00 (0.00)	2.00 (0.00)	0.98 (0.01)	4.02 (0.01)	0.79(0.04)	7.99 (0.05)
BIC	0.00 (0.00)	12.00 (0.00)	0.00 (0.00)	12.00 (0.00)	0.00 (0.00)	11.99 (0.01)

settings. For example, under settings where $s_0 = 2$ or 4, TPs of IC_0, $\text{IC}_{0.5}$ and IC_1 are larger than or equal to 93%. When $s_0 = 8$, expect for the last setting, TPs of the proposed information criteria are no less than 60% for all cases.

IC_0, $\text{IC}_{0.5}$ and IC_1 perform very similarly for small K. In the first three settings, TPs of these three information criteria are nearly the same for all cases. However, $\text{IC}_{0.5}$ and IC_1 are more robust than IC_0 for large K. This can be seen in the last scenario of Setting 6, where the TP of IC_0 is no more than 20%. Besides, in the last two settings, TP of IC_0 is smaller than $\text{IC}_{0.5}$ and IC_1 for all cases. These differences are due to the finite sample correction term $\tau_\alpha(n, K)$. As commented before, $\tau_{0.5}(n, K)$ and $\tau_1(n, K)$ increase the model complexity penalty term in $\text{IC}_{0.5}$ and IC_1 to avoid overfitting for large K.

4.3 Real data experiments

In this section, we apply the proposed information criteria to the "Social Evolution" dataset (Madan et al., 2012). This dataset comes from MIT's Human Dynamics Laboratory. It tracks everyday life of a whole undergraduate MIT dormitory from October 2008 to May 2009. We use the survey data, resulting in $n = 84$ participants and $K = 5$ binary relations. The five relations are: close relationship, political discussion, social interaction and two social media interaction.

We compute $\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i$ and $\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k$ for $s = \{1, \ldots, 12\}$ and select the number of latent factors using the proposed information criteria (8) and BIC (9). It turns out that IC_0, $\text{IC}_{0.5}$ and IC_1 all suggest the presence of 9 factors. In contrast, BIC selects 12 factors. To further evaluate the number of latent factors selected by the proposed information criteria, we consider the following cross-validation procedure. For any $s \in [1, \ldots, 12]$, we randomly select 80% of the observations and estimate $\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i$ and $\{\widehat{\boldsymbol{R}}_k^{(s)}\}$ by maximizing the observed

Table 3: AUC scores

s	1	2	3	4	5	6	7	8	9	10	11	12
AUC	0.7201	0.8341	0.8952	0.9095	0.9257	0.9364	0.9444	0.9486	0.9513	0.9518	0.9485	0.9467

likelihood function based on these training samples. Then we compute

$$\widehat{\pi}_{ijk} = \frac{\exp\{(\widehat{\boldsymbol{a}}_i^{(s)})^T \widehat{\boldsymbol{R}}_k^{(s)} \widehat{\boldsymbol{a}}_j^{(s)}\}}{1+\exp\{(\widehat{\boldsymbol{a}}_i^{(s)})^T \widehat{\boldsymbol{R}}_k^{(s)} \widehat{\boldsymbol{a}}_j^{(s)}\}}.$$

Based on these predicted probabilities, we calculate the area under the precision-recall curve (AUC) on the remaining 20% testing samples.

Reported in Table 3 are the AUC scores averaged over 100 replications. For any $s \in [1, \dots, 12]$, we denoted by AUC_s the corresponding AUC score. It can be seen from Table 3 that AUC_s first increases and then decreases as s increases. The maximum AUC score is achieved at $s = 10$. Observe that AUC_9 is very close to AUC_{10}, and it is larger than the remaining AUC scores. This demonstrates that the proposed information criteria select less latent factors while achieve better or similar link prediction results when compared to BIC.

5. Discussion

In this paper, we propose information criteria for selecting the number of latent factors in the RESCAL tensor factorization model and prove their model selection consistency. Although we focus on the logistic RESCAL model, the proposed information criteria can be applied to general tensor factorization models. More specifically, consider the following class of models:

$$Y_{ijk} = \mathrm{g}(\boldsymbol{a}_{i,0}^T \boldsymbol{R}_{k,0} \boldsymbol{b}_{j,0}) + e_{ijk}, \qquad \forall i, j \in \{1, \dots, n\}, k \in \{1, \dots, K\}, \tag{14}$$

with any of (or without) the following constraints:
(C1) $\boldsymbol{R}_{k,0}$ is diagonal;
(C2) $\boldsymbol{a}_{i,0} = \boldsymbol{b}_{i,0}$ for $i \in \{1, \dots, n\}$,
for some strictly increasing function g, $\boldsymbol{a}_{i,0}, \boldsymbol{b}_{i,0} \in \mathbb{R}^{s_0}$, $\boldsymbol{R}_{k,0} \in \mathbb{R}^{s_0 \times s_0}$ and some mean zero random errors $\{e_{ijk}\}_{ijk}$.

As commented in Section 2.1, such representation includes the RESCAL, CP and TUCKER-2 models. Specifically, it reduces to the TUCKER-2 model by setting g to be the identity function. If further (C1) holds, then (14) reduces to the CP model. When (C2) holds, (14) corresponds to the RESCAL model. Consider the following information criteria,

$$\text{IC}(s) = \ell_n(\{Y_{ijk}\}_{ijk}; \{\widehat{\boldsymbol{a}}_i\}_i, \{\widehat{\boldsymbol{R}}_k\}_k, \{\widehat{\boldsymbol{b}}_i\}_i) - s\kappa(n, K),$$

where ℓ_n stands for the likelihood function and $\widehat{\boldsymbol{a}}_i$, $\widehat{\boldsymbol{R}}_k$, $\widehat{\boldsymbol{b}}_i$ are the corresponding (constrained) MLEs. Similar to Theorem 3.1, we can show that with some properly chosen $\kappa(n, K)$, IC is consistent under this general setting.

Currently, we assume the tensor $\boldsymbol{Y}$ is completely observed. When some of the Y_{ijk}'s are missing, we can calculate $\widehat{\boldsymbol{a}}_i^{(s)}$'s and $\widehat{\boldsymbol{R}}_k^{(s)}$'s by maximizing the following observed likelihood function

$$\begin{aligned}\mathop{\arg\max}_{\substack{\boldsymbol{a}_1^{(s)},\dots,\boldsymbol{a}_n^{(s)}\in\Theta_a^{(s)}\\ \mathrm{vec}(\boldsymbol{R}_1^{(s)}),\dots,\mathrm{vec}(\boldsymbol{R}_K^{(s)})\in\Theta_r^{(s)}}} & \sum_{(i,j,k)\in N_{obs}}\left(Y_{ijk}(\boldsymbol{a}_i^{(s)})^T\boldsymbol{R}_k^{(s)}\boldsymbol{a}_j^{(s)}-\log[1+\exp\{(\boldsymbol{a}_i^{(s)})^T\boldsymbol{R}_k^{(s)}\boldsymbol{a}_j^{(s)}\}]\right),\\ \text{subject to} \qquad & [\boldsymbol{a}_1^{(s)},\dots,\boldsymbol{a}_s^{(s)}]=\boldsymbol{I}_s,\end{aligned}$$

where N_{obs} denotes the set of the observed responses. The above optimization problem can also be solved by a 3-block ADMM algorithm. Define the following class of information criteria,

$$\mathrm{IC}_{obs}(s)=\sum_{(i,j,k)\in N_{obs}}\left(Y_{ijk}(\widehat{\boldsymbol{a}}_i^{(s)})^T\widehat{\boldsymbol{R}}_k^{(s)}\widehat{\boldsymbol{a}}_j^{(s)}-\log[1+\exp\{(\widehat{\boldsymbol{a}}_i^{(s)})^T\widehat{\boldsymbol{R}}_k^{(s)}\widehat{\boldsymbol{a}}_j^{(s)}\}]\right)-\hat{p}s\kappa(n,K),$$

where $\hat{p}=|N_{obs}|/(n^2K)$ denotes the percentage of observed responses. Consistency of IC_{obs} can be similarly studied.

Appendix A. Proofs

In the following, we provide proofs of Lemma 1, Lemma 2 and Theorem 3.1. We define

$$\ell_0(\{\boldsymbol{a}_i\}_i,\{\boldsymbol{R}_k\}_k)=\mathrm{E}\ell_n(\boldsymbol{Y};\{\boldsymbol{a}_i\}_i,\{\boldsymbol{R}_k\}_k),$$

for any $\boldsymbol{a}_1,\dots,\boldsymbol{a}_n\in\mathbb{R}^s$, $\boldsymbol{R}_1,\dots,\boldsymbol{R}_K\in\mathbb{R}^{s\times s}$ and any integer $s\geq 1$. For any q-dimensional vector $\boldsymbol{q}$, let $\|\boldsymbol{q}\|_2$ denote its Euclidean norm. For any $m\times q$ matrix $\boldsymbol{Q}$, $\|\boldsymbol{Q}\|_2$ stands for the spectral norm of $\boldsymbol{Q}$ while $\|\boldsymbol{Q}\|_F$ denotes its Frobenius norm. Define $\theta_{ijk}=(\boldsymbol{a}_i^*)^T\boldsymbol{R}_k\boldsymbol{a}_j^*$ and $\hat{\theta}_{ijk}=\widehat{\boldsymbol{a}}_i^T\widehat{\boldsymbol{R}}_k\widehat{\boldsymbol{a}}_j$.

