Journal of Machine Learning Research 27 (2026) 1-135 Submitted 7/24; Revised 11/25; Published 5/26

Multi-relational Network Autoregression Model with Latent Group Structures

Yimeng Ren[1,2] YMREN@UST.HK
[1] *School of Data Science*
Fudan University
Shanghai, China
[2] *HKUST Business School*
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Xuening Zhu* XUENINGZHU@FUDAN.EDU.CN
School of Management
Fudan University
Shanghai, China

Ganggang Xu* GANGXU@BUS.MIAMI.EDU
Department of Management Science
University of Miami
Coral Gables, FL 33146, USA

Yanyuan Ma YANYUANMA@GMAIL.COM
Department of Statistics
The Pennsylvania State University
University Park, PA 16802, USA

Editor: Ji Zhu

Abstract

Multi-relational networks among entities are frequently observed in the era of big data. Quantifying the effects of multiple networks has attracted significant research interest recently. In this work, we model multiple network effects through an autoregressive framework for tensor-valued time series. To characterize the potential heterogeneity of the networks and handle the high dimensionality of the time series data simultaneously, we assume a separate group structure for entities in each network and estimate all group memberships in a data-driven fashion. Specifically, we propose a group tensor network autoregression (GTNAR) model, which assumes that within each network, entities in the same group share the same set of model parameters, and the parameters differ across networks. An iterative algorithm is developed to estimate the model parameters and the latent group memberships simultaneously. Theoretically, we show that the group-wise parameters and group memberships can be consistently estimated when the group numbers are correctly- or possibly over-specified. An information criterion for estimating the group number for each network is also provided to consistently select the group numbers. Lastly, we apply the GTNAR method to a Yelp dataset to illustrate its usefulness.

Keywords: Multi-relational networks, Latent group, Tensor-valued time series, Network autoregression.

* Corresponding authors.

1. Introduction

As the world becomes increasingly connected, studying network effects has become an important research topic across disciplines, including economics, finance, and many others. The primary focus of our study is to analyze the network effects inherent in high-dimensional time series observed over multiple networks. The existing literature has seen notable progress in the study of time series within a single network. For instance, Zhu et al. (2017) introduced a network autoregression model to investigate time series in large social networks. Armillotta and Fokianos (2023) extended the framework to the nonlinear network autoregression model, in which the multivariate observations could be both continuous and discrete. Chen et al. (2023) proposed a community network vector autoregression model for the high-dimensional time series. Fang et al. (2023) proposed a group network Hawkes process with a single network to model unit event occurrences. Chen et al. (2026) proposed a group network multivariate GARCH model that incorporates a latent group structure and an observed network adjacency matrix. Compared with conventional high-dimensional time series models (Walden and Serroukh, 2002; Leng and Tang, 2012; Zhou, 2014; Wang et al., 2019; Chen et al., 2020a; Chang et al., 2023; Wang and Tsay, 2023; Chen and Fan, 2023), the network autoregressive model distinguishes itself by providing enhanced parameter interpretability, offering deeper insights into the intricacies of network dynamics.

In real-world scenarios, entities within a population commonly establish connections across multiple networks, often referred to as multi-relational or multi-layer networks. Recently, there has been significant interest in investigating these networks collectively, as seen in works like Lei et al. (2020), Zhang et al. (2020), Jing et al. (2021), MacDonald et al. (2022), and Ma and Nandy (2023). While the majority of these studies focus on identifying community structures within multi-layer networks, there is also a considerable interest in quantifying the impacts of multi-relational network effects on various research objectives. For example, Emch et al. (2016) investigated the joint effects of spatial and social networks on disease transmission. Chen et al. (2017) proposed the utilization of network metrics from various social networks to predict the adoption of new products in marketing research. Besides, Corradini et al. (2021) investigated the influence of multi-dimensional social networks on negative reviews posted on Yelp. Although these models are valuable in empirical research, a critical gap remains in the availability of rigorous statistical models capable of providing valid statistical inferences about multiple network effects, and it is our intention to fill this gap.

The objective of our work is to investigate tensor-valued time series indexed across multi-relational networks. A straightforward way to analyze tensor-valued data is to stack it into vectors or to consider only one of its dimensions. However, this will destroy the intrinsic multidimensional structure and lack clear interpretations, and provide limited insights into network effects. This drawback has also been recognized in several recent studies on matrix- or tensor-valued time series (Zhou and Li, 2014; Wang et al., 2019; Chen et al., 2020a, 2021, 2022; Chen and Fan, 2023; Wang et al., 2024; Chen and Lam, 2024). Additionally, although stacking into vectors allows the implementation of techniques for high-dimensional VAR models, this still leads to a much larger number of parameters that need to be estimated, on the order of $\prod_{l=1}^{q}(N_l)^2$, where q is the dimension of tensor data, and N_l is the number of nodes in the lth dimension. Even with commonly used sparsity regularization,

the estimation variance of model parameters is likely to be much higher than that of the network VAR model when the data is generated by the latter (Zhu et al., 2023).

To address high dimensionality, several recent studies focus on factor models to capture low-rank structures (Zhou et al., 2023; Han et al., 2023; Wang et al., 2024). For example, Chen et al. (2022) introduced a general framework of factor models for tensor-valued time series. Han et al. (2024) propose a factor model and a high-order projection estimator to analyze high-dimensional dynamic tensor time series. Moreover, Wang et al. (2022) used the tensor decomposition technique to deal with the large transition matrices of the high-dimensional VAR model. Chen and Lam (2024) introduced a pre-averaging procedure, allowing for a spectrum of factor strengths, as well as the weak factors when both cross-sectional and serial correlations are present. Besides, Wang and Tsay (2023) proposed a general robust estimation procedure, which can address many high-dimensional VAR models with low-rank or sparse structure. However, existing works do not exploit network structure information across tensor dimensions and therefore do not provide direct statistical inference on these network effects.

On the other hand, it is remarkable that nodal heterogeneity widely exists in practice (Ke et al., 2015). The utilization of latent group structures to model heterogeneous data has a well-established history in panel data analysis. For instance, Ke et al. (2015) introduced a clustering algorithm in regression through data-driven segmentation to identify the groups, and Vogt and Linton (2017) shed light on the latent grouped structure in the non-parametric regression functions. Su et al. (2016) introduce a Classifier Lasso (C-Lasso) estimator for panel models. And more recently, Ando and Bai (2020) proposed a new estimation method for analyzing the quantile co-movement of large-scale financial time series data, which can identify latent group heterogeneity among the series. There have also been recent efforts to leverage latent group structures in modeling time series data within a single network, see, e.g., Zhu et al. (2023), Chen et al. (2023), Fang et al. (2023), Liu et al. (2024), and Chen et al. (2026). However, these works have predominantly focused on time series observed on a single network; hence, they cannot be directly used in modeling the tensor-valued time series data. Moreover, addressing the theoretical challenges involved in extending from a single network to multiple networks is non-trivial due to the interactions between different group structures, necessitating the development of new theoretical tools. To tackle this challenge, we divide the members of the network on each dimension into several latent subgroups. We assume that members within each subgroup share similar characteristics, carrying the same model coefficients.

The main contributions of our work can be summarized as follows. First, we introduce a highly interpretable network autoregression model for high-dimensional multivariate time series indexed by multiple networks, namely, the Group Tensor-valued Network Autoregression (GTNAR) model. Second, to account for network heterogeneity, we allow for separate group structures on related networks. Third, we establish estimation consistency for both model parameters and group memberships, not only when the numbers of groups are correct, but also when they are over-specified. Simultaneously estimating group memberships across multiple networks poses significant challenges compared to the existing literature, which typically considers at most one network (e.g., Fang et al., 2023). We develop new theoretical tools to address this challenge. Lastly, we develop a selection criterion that consistently selects the true group numbers and establishes asymptotic normality when the

group numbers are correctly specified. Our theoretical framework enables rigorous testing of multiple network effects, which are crucial across various research disciplines.

The remainder of this article is structured as follows. In Section 2, we first show a motivating example on the Yelp dataset to introduce a simple model form. Then, we introduce the general GTNAR model. Section 3 outlines the model estimation procedure and the selection method for the group numbers. Theoretical properties concerning parameter estimation, group membership estimation, and group number selection are discussed in Section 4. Two model extensions and corresponding theories are provided in Section 5. In Section 6, we present extensive simulation studies to illustrate the finite sample performance of GTNAR. Section 7 includes an application of GTNAR to the Yelp dataset. Finally, Section 8 provides concluding remarks. Additional technical proofs can be found in the Appendix.

2. Group Tensor-valued Network Autoregression

2.1 Notations and Tensor Algebra

Throughout the paper, we use the following notations. Denote $[n] = \{1, 2, \cdots, n\}$ for an integer n. For a vector $\mathbf{v} = (v_j : j \in [p])^\top \in \mathbb{R}^p$, let $\|\mathbf{v}\| = (\sum_{j=1}^p v_j^2)^{1/2}$. For a matrix $\mathbf{M} = (m_{ij}) \in \mathbb{R}^{n_1 \times n_2}$, let $\mathbf{M}_{i\cdot}$ be the ith row vector and $\mathbf{M}_{\cdot j}$ as the jth column vector of $\mathbf{M}$. In addition, let $\mathbf{M}^{(\mathcal{C},\cdot)} = (m_{ij} : i \in \mathcal{C}, j \in [n_2])$ and $\mathbf{M}^{(\cdot,\mathcal{C})} = (m_{ij} : i \in [n_1], j \in \mathcal{C})$, where $\mathcal{C}$ is an index set. For a symmetric matrix $\mathbf{M}$, define $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ as the smallest and largest eigenvalues, respectively. For a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_q}$, denote $\mathcal{X}^{(\mathcal{C}_1, \cdots, \mathcal{C}_q)}$ as the corresponding subset of the tensor with each dimension selected by the index sets $\mathcal{C}_l$s. Denote $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{n_1 n_3 \times n_2 n_4}$ as the Kronecker product between matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ and $\mathbf{B} \in \mathbb{R}^{n_3 \times n_4}$. For a matrix series $\{\mathbf{A}_l : l \in [q]\}$, denote $\mathbf{B} \otimes_{k \neq l} \mathbf{A}_k = \mathbf{B} \otimes \mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_{l-1} \otimes \mathbf{A}_{l+1} \otimes \cdots \otimes \mathbf{A}_q$. For a vector, matrix, or tensor $\mathbf{M}$, let $\|\mathbf{M}\|_{\max}$ denote its largest absolute entry. For a set $\mathcal{S}$, denote $|\mathcal{S}|$ as the cardinal number of $\mathcal{S}$. We denote $\mathbf{1}_p \in \mathbb{R}^p$ as a p-dimensional vector with all elements equal to one, and $\mathbf{I}_p$ is an identity matrix.

For a tensor $\mathcal{X} = (x_{i_1,\cdots,i_q})_{i_1 \in [n_1], \cdots, i_q \in [n_1]} \in \mathbb{R}^{n_1 \times \cdots \times n_q}$, denote the vectorization as $\mathbf{x} = \mathrm{vec}(\mathcal{X}) = (x_{1,1,\cdots,1}, x_{2,1,\cdots,1}, \cdots, x_{n_1,1,\cdots,1}, x_{1,2,1,\cdots,1}, \cdots, x_{n_1,n_2,\cdots,n_q})^\top \in \mathbb{R}^{\prod_l n_l}$. Denote its mode-$l$ matricization as $\mathcal{X}_{(l)} \in \mathbb{R}^{n_l \times \prod_{k \neq l} n_k}$, which is calculated by setting the lth tensor dimension as the matrix rows, and collapsing all other into its columns, for $l \in [q]$. For tensor $\mathcal{X} = (x_{i_1,\cdots,i_q}) \in \mathbb{R}^{n_1 \times \cdots \times n_q}$ and a matrix $\mathbf{M} = (m_{i,j}) \in \mathbb{R}^{q \times n_l}$, denote the mode-$l$ multiplication $\mathcal{X} \times_l \mathbf{M}$ as a tensor in $\mathbb{R}^{n_1 \times \cdots n_{l-1} \times q \times n_{l+1} \times \cdots \times n_q}$, and its $(i_1, \cdots, i_{l-1}, s, i_{l+1}, \cdots, i_q)$th element is calculated by $(\mathcal{X} \times_l \mathbf{M})_{i_1,\cdots,i_{l-1},s,i_{l+1},\cdots,i_q} = \sum_{i_l=1}^{n_l} x_{i_1,\cdots,i_l,\cdots i_q} m_{s,i_l}$. Subsequently, we denote $\mathcal{X} \times_{l=1}^q \mathbf{M}_l \stackrel{\text{def}}{=} (\mathcal{X} \times_1 \mathbf{M}_1) \times_2 \cdots \times_q \mathbf{M}_q$. For two tensors $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_q}$ and $\mathcal{Y} \in \mathbb{R}^{n_1 \times \cdots \times n_s}$ with $q \geq s$, their generalized inner product $\langle \mathcal{X}, \mathcal{Y} \rangle$ is a $(q-s)$ dimensional tensor, calculated by $\langle \mathcal{X}, \mathcal{Y} \rangle_{i_{s+1} \cdots i_q} = \sum_{i1=1}^{n_1} \cdots \sum_{i_s=1}^{n_s} \mathcal{X}_{i_1 \cdots i_s i_{s+1} \cdots i_q} \mathcal{Y}_{i_1 \cdots i_s}$. We use $\mathcal{A} \odot \mathcal{Y}$ to denote the element-wise product between two tensors ($\mathcal{A}$ and $\mathcal{Y}$) with the same dimensions. For a column vector $\mathbf{x}_l \in \mathbb{R}^{n_l}$, $\mathbf{x}_l \circ_{k \neq l} \mathbf{1}_{n_k} = \mathbf{1}_{n_1} \circ \cdots \circ \mathbf{x}_l \circ \cdots \circ \mathbf{1}_{n_q}$ is a q dimensional tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times \cdots \times n_q}$ with its $(i_1, \cdots, i_l, \cdots, i_q)$th element being x_{i_l}.

2.2 Motivating Example and A Simple Model

The objective of our work is to investigate time series indexed across multi-relational networks. To offer a more lucid representation of GTNAR, we demonstrate it through a motivating dataset collected from Yelp (https://www.yelp.com/). Yelp serves as a prominent review platform for various businesses, including restaurants, local retailers, entertainment establishments, and more. It also functions as a social platform where users can share information and their personal experiences. The dataset covers the period from 2010 to 2018 and is collected from five North American cities (i.e., Charlotte, Las Vegas, Phoenix, Scottsdale and Toronto), and comprises four main categories of information: user data (e.g., user registration time on Yelp), user-friend relationships, business information (including spatial location), and user reviews of businesses. For example, Figure 1 illustrates a user's review of a restaurant named "Esther's Kitchen" in Las Vegas. In this case, the user gave the restaurant a five-star rating, and their review received 18 tags from other users, including 8 "useful", 3 "funny", and 7 "cool" tags. Overall, the restaurant has accumulated 1611 reviews.

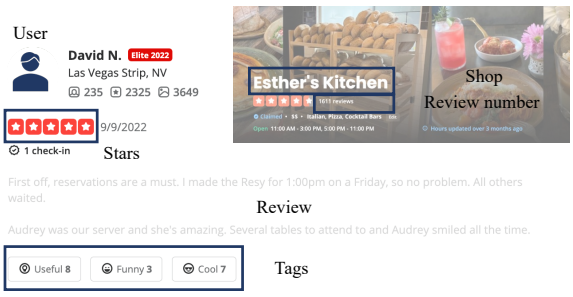
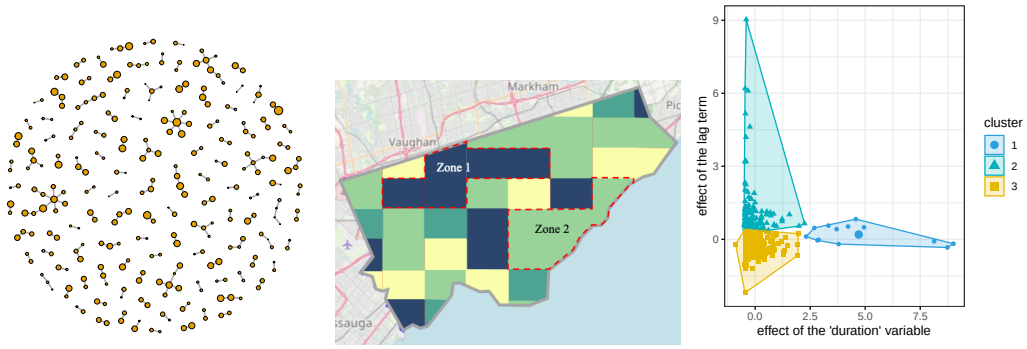


Figure 1: A review snapshot of the shop "Esther's Kitchen". It contains the user information, shop statistics, review text, and the tags assigned to this review.

Within the Toronto segment of the dataset, we have records for $N_1 = 462$ users who have provided reviews for restaurants in $N_2 = 56$ distinct locations, spanning $T = 36$ quarters. Our primary focus in this analysis concentrates on the variable denoted as $Y_{ij,t}$, representing the $\log(1+x)$-transformed number of reviews contributed by user i to restaurants in district j during the tth quarter. This variable forms a time series indexed by both the user ID (i) and the district ID (j). The first challenge we encounter when analyzing this dataset is that neither users nor districts can be considered isolated units. As a result, it becomes imperative to model the $Y_{ij,t}$'s in a collective manner. Specifically, users form a social network, while districts establish a spatial network that fosters substantial interactions among their respective network members. These interactions, in turn, significantly influence the outcome variable $Y_{ij,t}$ when considered jointly. For example, Tiwari and Richards (2016) found that peer social networks play a highly effective role in influencing restaurant preferences within social circles of friends. The social network analysis conducted on Yelp data by Fe (2023) suggests that social network friends are 64% more likely to visit the same restaurant when compared to non-friends. Sun and Paule (2017) discovered significant

spatial effects on ratings across various categories of Yelp venues, and Gan et al. (2021) investigated the spatial network effects on the tourism economy. However, these studies are primarily empirical and lack a strong statistical foundation. Furthermore, they tend to concentrate solely on the impact of a single network, rather than considering multiple network effects jointly. As a motivating example, Figure 2(a) depicts the social network of Toronto users with at least two friends, highlighting the observation that connected friends often have similar comment volumes. Meanwhile, Figure 2(b) presents the spatial network of Toronto, indicating that neighboring districts, such as zones 1 and 2, tend to exhibit similar comment volumes. Both network effects contribute jointly to the outcome variable. Therefore, the first challenge we intend to address is how to construct a multivariate time series model that can rigorously quantify the impact of multiple network effects for data similar to the Yelp review dataset.



(a) Social network (b) Spatial network (c) Estimated parameters

Figure 2: (a) The social network of Toronto users with degrees greater than one. Node sizes reflect the logarithm of the number of comments made in 2008. (b) The spatial network of Toronto districts during the last quarter of 2018. Colors represent four quantile intervals of the logarithm of the number of comments made in the 4th quarter of 2008. Darker colors indicate districts with higher comment activity. (c) Clustering results based on estimated regression coefficients, categorizing users into three distinct groups. Different colors and shapes visually represent each group.

The second challenge we encounter in the Yelp review dataset relates to the heterogeneity among members in both social and spatial networks. Previous research in network analysis (Newman et al., 2006) highlighted that users tend to naturally form clusters based on similar behaviors, a phenomenon consistent with Yelp's ecosystem, and users display distinct interaction patterns (e.g., frequent reviewers versus casual browsers) (Fe, 2023; Sun and Paule, 2017). There is also empirical research (Wedel and Kamakura, 2000; Fiebig et al., 2010), demonstrating that segmenting users enhances recommendation accuracy on platforms with varied preferences, which is particularly applicable to Yelp's diverse range of restaurants and businesses (Jagabathula et al., 2018). Similarly, restaurants located in different spatial regions, such as the central business district or other areas, may experience different levels of popularity, leading to varying degrees of spatial spillover effects (see, e.g., Koschinsky, 2009). Recent advances (e.g., Bonhomme and Manresa (2015); Su et al. (2016);

Liu et al. (2020)) in high-dimensional time series modeling demonstrated that subgrouping mitigates bias in parameter estimation and enhances estimation efficiency given limited sample sizes, which is also a critical feature for the Yelp data. As a motivating example, we conduct a preliminary regression analysis with aggregated users' comment volumes (log-transformed) as the response variable, denoted as $Y_{it} = \sum_j Y_{ij,t}$. Two covariates are included: the lagged response ($Y_{i(t-1)}$) and the user's duration since registration (i.e., the number of months since joining Yelp as of the $(t-1)$th quarter). The estimated regression coefficients obtained from all users are clustered into three groups using the k-means algorithm, and these clusters are visualized in Figure 2(c). An evident heterogeneous pattern is observed among the users' coefficients. To tackle this challenge, we adopt the approach proposed in Zhu et al. (2023), which involves dividing the members of both social and spatial networks into several subgroups. We assume that members of each subgroup share similar characteristics.