A.1 Proof of Lemma 1

Assume there exists some $\{\boldsymbol{a}_i\}_i$, $\{\boldsymbol{R}_k\}_k$ such that

$$\mathrm{g}(\boldsymbol{a}_{i,0}^T\boldsymbol{R}_{k,0}\boldsymbol{a}_{j,0})=\mathrm{g}(\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j),\quad \forall i,j,k.$$

Since $\mathrm{g}(\cdot)$ is strictly monotone, we have

$$\boldsymbol{a}_{i,0}^T\boldsymbol{R}_{k,0}\boldsymbol{a}_{j,0}=\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j,\quad \forall i,j,k,$$

or equivalently,

$$\begin{pmatrix}\boldsymbol{A}\boldsymbol{R}_1\\ \vdots\\ \boldsymbol{A}\boldsymbol{R}_K\end{pmatrix}\boldsymbol{A}^T=\begin{pmatrix}\boldsymbol{A}_0\boldsymbol{R}_{1,0}\\ \vdots\\ \boldsymbol{A}_0\boldsymbol{R}_{K,0}\end{pmatrix}\boldsymbol{A}_0^T.$$

where $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\dots,\boldsymbol{a}_n]^T$. Thus, it follows that

$$\begin{pmatrix}\boldsymbol{A}_0^T\boldsymbol{A}\boldsymbol{R}_1\\ \vdots\\ \boldsymbol{A}_0^T\boldsymbol{A}\boldsymbol{R}_K\end{pmatrix}\boldsymbol{A}^T=\begin{pmatrix}\boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{R}_{1,0}\\ \vdots\\ \boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{R}_{K,0}\end{pmatrix}\boldsymbol{A}_0^T.$$

By (A0), the matrix $\boldsymbol{A}_0^T\boldsymbol{A}_0$ is invertible. As a result, we have

$$\begin{pmatrix} (\boldsymbol{A}_0^T\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^T\boldsymbol{A}\boldsymbol{R}_1 \\ \vdots \\ (\boldsymbol{A}_0^T\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^T\boldsymbol{A}\boldsymbol{R}_K \end{pmatrix} \boldsymbol{A}^T = \begin{pmatrix} \boldsymbol{R}_{1,0} \\ \vdots \\ \boldsymbol{R}_{K,0} \end{pmatrix} \boldsymbol{A}_0^T.$$

In addition, the matrix $\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{R}_{k,0}$ is invertible under Condition (A0). Similarly, we can show

$$\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T(\boldsymbol{A}_0^T\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^T\boldsymbol{A}\boldsymbol{R}_1\right)\boldsymbol{A}^T = \left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{R}_{k,0}\right)\boldsymbol{A}_0^T,$$

and hence

$$\underbrace{\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{R}_{k,0}\right)^{-1}\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T(\boldsymbol{A}_0^T\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^T\boldsymbol{A}\boldsymbol{R}_k\right)}_{\boldsymbol{C}}\boldsymbol{A}^T = \boldsymbol{A}_0^T.$$

By Lemma 5.1 in Banerjee and Roy (2014), we have $\text{rank}(\boldsymbol{C}) \geq \text{rank}(\boldsymbol{A}_0) = s_0$. Therefore, $\boldsymbol{C}$ is invertible. It follows that

$$\boldsymbol{A} = \boldsymbol{A}_0(\boldsymbol{C}^T)^{-1}, \tag{15}$$

or equivalently,

$$\boldsymbol{a}_i = \boldsymbol{C}^{-1}\boldsymbol{a}_{i,0}, \quad \forall i = 1, \dots, n.$$

By (15), we obtain $\boldsymbol{A}_0(\boldsymbol{C}^T)^{-1}\boldsymbol{R}_k\boldsymbol{C}^{-1}\boldsymbol{A}_0^T = \boldsymbol{A}_0\boldsymbol{R}_{k,0}\boldsymbol{A}_0^T, \forall k$, and hence

$$\boldsymbol{A}_0^T\boldsymbol{A}_0(\boldsymbol{C}^T)^{-1}\boldsymbol{R}_k\boldsymbol{C}^{-1}\boldsymbol{A}_0^T\boldsymbol{A}_0 = \boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{R}_{k,0}\boldsymbol{A}_0^T\boldsymbol{A}_0, \quad \forall k = 1, \dots, K.$$

Since $\boldsymbol{A}_0^T\boldsymbol{A}_0$ is invertible, this further implies $(\boldsymbol{C}^T)^{-1}\boldsymbol{R}_k\boldsymbol{C}^{-1} = \boldsymbol{R}_{k,0}, \forall k$, or equivalently,

$$\boldsymbol{R}_k = \boldsymbol{C}^T\boldsymbol{R}_{k,0}\boldsymbol{C}, \quad \forall k = 1, \dots, K.$$

A.2 Proof of Lemma 2

To prove Lemma 2, we need the following lemma.

Lemma 4 (Mendelson et al. (2008), Lemma 2.3)**.** *Given $d \geq 1$, and $\varepsilon > 0$, we have*

$$N(\varepsilon, B_2^d, \|\cdot\|_2) \leq \left(1 + \frac{2}{\varepsilon}\right)^d,$$

where B_2^d is the unit ball in $\mathbb{R}^d$, and $N(\varepsilon, \cdot, \|\cdot\|_2)$ the covering number with respect to the Euclidean metric (see Definition 2.2.3 in van der Vaart and Wellner (1996) for details).

Under Assumption (A1), we have $\boldsymbol{a}_i^{(s)*} \in \Omega_a^{(s)}$ and $\mathrm{vec}(\boldsymbol{R}_k^{(s)*}) \in \Omega_w^{(s)}$ for all $s_0 \leq s \leq s_{\max}$. By definition, we have $\ell_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^{(s)*}\}_i, \{\boldsymbol{R}_k^{(s)*}\}_k) \leq \ell_n(\boldsymbol{Y}; \{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)$ and hence

$$\ell_n(\boldsymbol{Y}; \{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) \leq \ell_n(\boldsymbol{Y}; \{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k). \tag{16}$$

Since $\Omega_a^{(s)}$ and $\Omega_w^{(s)}$ are bounded for any $s \in \{1, \dots, s_{\max}\}$ and $s_{\max}$ is a bounded integer, we have

$$\max_i \|\widehat{\boldsymbol{a}}_i^{(s)}\|_2 \leq M_1, \quad \max_k \|\mathrm{vec}(\widehat{\boldsymbol{R}}_k^{(s)})\|_2 \leq M_1, \quad \forall s \in \{1, \dots, s_{\max}\}, \tag{17}$$

and

$$\max_i \|\boldsymbol{a}_i^*\|_2 \leq M_1, \quad \max_k \|\mathrm{vec}(\boldsymbol{R}_k^*)\|_2 \leq M_1. \tag{18}$$

for some constant $M_1 > 0$. Therefore,

$$\begin{aligned} & \max_{i,j,k} |(\widehat{\boldsymbol{a}}_i^{(s)})^T \widehat{\boldsymbol{R}}_k^{(s)} \widehat{\boldsymbol{a}}_j^{(s)}| \leq \max_i \|\widehat{\boldsymbol{a}}_i^{(s)}\|_2^2 \max_k \|\widehat{\boldsymbol{R}}_k^{(s)}\|_2 \\ \leq \quad & \max_i \|\widehat{\boldsymbol{a}}_i^{(s)}\|_2^2 \max_k \|\widehat{\boldsymbol{R}}_k^{(s)}\|_F \leq \max_i \|\widehat{\boldsymbol{a}}_i^{(s)}\|_2^2 \max_k \|\mathrm{vec}(\widehat{\boldsymbol{R}}_k^{(s)})\|_2 \leq M_1^3. \end{aligned}$$

Similarly, we can show $\max_{i,j,k} \|(\boldsymbol{a}_i^*)^T \boldsymbol{R}_k^* \boldsymbol{a}_j^*\|_2 \leq M_1^3$. Let $M_0 = M_1^3$, we obtain that

$$\max_{i,j,k} |(\widehat{\boldsymbol{a}}_i^{(s)})^T \widehat{\boldsymbol{R}}_k^{(s)} \widehat{\boldsymbol{a}}_j^{(s)}| \leq M_0, \quad \forall s \in \{s_0, \dots, s_{\max}\}, \quad \max_{i,j,k} |(\boldsymbol{a}_i^*)^T \boldsymbol{R}_k^* \boldsymbol{a}_j^*| \leq M_0. \tag{19}$$

We define $\theta_{ijk}^* = (\boldsymbol{a}_i^*)^T \boldsymbol{R}_k^* \boldsymbol{a}_j^*$ and $\hat{\theta}_{ijk}^{(s)} = (\widehat{\boldsymbol{a}}_i^{(s)})^T \widehat{\boldsymbol{R}}_k^{(s)} \widehat{\boldsymbol{a}}_j^{(s)}$. It follows from a second-order Taylor expansion that

$$\begin{aligned} & \mathrm{g}(\theta_{ijk}^*)\theta_{ijk}^* - \log\{1 + \exp(\theta_{ijk}^*)\} - \mathrm{g}(\theta_{ijk}^*)\hat{\theta}_{ijk}^{(s)} - \log\{1 + \exp(\hat{\theta}_{ijk}^{(s)})\} \\ = \quad & \frac{1}{2}(\theta_{ijk}^* - \hat{\theta}_{ijk}^{(s)})^2 \frac{\exp(\tilde{\theta}_{ijk}^{(s)})}{\{1 + \exp(\tilde{\theta}_{ijk}^{(s)})\}^2}, \end{aligned} \tag{20}$$

for some $\tilde{\theta}_{ijk}^{(s)}$ lying on the line segment joining θ_{ijk}^* and $\hat{\theta}_{ijk}^{(s)}$. By (19), we have for any i, j, k and $s \in \{s_0, \dots, s_{\max}\}$, $|\tilde{\theta}_{ijk}^{(s)}| \leq M_0$. This together with (20) gives that

$$\begin{aligned} & \mathrm{g}(\theta_{ijk}^*)\theta_{ijk}^* - \log\{1 + \exp(\theta_{ijk}^*)\} - \mathrm{g}(\theta_{ijk}^*)\hat{\theta}_{ijk}^{(s)} - \log\{1 + \exp(\hat{\theta}_{ijk}^{(s)})\} \\ \geq \quad & \frac{1}{2\{1 + \exp(M_0)\}^2}(\theta_{ijk}^* - \hat{\theta}_{ijk}^{(s)})^2, \end{aligned}$$

and hence

$$\begin{aligned} & \ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - \ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \\ = \quad & \sum_{ijk} \Big(\mathrm{g}(\theta_{ijk}^*)\theta_{ijk}^* - \log\{1 + \exp(\theta_{ijk}^*)\} - \mathrm{g}(\theta_{ijk}^*)\hat{\theta}_{ijk}^{(s)} - \log\{1 + \exp(\hat{\theta}_{ijk}^{(s)})\} \Big) \\ \geq \quad & \frac{1}{2\{1 + \exp(M_0)\}^2} \sum_{ijk} (\theta_{ijk}^* - \hat{\theta}_{ijk}^{(s)})^2. \end{aligned} \tag{21}$$