Let $\mathbf{A}^{(1)} = (a_{ij}^{(1)}) \in \mathbb{R}^{N_1 \times N_1}$ and $\mathbf{A}^{(2)} = (a_{ij}^{(2)}) \in \mathbb{R}^{N_2 \times N_2}$ represent the observed exogenous adjacency matrices characterizing the social and spatial networks, respectively. Specifically, $a_{ij}^{(1)} = 1$ implies that the ith user follows the jth user, while $a_{ij}^{(1)} = 0$ otherwise. Similarly, $a_{ij}^{(2)} = 1$ indicates that the ith district is a spatial neighbor of the jth district, while $a_{ij}^{(2)} = 0$ otherwise. We follow the convention by setting $a_{ii}^{(1)} = 0$ and $a_{jj}^{(2)} = 0$ for $1 \le i \le N_1$ and $1 \le j \le N_2$. We assume the existence of G_1 groups in the social network and G_2 groups in the spatial network. The group membership for the ith node in the social network is denoted as $g_i^{(1)}$ $(1 \le g_i^{(1)} \le G_1)$, and the group membership for the jth node in the spatial network is denoted as $g_j^{(2)}$ $(1 \le g_j^{(2)} \le G_2)$. We propose the following model with a two-way group structure:

$$
\begin{aligned}
Y_{ij,t} = \underbrace{\lambda_{g_i^{(1)}}^{(1)} \sum_{k=1}^{N_1} \frac{a_{ik}^{(1)}}{n_{1i}} Y_{kj,(t-1)}}_{\text{Social Network main effect}} + \underbrace{\lambda_{g_j^{(2)}}^{(2)} \sum_{k=1}^{N_2} \frac{a_{kj}^{(2)}}{n_{2j}} Y_{ik,(t-1)}}_{\text{Spatial Network main effect}} \\
+ \underbrace{\alpha_{g_i^{(1)} g_j^{(2)}} Y_{ij,(t-1)}}_{\text{Self-momentum}} + \underbrace{\mathbf{x}_{it}^{(1)\top} \boldsymbol{\zeta}_{g_i^{(1)}}^{(1)} + \mathbf{x}_{jt}^{(2)\top} \boldsymbol{\zeta}_{g_j^{(2)}}^{(2)}}_{\text{Covariate effects}} + \varepsilon_{ij,t},
\end{aligned} \tag{1}
$$

where $n_{1i} = \sum_{k=1}^{N_1} a_{ik}^{(1)}$, $n_{2j} = \sum_{k=1}^{N_2} a_{kj}^{(2)}$, $\mathbf{x}_{it}^{(1)} \in \mathbb{R}^{p_1}$ and $\mathbf{x}_{jt}^{(2)} \in \mathbb{R}^{p_2}$ are exogenous covariate vectors of finite dimensions associated with the ith user and jth district, respectively. Besides, $\varepsilon_{ij,t}$ represents independent and identically distributed (i.i.d.) white noise with $E(\varepsilon_{ij,t}) = 0$ and its variance $\mathrm{var}(\varepsilon_{ij,t}) = \sigma^2$. For identifiability, it is required that $\sum_{g^{(1)}=1}^{G_1} \zeta_{g^{(1)},1}^{(1)} = 0$ when both intercepts are included in $\mathbf{x}_{it}^{(1)}$ and $\mathbf{x}_{it}^{(2)}$, where $\zeta_{g_i^{(1)},1}^{(1)}$ (the first element of $\boldsymbol{\zeta}_{g_i^{(1)}}^{(1)}$) represents the intercept for $\mathbf{x}_{it}^{(1)}$.

The first term in (1), i.e., $\sum_{k=1}^{N_1} (a_{ik}^{(1)}/n_{1i}) Y_{kj,(t-1)}$, represents the average number of reviews ($\log(1+x)$-transformed) by user i's following friends on the restaurants in district j in the previous quarter. Consequently, $\lambda_{g_i^{(1)}}^{(1)}$ quantifies the influence of following friends on user i's attitude towards district j and encapsulates a social network main effect. On the other hand, the second term in (1), i.e., $\sum_{k=1}^{N_2} (a_{kj}^{(2)}/n_{2j}) Y_{ik,(t-1)}$, calculates the average

log number of reviews by the ith user on districts that are connected with the jth district in the previous quarter. Thus, $\lambda^{(2)}_{g^{(2)}_j}$ signifies how the user's review count towards district j is influenced by his/her reviews towards the districts connected with district j, and can be interpreted as the spatial network main effect. Additionally, $\alpha_{g^{(1)}_i g^{(2)}_j}$ represents the self-driven time effect for the (i,j)th time series, quantifying the momentum effect of the review activity by the ith user towards the jth district in the previous quarter. A higher value of $\alpha_{g^{(1)}_i g^{(2)}_j}$ suggests a greater level of loyalty of user i to district j. Finally, $\boldsymbol{\zeta}^{(1)}_{g^{(1)}_i} \in \mathbb{R}^{p_1}$ and $\boldsymbol{\zeta}^{(2)}_{g^{(2)}_j} \in \mathbb{R}^{p_2}$ are external covariate effects at the user and district levels, enhancing the model's capacity to account for user and district heterogeneity in the data. By evaluating the significance of $\lambda^{(1)}_{g^{(1)}}$'s and $\lambda^{(2)}_{g^{(2)}}$'s, one can examine the presence of social and spatial network main effects while controlling for other factors in Model (1).

Our framework for multiple networks further extends the two-network model (1) to multi-relational networks given by Model (2) in Section 2.3 below, which we refer to as the GTNAR model.

2.3 General Model in Tensor Form

Suppose there are q observed exogenous networks characterized by adjacency matrices $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(q)}$, with the lth matrix $\mathbf{A}^{(l)} = (a^{(l)}_{ij}) \in \mathbb{R}^{N_l \times N_l}$, where N_l is the number of nodes in the lth network, and q is finite. Specifically, $a^{(l)}_{ij} = 1$ implies that the ith node is connected with the jth node in the lth network, while $a^{(l)}_{ij} = 0$ otherwise. We follow the convention by setting $a^{(l)}_{i_l i_l} = 0$ for $1 \le i_l \le N_l$ and $1 \le l \le q$. We assume that there exists G_l groups in the lth network, and the group membership for the ith node in this network is denoted as $g^{(l)}_{i_l}$ $(1 \le g^{(l)}_{i_l} \le G_l)$. As a result, each network has its own latent group structure among the N_l network nodes, which is denoted by $\mathcal{G}_l = (g^{(l)}_i : 1 \le i \le N_l)$, $1 \le l \le q$. Under such a model framework, the response variables of interest constitute a tensor-valued time series, denoted by $\mathcal{Y}_t = (Y_{i_1 i_2 \ldots i_q, t}) \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_q}$, with the following model structure

$$Y_{i_1 i_2 \ldots i_q, t} = \sum_{l=1}^{q} \underbrace{\lambda^{(l)}_{g^{(l)}_{i_l}} \sum_{k=1}^{N_l} \frac{a^{(l)}_{i_l k}}{n_{l i_l}} Y_{i_1 \ldots i_{l-1} k i_{l+1} \ldots i_q, (t-1)}}_{\text{The } l\text{th Network main effect}} + \underbrace{\alpha_{g^{(1)}_{i_1} \ldots g^{(q)}_{i_q}} Y_{i_1 i_2 \ldots i_q, (t-1)}}_{\text{Self-momentum}}$$
$$+ \sum_{l=1}^{q} \underbrace{\mathbf{x}^{(l)\top}_{i_l t} \boldsymbol{\zeta}^{(l)}_{g^{(l)}_{i_l}}}_{\text{The } l\text{th covariate effects}} + \varepsilon_{i_1 i_2 \ldots i_q, t}, \quad (2)$$

where $n_{l i_l} = \sum_{k=1}^{N_1} a^{(l)}_{i_l k}$, $\mathbf{x}^{(l)}_{i_l t} \in \mathbb{R}^{p_l}$ represents exogenous covariates associated with the i_lth member in the lth network, and $\varepsilon_{i_1 i_2 \ldots i_q, t}$ denotes the white noise with $\mathrm{var}(\varepsilon_{i_1 \cdots i_q, t}) = \sigma^2$. In this work, we treat the number of layers (q) and the covariates dimensions ($p_l,\ l \in [q]$) as fixed constants. For indentifiability, assume that $\sum_{g^{(l)}=1}^{G_l} \zeta^{(1)}_{g^{(l)},1} = 0$ $(l \in [q-1])$. Define $\mathcal{G}_l = (g^{(l)}_{i_l} : 1 \le i_l \le N_l)^\top \in \mathbb{R}^{N_l}$ and denote $\mathcal{G} = \{\mathcal{G}_l : 1 \le l \le q\}$. Let $\mathcal{R}^{(l)}_g = \{i_l : g^{(l)}_{i_l} = g\}$

and further denote $N_{lg} = |\mathcal{R}_g^{(l)}|$. Denote $g^{-(l)} = (g^{(1)}, \cdots, g^{(l-1)}, g^{(l+1)}, \cdots, g^{(q)})$ as the group membership indices excluding the lth network, and denote the group index $\mathcal{I}_{g^{(1)},\cdots,g^{(q)}}$ as a function of $(g^{(1)}, \cdots, g^{(q)})$, which takes values from $\{1, \cdots, \prod_l G_l\}$.

Define $\mathbf{W}^{(l)} = (a_{ij}^{(l)}/n_{li})$ as the lth row-normalized adjacency matrix of $\mathbf{A}^{(l)}$. Then the GTNAR model (2) can be expressed in a tensor form as follows,

$$\mathcal{Y}_t = \sum_{l=1}^{q} (\mathcal{Y}_{t-1} \times_l \mathbf{W}^{(l)}) \times_l \mathbf{L}^{(l)} + \mathcal{A} \odot \mathcal{Y}_{t-1} + \sum_{l=1}^{q} \boldsymbol{\beta}_{X_l,t}^{(l)} \circ_{k \neq l} \mathbf{1}_{N_k} + \mathcal{E}_t, \tag{3}$$

where $\mathbf{L}^{(l)} = \mathrm{diag}(\lambda^{(l)}_{g_{i_l}^{(l)}} : 1 \le i_l \le N_l) \in \mathbb{R}^{N_l \times N_l}$, $\mathcal{A} = (\alpha_{g_{i_1}^{(1)} \cdots g_{i_q}^{(q)}} : 1 \le i_l \le N_l) \in \mathbb{R}^{N_1 \times \cdots \times N_q}$, $\boldsymbol{\beta}_{X_l,t}^{(l)} = (\mathbf{x}_{i_l t}^{(l)\top} \boldsymbol{\zeta}^{(l)}_{g_{i_l}^{(l)}} : 1 \le i_l \le N_l)^\top \in \mathbb{R}^{N_l}$ and $\mathcal{E}_t = (\varepsilon_{i_1,\cdots,i_q,t}) \in \mathbb{R}^{N_1 \times \cdots \times N_q}$.

Remark 1. (Matrix Autoregression When $q = 2$) *When $q = 2$, the model in (3) can be simplified as the matrix-valued autoregression model in (4) below,*

$$\mathbf{Y}_t = (\mathbf{L}^{(1)}\mathbf{W}^{(1)})\mathbf{Y}_{t-1} + \mathbf{Y}_{t-1}(\mathbf{W}^{(2)}\mathbf{L}^{(2)}) + \mathbf{A} \circ \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{X_1,t}\mathbf{1}_{N_2}^\top + \mathbf{1}_{N_1}\boldsymbol{\beta}_{X_2,t}^\top + \mathbf{E}_t, \tag{4}$$

where $\mathbf{Y}_t \in \mathbb{R}^{N_1 \times N_2}$ is the response matrix, $\mathbf{L}^{(l)} = \mathrm{diag}(\lambda^{(l)}_{g_{i_l}^{(l)}} : 1 \le i_l \le N_l) \in \mathbb{R}^{N_l \times N_l}$ for $l = 1, 2$, $\mathbf{A} = (\alpha_{g_{i_1}^{(1)} g_{i_2}^{(2)}} : 1 \le i_1 \le N_1, 1 \le i_2 \le N_2) \in \mathbb{R}^{N_1 \times N_2}$, and $\mathbf{E}_t \in \mathbb{R}^{N_1 \times N_2}$ is the random noise matrix.

Remark 2. *We remark that GTNAR is not limited to the motivating example and can be utilized in a wide range of applications. Specifically, GTNAR is designed for tensor-valued time series data, provided we can collect network relationships among the units. The network under consideration is not restricted to spatial networks, as we use in the motivating example. For instance, in financial data analysis, we can use the common shareholding relationship to examine the spillover effects of systematic risk (Feng et al., 2023). In the international trading market, regions can construct two adjacent networks from the export and import aspects, while the trading products categories form a similarity network (Alves et al., 2019). In the contextual recommendation system, the social network relationships among the users form a network on the first dimension, the similarities based on the item characteristics form the second one, and the proximities of different types of users' activities mode form the third network (Adomavicius et al., 2005; Wu et al., 2016). Other examples include supply chain linkages (Serpa and Krishnan, 2018), common interests in customer purchase (Carroni et al., 2020), and others, according to the application scenarios.*

Remark 3. *Applications of tensor-valued time series are widespread, spanning numerous fields. For instance, a contextual recommender system naturally generates a 3-dimensional tensor of users, items, and contexts, as it recommends items to users under varying contexts (such as the promotion types and the users' locations) over a period of time; each dimension can be further enriched by its corresponding network, such as a user social network, an item similarity network, and a contextual proximity network. Similarly, international trade data, capturing the export volumes of different product categories between regions over a period of time, forms another 3-dimensional tensor-valued time series, where geographical*

and product-similarity networks can be constructed on each dimension. In the domain of neuroimaging, functional Magnetic Resonance Imaging (fMRI) produces a series of 3D brain scans, creating a 3-dimensional tensor-valued time series where networks can be defined by the adjacency of brain regions. A fourth example is found in online live-streaming, where user activities (such as comments and shares) directed at various streamers and different topics form a 4-dimensional tensor. This type of tensor recorded along a period of time naturally forms a 4-dimensional tensor-valued time series. The practical and urgent need to analyze this type of complex, high-dimensional data, which is ubiquitous in modern socio-economic and scientific contexts, serves as the primary applied motivation for the GTNAR model.

Remark 4. (Parameter Dimension Reduction) *We remark that we reduce the parameter dimension by embedding the observed network weighting matrices* $\mathbf{W}^{(k)}$ *and a group structure. Take the case with* $q = 2$ *for example. In general, one needs to estimate* $O(N_1^2 + N_2^2)$ *parameters in model* (4)*. In contrast, with the imposed model structure, the number of estimated parameters reduces to* $O(G_1(p_1+1) + G_2(p_2+1) + G_1G_2)$ *when the group memberships are given, which is greatly reduced due to incorporating the weighting matrices* $\mathbf{W}^{(l)}$ *and memberships* $\mathcal{G}_1$ *and* $\mathcal{G}_2$*.*

2.4 Comparisons with Existing Literature

There are numerous works that are closely related to GTNAR. The most pertinent works encompass high-dimensional vector autoregression models (Davis et al., 2016; Wang et al., 2022; Miao et al., 2023), tensor-valued time series models (Hoff, 2015; Wang et al., 2019; Chen et al., 2021, 2022; Chen and Fan, 2023; Wang et al., 2024), grouped panel data models (Ke et al., 2015; Ando and Bai, 2016; Su et al., 2016), and the network autoregressive models with heterogeneous effects (Zhu and Pan, 2020; Zhu et al., 2023). In this subsection, we illustrate how the GTNAR model differs from the above existing models.

2.4.1 High-dimensional Vector Autoregression Models

Recently, there have been studies focusing on the high-dimensional vector autoregression (VAR) models (Davis et al., 2016; Wang et al., 2022; Miao et al., 2023). By stacking elements of the tensor $\mathcal{Y}_t = (Y_{i_1\cdots,i_q,t}) \in \mathbb{R}^{N_1\times\cdots\times N_q}$ into a vector, GTNAR model (3) can be re-written as a VAR model with dimension $N' = \prod_l N_l$,

$$\mathbf{y}_t = \Big[\sum_{l=1}^{q}\Big((\mathbf{L}^{(l)}\mathbf{W}^{(l)}) \otimes_{k\neq l} \mathbf{I}_{N_k}\Big) + \mathrm{diag}(\mathbf{a})\Big]\mathbf{y}_{t-1} + \sum_{l=1}^{q}\Big(\boldsymbol{\beta}^{(l)}_{X_l,t} \otimes_{k\neq l} \mathbf{1}_{N_k}\Big) + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{a} = \mathrm{vec}(\mathcal{A}) \in \mathbb{R}^{N_1\times\cdots\times N_q}$ and $\boldsymbol{\varepsilon}_t = \mathrm{vec}(\mathcal{E}_t)$. This allows for the application of existing techniques of high-dimensional VAR models. While this formulation imposes fewer assumptions on the model coefficients, it leads to a much larger number of parameters to be estimated. Specifically, the number of parameters is $O(\prod_l N_l^2)$, in comparison to $O(\sum_l G_l(p_l+1) + \prod_l G_l)$ (when fixing the group memberships) for the GTNAR model. Even with commonly used sparsity regularization, the estimation variance is likely to inflate compared to the network models under certain scenarios (Zhu et al., 2023), which is also investigated in our numerical studies. In addition, the "stacking" operation will destroy the

intrinsic data structure in the original tensor form, hence the coefficients estimated in the general VAR model lack clear interpretations and provide limited insights into the network effects. These drawbacks are also recognized in existing literatures (Chen et al., 2021; Wang et al., 2024).

2.4.2 Tensor-valued Time Series Models

This line of research contains two main categories, the tensor autoregressive framework (Ding and Dennis Cook, 2018; Chen et al., 2021; Wang et al., 2024) category, and the tensor factor model (Hoff, 2015; Wang et al., 2019; Chen et al., 2020b, 2022; Chen and Fan, 2023) category. Specifically, GTNAR should be classified into the first category. For ease of understanding, we compare our modeling approach with existing approaches when $q = 2$, which reduces to the matrix case in (4).

First, the tensor autoregressive (TAR) models characterize the dynamics of $\mathbf{Y}_t$ using its lagged information $\mathbf{Y}_{t-1}$. For example, the TAR model proposed by Wang et al. (2024) takes the following form, i.e.,

$$\mathcal{Y}_t = \langle \mathcal{A}, \mathcal{Y}_{t-1} \rangle + \mathcal{E}_t, \tag{5}$$

where $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_q \times n_1 \times \cdots \times n_q}$ is a $2q$ dimensional tensor, which has Tucker ranks $(r_1, \cdots, r_q)$ with $r_l = \text{rank}(\mathcal{A}_{(l)})$. They assume that the tensor $\mathcal{A}$ takes a low rank structure as $\mathcal{A} = \mathcal{C} \times_{l=1}^{2q} \mathbf{U}_l$, $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_{2q}}$ is the core low-rank tensor and $\mathbf{U}_l \in \mathbb{R}^{n_l \times r_l}$ for $l \in [2q]$. When $q = 2$, the model (5) can be rewritten as

$$\mathbf{Y}_t = \mathbf{U}_3(\mathcal{C}\mathbf{U}_1^\top \mathbf{Y}_{t-1}\mathbf{U}_2)\mathbf{U}_4^\top + \mathbf{E}_t. \tag{6}$$

In model (6), the number of parameters to be estimated is $O\{\prod_{l=1}^4 r_l + 2(N_1 + N_2)\}$. Compared to the existing TAR modeling approach, the GTNAR model (4) has two major differences. First, we embed the network weighting matrices $\mathbf{W}^{(l)}$ into the autoregression matrices to characterize the network dependence structure. To further reduce the number of estimated parameters, we consider a group structure with autoregression coefficients in $\mathbf{L}^{(1)}$ and $\mathbf{L}^{(2)}$, respectively. As we commented in Remark 4 for $q = 2$, GTNAR contains $O(G_1(p_1 + 1) + G_2(p_2 + 1) + G_1G_2)$ parameters in the network autoregression coefficients when the group memberships are fixed. In addition, we provide an interpretable modeling framework with the group-specific autoregression coefficients. Furthermore, we would like to remark that the autoregression matrices in model (4) cannot be expressed with a low rank form unless the network matrices $\mathbf{W}^{(l)}$ take a low rank structure. Second, in addition to the network autoregression terms, we use $\mathbf{A} \circ \mathbf{Y}_{t-1}$ to capture the self-momentum effect, which cannot be characterized by (5) with the low-rank assumption. One can require $\mathbf{A}$ to have a specific low rank structure $\mathbf{A} = \mathbf{M}^{(1)}\mathbf{G}\mathbf{M}^{(2)\top}$, where $\mathbf{M}^{(l)} = (\mathbf{e}_{g_i^{(l)}} : 1 \le i \le N_l)^\top \in \mathbb{R}^{N_l \times G_l}$ is a membership matrix for the l-th layer, and $\mathbf{G} = (\alpha_{gg'} : g \in [G_1], g' \in [G_2]) \in \mathbb{R}^{G_1 \times G_2}$. This relates GTNAR to the traditional low-rank models (Chen et al., 2021; Wang et al., 2024). We remark that the self-momentum effect contains cross-layer parameters in $\mathbf{G}$, which poses challenges in our theoretical analysis since the memberships from each layer cannot be analyzed separately.

The second category of tensor-valued time series models includes the tensor factor (TF) models considered by Wang et al. (2019), Chen et al. (2022), and Chen and Fan (2023).

The TF model assumes that the dynamics of $\mathcal{Y}_t$ can be driven by a low-dimensional latent dynamic tensor factor $\mathcal{F}_t$. For example, the TF model proposed by Chen et al. (2022) takes the form $\mathcal{Y}_t = \mathcal{F}_t(\prod_k \times_k)\mathbf{A}_k + \mathcal{E}_t$. We refer to Hoff (2015) and Chen et al. (2020b) for other TF models in various forms. Although the TF model can be used to directly conduct a dimension reduction for $\mathcal{Y}_t$, it cannot quantify the historical self-momentum effect for $\mathcal{Y}_t$, and it also lacks a sound interpretation regarding the latent group structure.

We also compare the numerical performance of GTNAR with that of the multilinear tensor model in Hoff (2015) to illustrate the advantages of our proposed model.

2.4.3 Grouped Panel Data Models

Utilizing latent group structures to model heterogeneous data has a well-established history in statistical models (Ke et al., 2015; Bester and Hansen, 2016; Liu et al., 2020). For instance, Bonhomme and Manresa (2015) and Bester and Hansen (2016) introduced grouped linear panel models with time-varying fixed effects and individual fixed effects, respectively. Su et al. (2016) introduced a Classifier Lasso (C-Lasso) estimator for panel models, and more recently, Liu et al. (2020) explored estimation and inference in the presence of possible over-specification of the group number. Specifically, Su et al. (2016) considered the grouped linear penal model $Y_{it} = \boldsymbol{\beta}_i^\top \mathbf{x}_{it} + \mu_i + \varepsilon_{it}$, where the parameter $\boldsymbol{\beta}_i$ takes values from G_0 group centers $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{G_0}\}$. However, the model format designed for vector-valued responses limits its applicability to complex tensor-valued data, thereby lacking generality. Compared to these studies, the GTNAR model further shows the additional advantage in quantifying the multiple heterogeneous network effects by incorporating different types of networks in tensor dimensions.