In the following, we provide an upper bound for

$$\begin{aligned} & \max_{s\in\{s_0,\dots,s_{\max}\}} \sup_{\substack{\boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s} \\ \max_i \|\boldsymbol{a}_i\|_2\le M_1 \\ \max_i \|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} |\ell_0(\{\boldsymbol{a}_i\}_i,\{\boldsymbol{R}_k\}_k)-\ell_n(\{\boldsymbol{a}_i\}_i,\{\boldsymbol{R}_k\}_k)| \\ = & \max_{s\in\{s_0,\dots,s_{\max}\}} \sup_{\substack{\boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s} \\ \max_i \|\boldsymbol{a}_i\|_2\le M_1 \\ \max_i \|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} \left|\sum_{ijk}(Y_{ijk}-\pi_{ijk}^*)\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j\right|, \end{aligned}$$

where $\pi_{ijk}^*=\exp(\theta_{ijk}^*)/\{1+\exp(\theta_{ijk}^*)\}$, $\forall i,j,k$.

Let $\varepsilon_{n,K}=M_1/(nK)^2$ and $\{\bar{\boldsymbol{a}}_1^{(s)},\dots,\bar{\boldsymbol{a}}_{N_{s,\varepsilon_{n,K}}}^{(s)}\}$ be a minimal $\varepsilon_{n,K}$-net of the vector space $(\{\boldsymbol{a}\in\mathbb{R}^s:\|\boldsymbol{a}\|_2\le M_1\},\|\cdot\|_2)$. It follows from Lemma 4 that

$$\begin{aligned} N_{s,\varepsilon_{n,K}} &= N(\varepsilon_{n,K},\{\boldsymbol{a}\in\mathbb{R}^s:\|\boldsymbol{a}\|_2\le M_1\},\|\cdot\|_2)=N(1/(nK)^2,B_2^s,\|\cdot\|_2) \\ &\le \left(1+2n^2K^2\right)^s\le(3nK)^{2s}. \end{aligned} \tag{22}$$

Let $\{\bar{\boldsymbol{R}}_1^{(s)},\dots,\bar{\boldsymbol{R}}_{N_{s^2,\varepsilon_{n,K}}}^{(s)}\}$ be a minimal $\varepsilon_{n,K}$-net of the vector space $(\{\boldsymbol{R}\in\mathbb{R}^{s\times s}:\|\mathrm{vec}(\boldsymbol{R})\|_2\le M_1\},\|\cdot\|_F)$. For any $s\times s$ matrices $\boldsymbol{Q}$, we have $\|\boldsymbol{Q}\|_F=\|\mathrm{vec}(\boldsymbol{Q})\|_2$. Hence, similar to (22), we can show that

$$N_{s^2,\varepsilon_{n,K}}\le(3nK)^{2s^2}. \tag{23}$$

Hence, for any $\boldsymbol{a}_i,\boldsymbol{a}_j\in\mathbb{R}^s$ and $\boldsymbol{R}_k\in\mathbb{R}^{s\times s}$ satisfying $\|\boldsymbol{a}_i\|_2,\|\boldsymbol{a}_j\|_2,\|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1$, there exists some $\bar{\boldsymbol{a}}_{l_i}^{(s)}$, $\bar{\boldsymbol{a}}_{l_j}^{(s)}$, $\bar{\boldsymbol{R}}_{t_k}^{(s)}$, such that

$$\|\boldsymbol{a}_i-\bar{\boldsymbol{a}}_{l_i}^{(s)}\|_2\le\varepsilon_{n,K},\quad \|\boldsymbol{a}_j-\bar{\boldsymbol{a}}_{l_j}^{(s)}\|_2\le\varepsilon_{n,K},\quad \|\boldsymbol{R}_k-\bar{\boldsymbol{R}}_{t_k}^{(s)}\|_F\le\varepsilon_{n,K}. \tag{24}$$

This further implies

$$\begin{aligned} & |(\boldsymbol{a}_i)^T\boldsymbol{R}_k\boldsymbol{a}_j-(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T\bar{\boldsymbol{R}}_{t_k}^{(s)}\bar{\boldsymbol{a}}_{l_j}^{(s)}|\le|(\boldsymbol{a}_i)^T\boldsymbol{R}_k\boldsymbol{a}_j-(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T\boldsymbol{R}_k\boldsymbol{a}_j| \\ + & |(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T\boldsymbol{R}_k\boldsymbol{a}_j-(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T\bar{\boldsymbol{R}}_{t_k}^{(s)}\boldsymbol{a}_j|+|(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T\bar{\boldsymbol{R}}_{t_k}^{(s)}\boldsymbol{a}_j-(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T\bar{\boldsymbol{R}}_{t_k}^{(s)}\bar{\boldsymbol{a}}_{l_j}^{(s)}| \\ \le & \|\boldsymbol{a}_i-\bar{\boldsymbol{a}}_{l_i}^{(s)}\|_2\|\boldsymbol{R}_k\|_2\|\boldsymbol{a}_j\|_2+\|\bar{\boldsymbol{a}}_{l_i}^{(s)}\|_2\|\boldsymbol{R}_k-\bar{\boldsymbol{R}}_{t_k}^{(s)}\|_2\|\boldsymbol{a}_j\|_2+\|\bar{\boldsymbol{a}}_i^{(s)}\|_2\|\bar{\boldsymbol{R}}_{t_k}^{(s)}\|_2\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}_{l_j}^{(s)}\|_2 \\ \le & \|\boldsymbol{a}_i-\bar{\boldsymbol{a}}_{l_i}^{(s)}\|_2\|\boldsymbol{R}_k\|_F\|\boldsymbol{a}_j\|_2+\|\bar{\boldsymbol{a}}_{l_i}^{(s)}\|_2\|\boldsymbol{R}_k-\bar{\boldsymbol{R}}_{t_k}^{(s)}\|_F\|\boldsymbol{a}_j\|_2+\|\bar{\boldsymbol{a}}_i^{(s)}\|_2\|\bar{\boldsymbol{R}}_{t_k}^{(s)}\|_F\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}_{l_j}^{(s)}\|_2 \\ \le & 3\varepsilon_{n,K}M_1^2\le3M_1^3/(nK)^2\le3M_0/(nK)^2. \end{aligned} \tag{25}$$

Therefore, we have

$$
\begin{aligned}
& \max_{s\in\{s_0,\ldots,s_{\max}\}} \sup_{\substack{\boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s}\\ \max_i\|\boldsymbol{a}_i\|_2\le M_1\\ \max_i\|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j\right| \qquad (26)\\
\le\ & \max_{s\in\{s_0,\ldots,s_{\max}\}} \max_{\substack{l_1,\ldots,l_n\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\ldots,t_K\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})(\bar{\boldsymbol{a}}^{(s)}_{l_i})^T\bar{\boldsymbol{R}}^{(s)}_{t_k}\bar{\boldsymbol{a}}^{(s)}_{l_j}\right| + \sum_{ijk}|Y_{ijk}-\pi^*_{ijk}|\frac{3M_0}{n^2K^2}\\
\le\ & \max_{s\in\{s_0,\ldots,s_{\max}\}} \max_{\substack{l_1,\ldots,l_n\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\ldots,t_K\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})(\bar{\boldsymbol{a}}^{(s)}_{l_i})^T\bar{\boldsymbol{R}}^{(s)}_{t_k}\bar{\boldsymbol{a}}^{(s)}_{l_j}\right| + O(1).
\end{aligned}
$$

Since Y_{ijk}'s are independent, we have

$$
\begin{aligned}
& \max_{s\in\{s_0,\ldots,s_{\max}\}} \mathrm{E} \max_{\substack{l_1,\ldots,l_n\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\ldots,t_K\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})(\bar{\boldsymbol{a}}^{(s)}_{l_i})^T\bar{\boldsymbol{R}}^{(s)}_{t_k}\bar{\boldsymbol{a}}^{(s)}_{l_j}\right| \qquad (27)\\
\le\ & K_0 \max_{s\in\{s_0,\ldots,s_{\max}\}}\left(\sigma_0\sqrt{\log N^n_{s,\varepsilon_{n,K}}N^K_{s^2,\varepsilon_{n,K}}}+\sqrt{\mathrm{E}M_2^2}(\log N^n_{s,\varepsilon_{n,K}}N^K_{s^2,\varepsilon_{n,K}})\right) \qquad (28)\\
\le\ & K_0 \max_{s\in\{s_0,\ldots,s_{\max}\}}\left(\sigma_0\sqrt{n\log N_{s,\varepsilon_{n,K}}+K\log N_{s^2,\varepsilon_{n,K}}}+\sqrt{\mathrm{E}M_2^2}(n\log N_{s,\varepsilon_{n,K}}+K\log N_{s^2,\varepsilon_{n,K}})\right)\\
=\ & O\left(\sigma_0\sqrt{(n+K)(\log n+\log K)}\right)+O\left(\sqrt{\mathrm{E}M_2^2}(n+K)(\log n+\log K)\right), \qquad (29)
\end{aligned}
$$

for some constant $K_0>0$, where

$$
\begin{aligned}
M_2 &= \max_{s\in\{s_0,\ldots,s_{\max}\}} \max_{\substack{l_1,\ldots,l_n\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\ldots,t_K\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \max_{ijk}|Y_{ijk}-\pi^*_{ijk}||(\bar{\boldsymbol{a}}^{(s)}_{l_i})^T\bar{\boldsymbol{R}}^{(s)}_{t_k}\bar{\boldsymbol{a}}^{(s)}_{l_j}|,\\
\sigma_0^2 &= \max_{s\in\{s_0,\ldots,s_{\max}\}} \max_{\substack{l_1,\ldots,l_n\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\ldots,t_K\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \sum_{ijk}\mathrm{E}|Y_{ijk}-\pi^*_{ijk}|^2|(\bar{\boldsymbol{a}}^{(s)}_{l_i})^T\bar{\boldsymbol{R}}^{(s)}_{t_k}\bar{\boldsymbol{a}}^{(s)}_{l_j}|^2.
\end{aligned}
$$

The inequality in (28) follows from Lemma A3 in Chernozhukov et al. (2013) and the equality in (29) is due to (22) and (23).

With some calculations, we can show that

$$
\begin{aligned}
M_2 &\le \max_{\substack{l_1,l_2\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ l_0\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \|\bar{\boldsymbol{a}}^{(s)}_{l_1}\|_2\|\bar{\boldsymbol{R}}^{(s)}_{l_0}\|_2\|\bar{\boldsymbol{a}}^{(s)}_{l_2}\|_2 \le \max_{\substack{l_1,l_2\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ l_0\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \|\bar{\boldsymbol{a}}^{(s)}_{l_1}\|_2\|\bar{\boldsymbol{R}}^{(s)}_{l_0}\|_F\|\bar{\boldsymbol{a}}^{(s)}_{l_2}\|_2\\
&\le \max_{\substack{l_1,l_2\in\{1,\ldots,N_{s,\varepsilon_{n,K}}\}\\ l_0\in\{1,\ldots,N_{s^2,\varepsilon_{n,K}}\}}} \|\bar{\boldsymbol{a}}^{(s)}_{l_1}\|_2\|\mathrm{vec}(\bar{\boldsymbol{R}}^{(s)}_{l_0})\|_2\|\bar{\boldsymbol{a}}^{(s)}_{l_2}\|_2 \le M_1^3 \le M_0, \qquad (30)
\end{aligned}
$$

and similarly

$$\begin{aligned}\sigma_0^2 &\le \max_{s\in\{s_0,\dots,s_{\max}\}} \max_{\substack{l_1,\dots,l_n\in\{1,\dots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\dots,t_K\in\{1,\dots,N_{s^2,\varepsilon_{n,K}}\}}} \sum_{ijk} \mathrm{E}|Y_{ijk}-\pi^*_{ijk}|^2 \|\bar{\boldsymbol{a}}_{l_i}^{(s)}\|_2^2 \|\mathrm{vec}(\bar{\boldsymbol{R}}_{t_k}^{(s)})\|_2^2 \|\bar{\boldsymbol{a}}_{l_j}^{(s)}\|_2^2 \\ &\le M_1^6 \sum_{ijk} \mathrm{E}|Y_{ijk}-\pi^*_{ijk}|^2 \le M_0^2 n^2 K.\end{aligned}$$

Combining these together with (27), we have

$$\begin{aligned}&\max_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \max_{\substack{l_1,\dots,l_n\in\{1,\dots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\dots,t_K\in\{1,\dots,N_{s^2,\varepsilon_{n,K}}\}}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)}\right| \\ =\;& O\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)}\right) + O\left((n+K)(\log n+\log K)\right).\end{aligned}$$

This together with (26) implies that

$$\begin{aligned}&\max_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \sup_{\substack{\boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s}\\ \max_i\|\boldsymbol{a}_i\|_2\le M_1\\ \max_i\|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j\right| \\ \le\;& O(1) + \max_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \max_{\substack{l_1,\dots,l_n\in\{1,\dots,N_{s,\varepsilon_{n,K}}\}\\ t_1,\dots,t_K\in\{1,\dots,N_{s^2,\varepsilon_{n,K}}\}}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)}\right| \\ =\;& O\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)}\right) + O\left((n+K)(\log n+\log K)\right).\end{aligned}$$

Since $s_{\max}$ is a bounded integer, we obtain

$$\begin{aligned}&\mathrm{E} \max_{s\in\{s_0,\dots,s_{\max}\}} \sup_{\substack{\boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s}\\ \max_i\|\boldsymbol{a}_i\|_2\le M_1\\ \max_i\|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j\right| \\ \le\;& \sum_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \sup_{\substack{\boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s}\\ \max_i\|\boldsymbol{a}_i\|_2\le M_1\\ \max_i\|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j\right| \\ \le\;& s_{\max} \max_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \sup_{\substack{\boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s}\\ \max_i\|\boldsymbol{a}_i\|_2\le M_1\\ \max_i\|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})\boldsymbol{a}_i^T \boldsymbol{R}_k \boldsymbol{a}_j\right| \\ =\;& O\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)}\right) + O\left((n+K)(\log n+\log K)\right).\end{aligned}$$

By Markov's inequality, we have that

$$\begin{aligned}
&\sup_{\substack{s\in\{s_0,\ldots,s_{\max}\}\\ \boldsymbol{a}_i\in\mathbb{R}^s,\boldsymbol{R}_k\in\mathbb{R}^{s\times s}\\ \max_i\|\boldsymbol{a}_i\|_2\le M_1\\ \max_i\|\mathrm{vec}(\boldsymbol{R}_k)\|_2\le M_1}} \left|\sum_{ijk}(Y_{ijk}-\pi^*_{ijk})\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j\right| \qquad (31)\\
=\;& O_p\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)}+(n+K)(\log n+\log K)\right).
\end{aligned}$$

Combining (17), (18) with (31), we obtain

$$\begin{aligned}
&\max\left(\max_s\left|\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)-\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\right|,\right. \qquad (32)\\
&\left.|\ell_0(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)-\ell_n(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)|\right)\\
=\;& O_p\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)}+(n+K)(\log n+\log K)\right).
\end{aligned}$$

Therefore, it follows from (21) that

$$\begin{aligned}
&\max_{s\in\{s_0,\ldots,s_{\max}\}}\sum_{ijk}(\theta^*_{ijk}-\hat{\theta}^{(s)}_{ijk})^2 \qquad (33)\\
\le\;& \max_{s\in\{s_0,\ldots,s_{\max}\}} 2\{1+\exp(M_0)\}^2\left(\ell_0(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)-\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\right)\\
\le\;& \max_{s\in\{s_0,\ldots,s_{\max}\}} 2\{1+\exp(M_0)\}^2\left|\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)-\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\right|\\
+\;& \max_{s\in\{s_0,\ldots,s_{\max}\}} 2\{1+\exp(M_0)\}^2\left|\ell_0(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)-\ell_n(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)\right|\\
+\;& \max_{s\in\{s_0,\ldots,s_{\max}\}} 2\{1+\exp(M_0)\}^2\left(\ell_n(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)-\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\right)\\
\le\;& \max_{s\in\{s_0,\ldots,s_{\max}\}} 2\{1+\exp(M_0)\}^2\left|\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)-\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\right|\\
+\;& \max_{s\in\{s_0,\ldots,s_{\max}\}} 2\{1+\exp(M_0)\}^2\left|\ell_0(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)-\ell_n(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)\right|\\
=\;& O_p\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)}+(n+K)(\log n+\log K)\right),
\end{aligned}$$

where the third inequality is due to that $\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\ge\ell_n(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k)$, for all $s\in\{s_0,\ldots,s_{\max}\}$.