2.4.4 Network Autoregression Model with Group Heterogeneity

There have also been recent efforts to leverage latent group structures in modeling time series data within a single network, as demonstrated by Zhu and Pan (2020); Zhu et al. (2023); Chen et al. (2023); and etc. However, these works have predominantly focused on time series observed on a single network, and to the best of our knowledge, our work is the first to tackle time series indexed by multiple networks with distinct group structures. Specifically, Zhu et al. (2023) considered an autoregression model for a vector formed response $\mathbf{y}_t$ as

$$\mathbf{y}_t = \mathbf{B}\mathbf{y}_{t-1} + \boldsymbol{\mu}_z + \boldsymbol{\varepsilon}_t, \tag{7}$$

where $\mathbf{B} \in \mathbb{R}^{N\times N}$ with its (i,j)th element being $b_{ij} = w_{ij}\beta_{g_i g_j}$ with $w_{ij} = a_{ij}/n_i$ for $i \neq j$ and $b_{ii} = \nu_{g_i}$ for $i = 1, \cdots, N$. In (7), $\boldsymbol{\mu}_z = (\mathbf{x}_1^\top \boldsymbol{\zeta}_{g_1}, \cdots, \mathbf{x}_N^\top \boldsymbol{\zeta}_{g_N})^\top$ represents the covariates term. The key difference between these two approaches lies in how they model network interactions. The goal of Zhu et al. (2023) is to capture interactions within a single network as flexibly as possible, where the effect between nodes i and j, namely, $w_{ij}\beta_{g_i g_j}$, depends jointly on their group memberships g_i and g_j. This makes the consistent estimation of group memberships substantially more challenging than in other existing group panel data models (Su et al., 2016; Liu et al., 2020), where group memberships do not interact in the parameter structure. To address this difficulty, Zhu et al. (2023) proposed a membership refinement procedure based on brute-force enumeration. While a straightforward extension of this approach to the multiple network tensor-valued data setting studied might be possible, we

adopt a completely different within-network effect model in (4), where in each network, the effect between node i and j takes the form $w_{ij}^{(l)}\lambda_{g_i^{(l)}}$, depending only on the membership of node i (but not j) within the lth network, $l = 1, \ldots, q$. This approach has the advantage of avoiding the refinement procedure of Zhu et al. (2023), while has its own complexity of handling multiple networks simultaneously. Thus, the theoretical challenges in Zhu et al. (2023) and in our work stem from different sources, and we employ very different strategies to resolve them. The GTNAR model is a new modeling approach that leverages the general grouping idea to address challenges unique to the multi-layer networks setting.

3. Model Estimation

We first introduce the necessary notations. For the network corresponding to the lth tensor dimension, there exist G_l groups for the network nodes, denoted by $g^{(l)} \in [G_l]$. For each group, the group-level parameters are denoted by $\boldsymbol{\theta}_{g^{(l)}}^{(l)} = (\lambda_{g^{(l)}}^{(l)}, \boldsymbol{\zeta}_{g^{(l)}}^{(l)\top})^\top \in \mathbb{R}^{p_l+1}$. The collection of layer-specific parameters is then denoted as $\boldsymbol{\theta}^{(l)} = (\boldsymbol{\theta}_{g^{(l)}}^{(l)} : g^{(l)} \in [G_l]) \in \mathbb{R}^{G_l(p_l+1)}$, leading to the collection of $\boldsymbol{\xi} = (\boldsymbol{\theta}^{(1)\top}, \cdots, \boldsymbol{\theta}^{(q)\top}, \mathrm{vec}(\boldsymbol{\alpha})^\top)^\top \in \mathbb{R}^{\sum_l G_l(p_l+1)+\prod_l G_l}$ for all q tensor dimensions. Finally, the self-momentum parameter is denoted as $\boldsymbol{\alpha} = (\alpha_{g^{(1)}\cdots g^{(q)}})_{g^{(l)}\in[G_l]} \in \mathbb{R}^{G_1\times\cdots\times G_q}$. In the following, we discuss estimation for the GTNAR model, which includes the group parameters $\boldsymbol{\xi}$ and memberships $\mathcal{G}$. We utilize an iterative algorithm to update $\boldsymbol{\xi}$ and $\mathcal{G}$. We first discuss the estimation of $\boldsymbol{\xi}$ when $\mathcal{G}$ is given, and then we introduce the iterative algorithm for joint estimation of $\{\boldsymbol{\xi}, \mathcal{G}\}$. For easy understanding, we first introduce the estimation procedure for the matrix model in (4) when $q = 2$. Then we present a general estimation procedure for a general q.

3.1 Estimation Procedure for Model (4) when $q = 2$

We first introduce the estimation of model (4) with two networks for clear illustration. We aim to minimize the following objective function

$$Q(\boldsymbol{\xi}, \mathcal{G}) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T}\Big(Y_{ij,t} - \lambda_{g_i^{(1)}}^{(1)}\sum_{k=1}^{N_1}\frac{a_{ik}^{(1)}}{n_{1i}}Y_{kj,(t-1)} - \lambda_{g_j^{(2)}}^{(2)}\sum_{k=1}^{N_2}\frac{a_{kj}^{(2)}}{n_{2j}}Y_{ik,(t-1)} - \alpha_{g_i^{(1)}g_j^{(2)}}Y_{ij,(t-1)} - \mathbf{x}_{it}^{(1)\top}\boldsymbol{\zeta}_{g_i^{(1)}}^{(1)} - \mathbf{x}_{jt}^{(2)\top}\boldsymbol{\zeta}_{g_j^{(2)}}^{(2)}\Big)^2. \tag{8}$$

To minimize (8), we solve

$$\frac{\partial Q(\boldsymbol{\xi}, \mathcal{G})}{\partial \boldsymbol{\theta}_{g^{(1)}}^{(1)}} = \mathbf{0}, \quad \frac{\partial Q(\boldsymbol{\xi}, \mathcal{G})}{\partial \boldsymbol{\theta}_{g^{(2)}}^{(2)}} = \mathbf{0}, \quad \frac{\partial Q(\boldsymbol{\xi}, \mathcal{G})}{\partial \alpha_{g^{(1)}g^{(2)}}} = 0.$$

By solving the equations above, we obtain the estimator when $q = 2$ as $\widehat{\boldsymbol{\theta}} = \mathbf{M}^{-1}\boldsymbol{\beta}$, where

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}^{(1)} & \mathbf{M}^{(12)} & \mathbf{M}^{(1\alpha)} \\ \mathbf{M}^{(12)\top} & \mathbf{M}^{(2)} & \mathbf{M}^{(2\alpha)} \\ \mathbf{M}^{(1\alpha)\top} & \mathbf{M}^{(2\alpha)\top} & \mathbf{M}^{\alpha} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \\ \mathbf{b}^{\alpha} \end{pmatrix},$$

and the specific expressions are given in Appendix B.

3.2 Estimation Procedure for General Model with General q

We now discuss the estimation of the GTNAR model (3) in general. To simultaneously estimate the model parameters and the group memberships, we aim to minimize the following least squares objective function:

$$Q(\boldsymbol{\xi},\mathcal{G}) = \sum_{i_1=1}^{N_1}\cdots\sum_{i_q=1}^{N_q}\sum_{t=1}^{T}\Big(Y_{i_1 i_2\ldots i_q,t} - \sum_{l=1}^{q}\lambda^{(l)}_{g^{(l)}_{i_l}}\sum_{k=1}^{N_l}\frac{a^{(l)}_{i_l k}}{n_{l i_l}}Y_{i_1\ldots i_{l-1}k i_{l+1}\ldots i_q,(t-1)} - \alpha_{g^{(1)}_{i_1}\ldots g^{(q)}_{i_q}}Y_{i_1 i_2\ldots i_q,(t-1)} - \sum_{l=1}^{q}\mathbf{x}^{(l)\top}_{i_l t}\boldsymbol{\zeta}^{(l)}_{g^{(l)}_{i_l}}\Big)^2. \tag{9}$$

We first discuss the estimation when the group memberships $\mathcal{G}$ are given. In this case, the minimization of (9) is equivalent to solving $\frac{\partial Q(\boldsymbol{\xi},\mathcal{G})}{\partial \boldsymbol{\theta}^{(l)}_{g^{(l)}}} = \mathbf{0}$, $\frac{\partial Q(\boldsymbol{\xi},\mathcal{G})}{\partial \alpha_{g^{(1)}\ldots g^{(q)}}} = 0$ for all $g^{(l)} \in [G_l]$ and $l \in [q]$. Recall that $N_{lg^{(l)}} = |\mathcal{R}^{(l)}_{g^{(l)}}|$ and let

$$\mathbb{X}^{(l)}_{g^{(1)}\ldots g^{(q)},t} = \Big(\mathrm{vec}\big\{(\mathcal{Y}_{t-1}\times_l \mathbf{W}^{(l)})_{g^{(1)}\ldots g^{(q)}}\big\}, \mathbf{1}_{N_{1g^{(1)}}}\otimes\cdots\otimes(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)}},\cdot)}\otimes\cdots\otimes\mathbf{1}_{N_{qg^{(q)}}}\Big), \tag{10}$$

where $\mathbf{X}^{(l)}_t = (\mathbf{x}^{(l)}_{1t},\cdots,\mathbf{x}^{(l)}_{N_l t})^\top \in \mathbb{R}^{N_l\times p_l}$, and that

$$(\mathcal{Y}_{t-1}\times_l \mathbf{W}^{(l)})_{g^{(1)}\ldots g^{(q)}} = \Big(\sum_{i_l=1}^{N_l} Y_{i_1,\cdots,i_l,\cdots,i_q,t-1}(a^{(l)}_{s,i_l}/n_{ls})\Big)_{i_1\in\mathcal{R}_{g^{(1)}},\cdots,s\in\mathcal{R}_{g^{(l)}},\cdots,i_q\in\mathcal{R}_{g^{(q)}}}.$$

Then one can verify that

$$\frac{\partial Q(\boldsymbol{\xi},\mathcal{G})}{\partial\boldsymbol{\theta}^{(l)}_{g^{(l)}}} = \Big(\sum_{t,g^{-(l)}}\mathbb{X}^{(l)\top}_{g^{(1)}\ldots g^{(q)},t}\mathbb{X}^{(l)}_{g^{(1)}\ldots g^{(q)},t}\Big)\boldsymbol{\theta}^{(l)}_{g^{(l)}} - \sum_{t,g^{-(l)}}\Big\{\mathbb{X}^{(l)\top}_{g^{(1)}\ldots g^{(q)},t}\Big(\mathbb{Y}_{g^{(1)}\ldots g^{(q)},t} - \mathbb{Y}_{g^{(1)}\ldots g^{(q)},(t-1)}\alpha_{g^{(1)}\ldots g^{(q)}} - \sum_{m\neq l}\mathbb{X}^{(m)}_{g^{(1)}\ldots g^{(q)},t}\boldsymbol{\theta}^{(m)}_{g^{(m)}}\Big)\Big\}, \tag{11}$$

where $\mathbb{Y}_{g^{(1)}\ldots g^{(q)},t} = \mathrm{vec}\Big\{\mathcal{Y}_t^{\left(\mathcal{R}^{(1)}_{g^{(1)}},\cdots,\mathcal{R}^{(q)}_{g^{(q)}}\right)}\Big\} \in \mathbb{R}^{N_{1g^{(1)}}\cdots N_{qg^{(q)}}}$, and the summation $\sum_{g^{-(l)}}$ is the simplified notation for $\sum_{g^{(1)},\cdots,g^{(l-1)},g^{(l+1)},\cdots g^{(q)}}$. Furthermore, it holds that

$$\frac{\partial Q(\boldsymbol{\xi},\mathcal{G})}{\partial\alpha_{g^{(1)}\ldots g^{(q)}}} = \sum_t\|\mathbb{Y}_{g^{(1)}\ldots g^{(q)},(t-1)}\|^2\alpha_{g^{(1)}\ldots g^{(q)}} - \sum_t\mathbb{Y}^\top_{g^{(1)}\ldots g^{(q)},(t-1)}\Big(\mathbb{Y}_{g^{(1)}\ldots g^{(q)},t} - \sum_{l=1}^{q}\mathbb{X}^{(l)}_{g^{(1)}\ldots g^{(q)},t}\boldsymbol{\theta}^{(l)}_{g^{(l)}}\Big). \tag{12}$$

Equations (11) and (12) define a system of linear equations, whose solutions have the form $\widehat{\boldsymbol{\xi}} = \mathbf{M}^{-1}\mathbf{b}$ with $\mathbf{M}\in\mathbb{R}^{\{\sum_l G_l(p_l+1)+\prod_l G_l\}\times\{\sum_l G_l(p_l+1)+\prod_l G_l\}}$ and $\mathbf{b}\in\mathbb{R}^{\sum_l G_l(p_l+1)+\prod_l G_l}$,

where

$$\mathbf{M}=\begin{pmatrix} \mathbf{M}^{(1)} & \mathbf{M}^{(12)} & \mathbf{M}^{(13)} & \cdots & \mathbf{M}^{(1q)} & \mathbf{M}^{(1\alpha)} \\ \mathbf{M}^{(21)} & \mathbf{M}^{(2)} & \mathbf{M}^{(23)} & \cdots & \mathbf{M}^{(2q)} & \mathbf{M}^{(2\alpha)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{M}^{(q1)} & \mathbf{M}^{(q2)} & \cdots & \cdots & \mathbf{M}^{(q)} & \mathbf{M}^{(q\alpha)} \\ \mathbf{M}^{(1\alpha)\top} & \mathbf{M}^{(2\alpha)\top} & \cdots & \cdots & \mathbf{M}^{(q\alpha)} & \mathbf{M}^{(\alpha)} \end{pmatrix}, \quad \mathbf{b}=\begin{pmatrix} \mathbf{b}^{(1)} \\ \cdots \\ \mathbf{b}^{(q)} \\ \mathbf{b}^{\alpha} \end{pmatrix}. \tag{13}$$

The specific terms in (13) are given in Appendix C. Subsequently, given the estimated parameters $\widehat{\boldsymbol{\xi}}$, we update the group memberships $\mathcal{G}_l$ from $l=1$ to q iteratively. Specifically, given $\boldsymbol{\xi}$ and $\mathcal{G}_{-l} \overset{\text{def}}{=} \{\mathcal{G} \setminus \mathcal{G}_l\}$, the $\mathcal{G}_l$ is updated by

$$\begin{aligned} \widehat{g}_{i_l}^{(l)} \in \arg\min_{g_{i_l}^{(l)} \in [G_l]} \sum_{i_{-l}} \sum_{t=1}^{T} \Big(& Y_{i_1 i_2 \ldots i_q, t} - \sum_{l=1}^{q} \lambda_{g_{i_l}^{(l)}}^{(l)} \sum_{k=1}^{N_l} \frac{a_{i_l k}^{(l)}}{n_{l i_l}} Y_{i_1 \ldots i_{l-1} k i_{l+1} \ldots i_q, (t-1)} \\ & - \alpha_{g_{i_1}^{(l)} \ldots g_{i_q}^{(q)}} Y_{i_1 i_2 \ldots i_q, (t-1)} - \sum_{l=1}^{q} \mathbf{x}_{i_l t}^{(l)\top} \boldsymbol{\zeta}_{g_{i_l}^{(l)}}^{(l)} \Big)^2, \end{aligned} \tag{14}$$

where $\sum_{i_{-l}} = \sum_{i_1=1}^{N_1} \cdots \sum_{i_{l-1}=1}^{N_{l-1}} \sum_{i_{l+1}=1}^{N_{l+1}} \cdots \sum_{i_q=1}^{N_q}$.

We summarize the algorithm in Algorithm A.2 in Appendix D, which consists of iterations of two major steps. The first step updates the estimated parameters given the group memberships, while the second step updates group memberships given the parameter estimates. We note that the update step in (14) requires input of memberships of other layers (i.e., $\mathcal{G}_{-l}$). Therefore, we adopt a sequential updating rule. During the kth iteration, for $l' < l$, we substitute $\mathcal{G}_{l'}$ by $\mathcal{G}_{l'}^{[k]}$, and for $l' > l$, we substitute $\mathcal{G}_{l'}$ by $\mathcal{G}_{l'}^{[k-1]}$ in the previous iteration. See (A.25) for detailed expression. Each step can be efficiently computed due to the simple analytical forms. This algorithm can be validated to converge to a local minimizer, of which the proof is given in Appendix E. We remark that whether it can converge to a global optimizer is a challenging question due to the nonconvexity of the objective function (9) (Murty and Kabadi, 1987; Lin et al., 2020). To increase the chance of finding the global optimizer, a set of initialization strategies is employed, which are described in detail in Algorithm A.3. We try multiple initial values and use the one with the minimum loss function as the best one, which guarantees a stable numerical performance. We leave the theoretical investigation for a global optimizer as an interesting future topic.

Remark 5. *In Algorithm A.2,* Step 2 *achieves dimension reduction due to the calculation of group memberships. Note that in each layer, the update equation* (A.25) *only involves subject i_l, and does not depend on other inner-layer subjects. Thus, the memberships update equation* (A.25) *can be computed in parallel for each inner-layer subject (i_l in the lth layer), greatly increasing the computational speed. We show our computational cost under each sample size setting in Figure A.7. Even at the maximum network size configuration ($N_1 = 300, N_2 = 250$), total computational costs remain no more than 20 seconds across all time horizons. More discussion can be found in Appendix L.4.*

3.3 Selection of Group Numbers

To simplify the notations, denote $\underline{G} = (G_1, \cdots, G_q)^\top$, and denote the corresponding true group numbers as $\underline{G}_0$. We focus our later theoretical analysis on the case that G_l is finite

for $1 \le l \le q$. Write $\widehat{\boldsymbol{\xi}}(\underline{G}), \widehat{\mathcal{G}}(\underline{G})$ as the estimators when the group numbers are specified as $\underline{G}$. Then, we estimate $\underline{G}_0$ by utilizing the following information criterion:

$$\mathrm{QIC}(\underline{G}) = \log\{Q(\widehat{\boldsymbol{\xi}}(\underline{G}), \widehat{\mathcal{G}}(\underline{G}))\} + \lambda(\underline{G}), \tag{15}$$

where $Q(\cdot,\cdot)$ is defined in (9), and $\lambda(\underline{G})$ is a penalty function. Then we estimate the group numbers by $\widehat{\underline{G}} \in \arg\min_{\underline{G}} \mathrm{QIC}(\underline{G})$. In practice, we specify $\lambda(\underline{G}) = \kappa(\sum_l G_l)$, in which κ is a tuning parameter. Our theoretical analysis shows that if $T^{-1/2}(m + \sum_l \log N_l) \ll \kappa \ll c_{\mathrm{gap}} c_\pi^q / (\prod_l G_l)$, the QIC can consistently select $\underline{G}_0$. Here c_{gap} and c_π are group structures related values depending on the strength of model signals, which will be defined in our theoretical analysis in Section 4. In our numerical study, we specify $\kappa = \{C(\log T)T^{1/8}\}^{-1}$ with $C = 40$, which achieves reliable finite sample performances in all of our numerical studies.

4. Theoretical Properties

4.1 Estimation Consistency

Define $\boldsymbol{\Theta}_{i_1\cdots i_q} = (\boldsymbol{\theta}^{(1)\top}_{g^{(1)}_{i_1}}, \cdots, \boldsymbol{\theta}^{(q)\top}_{g^{(q)}_{i_q}}, \alpha_{g^{(1)}_{i_1}\cdots g^{(l)}_{i_q}})^\top \in \mathbb{R}^{\sum_l (p_l+1)+1}$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_{i_1\cdots i_q} : i_l \in [N_l], 1 \le l \le q)$ as a tensor of dimension $N_1 \times \cdots \times N_q \times \{\sum_l (p_l+1)+1\}$. With $\boldsymbol{\Theta}$ we can rewrite the loss function (9) as follows

$$Q(\boldsymbol{\Theta}) = \sum_{i_1=1}^{N_1} \cdots \sum_{i_q=1}^{N_q} \sum_{t=1}^{T} \big(Y_{i_1\cdots i_q,t} - \mathcal{X}^\top_{i_1\cdots i_q,t} \boldsymbol{\Theta}_{i_1\cdots i_q}\big)^2 \stackrel{\mathrm{def}}{=} \sum_{i_1=1}^{N_1} \cdots \sum_{i_q=1}^{N_q} Q_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q}), \tag{16}$$

where

$$\begin{aligned} \mathcal{X}_{i_1\cdots i_q,t} \stackrel{\mathrm{def}}{=} \big(& \sum_{k=1}^{N_1} w^{(1)}_{i_1 k} Y_{k i_2\cdots i_q,(t-1)}, \mathbf{x}^{(1)\top}_{i_1 t}, \cdots, \\ & \sum_{k=1}^{N_q} w^{(q)}_{i_q k} Y_{i_1\cdots i_{(q-1)} k,(t-1)}, \mathbf{x}^{(q)\top}_{i_q t}, Y_{i_1\cdots i_q,(t-1)} \big)^\top \in \mathbb{R}^{\sum_l (p_l+1)+1}. \end{aligned} \tag{17}$$

Denote by $\widehat{\boldsymbol{\Theta}} = \big(\widehat{\boldsymbol{\Theta}}_{i_1\cdots i_q} = (\widehat{\boldsymbol{\theta}}^{(1)\top}_{\widehat{g}^{(1)}_{i_1}}, \cdots, \widehat{\boldsymbol{\theta}}^{(q)\top}_{\widehat{g}^{(q)}_{i_q}}, \widehat{\alpha}_{\widehat{g}^{(1)}_{i_1}\cdots\widehat{g}^{(q)}_{i_q}})^\top\big)$ the global minimizer of $Q(\boldsymbol{\Theta})$ with the estimated group memberships $\widehat{g}^{(l)}_{i_l}$ $(i_l \in [N_l], l \in [q])$, we define a pseudo distance as follows

$$d(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) = \sum_{l=1}^{q} \frac{1}{N_l} \sum_{i_l=1}^{N_l} \|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}} - \boldsymbol{\theta}^{(l)}_{g^{(l)}_{i_l}}\|^2 + \frac{1}{\prod_l N_l} \sum_{i_1=1}^{N_1} \cdots \sum_{i_q=1}^{N_q} |\widehat{\alpha}_{\widehat{g}^{(1)}_{i_1}\cdots\widehat{g}^{(q)}_{i_q}} - \alpha_{g^{(1)}_{i_1}\cdots g^{(q)}_{i_q}}|^2. \tag{18}$$

Intuitively, $d(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta})$ measures the average distance between $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}$. Next, we establish the consistency of the global optimizer $\widehat{\boldsymbol{\Theta}}$ of the loss function (9) in this pseudo distance, for which we require the following definition and assumptions.