Let $r_{n,K}=(n+K)^{-1/2}(\log n+\log K)^{-1/2}$. As $n\to\infty$, we have

$$\begin{aligned}
&r^2_{n,K}\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)}+(n+K)(\log n+\log K)\right)\\
=\;& \frac{n\sqrt{K}+\sqrt{(n+K)(\log n+\log K)}}{\sqrt{(n+K)(\log n+\log K)}}\ll\frac{n\sqrt{K}}{2\sqrt{n}}+1\ll\sqrt{nK}.
\end{aligned}$$

It follows from (33) that

$$\Pr\left(\max_{s\in\{s_0,\ldots,s_{\max}\}} r^2_{n,K}\sum_{ijk}(\theta^*_{ijk}-\hat{\theta}^{(s)}_{ijk})^2\ge\sqrt{nK}\right)\to 0. \qquad (34)$$

For any integer $m \ge 1$, define

$$\mathbb{S}_m^{(s)} = \Bigg\{ (\{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k) : 2^{m-1} < r_{n,K} \sqrt{\sum_{ijk} \{\theta_{ijk}^* - (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)}\}^2} \le 2^m,$$
$$\boldsymbol{a}_i^{(s)} \in \mathbb{R}^s, \forall i, \boldsymbol{R}_k^{(s)} \in \mathbb{R}^{s\times s}, \forall k, \max_i \|\boldsymbol{a}_i^{(s)}\|_2 \le M_1, \max_k \|\mathrm{vec}(\boldsymbol{R}_k^{(s)})\|_2 \le M_1 \Bigg\}.$$

For any $(\{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k) \in \mathbb{S}_m^{(s)}$, similar to (21), we can show that

$$\begin{aligned} \ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - \ell_0(\{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k) &\ge \frac{1}{2\{1+\exp(M_0)\}^2} \sum_{ijk} (\theta_{ijk}^* - (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)})^2 \\ &\ge \frac{2^{2m-3}}{\{1+\exp(M_0)\}^2 r_{n,K}^2}. \end{aligned} \tag{35}$$

For any $\{l_i\}_i$ and $\{t_k\}_k$ satisfying (24), it follows from (25) that

$$\begin{aligned} & \sum_{ijk} \{\theta_{ijk}^* - (\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)}\}^2 \\ \le\ & \sum_{ijk} \{\theta_{ijk}^* - (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)} + (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)} - (\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)}\}^2 \\ \le\ & 2 \sum_{ijk} \left(\{\theta_{ijk}^* - (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)}\}^2 + \{(\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)} - (\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)}\}^2 \right) \\ \le\ & \frac{2^{2m+1}}{r_{n,K}^2} + 2 \le \frac{2^{2(m+1)}}{r_{n,K}^2}. \end{aligned}$$

Let $\Lambda_m^{(s)} = \{(\{l_i\}_i, \{t_k\}_k) : \sum_{ijk} \{\theta_{ijk}^* - (\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)}\}^2 \le 2^{2(m+1)}/r_{n,K}^2\}$. Similar to (26) and (27), we can show there exists some constant $C_0 > 0$ such that

$$\begin{aligned} & \max_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \sup_{(\{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k) \in \mathbb{S}_m^{(s)}} \left| \sum_{ijk} (Y_{ijk} - \pi_{ijk}^*) \{(\boldsymbol{a}_i^*)^T \boldsymbol{R}_k^* \boldsymbol{a}_j^* - (\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)}\} \right| \\ \le\ & \max_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \max_{\substack{l_1,\dots,l_n \in \{1,\dots,N_{s,\varepsilon_{n,K}}\} \\ t_1,\dots,t_K \in \{1,\dots,N_{s^2,\varepsilon_{n,K}}\} \\ (\{l_i\}_i, \{t_k\}_k) \in \Lambda_m^{(s)}}} \left| \sum_{ijk} (Y_{ijk} - \pi_{ijk}^*) \{(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)} - \theta_{ijk}^*\} \right| + C_0 \\ \le\ & C_0 \left(\sigma_m^{(s)} r_{n,K}^{-1} + \sqrt{\mathrm{E} M_2^2} r_{n,K}^{-2} \right) + C_0, \end{aligned}$$

where

$$(\sigma_m^{(s)})^2 \le \max_{\substack{l_1,\dots,l_n \in \{1,\dots,N_{s,\varepsilon_{n,K}}\} \\ t_1,\dots,t_K \in \{1,\dots,N_{s^2,\varepsilon_{n,K}}\} \\ (\{l_i\}_i, \{t_k\}_k) \in \Lambda_m^{(s)}}} \sum_{ijk} \mathrm{E}|Y_{ijk} - \pi_{ijk}^*|^2 \{(\bar{\boldsymbol{a}}_{l_i}^{(s)})^T \bar{\boldsymbol{R}}_{t_k}^{(s)} \bar{\boldsymbol{a}}_{l_j}^{(s)} - \theta_{ijk}^*\}^2 \le 2^{2(m+1)}/r_{n,K}^2.$$

This together with (30) implies that

$$r_{n,K}^2 \max_{s\in\{s_0,\dots,s_{\max}\}} \mathrm{E} \sup_{(\{\boldsymbol{a}_i^{(s)}\}_i,\{\boldsymbol{R}_k^{(s)}\}_k)\in\mathbb{S}_m^{(s)}} \left|\sum_{ijk}(Y_{ijk}-\pi_{ijk}^*)\{(\boldsymbol{a}_i^*)^T\boldsymbol{R}_k^*\boldsymbol{a}_j^* - (\boldsymbol{a}_i^{(s)})^T\boldsymbol{R}_k^{(s)}\boldsymbol{a}_j^{(s)}\}\right| \le C_0 2^{(m+1)} + C_0M_0 + C_0. \tag{36}$$

The event $(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\in\mathbb{S}_m^{(s)}$ implies that

$$\sup_{(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\in\mathbb{S}_m^{(s)}} \ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \ge \ell_n(\{\boldsymbol{a}_i^*\}_i,\{\boldsymbol{R}_k^*\}_k),$$

and hence

$$\sup_{(\{\boldsymbol{a}_i^{(s)}\}_i,\{\boldsymbol{R}_k^{(s)}\}_k)\in\mathbb{S}_m^{(s)}} \left|\sum_{ijk}(Y_{ijk}-\pi_{ijk}^*)\{\theta_{ijk}^* - (\boldsymbol{a}_i^{(s)})^T\boldsymbol{R}_k^{(s)}\boldsymbol{a}_j^{(s)}\}\right| \ge \frac{2^{2m-3}}{\{1+\exp(M_0)\}^2 r_{n,K}^2},$$

by (35). Therefore, we obtain

$$\begin{aligned}
& \Pr\left((\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\in\mathbb{S}_m^{(s)}\right) \qquad (37)\\
\le\ & \Pr\left(\sup_{(\{\boldsymbol{a}_i^{(s)}\}_i,\{\boldsymbol{R}_k^{(s)}\}_k)\in\mathbb{S}_m^{(s)}} \left|\sum_{ijk}(Y_{ijk}-\pi_{ijk}^*)\{\theta_{ijk}^* - (\boldsymbol{a}_i^{(s)})^T\boldsymbol{R}_k^{(s)}\boldsymbol{a}_j^{(s)}\}\right| \ge \frac{2^{2m-3}}{\{1+\exp(M_0)\}^2 r_{n,K}^2}\right)\\
\le\ & \frac{C_0\{1+\exp(M_0)\}^2(2^{m+1}+M_0+1)}{2^{2m-3}},
\end{aligned}$$

for any integer $m\ge 1$, where the last inequality is due to (36) and Markov's inequality. For any integer $J_0\ge 1$, we have

$$\begin{aligned}
& \Pr\left(r_{n,K}\sqrt{\sum_{ijk}(\hat{\theta}_{ijk}^{(s)}-\theta_{ijk}^*)^2} \ge 2^{J_0}\right)\\
\le\ & \sum_{m>J_0, 2^m\le 2n^{1/4}K^{1/4}} \Pr\left((\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i,\{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\in\mathbb{S}_m^{(s)}\right) + \Pr\left(r_{n,K}^2\sum_{ijk}(\theta_{ijk}^*-\hat{\theta}_{ijk}^{(s)})^2 \ge \sqrt{nK}\right)\\
\le\ & \sum_{m>J_0} C_0\{1+\exp(M_0)\}^2\frac{2^{m+1}+M_0+1}{2^{2m-3}} + o(1)\\
\le\ & C_0\{1+\exp(M_0)\}^2\left(\frac{16}{2^{J_0}}+\frac{2(M_0+1)/3}{4^{J_0}}\right) + o(1),
\end{aligned}$$

where the second inequality is due to (37) and (34). As $J_0\to\infty$, we obtain

$$\Pr\left(r_{n,K}\sqrt{\sum_{ijk}(\hat{\theta}_{ijk}^{(s)}-\theta_{ijk}^*)^2} \ge 2^{J_0}\right) \to 0.$$

This implies $\sum_{ijk}(\hat{\theta}_{ijk}^{(s)}-\theta_{ijk}^*)^2 = O_p(1/r_{n,K}^2)$. The proof is hence completed.

A.3 Proof of Lemma 3

For any $s < s_0$, we have

$$\begin{aligned}
& \sum_{ijk} \left((\widehat{\boldsymbol{a}}_i^{(s)})^T \widehat{\boldsymbol{R}}_k^{(s)} \widehat{\boldsymbol{a}}_j^{(s)} - (\boldsymbol{a}_i^*)^T \boldsymbol{R}_k^* \boldsymbol{a}_j^* \right)^2 && (38) \\
\geq & \inf_{\substack{\boldsymbol{a}_1^{(s)},\ldots,\boldsymbol{a}_n^{(s)} \in \mathbb{R}^s \\ \boldsymbol{R}_1^{(s)},\ldots,\boldsymbol{R}_K^{(s)} \in \mathbb{R}^{s\times s}}} \sum_{ijk} \left((\boldsymbol{a}_i^{(s)})^T \boldsymbol{R}_k^{(s)} \boldsymbol{a}_j^{(s)} - (\boldsymbol{a}_i^*)^T \boldsymbol{R}_k^* \boldsymbol{a}_j^* \right)^2 \\
= & \inf_{\substack{\boldsymbol{A}^{(s)} \in \mathbb{R}^{n\times s} \\ \boldsymbol{R}_1^{(s)},\ldots,\boldsymbol{R}_K^{(s)} \in \mathbb{R}^{s\times s}}} \sum_{k=1}^K \|\boldsymbol{A}^{(s)} \boldsymbol{R}_k^{(s)} (\boldsymbol{A}^{(s)})^T - \boldsymbol{A}_0 \boldsymbol{R}_{k,0} \boldsymbol{A}_0^T\|_F^2 \\
= & \inf_{\substack{\boldsymbol{A}^{(s)} \in \mathbb{R}^{n\times s} \\ \boldsymbol{R}_1^{(s)},\ldots,\boldsymbol{R}_K^{(s)} \in \mathbb{R}^{s\times s}}} \left\| \begin{pmatrix} \boldsymbol{A}^{(s)} \boldsymbol{R}_1^{(s)} \\ \vdots \\ \boldsymbol{A}^{(s)} \boldsymbol{R}_K^{(s)} \end{pmatrix} (\boldsymbol{A}^{(s)})^T - \underbrace{\begin{pmatrix} \boldsymbol{A}_0 \boldsymbol{R}_{1,0} \\ \vdots \\ \boldsymbol{A}_0 \boldsymbol{R}_{K,0} \end{pmatrix}}_{\boldsymbol{B}_0} \boldsymbol{A}_0^T \right\|_F^2 \\
\geq & \inf_{\boldsymbol{A}^{(s)} \in \mathbb{R}^{n\times s}, \boldsymbol{B}^{(s)} \in \mathbb{R}^{nK\times s}} \left\| \boldsymbol{B}^{(s)} (\boldsymbol{A}^{(s)})^T - \boldsymbol{B}_0 \boldsymbol{A}_0^T \right\|_F^2 .
\end{aligned}$$