Definition 1. (K-Convex concentration) *Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector. If for every 1-Lipschitz convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$, we have $E|\varphi(\mathbf{x})| < \infty$ and for every $t > 0$, it holds that $P\Big(\Big|\varphi(\mathbf{x}) - E\{\varphi(\mathbf{x})\}\Big| \geq t\Big) \leq 2\exp(-t^2/K^2)$, then $\mathbf{x}$ is said to have the K-convex concentration property.*

Assumption 1. (Parameter Space) *The parameter satisfies that $\|\boldsymbol{\Theta}\|_{\max} < \infty$.*

Assumption 2. (Convexity) *Let $\boldsymbol{\Sigma}_{i_1\cdots i_q} = E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^{\top}_{i_1\cdots i_q,t})$. We assume that $\tau_{\min} \stackrel{\text{def}}{=} \min_{i_1,\cdots,i_q} \lambda_{\min}(\boldsymbol{\Sigma}_{i_1\cdots i_q})$ is a positive constant.*

Assumption 3. (Distribution of Noise Term) *Assume $\varepsilon_{i_1\cdots i_q,t}$ is i.i.d across $i_l \in [N_l]$ for all $l \in [q]$ and $t \in [T]$. In addition, $\varepsilon_{i_1\cdots i_q,t}$ is a zero-mean sub-Gaussian variable with a scale factor $0 < \nu < \infty$, i.e., $E\{\exp(u\varepsilon_{i_1\cdots i_q,t})\} \leq \exp(\nu^2u^2/2)$ for all $u \in \mathbb{R}$. Let $\varepsilon_{i_1\cdots i_q,t}$ be independent of $\{\mathcal{Y}_{s-1}, \mathbf{X}^{(1)}_s, \cdots, \mathbf{X}^{(q)}_s : s \leq t\}$, where $\mathbf{X}^{(l)}_s = (\mathbf{x}^{(l)}_{i_ls} : i_l \in [N_l])^{\top}$.*

Assumption 4. (Distribution of Covariates) *Recall that $\mathbf{x}^{(l)}_{i_lt} \in \mathbb{R}^{p_l}$ is the covariate vector of the i_lth subject in the lth layer at time t. Assume $E(\mathbf{x}^{(l)}_{i_lt}) = \mathbf{0}$ for all $i_l \in [N_l]$ and $t \in [T]$. Let $\boldsymbol{\eta}_l \in \mathbb{R}^{p_l}$ be a constant vector satisfying $\|\boldsymbol{\eta}_l\| \leq c$ for $l \in [q]$, where c is a positive constant. Define $\mathbf{x}^{(l)\eta}_t = (\mathbf{x}^{(l)\top}_{i_lt}\boldsymbol{\eta}_l : i_l \in [N_l])^{\top} \in \mathbb{R}^{N_l}$. Assume $\mathbf{x}^{(l)\eta} = (\mathbf{x}^{(l)\eta\top}_t : 0 \leq t \leq T)^{\top} \in \mathbb{R}^{N_l(T+1)}$ satisfies the K-convex concentration property for some constant K according to Definition 1.*

Assumption 5. (Stability) *Denote $\mathbb{Y}_0 \stackrel{\text{def}}{=} \text{vec}(\mathcal{Y}_0) = \mathbf{0}$, and assume that*

$$\max_{g^{(1)}\in[G_{1,0}],\cdots,g^{(q)}\in[G_{q,0}]} \Big|\sum_l \lambda^0_{g^{(l)}} + \alpha^0_{g^{(1)}\cdots g^{(q)}}\Big| \leq \kappa_{\max} < 1,$$

where $G_{l,0}$ is the true number of groups in the lth dimension, and $\kappa_{\max}$ is a positive constant.

Assumption 6. (Group Difference) *Assume $\min_{g^{(l)}_1 \neq g^{(l)}_2}\{\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}_1} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}_2}\|^2 + \max_{g^{-(l)}\in[G^0_{-l}]} |\alpha^0_{g^{(l)}_1 g^{-(l)}} - \alpha^0_{g^{(l)}_2 g^{-(l)}}|^2\} \geq c_{\text{gap}}$ holds for all $l \in [q]$, where $c_{\text{gap}} \gg T^{-1}(\sum_l \log N_l)^2$ as $T \to \infty$. Here $\{g^{-(l)} \in [G^0_{-l}]\}$ denotes $\{g^{(1)} \in [G_{1,0}], \cdots, g^{(l-1)} \in [G_{l-1,0}], g^{(l+1)} \in [G_{l+1,0}], \cdots, g^{(q)} \in [G_{q,0}]\}$.*

Assumption 7. (Group Proportion) *Let $\{g^{(l)0}_{i_l} : i_l \in [N_l]\}$ be non-random true membership sequences. Let $\pi^{(l)}_{g^{(l)},N_l} = \sum_{i_l} I(g^{(l)0}_{i_l} = g^{(l)})/N_l$ for $g^{(l)} \in [G_{l,0}]$. Assume that $\min_{l\in[q],g^{(l)}\in[G_{l,0}]} \pi^{(l)}_{g^{(l)},N_l} \geq c_\pi > 0$ for sufficiently large N_l, where c_π is a positive constant.*

Assumption 1 requires the parameter space to be bounded. Assumption 2 ensures the convexity of the element-wise objective function, i.e., $Q_{i_1\cdots i_1}(\boldsymbol{\Theta}_{i_1\cdots i_q})$, as a function of $\boldsymbol{\Theta}_{i_1\cdots i_q}$ for sufficiently large T. This condition guarantees the unique solution of the local objective function for each node. and is crucial for establishing the consistency result for the metric (18). These conditions are also commonly used in the literature, see, Self and Liang (1987), Fan and Li (2001), Zou (2006), etc., for example.

Assumptions 3–4 concern about the distributions of the error term and covariates, respectively. Specifically, Assumption 3 requires the error term $\varepsilon_{i_1\cdots i_q,t}$ to be *i.i.d.* sub-Gaussian variables, which is widely used in high-dimensional time series literature (Wang et al., 2013; Lugosi and Mendelson, 2019; Fan et al., 2021). We also provide a weighted least squares estimation procedure with group-specific variances, i.e., $\text{var}(\varepsilon_{i_1\cdots i_q,t}) = \sigma_{g_{i_1}^{(1)0}\cdots g_{i_q}^{(q)0}}$, in Section 5.2. Subsequently, Assumption 4 allows the covariates $\{\mathbf{x}_{i_l t}^{(l)}\}$ to be correlated but satisfying the K-convex concentration property according to Definition 1. This assumption is employed to establish Hanson-Wright type inequality for dependent variables (Adamczak, 2015). Although this is a high-level condition, there are a variety of random variables satisfying Definition 1, as discussed in the following Remark 6. We further comment that the Assumptions 3–4 together imply that $\mathbf{v}^\eta \stackrel{\text{def}}{=} (\mathbf{v}_t^\eta : 0 \le t \le T)^\top$ satisfies the K-convex concentration property for some constant K. Here $\mathbf{v}_t^\eta = (\mathbf{x}_t^{(1)\eta^\top}, \cdots, \mathbf{x}_t^{(q)\eta^\top}, \mathbb{E}_t^\top)^\top \in \mathbb{R}^{\sum_l N_l + \prod_l N_l}$, where $\mathbb{E}_t = \text{vec}(\mathcal{E}_t)$.

Remark 6. *As discussed by Adamczak (2015), there are a variety of random vectors* $\mathbf{x}$ *satisfying the* K*-convex concentration property in Definition 1. For example, (i) Any random vector* $\mathbf{x}$ *with its elements* x_i*s independent for all* i*, and* $|x_i| \le 1$ *a.s., satisfies Definition 1 (Talagrand, 1988); (ii) Any random vector* $\mathbf{x}$ *with its elements in a bounded interval and geometrically strongly mixing satisfies Definition 1 (Samson, 2000). We refer to Adamczak (2015) for more detailed discussions.*

Next, Assumption 5 ensures the stability of the tensor-valued time series data as T goes to infinity, as defined in Lütkepohl (2005). Assumptions 6 and 7 are imposed on certain group properties. Assumption 6 assumes there is a gap between the true parameters of two different groups within any network. The condition is an extension of the same type of conditions assumed by the group panel data models with groups assigned on one dimension (Su et al., 2016; Ando and Bai, 2016; Zhang et al., 2019; Liu et al., 2020). In Assumption 6, special care is paid to the self-momentum parameter $\alpha_{g^{(1)}\cdots g^{(q)}}$, indexed by the tensor dimension-wise group memberships, to ensure the parameter identifiability. Specifically, we require a min-max type condition for $\alpha^0_{g^{(1)}\cdots g^{(q)}}$ in Assumption 6. Furthermore, instead of assuming $c_{\text{gap}} > c > 0$ by a positive constant c in existing literature (Bonhomme and Manresa, 2015; Su et al., 2016; Zhang et al., 2019; Liu et al., 2020), we allow $c_{\text{gap}} \to 0$ to study how this signal strength affects the theoretical properties. Lastly, Assumption 7 assumes a lower bound of group proportions. In the following, we establish the consistency of the pseudo distance.

Theorem 2. *Suppose* $G_l \ge G_{l,0}$ *for all* $l \in [q]$*, where* $G_{l,0}$ *is the true number of groups. In addition, assume Assumptions 1–5 hold. We have that* $d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}^0) = O_p\{(\sum_l \log N_l)^2 T^{-1}\}$*, where* $\mathbf{\Theta}^0 = (\mathbf{\Theta}^0_{i_1\cdots i_q} = (\boldsymbol{\theta}^{(1)0\top}_{g_{i_1}^{(1)0}}, \cdots, \boldsymbol{\theta}^{(q)0\top}_{g_{i_q}^{(q)0}}, \alpha^0_{g_{i_1}^{(1)0}\cdots g_{i_q}^{(q)0}})^\top)$ *is the true parameter for* $\mathbf{\Theta}$*.*

Theorem 2 implies that as long as we have $T \gg (\sum_l \log N_l)^2$, $\widehat{\mathbf{\Theta}}$ is a consistent estimator for $\mathbf{\Theta}^0$ in the metric when G_l is possibly over-specified, i.e., $G_l \ge G_{l,0}$. This allows N_l to grow exponentially fast with $\sqrt{T}$, which is a mild condition compared to existing literatures (Su et al., 2016). As a result, it allows for moderately large N_l with respect to $\sqrt{T}$. Subsequently, we show in Theorem 3 that the QIC can consistently select the true group numbers.

Theorem 3. *Under Assumptions 1–7, and assume* $\kappa = \lambda(\underline{G})/(\sum_l G_l)$ *satisfies*

$$T^{-1}(\sum_l \log N_l)^2 \ll \kappa \ll c_{\text{gap}}/(\prod_l G_l). \tag{19}$$

Then we have $P(\widehat{G}_1 = G_{1,0}, \cdots, \widehat{G}_q = G_{q,0}) \to 1$ *as* $\min\{N_1, \cdots, N_q, T\} \to \infty$ *and* $T^{-1}(\sum_l \log N_l)^2 \to 0$.

Theorem 3 implies that if we set κ to satisfy (19), then we can consistently estimate the true group numbers. Denote $G_{-l} = (G_k : k \neq l)^\top \in \mathbb{R}^{q-1}$, we need $\kappa \gg T^{-1}(\sum_l \log N_l)^2$ to ensure that $\text{QIC}(\underline{G}) > \text{QIC}(\underline{G}_0)$ for the over-fitting case, i.e., $G_l > G_{l,0}$ and $G_{-l} \geq G_{-l,0}$ for any $l \in [q]$. Here $G_{-l} \geq G_{-l,0}$ means $G_k \geq G_{k,0}$ for any $k \neq l$. Conversely, we need $\kappa \ll c_{\text{gap}}/(\prod_l G_l)$ to guarantee that $\text{QIC}(\underline{G}) > \text{QIC}(\underline{G}_0)$ for the under-fitting case, i.e., there exists an l with $G_l < G_{l,0}$. When both conditions are met, we can obtain $\widehat{G}_l = G_{l,0}$ for all $l \in [q]$ with a probability approaching 1. In the next subsection, we further discuss the results of node-wise parameter estimation and the strong group membership consistency.

4.2 Membership Estimation Consistency and Asymptotic Normality

As we stated before, the metric in (18) measures the average estimation error between $\widehat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^0$. Therefore, the result in Theorem 2 is not sufficient to imply the parameter consistency for each node. To this end, we derive the following node-wise parameter consistency result, which will be crucial to building the strong membership estimation consistency later.

Proposition 4. *Under Assumptions 1–5, when* $G_l \geq G_{l,0}$ *for all* $l \in [q]$*, we have*

$$\begin{aligned}&\sup_{i_l} \left\{ \big\| \widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}} \big\|^2 + \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l} \sum_{i_m=1}^{N_m} \big| \widehat{\alpha}_{\widehat{g}^{(1)}_{i_1} \cdots \widehat{g}^{(q)}_{i_q}} - \alpha^0_{g^{(1)0}_{i_1} \cdots g^{(q)0}_{i_q}} \big|^2 \right\} \\ &= O_p\Big\{ T^{-1}(\sum_l \log N_l)^2 \Big\}. \end{aligned} \tag{20}$$

Recall that $\mathcal{G}_l = (g^{(l)}_{i_l} : 1 \leq i_l \leq N_l)^\top$, and let $\mathcal{G}^0_l$ denote the corresponding true memberships. Equation (20) establishes uniform node-wise parameter estimation consistency, which is crucial for achieving the following strong consistency of membership estimation for $\widehat{\mathcal{G}}_l$ when $G_l \geq G_{l,0}$ for all $l \in [q]$. Denote the node set $\widehat{\mathcal{R}}^{(l)}_{\widetilde{g}^{(l)}} = \{i_l : \widehat{g}^{(l)}_{i_l} = \widetilde{g}^{(l)}\}$ for $\widetilde{g}^{(l)} \in [G_l]$ and $\mathcal{R}^{(l)0}_{g^{(l)}} = \{i_l : g^{(l)0}_{i_l} = g^{(l)}\}$ for $g^{(l)} \in [G_{l,0}]$, where $\widehat{g}^{(l)}_{i_l}$ is obtained by (14) when $\widehat{\boldsymbol{\xi}}$ is specified.

Theorem 5. *(Strong Consistency of Membership Estimation) Under Assumptions 1–7, and suppose* $G_l \geq G_{l,0}$ *for all* $l \in [q]$*. Then for each estimated group* $\widetilde{g}^{(l)} \in [G_l]$*, there exists a true group* $g^{(l)} \in [G_{l,0}]$*, such that* $\widehat{\mathcal{R}}^{(l)}_{\widetilde{g}^{(l)}} \subset \mathcal{R}^{(l)0}_{g^{(l)}}$ *with probability tending to 1.*

As implied by Theorem 5, the true groups are split into subgroups instead of joining into new groups when $G_l \geq G_{l,0}$ for all $l \in [q]$, as shown in Figure A.6. Define $\widehat{\mathcal{G}}_l = (\widehat{g}^{(l)}_{i_l} : i_l \in [N_l])^\top \in \mathbb{R}^{N_l}$ as the estimated membership vectors for all $l \in [q]$. Particularly, when $G_l = G_{l,0}$, we can show that $\widehat{\mathcal{G}}_l = \mathcal{G}^0_l$ holds with probability tending to

1 under certain label permutations. We would like to remark that establishing the group consistency result is non-trivial due to several challenges under our framework. As shown in the GTNAR model (3), group memberships from different dimensions are involved and tangled in an additive form. In addition, the network dependence structure creates extra difficulty in establishing the node-wise membership consistency result. To establish the membership consistency, for instance, in Zhu et al. (2023), a refinement procedure is further required for obtaining this property. In Su et al. (2023), a two step estimation procedure is used to achieve group membership estimation consistency when the true group number is known. Specifically, their two-step procedure establishes node-wise parameter convergence and consistent group membership identification sequentially, whereas we allow simultaneous estimation of group membership and corresponding group-wise parameters within a unified framework. In contrast to their work, a further investigation of our additive model form enables us to address these challenges and establish the group membership consistency result for stochastically estimated $\widehat{\mathcal{G}}_l$ directly, which is critically important for later statistical inference. Furthermore, let $\widehat{\boldsymbol{\xi}}^{\text{or}}$ be the oracle estimator when the true group memberships $\mathcal{G}_l^0$s are known for all $l \in [q]$. Then the oracle property holds that $\widehat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^{\text{or}}$ with probability tending to 1, where $\widehat{\boldsymbol{\xi}}$ is the estimator for $\boldsymbol{\xi}^0 = (\boldsymbol{\theta}^{(1)0\top}, \cdots, \boldsymbol{\theta}^{(q)0\top}, \text{vec}(\boldsymbol{\alpha}^0)^\top)^\top$. The results are presented in the following Corollary.

Corollary 6. *Under Assumptions 1–7, and assume $G_l = G_{l,0}$ for all $l \in [q]$. Then under label permutations, we have*

$$\lim_{\min\{N_1,\cdots,N_q,T\}\to\infty} P\left(\widehat{\mathcal{G}}_1 = \mathcal{G}_1^0, \cdots, \widehat{\mathcal{G}}_q = \mathcal{G}_q^0\right) \to 1, \tag{21}$$

$$\lim_{\min\{N_1,\cdots,N_q,T\}\to\infty} P\left(\widehat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^{\text{or}}\right) \to 1. \tag{22}$$

The results in Corollary 6 imply that $\widehat{\boldsymbol{\xi}}$ is asymptotically equivalent to $\widehat{\boldsymbol{\xi}}^{\text{or}}$. Therefore, to derive the asymptotic distribution of $\widehat{\boldsymbol{\xi}}$, it is sufficient to investigate the case for $\widehat{\boldsymbol{\xi}}^{\text{or}}$. We establish the asymptotic normality result in the following theorem.

Theorem 7. *Assume Assumptions 1–7, $G_l = G_{l,0}$ and that there exists n, such that $c_1 n \le \min_l N_l \le \max_l N_l \le c_2 n$ for some constants $c_1, c_2 > 0$. Define $\mathbf{M}_{nT}^0 = n^{-q}T^{-1}E(\mathbf{M})$ and assume $\mathbf{M}^0 = \lim_{\min\{n,T\}\to\infty} \mathbf{M}_{nT}^0$ exists, where $\mathbf{M}$ is given in (13). Assume $\lambda_{\min}(\mathbf{M}^0) \ge \tau > 0$ for a positive constant τ. Let $k = \sum_l \{G_l(p_l+1)\} + \prod_l G_l$. Then for any $\boldsymbol{\eta} \in \mathbb{R}^k$ with $\|\boldsymbol{\eta}\| = 1$ we have*

$$n^{q/2}T^{1/2}\boldsymbol{\eta}^\top(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0) \to_d N\left(\mathbf{0}, \sigma^2\boldsymbol{\eta}^\top(\mathbf{M}^0)^{-1}\boldsymbol{\eta}\right). \tag{23}$$

Theorem 7 establishes the asymptotic normality of the estimator. Specifically, the convergence rates of $\widehat{\boldsymbol{\theta}}^{(l)}$ is $\sqrt{n^q T}$ for all $l \in [q]$. Using (23), we can conduct the statistical inference.

5. Model Extensions

The GTNAR model can be flexibly extended to accommodate an interactive model setting and group-specific error variances. In this section, we present the corresponding estimation and establish its statistical guarantees.

5.1 Mixed GTNAR Model

GTNAR model can be extended to incorporating the interactive network effects as

$$\begin{aligned}\mathcal{Y}_t = &(\mathcal{Y}_{t-1} \times_{l=1}^{q} \mathbf{W}^{(l)}) \times_{l=1}^{q} \mathbf{\Gamma}^{(l)} \\ &+ \sum_{l} (\mathcal{Y}_{t-1} \times_l \mathbf{W}^{(l)}) \times_l \mathbf{L}^{(l)} + \mathcal{A} \odot \mathcal{Y}_{t-1} + \sum_{l=1}^{q} \boldsymbol{\beta}^{(l)}_{X_l,t} \circ_{k\neq l} \mathbf{1}_{N_k} + \mathcal{E}_t, \end{aligned} \tag{24}$$

which we refer to as the Mixed GTNAR model. The first term represents the interactive network effects with $\mathbf{\Gamma}^{(l)} = \mathrm{diag}\big(\gamma^{(l)}_{g^{(l)}_{i_l}} : i_l \in [N_l]\big) = \mathrm{diag}(\boldsymbol{\gamma}^{(l)})$. Specifically, the interactive network coefficients are involved in (24) in a multiplicative form across different layers. Take the case of $q = 2$ as an example, model (24) can be expressed as

$$\begin{aligned}\mathbf{Y}_t = &(\mathbf{\Gamma}^{(1)}\mathbf{W}^{(1)})\mathbf{Y}_{t-1}(\mathbf{W}^{(2)}\mathbf{\Gamma}^{(2)}) + (\mathbf{L}^{(1)}\mathbf{W}^{(1)})\mathbf{Y}_{t-1} + \mathbf{Y}_{t-1}(\mathbf{W}^{(2)}\mathbf{L}^{(2)}) \\ &+ \mathbf{A} \odot \mathbf{Y}_{t-1} + \boldsymbol{\beta}^{(1)}_{X_1,t}\mathbf{1}^{\top}_{N_2} + \mathbf{1}_{N_1}\boldsymbol{\beta}^{(2)\top}_{X_2,t} + \mathbf{E}_t. \end{aligned} \tag{25}$$

Note that the multiplication form in $\mathbf{\Gamma}^{(1)}\mathbf{W}^{(1)}\mathbf{Y}_{t-1}\mathbf{W}^{(2)}\mathbf{\Gamma}^{(2)}$ causes the parameter identification issue. Following the convention (Chen et al., 2021), we set $\|\mathbf{\Gamma}^{(1)}\|_F = 1$ to guarantee the interactive network effects $\mathbf{\Gamma}^{(1)}$ and $\mathbf{\Gamma}^{(2)}$ to be identifiable with sign flips. To estimate model (25), we apply an iterative least squares method, and the detailed estimation algorithms are given in Appendix G.1. In the following, we present the theoretical properties for the estimator of model (24). Note that one can re-write the first interactive term in (25) as $(\boldsymbol{\gamma}^{(1)}\boldsymbol{\gamma}^{(2)\top}) \odot (\mathbf{W}^{(1)}\mathbf{Y}_{t-1}\mathbf{W}^{(2)})$, hence $\boldsymbol{\gamma}^{(1)}\boldsymbol{\gamma}^{(2)\top}$ plays a similar role as $\mathbf{A}$. As a result, we can borrow the proof idea for dealing with the autoregressive matrix $\mathbf{A}$ in our original model (4) with $q = 2$ to establish the consistency result for the Mixed GTNAR model. We first introduce some necessary conditions. For notational simplicity, denote $\mathbf{\Theta}_{ij} = (\boldsymbol{\theta}^{(1)\top}_{g^{(1)}_i}, \boldsymbol{\theta}^{(2)\top}_{g^{(2)}_j}, \alpha_{g^{(1)}_i g^{(2)}_j}, \gamma_{g^{(1)}_i}\gamma_{g^{(2)}_j})^{\top} \in \mathbb{R}^{p_1+p_2+4}$, and denote $\mathbf{\Theta} = (\mathbf{\Theta}_{ij} : i \in [N_1], j \in [N_2]) \in \mathbb{R}^{N_1\times N_2\times(p_1+p_2+4)}$. Let $\mathbf{\Theta}^0 = (\mathbf{\Theta}^0_{ij} = (\boldsymbol{\theta}^{(1)0\top}_{g^{(1)}_i}, \boldsymbol{\theta}^{(2)0\top}_{g^{(2)}_j}, \alpha^0_{g^{(1)}_i g^{(2)}_j}, \gamma^0_{g^{(1)0}_i}\gamma^0_{g^{(2)0}_j})^{\top})$ be the true parameter for $\mathbf{\Theta}$. Define the pseudo distance as

$$\begin{aligned} d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}) = &\frac{1}{N_1}\sum_{i=1}^{N_1} \|\widehat{\boldsymbol{\theta}}^{(1)}_{\widehat{g}^{(1)}_i} - \boldsymbol{\theta}^{(1)}_{g^{(1)}_i}\|^2 + \frac{1}{N_2}\sum_{j=1}^{N_2} \|\widehat{\boldsymbol{\theta}}^{(2)}_{\widehat{g}^{(2)}_j} - \boldsymbol{\theta}^{(2)}_{g^{(2)}_j}\|^2 \\ &+ \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=2}^{N_2}\Big\{|\widehat{\alpha}_{\widehat{g}^{(1)}_i\widehat{g}^{(2)}_j} - \alpha_{g^{(1)}_i g^{(2)}_j}|^2 + |\widehat{\gamma}^{(1)}_{\widehat{g}^{(1)}_i}\widehat{\gamma}^{(2)}_{\widehat{g}^{(2)}_j} - \gamma^{(1)}_{g^{(1)}_i}\gamma^{(2)}_{g^{(2)}_j}|^2\Big\}. \end{aligned} \tag{26}$$

In the above distance, we note that the interactive effect $\widehat{\gamma}^{(1)}_{\widehat{g}^{(1)}_i}\widehat{\gamma}^{(2)}_{\widehat{g}^{(2)}_j}$ plays a similar role as $\widehat{\alpha}_{\widehat{g}^{(1)}_i\widehat{g}^{(2)}_j}$ in $d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta})$. Hence, we can borrow the proof idea of Theorem 5 to establish the estimation consistency result with $d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta})$ for model (25). The required assumptions are given in Appendix G.2. We state the formal theoretical results in the following.

Theorem 8. *Under Assumptions 2–4 and A.11–A.13, the following conclusions hold.*
(i) Suppose $G_l \geq G_{l,0}$ for $l = 1, 2$, then $d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}^0) = O_p\{T^{-1}(\log(N_1N_2))^2\}$.
(ii) Assume $\kappa_1 = \lambda(\underline{G})/(G_1 + G_2)$ satisfies

$$T^{-1}(\log(N_1N_2))^2 \ll \kappa_1 \ll c_{\text{gap}}/(G_1G_2). \tag{27}$$

Then $P(\widehat{G}_1 = G_{1,0}, \widehat{G}_2 = G_{2,0}) \to 1$ as $\{N_1, N_2, T\} \to \infty$ and $T^{-1}(\log(N_1N_2))^2 \to 0$.
(iii) Further suppose Assumption 7 and A.14 hold, and $G_l \geq G_{l,0}$ for $l = 1, 2$. Then for any estimated group $\widehat{g}^{(1)} \in [G_1]$ and $\widehat{g}^{(2)} \in [G_2]$, there exists true groups $g^{(1)} \in [G_{1,0}]$ and $g^{(2)} \in [G_{2,0}]$, such that

$$P\Big\{\widehat{\mathcal{R}}^{(1)}_{\widehat{g}^{(1)}} \in \mathcal{R}^{(1)0}_{g^{(1)0}}\Big\} \to 1, \quad P\Big\{\widehat{\mathcal{R}}^{(2)}_{\widehat{g}^{(2)}} \in \mathcal{R}^{(2)0}_{g^{(2)0}}\Big\} \to 1,$$

where $\widehat{\mathcal{R}}^{(l)}_{\widehat{g}^{(l)}} = \{i_l : \widehat{g}^{(l)}_{i_l} = \widehat{g}^{(l)}\}$ for $l = 1, 2$.

In Theorem 8 (i), we first establish the estimation consistency based on the pseudo distance $d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta})$ when the group numbers are possibly over-specified. It shows that when the interactive model is correctly specified or over-specified, the parameter $\widehat{\mathbf{\Theta}}$ is a consistent estimator for $\mathbf{\Theta}^0$ as long as $T \gg (\log(N_1N_2))^2$. Next in (ii), we show that the QIC selection method in (A.30) leads to consistent group number estimators. When the tuning parameter κ_1 satisfies the condition (27), we can consistently estimate the group numbers. The lower bound and upper bound for κ_1 are applied for the results in over- and under-specified situations, respectively. Subsequently, we establish the strong group membership estimation consistency in (iii). The proof of Theorem 8 is given in Appendix G.4. Since the interactive parameter $\gamma^{(1)}_{g_i^{(1)}}\gamma^{(2)}_{g_j^{(2)}}$ plays a similar role as $\alpha_{g_i^{(1)}g_j^{(2)}}$ in model (25), the proofs exhibit analogous roadmap to the proofs of Theorem 2–5 in Section 4. However, the new interactive term in $\mathcal{X}_{ij,t}$ defined in (A.31) requires additional technical Lemmas 13–16 and Lemma 18, which are all crucial to the proof of the theorem.

To demonstrate the finite sample performance, we conduct several simulation experiments, which can be found in Appendix G.3.

5.2 Weighted Least Squares Estimation with Group-specific Error Variances

Next, we design a weighted least squares estimator for the case where the error terms have group-specific variances when $q = 2$. Several simulation studies are presented in Appendix B.4.5.

5.2.1 Estimation Procedure

In the case where the error term has group-specific variance, the GTNAR model can be modified as

$$\begin{aligned} Y_{ij,t} = {} & \lambda^{(1)}_{g_i^{(1)}} \sum_{k=1}^{N_1} w_{1ik} Y_{kj,(t-1)} + \lambda^{(2)}_{g_j^{(2)}} \sum_{k=1}^{N_2} Y_{ik,(t-1)} w_{2kj} + \alpha_{g_i^{(1)}g_j^{(2)}} Y_{ij,(t-1)} \\ & + \mathbf{x}^{(1)\top}_{it} \boldsymbol{\zeta}^{(1)}_{g_i^{(1)}} + \mathbf{x}^{(2)\top}_{jt} \boldsymbol{\zeta}^{(2)}_{g_j^{(2)}} + \varepsilon_{ij,t}, \end{aligned} \tag{28}$$

where $\varepsilon_{ij,t}$ satisfies that $E(\varepsilon_{ij,t}) = 0$ and $\text{var}(\varepsilon_{ij,t}) = \sigma^2_{g_i^{(1)0} g_j^{(2)0}}$, and $g_i^{(1)0}$ are $g_j^{(2)0}$ are true memberships. Under this situation, a weighted least squares (WLS) estimation procedure can be introduced (Carroll and Cline, 1988; Shao, 1989). Specifically, similar to the calculation for (A.7), a weighted version for the corresponding terms can be written as

$$\widetilde{\mathbf{M}} = \begin{pmatrix} \widetilde{\mathbf{M}}^{(1)} & \widetilde{\mathbf{M}}^{(12)} & \widetilde{\mathbf{M}}^{(1\alpha)} \\ \widetilde{\mathbf{M}}^{(12)\top} & \widetilde{\mathbf{M}}^{(2)} & \widetilde{\mathbf{M}}^{(2\alpha)} \\ \widetilde{\mathbf{M}}^{(1\alpha)\top} & \widetilde{\mathbf{M}}^{(2\alpha)\top} & \widetilde{\mathbf{M}}^{\alpha} \end{pmatrix}, \quad \widetilde{\mathbf{b}} = \begin{pmatrix} \widetilde{\mathbf{b}}^{(1)} \\ \widetilde{\mathbf{b}}^{(2)} \\ \widetilde{\mathbf{b}}^{\alpha} \end{pmatrix}. \tag{29}$$

Take the first term $\widetilde{\mathbf{M}}^{(1)}$ as an example, we set

$$\widetilde{\mathbf{M}}^{(1)} = \text{diag}\{\widetilde{\mathbf{M}}^{(1)}_{g^{(1)}} : g^{(1)} \in [G_1]\} \in \mathbb{R}^{G_1(p_1+1)\times G_1(p_1+1)}, \tag{30}$$

$$\widetilde{\mathbf{M}}^{(1)}_{g^{(1)}} = \sum_{t,g^{(2)}} \sigma^{-2}_{g^{(1)}g^{(2)}} \mathbb{X}^{\top}_{g^{(1)}g^{(2)},t} \mathbb{X}_{g^{(1)}g^{(2)},t}, \tag{31}$$

with additional weights $\sigma^{-2}_{g^{(1)}g^{(2)}}$. Other terms in (29) can be calculated similarly. Then, the WLS estimator is obtained as $\widetilde{\boldsymbol{\theta}} = \widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{b}}$.

However, there are still unknown parameters (e.g., $\sigma^2_{g^{(1)}g^{(2)}}$) in $\widetilde{\boldsymbol{\theta}}$. Therefore, to obtain a feasible WLS estimator, we first estimate $\sigma^2_{g^{(1)}g^{(2)}}$ as follows

$$\widehat{\sigma}^2_{g^{(1)}g^{(2)}} = \frac{1}{N_{1g^{(1)}} N_{2g^{(2)}} T} \sum_{i\in\widehat{\mathcal{R}}^{(1)}_{g^{(1)}}} \sum_{j\in\widehat{\mathcal{R}}^{(2)}_{g^{(2)}}} \sum_{t} (Y_{ij,t} - \mathcal{X}^{\top}_{ij,t}\widehat{\boldsymbol{\Theta}}_{ij})^2 \tag{32}$$

and $\widehat{\boldsymbol{\Theta}}_{ij}$ is the estimator obtained by Algorithm A.2. This allows us to obtain the WLS estimator as

$$\widetilde{\boldsymbol{\theta}}^{w} = \widetilde{\mathbf{M}}^{-1}\widetilde{\boldsymbol{\delta}}, \tag{33}$$

where $\widetilde{\mathbf{M}}$ and $\widetilde{\boldsymbol{\delta}}$ are obtained by substituting the estimators $\widehat{\sigma}^2_{g^{(1)}g^{(2)}}$ and $\widehat{\boldsymbol{\Theta}}$ in $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{b}}$ of (29). In the next subsection, we establish the theoretical properties for $\widetilde{\boldsymbol{\theta}}^{w}$.

5.2.2 Theoretical Properties

We first show that when the group-specific error variances exist, the membership estimation consistency in Theorem 5 still holds. Technically, we need the Assumption A.10 in Appendix B.4.1 instead of the Assumption 3 in Section 4. We state the formal theoretical results in Theorem 9.

Theorem 9. *Assume the group-specific error variances exist and the model is formed as* (28)*. Suppose that Assumptions 1, 2, 4–7, A.10 hold. When* $G_1 = G_{1,0}, G_2 = G_{2,0}$*, we have,*
(i) under label permutation,

$$\lim_{\min(N_1,N_2,T)\to\infty} P\Big(\widehat{\mathcal{G}}^{(1)} = \mathcal{G}^{(1)0}, \widehat{\mathcal{G}}^{(2)} = \mathcal{G}^{(2)0}\Big) \to 1, \tag{34}$$

$$\lim_{\min(N_1,N_2,T)\to\infty} P\Big(\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}^{or}\Big) \to 1. \tag{35}$$

Further assume that there exists n, such that $c_1 n \le \min_l N_l \le \max_l N_l \le c_2 n$ and suppose that $\sqrt{T} \gg \log(G_1 G_2 n^2)$,
(ii) it holds that

$$\sup_{g^{(1)},g^{(2)}} \Big| \frac{1}{\widehat{\sigma}^2_{g^{(1)}g^{(2)}}} - \frac{1}{\sigma^2_{g^{(1)}g^{(2)}}} \Big| = o_p(1).$$

(iii) Assume $\lambda_{\min}(\mathbf{M}^0) \ge \tau > 0$ for a positive constant τ. Let $s = G_1(p_1+1) + G_2(p_2+1) + G_1G_2$, then for any $\boldsymbol{\eta} \in \mathbb{R}^s$ with $\|\boldsymbol{\eta}\| = 1$, we have

$$(n\sqrt{T})^{-1}\boldsymbol{\eta}^\top(\widetilde{\boldsymbol{\theta}}^w - \boldsymbol{\theta}^0) \to_d N(\mathbf{0}, \boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0)^{-1}\boldsymbol{\eta}), \tag{36}$$

where $\widetilde{\mathbf{M}}^0 = \lim_{n,T\to\infty}(n^2T)^{-1}E(\widetilde{\mathbf{M}})$, and $\widetilde{\mathbf{M}}$ is defined in (29).

The proof of Theorem 9 is provided in Appendix B.4.2–B.4.4. Conclusion (i) shows that in spite of the group-specific error variance, the group membership maintains strong consistency under nearly the same conditions. By conclusion (i), we have $\widehat{\mathcal{G}}^{(1)} = \mathcal{G}^{(1)0}$ and $\widehat{\mathcal{G}}^{(2)} = \mathcal{G}^{(2)0}$ with probability tending to 1, hence we next treat the group memberships as known, denoted as $\mathcal{R}^{(1)}_{g^{(1)}}$ and $\mathcal{R}^{(2)}_{g^{(2)}}$ in the technical proof. Note that we substitute $\sigma^{-2}_{g^{(1)}g^{(2)}}$ in (31) to be its estimator $\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}$. Hence, before we show the theoretical convergence properties of $\widetilde{\boldsymbol{\theta}}^w$, we present the estimation consistency for $\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}$ in (ii), which is a critical result for later analysis. Conclusion (ii) shows that when $\sqrt{T} \gg \log(G_1G_2n^2)$, the sample weight $\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}$ is a consistent estimator of $\sigma^{-2}_{g^{(1)}g^{(2)}}$. Next, we establish the asymptotic normality for the WLS estimator $\widetilde{\boldsymbol{\theta}}^w = \widetilde{\mathbf{M}}^{-1}\widetilde{\boldsymbol{\delta}}$ in (iii). In practice, one can calculate the estimator for $\widetilde{\mathbf{M}}^0$ as $\widetilde{\mathbf{M}}_{nT} = (n^2T)^{-1}\widetilde{\mathbf{M}}$, where $\widetilde{\mathbf{M}}$ is obtained by plugging in the error variance estimator $\widehat{\sigma}_{g^{(1)}g^{(2)}}$ calculated by (32) in $\widetilde{\mathbf{M}}$ defined in (29). In addition, we evaluate the finite sample performance of the WLS estimator $\widetilde{\boldsymbol{\theta}}^w$ in Appendix B.4.5.

6. Simulation Study

6.1 Model Settings

To evaluate the finite sample performance of the GTNAR method, we conduct several simulation studies in this section for $q = 2$ (i.e., two networks). Three different scenarios for $G_{l,0}$ are considered. In each scenario, the node memberships $g^{(l)}_{i_l}$s are sampled from the multinomial distribution with probability $\boldsymbol{\pi}_l = \{\pi^{(l)}_{g^{(l)}} = G^{-1}_{l,0} : g^{(l)} = 1, \cdots, G_{l,0}\}$, $l \in [q]$. The dimension of exogenous covariates is set as $p_l = 3$, and the corresponding true parameters are shown in Table 1. For all scenarios, the covariates $\mathbf{x}^{(l)}_{i_l t}, l \in [q]$ are generated from multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_{p_l})$. Subsequently, we generate the noise term $\varepsilon_{i_1 i_2, t}$ from $N(0,1)$ independently. For each scenario, the following two network structures are considered.

Table 1: True parameters for different scenarios.

Scenario 1		Scenario 2		Scenario 3	
$G_{1,0}$	$G_{2,0}=2$	$G_{1,0}=3$	$G_{2,0}=2$	$G_{1,0}=3$	$G_{2,0}=3$
$\lambda^{(1)}_{g^{(1)}}$ (0.15, 0.2)	$\lambda^{(2)}_{g^{(2)}}$ (0.25, 0.4)	$\lambda^{(1)}_{g^{(1)}}$ (0.15, 0.2, 0.3)	$\lambda^{(2)}_{g^{(2)}}$ (0.25, 0.3)	$\lambda^{(1)}_{g^{(1)}}$ (0.15, 0.2, 0.3)	$\lambda^{(2)}_{g^{(2)}}$ (0.25, 0.3, 0.4)
$\boldsymbol{\zeta}^{(1)}_{g^{(1)}}$ $\begin{pmatrix} 0.2 & 0.25 & -0.3 \\ 0.15 & 0.35 & -0.35 \end{pmatrix}$	$\boldsymbol{\delta}^{(2)}_{g^{(2)}}$ $\begin{pmatrix} 0.25 & -0.3 & 0.35 \\ 0.2 & -0.25 & 0.32 \end{pmatrix}$	$\boldsymbol{\zeta}^{(1)}_{g^{(1)}}$ $\begin{pmatrix} 0.2 & 0.25 & -0.3 \\ 0.15 & 0.35 & -0.35 \\ 0.24 & 0.3 & -0.32 \end{pmatrix}$	$\boldsymbol{\delta}^{(2)}_{g^{(2)}}$ $\begin{pmatrix} 0.25 & -0.3 & 0.35 \\ 0.2 & -0.25 & 0.32 \end{pmatrix}$	$\boldsymbol{\zeta}^{(1)}_{g^{(1)}}$ $\begin{pmatrix} 0.2 & 0.25 & -0.3 \\ 0.15 & 0.35 & -0.35 \\ 0.24 & 0.30 & -0.32 \end{pmatrix}$	$\boldsymbol{\delta}^{(2)}_{g^{(2)}}$ $\begin{pmatrix} 0.25 & -0.3 & 0.35 \\ 0.2 & -0.25 & 0.32 \\ 0.1 & -0.2 & 0.2 \end{pmatrix}$
$\alpha_{g^{(1)}g^{(2)}}$ $\begin{pmatrix} -0.2 & 0.3 \\ -0.18 & 0.35 \end{pmatrix}$		$\alpha_{g^{(1)}g^{(2)}}$ $\begin{pmatrix} -0.2 & 0.3 \\ -0.18 & 0.35 \\ -0.15 & 0.28 \end{pmatrix}$		$\alpha_{g^{(1)}g^{(2)}}$ $\begin{pmatrix} -0.2 & 0.3 & 0.4 \\ -0.18 & 0.35 & 0.4 \\ -0.15 & 0.28 & 0.2 \end{pmatrix}$	

Example 1. (Stochastic Block Model, SBM) The first type of network is the stochastic block model, in which nodes in the same block (group) are assigned with a higher probability to be connected. In the lth network, based on the group memberships $g^{(l)}_{i_l}$s and following the setting of Nowicki and Snijders (2001), we set $P(a^{(l)}_{ij}=1)=20/N_l$ when the ith and the jth node are in the same group, and otherwise we set $P(a^{(l)}_{ij}=1)=2/N_l$.

Example 2. (Power-Law Distribution Network) The second type of network is generated from a power-law distribution following Clauset et al. (2009). For the ith node in the lth network, its in-degree $d^{(l)}_i=\sum_{j=1}^N a^{(l)}_{ji}$ is assumed to be power-law distributed. Specifically, we first generate $\widetilde{d}^{(l)}_i$ with a probability $P(\widetilde{d}^{(l)}_i=k)\propto k^{-2.5}$, and set $d^{(l)}_i=4\widetilde{d}^{(l)}_i$. Then, $d^{(l)}_i$ followers of the ith node are randomly selected to construct the adjacency matrix. As a result, the adjacency matrix is not symmetric, which implies a directed network.

6.2 Performance Measure and Simulation Results

We first introduce the model performance measure and then present the simulation results. We set the network sizes $(N_1,N_2)\in\{(100,80),(200,150),(300,250)\}$. The time length is set to be $T\in\{20,40\}$. For each scenario, we repeat the experiments for $R=500$ times. The networks are fixed throughout all replicates under one setting. In the initialization, we use 3 trials for each clustering type, as described in Algorithm A.3 of Appendix F. Denote the estimated parameters in the rth replicate as $\widehat{\lambda}^{(1)[r]}_{g^{(1)}}, \widehat{\lambda}^{(2)[r]}_{g^{(2)}}, \widehat{\boldsymbol{\zeta}}^{(1)[r]}_{g^{(1)}}, \widehat{\boldsymbol{\zeta}}^{(2)[r]}_{g^{(2)}}, \widehat{\alpha}^{[r]}_{g^{(1)}g^{(2)}}$ and the estimated group number as $\widehat{G}^{[r]}_1$ and $\widehat{G}^{[r]}_2$.