Define

$$(\widehat{\boldsymbol{A}}^{(s)}, \widehat{\boldsymbol{B}}^{(s)}) = \arg \min_{\substack{\boldsymbol{A}^{(s)} \in \mathbb{R}^{n\times s} \\ \boldsymbol{B}^{(s)} \in \mathbb{R}^{nK\times s}}} \|\boldsymbol{B}^{(s)} (\boldsymbol{A}^{(s)})^T - \boldsymbol{B}_0 \boldsymbol{A}_0^T\|_F^2.$$

The above minimizers are not unique. Notice that $\text{rank}(\boldsymbol{B}_0 \boldsymbol{A}_0^T) \leq \text{rank}(\boldsymbol{A}_0^T) \leq s_0$. Assume $\boldsymbol{B}_0 \boldsymbol{A}_0^T$ has the following singular value decomposition,

$$\boldsymbol{B}_0 \boldsymbol{A}_0^T = n \boldsymbol{U}_n \boldsymbol{\Lambda}_n \boldsymbol{V}_n^T,$$

for some $\boldsymbol{U}_n \in \mathbb{R}^{nK\times s_0}, \boldsymbol{V}_n \in \mathbb{R}^{n\times s_0}$ such that $\boldsymbol{U}_n^T \boldsymbol{U}_n = \boldsymbol{V}_n^T \boldsymbol{V}_n = \boldsymbol{I}_{s_0}$, and some diagonal matrix

$$\boldsymbol{\Lambda}_n = \text{diag}(\lambda_n^{(1)}, \lambda_n^{(2)}, \ldots, \lambda_n^{(s_0)})$$

such that $|\lambda_n^{(1)}| \geq |\lambda_n^{(2)}| \geq \cdots \geq |\lambda_n^{(s_0)}|$. Then one solution is given by

$$\widehat{\boldsymbol{A}}^{(s)} = n \boldsymbol{V}_n \boldsymbol{\Lambda}_n \boldsymbol{U}_n^T \boldsymbol{U}_n^{(s)}, \quad \widehat{\boldsymbol{B}}^{(s)} = \boldsymbol{U}_n^{(s)},$$

where $\boldsymbol{U}_n^{(s)}$ is the submatrix of $\boldsymbol{U}_n$ formed by its first s columns.

Since $\boldsymbol{U}_n^T \boldsymbol{U}_n = \boldsymbol{I}_{s_0}$, we have

$$\begin{aligned}
\widehat{\boldsymbol{B}}^{(s)} (\widehat{\boldsymbol{A}}^{(s)})^T &= n \boldsymbol{U}_n^{(s)} (\boldsymbol{U}_n^{(s)})^T \boldsymbol{U}_n \boldsymbol{\Lambda}_n \boldsymbol{V}_n^T = n \boldsymbol{U}_n^{(s)} (\boldsymbol{I}_s, \boldsymbol{O}_{s,s_0-s}) \boldsymbol{\Lambda}_n \boldsymbol{V}_n^T \\
&= n \boldsymbol{U}_n \begin{pmatrix} \boldsymbol{I}_s & \boldsymbol{O}_{s,s_0-s} \\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{O}_{s_0-s,s_0-s} \end{pmatrix} \boldsymbol{\Lambda}_n \boldsymbol{V}_n^T,
\end{aligned}$$

and hence

$$\begin{aligned}\widehat{\boldsymbol{B}}^{(s)}(\widehat{\boldsymbol{A}}^{(s)})^T - \boldsymbol{B}_0\boldsymbol{A}_0^T &= n\boldsymbol{U}_n\begin{pmatrix}\boldsymbol{I}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{O}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{\Lambda}_n\boldsymbol{V}_n^T - n\boldsymbol{U}_n\boldsymbol{I}_{s_0}\boldsymbol{\Lambda}_n\boldsymbol{V}_n^T\\ &= n\boldsymbol{U}_n\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{\Lambda}_n\boldsymbol{V}_n^T.\end{aligned}$$

This together with (38) implies that

$$\begin{aligned}&\sum_{ijk}\left((\widehat{\boldsymbol{a}}_i^{(s)})^T\widehat{\boldsymbol{R}}_k^{(s)}\widehat{\boldsymbol{a}}_j^{(s)} - (\boldsymbol{a}_i^*)^T\boldsymbol{R}_k^*\boldsymbol{a}_j^*\right)^2 \ge \|\widehat{\boldsymbol{B}}^{(s)}(\widehat{\boldsymbol{A}}^{(s)})^T - \boldsymbol{B}_0\boldsymbol{A}_0^T\|_F^2\\
\ge\ & n^2\text{trace}\left(\boldsymbol{U}_n\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{\Lambda}_n\boldsymbol{V}_n^T\boldsymbol{V}_n\boldsymbol{\Lambda}_n\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{U}_n^T\right)\\
=\ & n^2\text{trace}\left(\boldsymbol{U}_n\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{\Lambda}_n^2\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{U}_n^T\right) \qquad (39)\\
=\ & n^2\text{trace}\left(\boldsymbol{\Lambda}_n\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{U}_n^T\boldsymbol{U}_n\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{\Lambda}_n\right)\\
=\ & n^2\text{trace}\left(\boldsymbol{\Lambda}_n\begin{pmatrix}\boldsymbol{O}_s & \boldsymbol{O}_{s,s_0-s}\\ \boldsymbol{O}_{s_0-s,s} & \boldsymbol{I}_{s_0-s,s_0-s}\end{pmatrix}\boldsymbol{\Lambda}_n\right) \ge n^2(\lambda_n^{(s_0)})^2, \qquad (40)\end{aligned}$$

where (39) is due to that $\boldsymbol{V}_n^T\boldsymbol{V}_n = \boldsymbol{I}_{s_0}$ and the first equality in (40) is due to that $\boldsymbol{U}_n^T\boldsymbol{U}_n = \boldsymbol{I}_{s_0}$.

To summarize, we've shown

$$\sum_{ijk}\left((\widehat{\boldsymbol{a}}_i^{(s)})^T\widehat{\boldsymbol{R}}_k^{(s)}\widehat{\boldsymbol{a}}_j^{(s)} - (\boldsymbol{a}_i^*)^T\boldsymbol{R}_k^*\boldsymbol{a}_j^*\right)^2 \ge n^2(\lambda_n^{(s_0)})^2. \qquad (41)$$

In the following, we provide a lower bound for $(\lambda_n^{(s_0)})^2$. By definition, $(\lambda_n^{(s_0)})^2$ is the s_0-th largest eigenvalue of

$$\frac{1}{n^2}\boldsymbol{A}_0\boldsymbol{B}_0^T\boldsymbol{B}_0\boldsymbol{A}_0^T = \frac{1}{n^2}\sum_{k=1}^K\boldsymbol{A}_0\boldsymbol{R}_{k,0}^T\boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{R}_{k,0}\boldsymbol{A}_0^T.$$

We first provide an lower bound for $\lambda_{\min}(\sum_{k=1}^K\boldsymbol{R}_{k,0}^T\boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{R}_{k,0}/n)$. Let $\boldsymbol{\Sigma}_A = \boldsymbol{A}_0^T\boldsymbol{A}_0/n$. Consider the following eigenvalue decomposition:

$$\boldsymbol{\Sigma}_A = \boldsymbol{U}_A\boldsymbol{\Lambda}_A\boldsymbol{U}_A^T,$$

for some orthogonal matrix $\boldsymbol{U}_A$ and some diagonal matrix $\boldsymbol{\Lambda}_A$. Under Assumption (A2), the matrix

$$\boldsymbol{\Sigma}_A - \bar{c}\boldsymbol{I}_{s_0}$$

is positive semidefinite. As a result, the matrix

$$\sum_{k=1}^K\boldsymbol{R}_{k,0}^T(\boldsymbol{\Sigma}_A - \bar{c}\boldsymbol{I}_{s_0})\boldsymbol{R}_{k,0}$$

is positive semidefinite. Therefore, we have

$$\begin{aligned}\lambda_{\min}\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{\Sigma}_A\boldsymbol{R}_{k,0}\right) &= \inf_{\boldsymbol{a}_0\in\mathbb{R}^{s_0},\|\boldsymbol{a}_0\|_2=1} \boldsymbol{a}_0^T\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{\Sigma}_A\boldsymbol{R}_{k,0}\right)\boldsymbol{a}_0 \qquad (42)\\ = \inf_{\boldsymbol{a}_0\in\mathbb{R}^{s_0},\|\boldsymbol{a}_0\|_2=1} &\left\{\boldsymbol{a}_0^T\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T(\boldsymbol{\Sigma}_A-\bar{c}\boldsymbol{I}_{s_0})\boldsymbol{R}_{k,0}\right)\boldsymbol{a}_0+\bar{c}\boldsymbol{a}_0^T\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{R}_{k,0}\right)\boldsymbol{a}_0\right\}\\ \geq \bar{c}\inf_{\boldsymbol{a}_0\in\mathbb{R}^{s_0},\|\boldsymbol{a}_0\|_2=1} &\boldsymbol{a}_0^T\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{R}_{k,0}\right)\boldsymbol{a}_0 = \bar{c}\lambda_{\min}\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{R}_{k,0}\right)=\bar{c}\bar{K}.\end{aligned}$$