6.2.1 Estimation when $G_1=G_{1,0}$ and $G_2=G_{2,0}$

We first evaluate the estimation accuracy when the group numbers are correctly specified. Take $\boldsymbol{\lambda}^{(1)}=(\lambda^{(1)}_1,\cdots,\lambda^{(1)}_{G_1})^\top$ for example. Denote $\widehat{\boldsymbol{\lambda}}^{(1)[r]}$ as the estimator of $\boldsymbol{\lambda}^{(1)0}=(\lambda^{(1)0}_1,\cdots,\lambda^{(1)0}_{G_1})^\top$ in the rth replicate. To evaluate the estimation accuracy, we calculate the root mean squared error (RMSE) as $\text{RMSE}_{\boldsymbol{\lambda}^{(1)}}=\{R^{-1}\sum_{r=1}^R(\|\widehat{\boldsymbol{\lambda}}^{(1)[r]}-\boldsymbol{\lambda}^{(1)0}\|^2)\}^{1/2}$. Next, to gauge the performance of the statistical inference, we construct the 95% confidence interval for each parameter. For example, denote the estimated standard error of $\lambda^{(1)}_{g^{(1)}}$ as $\widehat{\text{SE}}^{[r]}_{\lambda^{(1)}_{g^{(1)}}}$ for the rth replicate, then the 95% confidence interval for $\widehat{\lambda}^{(1)[r]}_{g^{(1)}}$ is con-

structed as $\mathrm{CI}^{[r]}_{\lambda^{(1)}_{g^{(1)}}} = (\widehat{\lambda}^{(1)[r]}_{g^{(1)}} - 1.96 \times \widehat{\mathrm{SE}}^{[r]}_{\lambda^{(1)}_{g^{(1)}}}, \widehat{\lambda}^{(1)[r]}_{g^{(1)}} + 1.96 \times \widehat{\mathrm{SE}}^{[r]}_{\lambda^{(1)}_{g^{(1)}}})$. Here $\widehat{\mathrm{SE}}^{[r]}_{\lambda^{(1)}_{g^{(1)}}}$ is obtained by Theorem 7, plugging in the variance estimator $\widehat{\sigma}^2$, which can be obtained by the procedure in Appendix B.2. Subsequently, the coverage probability (CP) is formed as $\mathrm{CP}_{\lambda^{(1)}_{g^{(1)}}} = R^{-1}\sum_{r=1}^{R} I(\lambda^{(1)0}_{g^{(1)}} \in \mathrm{CI}^{[r]}_{\lambda^{(1)}_{g^{(1)}}})$. We calculate the CPs for other parameters similarly. For comparison, we also calculate the RMSE and CP values for the oracle estimators under the true group memberships (denoted as $\widehat{\lambda}^{(1)\mathrm{or}}_{g^{(1)}}, \widehat{\lambda}^{(2)\mathrm{or}}_{g^{(2)}}, \widehat{\boldsymbol{\zeta}}^{(1)\mathrm{or}}_{g^{(1)}}, \widehat{\boldsymbol{\zeta}}^{(2)\mathrm{or}}_{g^{(2)}}, \widehat{\alpha}^{\mathrm{or}}_{g^{(1)}g^{(2)}}$ accordingly). Lastly, to evaluate the group memberships estimation, we calculate the mis-clustering rates for two network groups as $\widehat{\eta}_l = (N_l R)^{-1}\sum_{r=1}^{R}\sum_{i_l} I(\widehat{g}^{(l)}_{i_l} \neq g^{(l)0}_{i_l})$ for $l = 1, 2$, where $\widehat{g}^{(l)}_{i_l}$ is the estimated group membership of the i_lth node. Here the mis-clustering rates are calculated after proper group permutations.

The simulation results of $G_{1,0} = G_{2,0} = 3$ are shown in Table 2. The first finding across all combinations is that once the group numbers are specified as the true values in advance, our iterative method can estimate the true group memberships with high accuracy, especially when the sample size is large. As N_1, N_2 or T increase, the mis-clustering rates for both network groups approach zero. Furthermore, we note that the RMSEs decrease either when the network sizes N_1 and N_2 increase or the time length T increases, and they approach the oracle RMSEs when the sample sizes are large. Next, we inspect the statistical inference results. We observe that in Table 2, the CPs are slightly smaller when the sample sizes are not very large, but they grow up to around 0.95 as N_1, N_2 and T increase. This guarantees that even under the scenario with a large number of parameters to be estimated, GTNAR can still perform well in terms of both estimation and inference for sufficiently large sample sizes. The patterns are similar in the other two scenarios.

6.2.2 Estimation when $G_l \geq G_{l,0}$ for $l = 1, 2$

We next consider the case of estimation without specifying the true group numbers in advance. Specifically, we estimate the group numbers by QIC in Section 3.3, where the tuning parameter is set to be $\kappa = 1/\{40\log(T)T^{1/8}\}$. Let the true group numbers be $G_{1,0} = G_{2,0} = 3$, and the corresponding true parameters are shown in Table 1. To evaluate the estimation accuracy, we calculate the RMSE for each parameter as explained below. Take $\boldsymbol{\lambda}^{(1)}$ for example, define $\mathrm{RMSE}_{\boldsymbol{\lambda}^{(1)},all} = \{(RN_1)^{-1}\sum_{r=1}^{R}\sum_{i_1=1}^{N_1}(\widehat{\lambda}^{(1)[r]}_{\widehat{g}^{(1)}_{i_1}} - \lambda^{(1)0}_{g^{(1)0}_{i_1}})^2\}^{1/2}$ as the RMSE for all nodes. RMSE for other parameters is calculated similarly. For the group memberships, the mis-clustering rates are calculated following the idea of Zhu et al. (2023). Recall that we have $\widehat{\mathcal{R}}^{(l)}_{g^{(l)}} = \{i_l : \widehat{g}^{(l)}_{i_l} = g^{(l)}\}$, where $\widehat{g}^{(l)}_{i_l}$ is denoted as the estimated group membership for the i_lth in the lth network. Note that G_l is not necessarily equal to $G_{l,0}$. In this case, we define the mappings from the estimated group memberships to the true group memberships $\chi_l : \{1, \cdots, G_l\} \to \{1, \cdots, G_{l,0}\}$ as $\chi_l(g^{(l)}) = \mathrm{argmax}_{g^{(l)\prime} \in \{1,\cdots,G_{l,0}\}}\sum_{i_l=1}^{N_l} I\big(i_l \in \widehat{\mathcal{R}}^{(l)}_{g^{(l)}}, g^{(l)0}_{i_l} = g^{(l)\prime}\big)$ for $g^{(l)} \in \{1, \cdots, G_l\}$. Thus, the mapping $\chi_l(g^{(l)})$ maps group $g^{(l)}$ to the true membership $g^{(l)\prime}$ where the majority of nodes in $\widehat{\mathcal{R}}^{(l)}_{g^{(l)}}$ belong to. Then, for the row group memberships, the mis-clustering rate in the rth replicate is defined as $\widehat{\xi}^{[r]}_l = N_l^{-1}\sum_{g^{(l)}=1}^{G_l}\sum_{i_l=1}^{N_l} I\big(i_l \in \widehat{\mathcal{R}}^{(l)[r]}_{g^{(l)}}, g^{(l)0}_{i_l} \neq \chi_l(g^{(l)})\big)$, where

Table 2: RMSEs (×1000) of estimated parameters under scenario 3 ($G_{1,0} = G_{2,0} = 3$) with 500 replications. The performances are evaluated for different sample sizes N_1, N_2 and the time length T. Results under two network structures are provided. The corresponding CPs are shown in parentheses.

Network	G_1	G_2	N_1	N_2	T	$\widehat{\boldsymbol{\lambda}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}^{(2)}$	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\lambda}}^{(1)\text{or}}$	$\widehat{\boldsymbol{\lambda}}^{(2)\text{or}}$	$\widehat{\boldsymbol{\zeta}}^{(1)\text{or}}$	$\widehat{\boldsymbol{\zeta}}^{(2)\text{or}}$	$\widehat{\boldsymbol{\alpha}}^{\text{or}}$	$\widehat{\eta}_1$	$\widehat{\eta}_2$
SBM	3	3	100	80	20	25.6 (0.705)	12.5 (0.907)	59.1 (0.595)	15.9 (0.920)	58.6 (0.731)	11.7 (0.947)	10.8 (0.938)	13.6 (0.945)	12.9 (0.949)	19.1 (0.949)	0.1634	0.0113
					40	9.1 (0.911)	7.5 (0.945)	14.4 (0.871)	9.1 (0.948)	15.2 (0.913)	7.8 (0.951)	7.4 (0.945)	9.4 (0.938)	9.1 (0.949)	12.8 (0.947)	0.0285	0.0001
			200	150	20	9.1 (0.825)	5.6 (0.925)	17.4 (0.805)	7.0 (0.940)	14.5 (0.857)	6.1 (0.932)	5.3 (0.939)	6.8 (0.940)	6.9 (0.942)	9.6 (0.945)	0.0524	0.0006
					40	4.2 (0.942)	3.8 (0.939)	5.4 (0.940)	4.7 (0.949)	7.0 (0.941)	4.0 (0.948)	3.8 (0.941)	4.7 (0.950)	4.6 (0.950)	6.6 (0.948)	0.0033	0.0001
			300	250	20	4.4 (0.916)	3.5 (0.935)	6.1 (0.906)	4.2 (0.941)	6.9 (0.920)	3.7 (0.948)	3.5 (0.936)	4.2 (0.946)	4.2 (0.941)	5.9 (0.944)	0.0092	0.0001
					40	3.0 (0.930)	2.3 (0.948)	4.2 (0.927)	2.9 (0.952)	4.8 (0.936)	2.5 (0.947)	2.3 (0.950)	2.9 (0.948)	2.9 (0.952)	4.1 (0.948)	0.0006	0
Power-Law	3	3	100	80	20	17.5 (0.769)	9.2 (0.931)	40.2 (0.687)	17.6 (0.923)	34.7 (0.790)	11.0 (0.938)	8.6 (0.941)	12.9 (0.948)	13.2 (0.944)	17.8 (0.947)	0.1361	0.0156
					40	11.4 (0.863)	5.7 (0.947)	20.0 (0.831)	9.3 (0.937)	19.5 (0.886)	7.2 (0.945)	5.6 (0.943)	9.5 (0.941)	9.3 (0.937)	13.2 (0.940)	0.0411	0
			200	150	20	7.7 (0.843)	4.6 (0.939)	14.1 (0.840)	6.8 (0.945)	12.1 (0.883)	5.4 (0.937)	4.5 (0.945)	6.7 (0.947)	6.7 (0.946)	8.7 (0.950)	0.0408	0.0001
					40	4.3 (0.933)	2.1 (0.948)	5.7 (0.927)	4.7 (0.948)	6.6 (0.940)	4.0 (0.946)	2.1 (0.949)	4.8 (0.940)	4.7 (0.948)	6.2 (0.948)	0.0039	0
			300	250	20	4.7 (0.890)	2.4 (0.937)	7.8 (0.892)	5.4 (0.938)	10.7 (0.907)	3.3 (0.939)	2.1 (0.943)	4.2 (0.943)	4.3 (0.944)	5.5 (0.945)	0.0170	0.0001
					40	2.7 (0.942)	1.9 (0.936)	3.2 (0.950)	2.9 (0.948)	4.1 (0.947)	2.6 (0.944)	1.9 (0.935)	2.9 (0.952)	2.9 (0.948)	4.0 (0.949)	0.0010	0

$\widehat{\mathcal{R}}^{(l)[r]}_{g^{(l)}}$ is the estimated node set belonging to the group $g^{(l)}$ in the rth replicate. Then, the overall group memberships error rate is calculated as $\widehat{\xi}_l = R^{-1}\sum_r \widehat{\xi}_l^{[r]}$. The results are shown in Table 3.

We discuss the results shown in Table 3 from two aspects. On one hand, when the group number is under-specified ($G_1 = 2, G_2 = 2$), the node-wise RMSEs are large and usually do not decrease when the N_1, N_2 and T grow. Besides, the error rates $\widehat{\xi}_1$ and $\widehat{\xi}_2$ are around 0.3, indicating a low accuracy in estimating node memberships. These results are expected since a non-ignorable estimation bias exists in an under-fitted model. On the other hand, when the group numbers G_ls are correctly ($G_1 = 3, G_2 = 3$) or over-specified ($G_1 = 4, G_2 = 4$), the RMSE values are generally much lower. This is consistent with our theoretical analysis in Theorem 2.

7. Real Data Applications

7.1 Data Description

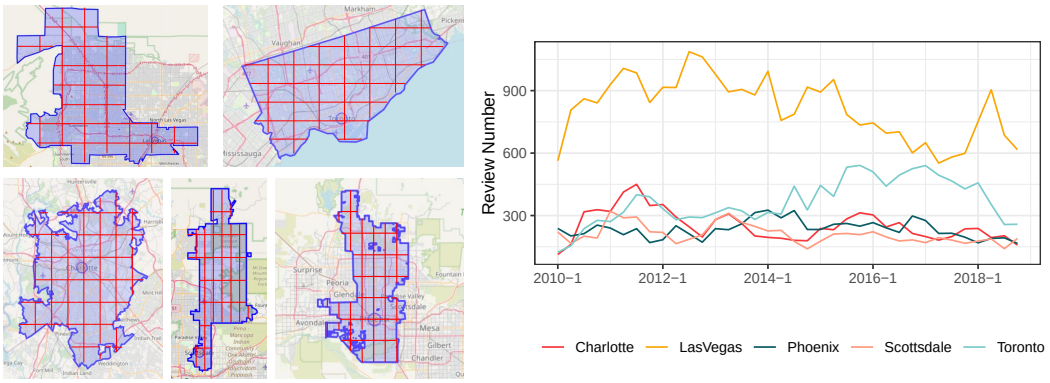
The Yelp dataset spans from 2010 to 2018 and covers five North American cities: Charlotte, Las Vegas, Phoenix, Scottsdale, and Toronto. The observation period is divided into $T = 36$ quarters. To ensure data quality, we filter the dataset to retain active users who have provided more than 5 reviews over this time span. We further divide each city into districts, as illustrated in Figure 3(a). Our response variable, denoted as $Y_{ij,t}$, represents the $\log(1+x)$ transformed number of reviews by user i on district j during the tth quarter. Here $Y_{ij,t}$ is treated as a continuous variable in our real data analysis. To visualize temporal trends in review activity, we calculate the quarterly average responses for each city and depict them in Figure 3(b). Different patterns emerge from this analysis. For instance, Las Vegas stands out as the city with the most reviews, reflecting its bustling business environment. Charlotte, Phoenix, and Scottsdale exhibit relatively similar and stable review trends. In contrast, Toronto shows a noticeable increase in review volume after 2015, likely due to Yelp's expansion in the Toronto area during that period.

Next, we construct the adjacency matrices for users ($\mathbf{A}^{(1)}$) and districts ($\mathbf{A}^{(2)}$) as follows. The user network is built based on the friend list information. Specifically, if user j is on the friend list of user i on Yelp, then we set $a^{(1)}_{ij} = 1$. Otherwise we set $a^{(1)}_{ij} = 0$. The spatial network is built based on the geographical adjacent relationship. Specifically, we set $a^{(2)}_{ij} = 1$ if the district j is adjacent to district i.

Lastly, to characterize the dynamic patterns of the responses, we collect a number of covariates for users and districts, respectively. For user i in quarter t, we consider the following five covariates: (1) the number of months after joining Yelp by the start of the quarter t ($x^{(1)}_{it,\text{dur}}$), (2) whether the user is VIP by the start of the quarter t ($x^{(1)}_{it,\text{vip}}$), (3) average tags (i.e., "useful", "funny" and "cool") the user i obtains for his/her reviews during the last quarter ($x^{(1)}_{it,\text{use}}, x^{(1)}_{it,\text{fun}}, x^{(1)}_{it,\text{cool}}$). Next, for the jth district in quarter t, we consider two covariates: (1) the average "stars" ($x^{(2)}_{jt,\text{star}}$), and (2) the average review number ($x^{(2)}_{jt,\text{num}}$) obtained by the jth district during the $(t-1)$th quarter. These two covariates are indicative of the average popularity levels in the preceding time period. We standardize all continuous covariates to the range $[0, 1]$ for subsequent analysis. In Figure A.11 in Appendix M.1, we

Table 3: Simulation results for the two network examples with pre-specified group numbers as well as the QIC selection group numbers $\widehat{G}_l$. The true group numbers are set as $G_{1,0} = G_{2,0} = 3$. The node-wise RMSEs of different estimators are denoted as $\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(1)}, \widehat{\boldsymbol{\lambda}}_{\text{all}}^{(2)}, \boldsymbol{\zeta}_{\text{all}}^{(1)}, \boldsymbol{\zeta}_{\text{all}}^{(2)}, \boldsymbol{\alpha}_{\text{all}}$.

N_1	N_2	T	G_1	G_2	Scenario 1 (SBM)							Scenario 2 (Power-Law)						
					$\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}_{\text{all}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}_{\text{all}}^{(2)}$	$\widehat{\boldsymbol{\alpha}}_{\text{all}}$	$\widehat{\xi}_1$	$\widehat{\xi}_2$	$\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}_{\text{all}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}_{\text{all}}^{(2)}$	$\widehat{\boldsymbol{\alpha}}_{\text{all}}$	$\widehat{\xi}_1$	$\widehat{\xi}_2$
100	80	20	Oracle		0.0065	0.0059	0.0071	0.0070	0.0061	-	-	0.0072	0.0039	0.0071	0.0072	0.0060	-	-
			2	2	0.0285	0.0419	0.0427	0.0583	0.0368	0.3219	0.3125	0.0237	0.0438	0.0382	0.0647	0.0367	0.2705	0.3705
			3	3	0.0207	0.0096	0.0251	0.0099	0.0174	0.1403	0.0142	0.0196	0.0129	0.0263	0.0114	0.0169	0.1487	0.0208
			4	4	0.0192	0.0099	0.0181	0.0097	0.0172	0.0720	0.0042	0.0201	0.0134	0.0223	0.0101	0.0176	0.1046	0.0077
			$\widehat{G}_1$	$\widehat{G}_2$	0.0278	0.0407	0.0416	0.0565	0.0358	-	-	0.0218	0.0374	0.0339	0.0554	0.0323	-	-
		40	Oracle		0.0044	0.0043	0.0050	0.0050	0.0043	-	-	0.0046	0.0040	0.0050	0.0049	0.0040	-	-
			2	2	0.0216	0.0436	0.0430	0.0503	0.0327	0.3100	0.2625	0.0210	0.0428	0..0403	0.0607	0.0361	0.2800	0.3375
			3	3	0.0082	0.0070	0.0104	0.0063	0.0078	0.0366	0.0061	0.0069	0.0040	0.0085	0.0049	0.0053	0.0233	0
			4	4	0.0081	0.0061	0.0084	0.0065	0.0084	0.0136	0.0006	0.0076	0.0066	0.0078	0.0067	0.0082	0.0101	0.0015
			$\widehat{G}_1$	$\widehat{G}_2$	0.0086	0.0112	0.0121	0.0121	0.0100	-	-	0.0069	0.0040	0.0085	0.0049	0.0053	-	-
200	150	20	Oracle		0.0034	0.0031	0.0036	0.0037	0.0031	-	-	0.0036	0.0021	0.0036	0.0037	0.0030	-	-
			2	2	0.0221	0.0431	0.0455	0.0592	0.0351	0.3505	0.3133	0.0217	0.0420	0.0416	0.0559	0.0351	0.3002	0.3200
			3	3	0.0091	0.0037	0.0112	0.0040	0.0068	0.0504	0.0015	0.0076	0.0022	0.0089	0.0037	0.0054	0.0342	0.0001
			4	4	0.0106	0.0052	0.0089	0.0054	0.0096	0.0296	0.0003	0.0086	0.0047	0.0082	0.0050	0.0077	0.0240	0.0004
			$\widehat{G}_1$	$\widehat{G}_2$	0.0178	0.0334	0.0354	0.0457	0.0275	-	-	0.0076	0.0026	0.0089	0.0042	0.0056	-	-
		40	Oracle		0.0024	0.0021	0.0026	0.0026	0.0022	-	-	0.0023	0.0011	0.0025	0.0026	0.0020	-	-
			2	2	0.0199	0.0405	0.0383	0.0539	0.0343	0.2600	0.3133	0.0218	0.0409	0.0437	0.0554	0.0350	0.3250	0.3133
			3	3	0.0029	0.0021	0.0034	0.0026	0.0026	0.0028	0.0001	0.0026	0.0011	0.0030	0.0026	0.0022	0.0027	0
			4	4	0.0044	0.0038	0.0043	0.0040	0.0055	0.0047	0.0017	0.0049	0.0042	0.0045	0.0042	0.0057	0.0063	0.0032
			$\widehat{G}_1$	$\widehat{G}_2$	0.0030	0.0027	0.0035	0.0034	0.0029	-	-	0.0026	0.0011	0.0030	0.0026	0.0022	-	-



(a) Block map

(b) Average review volumes

Figure 3: (a) Geographical maps with split districts in each city. The city from the top left panel to the bottom right panel shows the map of Las Vegas, Toronto, Charlotte, Scottsdale, and Phoenix, respectively. (b) Average number of reviews from 2010-Q1 to 2018-Q4 in five cities.

visualize the relationship between these user-related covariates and the response variable. The plot reveals that users who receive more tags for their reviews tend to be motivated to contribute more reviews. Notably, VIP users in Scottsdale and Toronto tend to write more reviews, whereas VIP users in Charlotte exhibit comparatively less activity. Subsequently, we apply the GTNAR model to each of the five cities, enabling us to analyze and understand the distinctive group patterns within each urban area.

7.2 Estimation Results

There has been a lot of research finding that the activeness of users on the platform can lead to higher profit (Forman et al., 2008; Pansari and Kumar, 2017), hence finding out the key factors that positively affect the review volumes can help business owners make personalized recommendations to users and develop a commercial strategy. We employ QIC for the selection of group numbers, and show the results of Charlotte, Las Vegas, and Phoenix in Table 4. The results of Scottsdale and Toronto are provided in Table A.16 in Appendix M.1. The numbers of user groups and district groups vary across the five cities, indicating different levels of heterogeneity among them. For instance, consider the results for Phoenix, where there are estimated 3 user groups and 2 district groups.