By the eigenvalue decomposition, we have

$$\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{\Sigma}_A\boldsymbol{R}_{k,0}=\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}\boldsymbol{U}_{RA},$$

for some orthogonal matrix $\boldsymbol{U}_{RA}\in\mathbb{R}^{s_0\times s_0}$ and some diagonal matrix $\boldsymbol{\Lambda}_{RA}\in\mathbb{R}^{s_0\times s_0}$. It follows from (42) that all the diagonal elements in $\boldsymbol{\Lambda}_{RA}$ are positive. Let $\boldsymbol{\Lambda}_{RA}^{1/2}$ be the diagonal matrix such that $\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{\Lambda}_{RA}^{1/2}=\boldsymbol{\Lambda}_{RA}$. Apparently, the diagonal elements in $\boldsymbol{\Lambda}_{RA}^{1/2}$ are nonzero. Notice that

$$\frac{1}{n^2}\sum_{k=1}^K \boldsymbol{A}_0\boldsymbol{R}_{k,0}^T\boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{R}_{k,0}\boldsymbol{A}_0^T=\frac{1}{n}\boldsymbol{A}_0\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}\boldsymbol{A}_0^T.$$

The s_0 largest eigenvalues in $\frac{1}{n^2}\sum_{k=1}^K \boldsymbol{A}_0\boldsymbol{R}_{k,0}^T\boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{R}_{k,0}\boldsymbol{A}_0^T$ corresponds to the smallest eigenvalue in

$$\frac{1}{n}\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}\boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}.$$

Similar to (42), we can show that

$$\begin{aligned}\lambda_{\min}\left(\frac{1}{n}\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}\boldsymbol{A}_0^T\boldsymbol{A}_0\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}\right) &\geq \bar{c}\lambda_{\min}\left(\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}^{1/2}\boldsymbol{U}_{RA}\right)\\ = \bar{c}\lambda_{\min}(\boldsymbol{U}_{RA}^T\boldsymbol{\Lambda}_{RA}\boldsymbol{U}_{RA}) &\geq \bar{c}^2\lambda_{\min}\left(\sum_{k=1}^K \boldsymbol{R}_{k,0}^T\boldsymbol{\Sigma}_A\boldsymbol{R}_{k,0}\right).\end{aligned}$$

Combining this together with (42), we obtain that

$$(\lambda_{n,k}^{(s_0)})^2\geq\bar{c}^2\bar{K}.$$

It follows from (41) that

$$\sum_{ijk}\left((\widehat{\boldsymbol{a}}_i^{(s)})^T\widehat{\boldsymbol{R}}_k^{(s)}\widehat{\boldsymbol{a}}_j^{(s)}-(\boldsymbol{a}_i^*)^T\boldsymbol{R}_k^*\boldsymbol{a}_j^*\right)^2\geq n^2\bar{c}^2\bar{K}.$$

This completes the proof.

A.4 Proof of Theorem 3.1

It suffices to show

$$\Pr\left(\mathrm{IC}(s_0) > \max_{1\le s<s_0} \mathrm{IC}(s)\right) \to 1, \tag{43}$$

and

$$\Pr\left(\mathrm{IC}(s_0) > \max_{s_0\le s\le s_{\max}} \mathrm{IC}(s)\right) \to 1. \tag{44}$$

We first show (43). Combining Lemma 3 with (21), we obtain that

$$2\ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \ge \frac{1}{\{1+\exp(M_0)\}^2}(\theta_{ijk}^* - \hat{\theta}_{ijk}^{(s)})^2 \ge \frac{n^2\bar{c}^2\bar{K}}{\{1+\exp(M_0)\}^2},$$

for some constant $M_0>0$. Hence, for sufficiently large n, we have for any $s\in\{1,\dots,s_0-1\}$,

$$2\ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \ge c_0 n^2 \bar{K},$$

for some constant $c_0>0$.

Combining this with (32), we obtain that

$$\begin{aligned}
& 2\ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \qquad\qquad (45)\\
\ge\ & 2\ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - 2\left|\ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - \ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k)\right|\\
-\ & 2\left|\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - \ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k)\right| \ge c_0 n^2\\
-\ & O_p\left(n\sqrt{K}\sqrt{(n+K)(\log n+\log K)} + (n+K)(\log n + \log K)\right).
\end{aligned}$$

Under the given conditions, we have $K\sim n^{l_0}$ for some $0\le l_0\le 1$. When $0\le l_0<1$, we have that

$$\begin{aligned}
& n\sqrt{K}\sqrt{(n+K)(\log n+\log K)} + (n+K)(\log n+\log K)\\
=\ & O(n^{1+l_0/2}\sqrt{n\log n}) + O(n\log n) \ll n^2\bar{K},
\end{aligned} \tag{46}$$

under the condition that $\liminf \bar{K}>0$. When $l_0=1$, (46) still holds, under the condition that $\liminf \bar{K} > \sqrt{\log n}$. Therefore,

$$n\sqrt{K}\sqrt{n\log n + K\log K} + (n\log n + K\log K) \ll n^2\bar{K}.$$

By (45), we have with probability tending to 1 that

$$2\ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \ge \frac{c_0}{2}n^2\bar{K}, \tag{47}$$

for all $1\le s<s_0$. By definition, we have

$$\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s_0)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s_0)}\}_k) \ge \ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k).$$

This together with (47) gives that for all $1 \leq s < s_0$,

$$2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s_0)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s_0)}\}_k) - 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \geq \frac{c_0}{2} n^2 \bar{K}, \tag{48}$$

with probability tending to 1.

Under the given conditions, we have $\kappa(n, K) \ll n^2 \bar{K}$. Under the event defined in (48), we have that

$$\begin{aligned} \mathrm{IC}(s_0) - \mathrm{IC}(s) &= 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s_0)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s_0)}\}_k) - 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - (s_0 - s)\kappa(n, K) \\ &\geq \frac{c_0}{2} n^2 \bar{K} - s_0 \kappa(n, K) \gg 0, \end{aligned}$$

since s_0 is fixed. This proves (43).

Now we show (44). Similar to (37), we can show that for any $s_0 \leq s \leq s_{\max}$,

$$\sup_{\substack{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^s \\ \boldsymbol{R}_1, \ldots, \boldsymbol{R}_K \in \mathbb{R}^{s \times s} \\ d(\{\boldsymbol{a}_i\}_i, \{\boldsymbol{R}_k\}_k; \{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) = O\left(\frac{r_{n,K}^{-1}}{n\sqrt{K}}\right)}} \left| \ell_n(\{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k) - \ell_0(\{\boldsymbol{a}_i^{(s)}\}_i, \{\boldsymbol{R}_k^{(s)}\}_k) \right| = O_p\left(\frac{1}{r_{n,K}^2}\right),$$

where $r_{n,K} = (n + K)^{-1/2}(\log n + \log K)^{-1/2}$. By Lemma 2, we obtain that for all $s_0 \leq s \leq s_{\max}$,

$$\left| 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - 2\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \right| = O_p\left(\frac{1}{r_{n,K}^2}\right), \tag{49}$$

and

$$|2\ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k)| = O_p\left(\frac{1}{r_{n,K}^2}\right). \tag{50}$$

Since $\ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) \geq \ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k)$, it follows from (49) and (50) that

$$\begin{aligned} & 2\ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \\ \geq\ & 2\ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \\ -\ & \left| 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) - 2\ell_0(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \right| \\ -\ & |2\ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k) - 2\ell_0(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k)| \geq O_p(r_{n,K}^{-2}). \end{aligned}$$

Observe that

$$\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s_0)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s_0)}\}_k) \geq \ell_n(\{\boldsymbol{a}_i^*\}_i, \{\boldsymbol{R}_k^*\}_k),$$

we have

$$2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s_0)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s_0)}\}_k) - 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) \geq O_p(r_{n,K}^{-2}), \tag{51}$$

for all $s_0 < s \leq s_{\max}$. Therefore,

$$\begin{aligned}\mathrm{IC}(s_0) - \mathrm{IC}(s) &\geq 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s_0)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s_0)}\}_k) - 2\ell_n(\{\widehat{\boldsymbol{a}}_i^{(s)}\}_i, \{\widehat{\boldsymbol{R}}_k^{(s)}\}_k) + (s - s_0)\kappa(n, K) \\ &\geq (s - s_0)\kappa(n, K) - O_p(r_{n,K}^{-2}).\end{aligned}$$

Under the condition $\kappa(n, K) \gg (n + K)(\log n + \log K)$, we have that

$$\Pr(\mathrm{IC}(s_0) > \mathrm{IC}(s)) \to 1,$$

for all $s_0 < s \leq s_{\max}$. Since $s_{\max}$ is bounded, (44) is proven. This completes the proof.