Notably, the spatial (column) network effects are consistently positive, suggesting a favorable effect from neighboring districts. Such an observation is consistent with the findings in the literature (e.g. Sun and Paule, 2017). Furthermore, we can also observe that within the two group-wise spatial effects, $\widehat{\lambda}_2^{(2)} = 0.270$ is larger than that of Group 1 ($\widehat{\lambda}_1^{(2)} = 0.05$), signifying a stronger neighbor effect. In other words, if the second group of districts' neighbors obtain more reviews in the last period, then it is likely that these districts would receive more reviews in this period. We further visualize the districts by estimated groups on the left panel of Figure 4, where the gray districts are from Group 2, and the white districts are from Group 1. Subsequently, we mark the shops on the map, which are shown by red

points in the right panel of Figure 4. By scrutinizing the locations, we find that the shops in Group 2 are mainly located in the central business districts. The average number of shops per district in Group 2 is 25.16, whereas that in Group 1 is 6.57, showing a higher density of shops in the gray areas (Group 2). Our estimation results reveal that shops in Group 2 exhibit larger spillover effects on their neighbors. This finding is consistent with the existing research, which also reveals a stronger spillover effect for geometrically clustered units (Arzaghi and Henderson, 2008; Rossi-Hansberg et al., 2010; Vitorino, 2012). Performing a

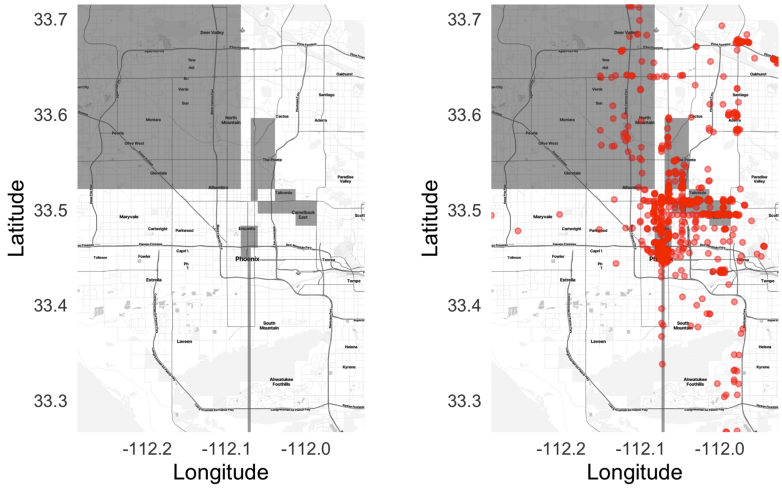


Figure 4: The left panel shows the two groups of districts in Phoenix (Group 2 is marked as gray). The right panel marks all shops as red points.

similar investigation on the user dimension, one can find that the social network effect in Group 1 appears to be significantly negative ($\widehat{\lambda}_1^{(1)} = -0.02$), whereas corresponding coefficients in the other two groups are positive. This implies that user activities in Group 1 are influenced oppositely by their friends' behaviors, whereas in the other groups, users are still positively influenced by their friends. By scrutinizing deeper into users in Group 1, we found that their average duration after joining Yelp is longer than in Groups 2 and 3. This finding aligns with existing works showing that the social effects would decrease as the users' membership duration prolongs (Nitzan and Libai, 2011; Aral and Walker, 2011). Further, the social network effect is the largest in Group 3, i.e., $\widehat{\lambda}_3^{(1)} = 0.024$. The network in-degree and out-degree of users in this group are higher than in Groups 1 and 2, which is visualized in the left panel of Figure 5. The phenomenon of larger network degrees associated with larger social network effects has been extensively validated in literature (Hill et al., 2006; Hinz et al., 2011; Susarla et al., 2012). Therefore, the business owners can implement a precision marketing strategy targeting users in Group 3, who have stronger network effects.

Besides, the estimated $\widehat{\boldsymbol{\alpha}}$ values are all positive, indicating a positive self-motivated effect overall, whose values are visualized in the right panel of Figure 5. Specifically, the momentum effect on the review number is the largest for the Group 1 users' evaluation on

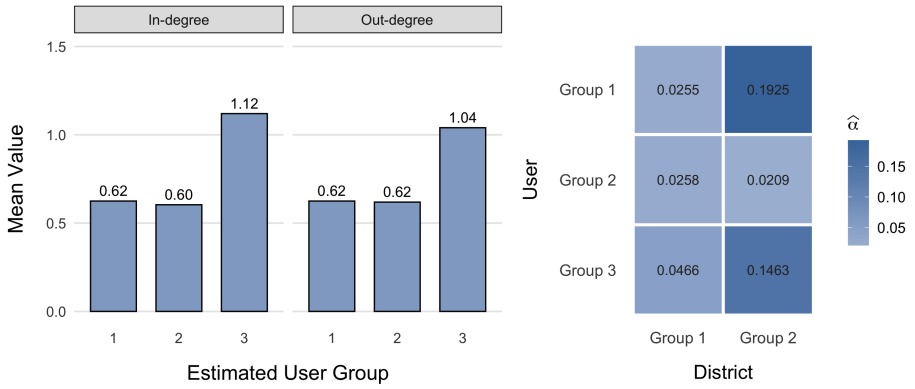


Figure 5: Left panel: average in-degree and out-degree of users in the three estimated groups in Phoenix. Right panel: heatmap of estimated momentum effect $\widehat{\boldsymbol{\alpha}}$ in Phoenix.

the Group 2 districts, i.e., $\widehat{\alpha}_{12} = 0.193$. This implies that users in Group 1 have higher self-driven persistency in the historical reviews than the other groups, while they exhibit smaller social network effects.

8. Concluding Remarks

In this work, we introduce a novel Group Tensor Network Autoregression (GTNAR) model designed for time series data indexed by multi-relational networks. By leveraging network structures on each dimension, GTNAR establishes a unique framework for analyzing tensor-valued time series data. This model presents a valuable tool for analyzing data collected in complex network environments, shedding light on various network effects. From a modeling perspective, several intriguing future topics emerge. It would be interesting to consider the high-dimensional covariates in this framework. Furthermore, it is worth exploring models that consider multiple network effects in a multiplicative form, as investigated by Chen et al. (2021). Besides, introducing a hidden factor structure into the GTNAR model can potentially offer more insights into high-dimensional data, capturing more underlying information within the tensor-valued time series, and thus represents an interesting future research topic. Lastly, it is also important to further investigate how to model categorical responses in our modeling framework. From the theoretical perspective, although the convergence rate established in Theorem 2 is typical in existing research, how to obtain a minimax rate needs further investigation.

Acknowledgments

Yimeng Ren and Xuening Zhu's research are supported by the National Natural Science Foundation of China (nos. 72573038, 12331009), MOE Laboratory for National Devel-

Table 4: Estimation results for Charlotte, Las Vegas, and Phoenix. The p-values are shown in parentheses. Take Phoenix for example, $\boldsymbol{\lambda}^{(1)}$ is clustered into three groups, with $\lambda^{(1)}_{g^{(1)}}$ representing the social network effect on the users in the $g^{(1)}$th group from their friends' behavior. $\boldsymbol{\lambda}^{(2)}$ is clustered into two groups, with $\lambda^{(2)}_{g^{(2)}}$ meaning the spatial network effect on the districts in the hth group from their adjacent neighbors. $\boldsymbol{\zeta}^{(1)}$ and $\boldsymbol{\zeta}^{(2)}$ are the covariates effects, and we take $\zeta^{(1)}_{\text{use}}$ to illustrate. It has also been clustered into three groups, with the coefficient in the $g^{(1)}$th group meaning the effect on the users in the $g^{(1)}$th group from the average "useful" tag. $\boldsymbol{\alpha}$ is clustered into three groups by row and into two groups by column, where $\alpha_{g^{(1)}g^{(2)}}$ means the self-momentum effect of the users in the $g^{(1)}$th group by row and $g^{(2)}$th group by column.

Parameters	Charlotte ($N_1 = 240,\ N_2 = 60$)						LasVegas ($N_1 = 826,\ N_2 = 64$)					Phoenix($N_1 = 323,\ N_2 = 63$)				
	$\lambda^{(1)}_{g^{(1)}}$			$\lambda^{(2)}_{g^{(2)}}$			$\lambda^{(1)}_{g^{(1)}}$			$\lambda^{(2)}_{g^{(2)}}$		$\lambda^{(1)}_{g^{(1)}}$			$\lambda^{(2)}_{g^{(2)}}$	
	0.004	0.110	0.047	0.035	0.475	0.286	0.068	0.008	0.054	0.150	0.326	-0.020	0.007	0.024	0.050	0.270
	(0.017)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.101)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
	$\zeta^{(1)}_{g^{(1)}}$			$\zeta^{(2)}_{g^{(2)}}$			$\zeta^{(1)}_{g^{(1)}}$			$\zeta^{(2)}_{g^{(2)}}$		$\zeta^{(1)}_{g^{(1)}}$			$\zeta^{(2)}_{g^{(2)}}$	
Intercept	0.001	-0.004	0.003	-0.004	0.041	-10^{-4}	0.011	-0.005	-0.006	0.010	0.003	0.058	-0.052	-0.006	0.050	0.021
	(0.328)	(<0.001)	(0.007)	(<0.001)	(<0.001)	(0.796)	(<0.001)	(<0.001)	(0.586)	(<0.001)	(0.011)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
$\zeta^{(1)}_{\text{dur}}$ / $\zeta^{(2)}_{\text{star}}$	-0.001	0.050	0.015	0.002	-0.012	0.004	0.059	0.001	0.019	10^{-4}	0.001	-0.093	0.001	0.002	0.003	0.004
	(0.162)	(<0.001)	(<0.001)	(<0.001)	(0.078)	(0.005)	(<0.001)	(0.194)	(<0.001)	(0.170)	(0.744)	(<0.001)	(0.150)	(0.148)	(<0.001)	(0.010)
$\zeta^{(1)}_{\text{vip}}$ / $\zeta^{(2)}_{\text{num}}$	0.001	-0.037	-0.002	0.007	0.008	0.003	0.028	0.000	-0.010	0.016	0.038	0.008	10^{-4}	10^{-4}	0.002	10^{-4}
	(0.012)	(<0.001)	(0.044)	(<0.001)	(0.001)	(0.233)	(<0.001)	(0.093)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.054)	(0.369)	(<0.001)	(0.938)
$\zeta^{(1)}_{\text{use}}$	0.017	0.104	0.052				1.115	0.011	0.143			0.385	0.011	0.071		
	(<0.001)	(<0.001)	(<0.001)				(<0.001)	(<0.001)	(<0.001)			(<0.001)	(<0.001)	(<0.001)		
$\zeta^{(1)}_{\text{fun}}$	-0.001	-0.020	10^{-4}				-0.589	-0.001	-0.289			0.064	-0.008	-0.033		
	(0.806)	(0.002)	(0.995)				(<0.001)	(0.795)	(<0.001)			(0.080)	(0.002)	(<0.001)		
$\zeta^{(1)}_{\text{cool}}$	-0.001	0.003	-0.017				-0.644	-0.008	0.217			0.208	0.010	-0.011		
	(0.908)	(0.621)	(<0.001)				(<0.001)	(0.043)	(<0.001)			(<0.001)	(0.005)	(0.007)		
	$\boldsymbol{\alpha}^\top \in \mathbb{R}^{G_1 \times G_2}$						$\boldsymbol{\alpha}^\top \in \mathbb{R}^{G_1 \times G_2}$					$\boldsymbol{\alpha}^\top \in \mathbb{R}^{G_1 \times G_2}$				
	0.019		0.051		0.058		0.288		0.021		0.174	0.026		0.026		0.047
	(0.002)		(<0.001)		(<0.001)		(<0.001)		(<0.001)		(<0.001)	(<0.001)		(<0.001)		(<0.001)
	0.005		0.374		0.224		0.478		0.036		0.330	0.193		0.021		0.146
	(0.218)		(<0.001)		(<0.001)		(<0.001)		(<0.001)		(<0.001)	(<0.001)		(<0.001)		(<0.001)
	0.017		0.200		0.115											
	(<0.001)		(<0.001)		(<0.001)											

opment and Intelligent Governance, Fudan University. Yanyuan Ma's work is partially supported by grants from NIH.

References

Radoslaw Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information systems (TOIS)*, 23(1):103–145, 2005.

Luiz GA Alves, Giuseppe Mangioni, Isabella Cingolani, Francisco Aparecido Rodrigues, Pietro Panzarasa, and Yamir Moreno. The nested structural organization of the worldwide trade multi-layer network. *Scientific reports*, 9(1):2866, 2019.

Tomohiro Ando and Jushan Bai. Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191, 2016.

Tomohiro Ando and Jushan Bai. Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity. *Journal of the American Statistical Association*, 115(529):266–279, 2020.

Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, 57(9):1623–1639, 2011.

Mirko Armillotta and Konstantinos Fokianos. Nonlinear network autoregression. *The Annals of Statistics*, 51(6):2526–2552, 2023.

Mohammad Arzaghi and J Vernon Henderson. Networking off madison avenue. *The Review of Economic Studies*, 75(4):1011–1038, 2008.

Marc Auboin and Michele Ruta. The relationship between exchange rates and international trade: a literature review. *World Trade Review*, 12(3):577–605, 2013.

Sumanta Basu, George Michailidis, et al. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

Nicolas Berman and Jérôme Héricourt. Financial factors and the margins of trade: Evidence from cross-country firm-level data. *Journal of Development Economics*, 93(2):206–217, 2010.

Andrew B Bernard and J Bradford Jensen. Exceptional exporter performance: cause, effect, or both? *Journal of international economics*, 47(1):1–25, 1999.

Andrew B Bernard and J Bradford Jensen. Why some firms export. *Review of economics and Statistics*, 86(2):561–569, 2004.

C Alan Bester and Christian B Hansen. Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 190(1):197–208, 2016.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

Raymond J Carroll and Daren BH Cline. An asymptotic theory for weighted least-squares with weights estimated by replication. *Biometrika*, 75(1):35–43, 1988.

Elias Carroni, Paolo Pin, and Simone Righi. Bring a friend! privately or publicly? *Management Science*, 66(5):2269–2290, 2020.

Jinyuan Chang, Jing He, Lin Yang, and Qiwei Yao. Modelling matrix time series via a tensor cp-decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):127–148, 2023.

Elynn Y Chen and Jianqing Fan. Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, 118(542):1038–1055, 2023.

Elynn Y Chen, Ruey S Tsay, and Rong Chen. Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, 115(530):775–793, 2020a.

Elynn Y Chen, Ruey S Tsay, and Rong Chen. Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, 2020b.

Elynn Y Chen, Jianqing Fan, and Xuening Zhu. Community network auto-regression for high-dimensional time series. *Journal of Econometrics*, 235(2):1239–1256, 2023.

Jian Chen, Weidong Ma, and Ganggang Xu. Group network multivariate garch. *Available at SSRN 6760338*, 2026.

Rong Chen, Han Xiao, and Dan Yang. Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1):539–560, 2021.

Rong Chen, Dan Yang, and Cun-Hui Zhang. Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116, 2022.

Weilin Chen and Clifford Lam. Rank and factor loadings estimation in time series tensor factor model by pre-averaging. *The Annals of Statistics*, 52(1):364–391, 2024.

Xi Chen, Ralf Van Der Lans, and Tuan Q Phan. Uncovering the importance of relationship characteristics in social networks: Implications for seeding strategies. *Journal of Marketing Research*, 54(2):187–201, 2017.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

Enrico Corradini, Antonino Nocera, Domenico Ursino, and Luca Virgili. Investigating negative reviews and detecting negative influencers in yelp through a multi-dimensional social network based model. *International Journal of Information Management*, 60:102377, 2021.

Richard A Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096, 2016.

Shanshan Ding and R Dennis Cook. Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(2):387–408, 2018.

Michael Emch, Elisabeth D Root, Sophia Giebultowicz, Mohammad Ali, Carolina Perez-Heydrich, and Mohammad Yunus. Integration of spatial and social network analysis in disease transmission studies. In *Geographies of Health, Disease and Well-being*, pages 130–141. Routledge, 2016.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Yuan Ke, and Yuan Liao. Augmented factor models with applications to validating market risk factors and forecasting bond risk premia. *Journal of Econometrics*, 222(1):269–294, 2021.

Guanhua Fang, Ganggang Xu, Haochen Xu, Xuening Zhu, and Yongtao Guan. Group network hawkes process. *Journal of the American Statistical Association*, pages 1–17, 2023.

Hao Fe. Social networks and consumer behavior: evidence from yelp. *Journal of Economic Behavior & Organization*, 209:1–14, 2023.

Yusen Feng, Gang-Jin Wang, You Zhu, and Chi Xie. Systemic risk spillovers and the determinants in the stock markets of the belt and road countries. *Emerging Markets Review*, 55:101020, 2023.

Denzil G Fiebig, Michael P Keane, Jordan Louviere, and Nada Wasi. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing science*, 29(3):393–421, 2010.

Chris Forman, Anindya Ghose, and Batia Wiesenfeld. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3):291–313, 2008.

Caroline L Freund and Diana Weinhold. The effect of the internet on international trade. *Journal of international economics*, 62(1):171–189, 2004.

Chang Gan, Mihai Voda, Kai Wang, Lijun Chen, and Jun Ye. Spatial network structure of the tourism economy in urban agglomeration: A social network analysis. *Journal of Hospitality and Tourism Management*, 47:124–133, 2021.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Yuefeng Han, Rong Chen, Cun-Hui Zhang, and Qiwei Yao. Simultaneous decorrelation of matrix time series. *Journal of the American Statistical Association*, pages 1–13, 2023.

Yuefeng Han, Dan Yang, Cun-Hui Zhang, and Rong Chen. Cp factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae036, 2024.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3): 1079–1083, 1971.

Shawndra Hill, Foster Provost, and Chris Volinsky. Network-Based Marketing: Identifying Likely Adopters via Consumer Networks. *Statistical Science*, 21(2):256–276, 2006.

Oliver Hinz, Bernd Skiera, Christian Barrot, and Jan U Becker. Seeding strategies for viral marketing: an empirical comparison. *Journal of Marketing*, 75(6):55–71, 2011.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169, 2015.

Peter Hooper and Steven W Kohlhagen. The effect of exchange rate uncertainty on the prices and volume of international trade. *Journal of international Economics*, 8(4):483–511, 1978.

IMF. *Direction of Trade Statistics, International Monetary Fund*, 2017.

Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. A model-based embedding technique for segmenting customers. *Operations Research*, 66 (5):1247–1267, 2018.

Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49 (6):3181–3205, 2021.

Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

Julia Koschinsky. Spatial heterogeneity in spillover effects of assisted and unassisted rental housing. *Journal of Urban Affairs*, 31(3):319–347, 2009.

Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 2020.

Chenlei Leng and Cheng Yong Tang. Sparse matrix graphical models. *Journal of the American Statistical Association*, 107(499):1187–1200, 2012.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on learning theory*, pages 2738–2779. PMLR, 2020.

Ruiqi Liu, Zuofeng Shang, Yonghui Zhang, and Qiankun Zhou. Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics*, 215(2): 574–590, 2020.

Wenyang Liu, Ganggang Xu, Jianqing Fan, and Xuening Zhu. Two-way homogeneity pursuit for quantile network vector autoregression. *arXiv preprint arXiv:2404.18732*, 2024.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Zongming Ma and Sagnik Nandy. Community detection with contextual multilayer networks. *IEEE Transactions on Information Theory*, 69(5):3203–3239, 2023.

Peter W MacDonald, Elizaveta Levina, and Ji Zhu. Latent space models for multiplex networks with shared structure. *Biometrika*, 109(3):683–706, 2022.

Ke Miao, Peter CB Phillips, and Liangjun Su. High-dimensional vars with common factors. *Journal of Econometrics*, 233(1):155–183, 2023.

Katta G. Murty and Santosh N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Math. Program.*, 39(2):117–129, June 1987. ISSN 0025-5610.

Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts. *The structure and dynamics of networks*. Princeton University Press, 2006.

William B Nicholson, Ines Wilms, Jacob Bien, and David S Matteson. High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21(166):1–52, 2020.

Irit Nitzan and Barak Libai. Social effects on customer retention. *Journal of Marketing*, 75 (6):24–38, 2011.

Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

Anita Pansari and Vera Kumar. Customer engagement: the construct, antecedents, and consequences. *Journal of the academy of marketing science*, 45(3):294–311, 2017.

Stephen Redding and Anthony J Venables. Economic geography and international inequality. *Journal of international Economics*, 62(1):53–82, 2004.

Esteban Rossi-Hansberg, Pierre-Daniel Sarte, and Raymond Owens III. Housing externalities. *Journal of political Economy*, 118(3):485–535, 2010.

Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

Juan Camilo Serpa and Harish Krishnan. The impact of supply chains on firm-level productivity. *Management Science*, 64(2):511–532, 2018.

Jun Shao. Asymptotic distribution of the weighted least squares estimator. *Annals of the Institute of Statistical Mathematics*, 41:365–382, 1989.

Liangjun Su, Zhentao Shi, and Peter CB Phillips. Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264, 2016.

Liangjun Su, Wuyi Wang, and Xingbai Xu. Identifying latent group structures in spatial dynamic panels. *Journal of Econometrics*, 235(2):1955–1980, 2023.

Yeran Sun and Jorge David Gonzalez Paule. Spatial analysis of users-generated ratings of yelp venues. *Open Geospatial Data, Software and Standards*, 2:1–9, 2017.

Anjana Susarla, Jeong-Ha Oh, and Yong Tan. Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research*, 23(1):23–41, 2012.

Michel Talagrand. An isoperimetric theorem on the cube and the kintchine-kahane inequalities. *Proceedings of the American Mathematical Society*, 104(3):905–909, 1988.

Nuray Terzi. The impact of e-commerce on international trade and employment. *Procedia-social and behavioral sciences*, 24:745–753, 2011.

Ashutosh Tiwari and Timothy J Richards. Social networks and restaurant ratings. *Agribusiness*, 32(2):153–174, 2016.

Maria Ana Vitorino. Empirical entry games with complementarities: an application to the shopping center industry. *Journal of Marketing Research*, 49(2):175–191, 2012.

Michael Vogt and Oliver Linton. Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):5–27, 2017.

AT Walden and A Serroukh. Wavelet analysis of matrix–valued time–series. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 458(2017):157–179, 2002.

Di Wang and Ruey S Tsay. Rate-optimal robust estimation of high-dimensional vector autoregressive models. *The Annals of Statistics*, 51(2):846–877, 2023.

Di Wang, Yao Zheng, Heng Lian, and Guodong Li. High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, 117(539):1338–1356, 2022.