Appendix B. Discussion of Condition (A0)

In this section, we show the necessity of (A0) when the matrices $\boldsymbol{R}_{1,0}, \ldots, \boldsymbol{R}_{K,0}$ are symmetric. More specifically, when (A0) doesn't hold, we show there exist some $0 \leq s < s_0$, $\overline{\boldsymbol{a}}_{1,0}, \ldots, \overline{\boldsymbol{a}}_{n,0} \in \mathbb{R}^s$ and $\overline{\boldsymbol{R}}_{1,0}, \ldots, \overline{\boldsymbol{R}}_{K,0} \in \mathbb{R}^{s \times s}$ such that

$$\overline{\boldsymbol{a}}_{i,0}^T \overline{\boldsymbol{R}}_{k,0} \overline{\boldsymbol{a}}_{j,0} = \boldsymbol{a}_{i,0}^T \boldsymbol{R}_{k,0} \boldsymbol{a}_{j,0}, \quad \forall 1 \leq i, j \leq n, 1 \leq k \leq K. \tag{52}$$

Let's first consider the case when $\mathrm{rank}(\boldsymbol{A}_0) = s$ for some $s < s_0$. Thus, it follows that

$$\boldsymbol{A}_0 = \overline{\boldsymbol{A}}_0 \boldsymbol{C},$$

for some $\overline{\boldsymbol{A}}_0 \in \mathbb{R}^{n \times s}$ and $\boldsymbol{C} \in \mathbb{R}^{s \times s_0}$. Set $\overline{\boldsymbol{a}}_{i,0}$ to be the ith row of $\overline{\boldsymbol{A}}_0$ and $\overline{\boldsymbol{R}}_{k,0} = \boldsymbol{C}\boldsymbol{R}_{k,0}\boldsymbol{C}^T$. (52) is thus satisfied. In addition, the new matrix $\overline{\boldsymbol{A}}_0$ shall have full column rank.

Let $\boldsymbol{R}_0 = (\boldsymbol{R}_{1,0}^T, \ldots, \boldsymbol{R}_{K,0}^T)^T$. Consider the case when $\mathrm{rank}(\boldsymbol{R}_0) = s$ for some $s < s_0$. It follows from the singular value decomposition that

$$\boldsymbol{R}_0 = \boldsymbol{U}_0 \boldsymbol{\Lambda}_0 \boldsymbol{V}_0^T, \tag{53}$$

for some diagonal matrix $\boldsymbol{\Lambda}_0 \in \mathbb{R}^{s \times s}$, and some matrices $\boldsymbol{U}_0 \in \mathbb{R}^{Ks_0 \times s}$, $\boldsymbol{V}_0 \in \mathbb{R}^{s_0 \times s}$ that satisfy $\boldsymbol{U}_0^T \boldsymbol{U}_0 = \boldsymbol{V}_0^T \boldsymbol{V}_0 = \boldsymbol{I}_s$. Denoted by $\boldsymbol{U}_{k,0}$ the submatrix of $\boldsymbol{U}_0$ formed by rows in $\{(k-1)s_0 + 1, (k-1)s_0 + 2, \cdots, ks_0\}$ and columns in $\{1, 2, \ldots, s\}$. It follows from (53) that

$$\boldsymbol{R}_{k,0} = \boldsymbol{U}_{k,0} \boldsymbol{\Lambda}_0 \boldsymbol{V}_0^T, \quad \forall 1 \leq k \leq K.$$

Since $\boldsymbol{R}_{k,0}$ is symmetric, we have $\boldsymbol{U}_{k,0} \boldsymbol{\Lambda}_0 \boldsymbol{V}_0^T = \boldsymbol{V}_0 \boldsymbol{\Lambda}_0 \boldsymbol{U}_{k,0}^T = \boldsymbol{R}_{k,0}$. Notice that $\boldsymbol{V}_0^T \boldsymbol{V}_0 = \boldsymbol{I}_s$. It follows that $\boldsymbol{U}_{k,0} \boldsymbol{\Lambda}_0 = \boldsymbol{V}_0 \boldsymbol{\Lambda}_0 \boldsymbol{U}_{k,0}^T \boldsymbol{V}_0$. Therefore, we have

$$\boldsymbol{R}_{k,0} = \boldsymbol{V}_0 \boldsymbol{\Lambda}_0 \boldsymbol{U}_{k,0}^T \boldsymbol{V}_0 \boldsymbol{V}_0^T. \tag{54}$$

Define $\overline{\boldsymbol{a}}_{i,0} = \boldsymbol{V}_0^T \boldsymbol{a}_{i,0}, \forall 1 \leq i \leq n$ and $\overline{\boldsymbol{R}}_{k,0} = \boldsymbol{\Lambda}_0 \boldsymbol{U}_{k,0}^T \boldsymbol{V}_0$. In view of (54), it is immediate to see that (52) holds. Since $\boldsymbol{V}_0^T \boldsymbol{V}_0 = \boldsymbol{I}_s$, we have $\overline{\boldsymbol{R}}_{k,0} = \boldsymbol{V}_0^T \boldsymbol{V}_0 \boldsymbol{\Lambda}_0 \boldsymbol{U}_{k,0}^T \boldsymbol{V}_0 = \boldsymbol{V}_0^T \boldsymbol{R}_{k,0} \boldsymbol{V}_0$. As a result, $\overline{\boldsymbol{R}}_{1,0}, \ldots, \overline{\boldsymbol{R}}_{K,0}$ are also symmetric. Suppose $\overline{\boldsymbol{R}}_0 = (\overline{\boldsymbol{R}}_{1,0}^T, \cdots, \overline{\boldsymbol{R}}_{K,0}^T)^T$ doesn't have full column rank. Using the same arguments, we can find some $s' < s$, $\widetilde{\boldsymbol{a}}_{1,0}, \cdots, \widetilde{\boldsymbol{a}}_{n,0} \in \mathbb{R}^{s'}$, $\widetilde{\boldsymbol{R}}_{1,0}, \cdots, \widetilde{\boldsymbol{R}}_{K,0} \in \mathbb{R}^{s' \times s'}$ such that

$$\widetilde{\boldsymbol{a}}_{i,0}^T \widetilde{\boldsymbol{R}}_{k,0} \widetilde{\boldsymbol{a}}_{j,0} = \boldsymbol{a}_{i,0}^T \boldsymbol{R}_{k,0} \boldsymbol{a}_{j,0}, \quad \forall 1 \leq i, j \leq n, 1 \leq k \leq K. \tag{55}$$

We may repeat this procedure until we find some $\widetilde{\boldsymbol{a}}_{i,0}$'s, $\widetilde{\boldsymbol{R}}_{k,0}$'s satisfy (55) and that the matrix $(\widetilde{\boldsymbol{R}}_{1,0}^T, \cdots, \widetilde{\boldsymbol{R}}_{K,0}^T)^T$ has full column rank.

References

Genevera Allen. Sparse higher-order principal components analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 27–36, 2012.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002. ISSN 0012-9682. doi: 10.1111/1468-0262.00273. URL `http://dx.doi.org/10.1111/1468-0262.00273`.

Sudipto Banerjee and Anindya Roy. *Linear algebra and matrix analysis for statistics*. CRC Press, 2014.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Testing many moment inequalities. *arXiv preprint arXiv:1312.7614*, 2013.

Eric C. Chi and Tamara G. Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM J. Matrix Anal. Appl.*, 33(4):1272–1299, 2012. ISSN 0895-4798. URL `https://doi.org/10.1137/110859063`.

Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(3):531–552, 2013. ISSN 1369-7412. doi: 10.1111/rssb.12001. URL `http://dx.doi.org/10.1111/rssb.12001`.

Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Patrick Alken, Michael Booth, Fabrice Rossi, and Rhys Ulerich. *GNU Scientific Library Reference Manual (Version 2.1)*, 2015. URL `http://www.gnu.org/software/gsl/`.

Lise Getoor and Lilyana Mihalkova. Learning statistical models from relational data. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 1195–1198. ACM, 2011.

Richard A Harshman. Models for analysis of asymmetrical relationships among n objects or stimuli. In *Paper presented at the First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology, Hamilton, Ontario, August*, 1978.

Richard A Harshman and Margaret E Lundy. Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72, 1994.

Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

Stanley Kok and Pedro Domingos. Statistical predicate invention. In *Proceedings of the 24th international conference on Machine learning*, pages 433–440. ACM, 2007.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009. ISSN 0036-1445. URL `https://doi.org/10.1137/07070111X`.

Anmol Madan, Manuel Cebrian, Sai Moturu, Katayoun Farrahi, et al. Sensing the" health state" of a community. *IEEE Pervasive Computing*, 11(4):36–45, 2012.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3): 277–289, 2008.

Maximilian Nickel. *Tensor factorization for relational learning*. PhD thesis, the Ludwig-Maximilians-University of Munich, 2013.

Maximilian Nickel and Volker Tresp. Logistic tensor factorization for multi-relational data. *arXiv preprint arXiv:1306.2084*, 2013.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 809–816, 2011.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.*, 96(455):1077–1087, 2001. ISSN 0162-1459. doi: 10.1198/016214501753208735. URL `http://dx.doi.org/10.1198/016214501753208735`.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62 (1):107–136, 2006.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978. ISSN 0090-5364. URL `http://links.jstor.org/sici?sici=0090-5364(197803)6:2<461:ETDOAM>2.0.CO;2-5&origin=MSN`.

Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):899–916, 2017. ISSN 1369-7412. URL `https://doi.org/10.1111/rssb.12190`.

Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. doi: 10.1007/978-1-4757-2545-2. URL `http://dx.doi.org/10.1007/978-1-4757-2545-2`. With applications to statistics.

Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2017.

Yun Yang and David B. Dunson. Bayesian conditional tensor factorizations for high-dimensional classification. *J. Amer. Statist. Assoc.*, 111(514):656–669, 2016. ISSN 0162-1459. URL `https://doi.org/10.1080/01621459.2015.1029129`.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. A consistent information criterion for support vector machines in diverging model spaces. *J. Mach. Learn. Res.*, 17:Paper No. 16, 26, 2016. ISSN 1532-4435.