Di Wang, Yao Zheng, and Guodong Li. High-dimensional low-rank tensor autoregressive time series modeling. *Journal of Econometrics*, 238(1):105544, 2024.

Dong Wang, Xialu Liu, and Rong Chen. Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208(1):231–248, 2019.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics*, 41(5):2505–2536, 2013.

Michel Wedel and Wagner A Kamakura. *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media, 2000.

Shu Wu, Qiang Liu, Liang Wang, and Tieniu Tan. Contextual operation for recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2000–2012, 2016.

Zhonghua Yin, Fang Wang, and Jianbang Gan. Spatial spillover effects of global forest product trade. *Forest Policy and Economics*, 113:102112, 2020. ISSN 1389-9341.

Xuefei Zhang, Songkai Xue, and Ji Zhu. A flexible latent space model for multilayer networks. In *International Conference on Machine Learning*, pages 11288–11297. PMLR, 2020.

Yingying Zhang, Huixia Judy Wang, and Zhongyi Zhu. Quantile-regression-based clustering for panel data. *Journal of Econometrics*, 213(1):54–67, 2019.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483, 2014.

Jie Zhou, Will Wei Sun, Jingfei Zhang, and Lexin Li. Partially observed dynamic tensor response regression. *Journal of the American Statistical Association*, 118(541):424–439, 2023.

Shuheng Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.

Xuening Zhu and Rui Pan. Grouped network vector autoregression. *Statistica Sinica*, 30 (3):1437–1462, 2020.

Xuening Zhu, Rui Pan, Guodong Li, Yuewen Liu, and Hansheng Wang. Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123, 2017.

Xuening Zhu, Ganggang Xu, and Jianqing Fan. Simultaneous estimation and group identification for network vector autoregressive model with heterogeneous nodes. *Journal of Econometrics*, page 105564, 2023.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

In this part, the required notations are provided in Appendix A. More estimation details for $q = 2$ and general q are provided in Appendix B–D. An numerical convergence analysis for the algorithm is shown in Appendix E. The initialization procedure and of our main algorithm is provided in Appendix F. We provide additional discussion about the interactive model in Appendix G, followed by discussion on future extensions in Appendix H. We provide the technical details and proofs for the main text in Section I, and several useful lemmas are given in Section J–K. Besides, we provide a number of additional simulation studies in Appendix L. Lastly, some additional real data analysis are given in Appendix M and Appendix N.

Appendix A. Notations

Define $Q^*(\mathbf{\Theta}) = E\{Q(\mathbf{\Theta})\}$ and $Q^*_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) = E\{Q_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q})\}$, where $Q(\mathbf{\Theta})$ and $Q_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q})$ are as defined in (16) in the main text. Denote $\mathcal{G}_{-l} = \{\mathcal{G}_m : m \neq l\}$ and $\mathbf{i}_{-l} = (i_m : m \neq l)^\top \in \mathbb{R}^{q-1}$. Further, denote $\boldsymbol{\theta}^{-(l)} = \{(\boldsymbol{\theta}^{(m)}_{g^{(m)}} : g^{(m)} \in [G_m]) \in \mathbb{R}^{G_m(p_m+1)} : m \neq l, 1 \leq m \leq q\}$, $\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}} = (\boldsymbol{\theta}^{(l)\top}_{g^{(l)}_{i_l}}, \mathrm{vec}(\boldsymbol{\alpha}_{\cdot g^{(l)}_{i_l}\cdot})^\top)^\top$, $\boldsymbol{\xi}^{-(l)}_{g^{-(l)}} = (\boldsymbol{\theta}^{(m)\top}_{g^{(m)}}, \mathrm{vec}(\boldsymbol{\alpha}_{\cdot g^{(m)}\cdot})^\top : m \neq l)^\top$, and $g^{-(l)}_{\mathbf{i}_{-l}} = (g^{(m)}_{i_m} : m \neq l)$. Using these notations, we define

$$Q_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}}; \boldsymbol{\xi}^{-(l)}_{g^{-(l)}}, \mathcal{G}_{-l}) = \sum_{m\neq l}\sum_{i_m=1}^{N_m}\sum_{t=1}^{T}\Big\{Y_{i_1\cdots i_q,t} - \sum_{l=1}^{q}\lambda^{(l)}_{g^{(l)}_{i_l}}\sum_{k=1}^{N_l} w^{(l)}_{i_l k} Y_{i_1\cdots i_{l-1}k i_{l+1}\cdots i_q,(t-1)} - \alpha_{g^{(1)}_{i_1}\cdots g^{(q)}_{i_q}} Y_{i_1\cdots i_q,(t-1)} - \sum_{l=1}^{q}\mathbf{x}^{(l)\top}_{i_l t}\boldsymbol{\zeta}^{(l)}_{g^{(l)}_{i_l}}\Big\}^2 \tag{A.1}$$

and $Q^*_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}}; \boldsymbol{\xi}^{-(l)}_{g^{-(l)}}, \mathcal{G}_{-l}) = E\{Q_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}}; \boldsymbol{\xi}^{-(l)}_{g^{-(l)}}, \mathcal{G}_{-l})\}$. By the definition, it readily follows that $Q(\boldsymbol{\xi}, \mathcal{G}) = \sum_{i_l=1}^{N_l} Q_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}}; \boldsymbol{\xi}^{-(l)}_{g^{-(l)}}, \mathcal{G}_{-l})$, where the $Q(\boldsymbol{\xi}, \mathcal{G})$ is the objective function defined in (9). For the tensor $\boldsymbol{\alpha} \in \mathbb{R}^{G_1\times\cdots\times G_q}$, we use $\boldsymbol{\alpha}_{\cdot g^{(l)}_{i_l}\cdot} \in \mathbb{R}^{G_1\times\cdots G_{l-1}\times G_{l+1}\times\cdots\times G_q}$ to denote its subset by specifying the lth dimension being equal to $g^{(l)}_{i_l}$. Similarly, for the parameter tensor $\mathbf{\Theta} \in \mathbb{R}^{N_1\times\cdots\times N_q\times m}$, when specifying the lth dimension as i_l, we obtain the subset tensor $\mathbf{\Theta}_{\cdot i_l\cdot} \in \mathbb{R}^{N_1\times\cdots\times N_{l-1}\times N_{l+1}\times\cdots\times N_q\times m}$. We define the pseudo distance

$$\begin{aligned} d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot}, \mathbf{\Theta}_{\cdot i_l\cdot}) &= \frac{1}{\prod_{m\neq l} N_m}\sum_{m\neq l}\sum_{i_m=1}^{N_m}\Big\|\widehat{\mathbf{\Theta}}_{i_1\cdots i_q} - \mathbf{\Theta}_{i_1\cdots i_q}\Big\|^2 \\ &= \sum_{m\neq l}\frac{1}{N_m}\sum_{i_m}\|\widehat{\boldsymbol{\theta}}^{(m)}_{\widehat{g}^{(m)}_{i_m}} - \boldsymbol{\theta}^{(m)}_{g^{(m)}_{i_m}}\|^2 + \|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}} - \boldsymbol{\theta}^{(l)}_{g^{(l)}_{i_l}}\|^2 \\ &\quad + \frac{1}{\prod_{m\neq l} N_m}\sum_{m\neq l}\sum_{i_m=1}^{N_m}|\widehat{\alpha}_{\widehat{g}^{(1)}_{i_1}\cdots\widehat{g}^{(q)}_{i_q}} - \alpha_{g^{(1)}_{i_1}\cdots g^{(q)}_{i_q}}|^2. \end{aligned} \tag{A.2}$$

For the parameter $\mathbf{\Theta}_{i_1\cdots i_q}$, define the element-wise pseudo distance as

$$d_{i_1\cdots i_q}(\widehat{\mathbf{\Theta}}_{i_1\cdots i_q}, \mathbf{\Theta}_{i_1\cdots i_q}) = \sum_{l=1}^{q} \|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}} - \boldsymbol{\theta}^{(l)}_{g^{(l)}_{i_l}}\|^2 + |\widehat{\alpha}_{\widehat{g}^{(1)}_{i_1}\cdots\widehat{g}^{(q)}_{i_q}} - \alpha_{g^{(1)}_{i_1}\cdots g^{(q)}_{i_q}}|^2. \tag{A.3}$$

Let $\mathbb{Y}_t = \mathrm{vec}(\mathcal{Y}_t) \in \mathbb{R}^{\prod_l N_l}$ and $\mathbb{E}_t = \mathrm{vec}(\mathcal{E}_t) \in \mathbb{R}^{\prod_l N_l}$. Then (3) can be rewritten as $\mathbb{Y}_t = \mathbf{B}_0\mathbb{Y}_{t-1} + \mathbf{c}_t + \mathbb{E}_t$, where

$$\mathbf{B}_0 = \sum_l \mathbf{I}_{N_1} \otimes \cdots \otimes \mathbf{I}_{N_{l-1}} \otimes (\mathbf{L}_{l,0}\mathbf{W}^{(l)}) \otimes \mathbf{I}_{N_{l+1}} \otimes \cdots \otimes \mathbf{I}_{N_q} + \mathrm{diag}\{\mathrm{vec}(\mathcal{A}_0)\} \in \mathbb{R}^{\prod_l N_l \times \prod_l N_l}, \tag{A.4}$$

$$\mathbf{c}_t = \sum_{l=1}^{q} \mathrm{vec}(\mathbf{1}_{N_1} \circ \cdots \circ \mathbf{1}_{N_{l-1}} \circ \boldsymbol{\beta}^{(l)0}_{X_l,t} \circ \mathbf{1}_{N_{l+1}} \circ \cdots \circ \mathbf{1}_{N_q}) \stackrel{\text{def}}{=} \sum_l \mathbf{c}^{(l)}_t \in \mathbb{R}^{\prod_l N_l}, \tag{A.5}$$

and $\mathbf{L}_{l,0}$, $\mathcal{A}_0$, $\boldsymbol{\beta}^{(l)0}_{X_l,t}$ are the true values of the corresponding terms for $l \in [q]$, where $\boldsymbol{\beta}^{(l)}_{X_l,t} = (\mathbf{x}^{(l)\top}_{i_l t}\boldsymbol{\zeta}^{(l)}_{g^{(l)}_{i_l}} : 1 \le i_l \le N_l)^\top \in \mathbb{R}^{N_l}$ is defined in the main text. Recall that $E(\mathbf{c}_t) = \mathbf{0}$ because $E(\mathbf{x}^{(l)}_{i_l t}) = \mathbf{0}$ for all $i_l \in [N_l]$ and $l \in [q]$, and by Assumption 5 that $\mathbb{Y}_0 = \mathbf{0}$, we have

$$\mathbb{Y}_t = \sum_{k=0}^{t} \mathbf{B}_0^k \mathbf{c}_{t-k} + \sum_{k=0}^{t} \mathbf{B}_0^k \mathbb{E}_{t-k} \stackrel{\text{def}}{=} \mathbb{Y}^c_t + \mathbb{Y}^e_t. \tag{A.6}$$

Let $\mathbf{\Gamma} = \mathrm{cov}(\mathbb{Y}_t)$. Denote $\tau_{\max} = \max_{i_1,\cdots,i_q} \lambda_{\max}(\mathbf{\Sigma}_{i_1\cdots i_q})$ and $\mathbf{\Sigma}_{i_1\cdots i_q}$ is defined in Assumption 2.

General Notations. For a matrix $\mathbf{M} = (m_{ij}) \in \mathbb{R}^{m\times n}$, let $|\mathbf{M}|_e = (|m_{ij}|)_{i\in[m],j\in[n]}$. For two matrices $\mathbf{M}_1 = (m_{1ij}) \in \mathbb{R}^{m\times n}$ and $\mathbf{M}_2 = (m_{2ij}) \in \mathbb{R}^{m\times n}$, $\mathbf{M}_1 \preccurlyeq \mathbf{M}_2$ means that $m_{1ij} \le m_{2ij}$ for all $i \in [m]$ and $j \in [n]$. In addition, define $\|\mathbf{M}\|$ as the largest singular value of $\mathbf{M}$, $\|\mathbf{M}\|_F = \mathrm{tr}(\mathbf{M}\mathbf{M}^\top)^{1/2}$ as the Frobenius norm, $\|\mathbf{M}\|_\infty = \max_i \sum_j |m_{ij}|$, and $\|\mathbf{M}\|_{\max}$ as its largest absolute value. For a vector $\mathbf{v}$, denote the norm $\|\mathbf{v}\|_1 = \sum_i |v_i|$. Denote $\mathbf{e}^{(n)}_i$ as a vector of length n, whose ith element equals to 1 while others are 0.

Appendix B. Additional Discussion for the Case when $q = 2$

B.1 Estimation Details

As shown in Section 3.1, the estimator when $q = 2$ could be written as

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}^{(1)} & \mathbf{M}^{(12)} & \mathbf{M}^{(1\alpha)} \\ \mathbf{M}^{(12)\top} & \mathbf{M}^{(2)} & \mathbf{M}^{(2\alpha)} \\ \mathbf{M}^{(1\alpha)\top} & \mathbf{M}^{(2\alpha)\top} & \mathbf{M}^{\alpha} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \\ \mathbf{b}^{\alpha} \end{pmatrix}, \tag{A.7}$$

where the expressions are given as

$$\mathbf{M}^{(1)}_{g^{(1)}} = \sum_{t,g^{(2)}} \mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)}t}\mathbb{X}^{(1)}_{g^{(1)}g^{(2)}t}, \quad \mathbf{M}^{(1)} = \mathrm{diag}\left\{\mathbf{M}^{(1)}_{g^{(1)}} : g^{(1)} \in [G_1]\right\} \in \mathbb{R}^{G_1(p_1+1)\times G_1(p_1+1)},$$

$$\mathbf{M}^{(12)}_{g^{(1)}g^{(2)}} = \sum_t \mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)}t}\mathbb{X}^{(2)}_{g^{(1)}g^{(2)}t},$$

$$\mathbf{M}^{(12)} = \left(\mathbf{M}^{(12)}_{g^{(1)}g^{(2)}} : g^{(1)} \in [G_1], g^{(2)} \in [G_2]\right) \in \mathbb{R}^{G_1(p_1+1)\times G_2(p_2+1)},$$

$$\mathbf{M}^{(1\alpha)}_{g^{(1)}\mathcal{I}_{g^{(1)\prime}g^{(2)}}} = \sum_t \mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)}t}\mathbb{Y}_{g^{(1)}g^{(2)},(t-1)} I\left(g^{(1)} = g^{(1)\prime}\right), \quad \mathcal{I}_{g^{(1)\prime}g^{(2)}} = (g^{(2)}-1)G_1 + g^{(1)\prime},$$

$$\mathbf{M}^{(1\alpha)} = \left(\mathbf{M}^{(1\alpha)}_{g^{(1)}\mathcal{I}_{g^{(1)\prime}g^{(2)}}} : g^{(1)} \in [G_1], \mathcal{I}_{g^{(1)\prime}g^{(2)}} \in [G_1G_2]\right) \in \mathbb{R}^{G_1(p_1+1)\times G_1G_2},$$

$$\mathbf{M}^{(2)}_{g^{(2)}} = \sum_{t,g^{(1)}} \mathbb{X}^{(2)\top}_{g^{(1)}g^{(2)}t}\mathbb{X}^{(2)}_{g^{(1)}g^{(2)}t}, \quad \mathbf{M}^{(2)} = \operatorname{diag}\left\{\mathbf{M}^{(2)}_{g^{(2)}} : g^{(2)} \in [G_2]\right\} \in \mathbb{R}^{G_2(p_2+1)\times G_2(p_2+1)},$$

$$\mathbf{M}^{(2)\alpha}_{g^{(2)}\mathcal{I}_{g^{(1)}g^{(2)\prime}}} = \sum_t \mathbb{X}^{(2)\top}_{g^{(1)}g^{(2)}t}\mathbb{Y}_{g^{(1)}g^{(2)},(t-1)} I\left(g^{(2)} = g^{(2)\prime}\right),$$

$$\mathbf{M}^{(2)\alpha} = \left(\mathbf{M}^{(2)\alpha}_{g^{(1)}\mathcal{I}_{g^{(1)}g^{(2)\prime}}} : g^{(2)} \in [G_2], \mathcal{I}_{g^{(1)}g^{(2)\prime}} \in [G_1G_2]\right) \in \mathbb{R}^{G_2(p_2+1)\times G_1G_2},$$

$$\mathbf{M}^{\alpha}_{\mathcal{I}_{g^{(1)}g^{(2)}}\mathcal{I}_{g^{(1)\prime}g^{(2)\prime}}} = \sum_t \left\|\mathbb{Y}_{g^{(1)}g^{(2)},(t-1)}\right\|^2 I\left(g^{(1)} = g^{(1)\prime}, g^{(2)} = g^{(2)\prime}\right),$$

$$\mathbf{M}^{\alpha} = \left(\mathbf{M}^{\alpha}_{\mathcal{I}_{g^{(1)}g^{(2)}}\mathcal{I}_{g^{(1)\prime}g^{(2)\prime}}} : \mathcal{I}_{g^{(1)}g^{(2)}} \in [G_1G_2], \mathcal{I}_{g^{(1)\prime}g^{(2)\prime}} \in [G_1G_2]\right) \in \mathbb{R}^{G_1G_2\times G_1G_2},$$

$$\mathbf{b}^{(1)}_{g^{(1)}} = \sum_{t,g^{(2)}} \mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)}t}\mathbb{Y}_{g^{(1)}g^{(2)},t}, \quad \mathbf{b}^{(1)} = \left(\mathbf{b}^{\top}_{g^{(1)}} : g^{(1)} \in [G_1]\right)^{\top} \in \mathbb{R}^{G_1(p_1+1)},$$

$$\mathbf{b}^{(2)}_{g} = \sum_{t,g^{(1)}} \mathbb{X}^{(2)\top}_{g^{(1)}g^{(2)}t}\mathbb{Y}_{g^{(1)}g^{(2)},t}, \quad \mathbf{b}^{(2)} = \left(\mathbf{b}^{(2)\top}_{g^{(2)}} : g^{(2)} \in [G_2]\right)^{\top} \in \mathbb{R}^{G_2(p_2+1)},$$

$$\mathbf{b}^{\alpha}_{\mathcal{I}_{g^{(1)}g^{(2)}}} = \sum_t \mathbb{Y}^{\top}_{g^{(1)}g^{(2)},(t-1)}\mathbb{Y}_{g^{(1)}g^{(2)},t}, \quad \mathbf{b}^{\alpha} = \left(\mathbf{b}^{\alpha}_{\mathcal{I}_{g^{(1)}g^{(2)}}} : \mathcal{I}_{g^{(1)}g^{(2)}} \in [G_1G_2]\right)^{\top} \in \mathbb{R}^{G_1G_2}.$$

In the above expressions,

$$\mathbb{X}^{(1)}_{g^{(1)}g^{(2)},t} = \left(\operatorname{vec}(\mathbf{W}_1^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\mathbf{Y}_{t-1}^{(\cdot,\mathcal{R}^{(2)}_{g^{(2)}})}), \mathbf{1}_{N_{2g^{(2)}}} \otimes (\mathbf{X}^{(1)}_t)^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\right) \in \mathbb{R}^{(N_{1g^{(1)}}N_{2g^{(2)}})\times(p_1+1)}, \tag{A.8}$$

$$\mathbb{X}^{(2)}_{g^{(1)}g^{(2)},t} = \left(\operatorname{vec}(\mathbf{Y}_{t-1}^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\mathbf{W}_2^{(\cdot,\mathcal{R}^{(2)}_{g^{(2)}})}), (\mathbf{X}^{(2)}_t)^{(\mathcal{R}^{(2)}_{g^{(2)}},\cdot)} \otimes \mathbf{1}_{N_{1g^{(1)}}}\right) \in \mathbb{R}^{(N_{1g^{(1)}}N_{2g^{(2)}})\times(p_2+1)}, \tag{A.9}$$

and $\mathbf{X}^{(l)}_t = (\mathbf{x}^{(l)}_{1t}, \cdots, \mathbf{x}^{(l)}_{N_1 t})^{\top} \in \mathbb{R}^{N_l\times p_l}$, $N^{(l)}_{lg^{(l)}} = |\mathcal{R}^{(l)}_{g^{(l)}}|$ for $l = 1, 2$, and $\mathbb{Y}_{g^{(1)}g^{(2)},t} = \operatorname{vec}(\mathcal{Y}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\mathcal{R}^{(2)}_{g^{(2)}})}) \in \mathbb{R}^{|\mathcal{R}^{(1)}_{g^{(1)}}||\mathcal{R}^{(2)}_{g^{(2)}}|}$. We then provide the iterative update algorithm when $q = 2$ in the following Algorithm A.1.

We remark here that the technical conditions in Section 4 can be simplified in this matrix case for easy understanding. For example, the covariates Assumption 4 can be rewritten as follow.

Assumption A.8. (Distribution of Covariates of Matrix Case) *Recall that* $\mathbf{x}^{(l)}_{i_l t} \in \mathbb{R}^{p_l}$ *is the covariate vector of the* i_l*th subject in the* l*th layer at time* t*. Assume* $E(\mathbf{x}^{(1)}_{it}) = \mathbf{0}$

Algorithm A.1 Estimation of the GTNAR Model When $q = 2$

1: **Input:** $\{\mathcal{Y}_t, \mathbf{X}_t^{(l)}, \mathbf{W}^{(l)}, G_l\}$ for $l = 1, 2$.

2: Obtain initial group memberships $\mathcal{G}_l^{[0]}$ according to Algorithm A.3 in Appendix F. Let $\{\boldsymbol{\xi}^{[k]}, \mathcal{G}_l^{[k]}\}$ be the estimators and memberships in the kth iteration.

3: Repeat Step 1 and Step 2 for $k = 1, 2, \cdots$ until convergence.

Step 1. Given $\{\mathcal{G}_l^{[k-1]}\}$ for all layers $l = 1, 2$, calculate $\boldsymbol{\xi}^{[k-1]} = (\mathbf{M}^{[k-1]})^{-1}\mathbf{b}^{[k-1]}$, where $\mathbf{M}^{[k-1]}$ and $\mathbf{b}^{[k-1]}$ are obtained from (13) with $\mathcal{G}_l^{[k-1]}$s specified.

Step 2. Given $\boldsymbol{\xi}^{[k-1]}$, update the memberships in layer $l = 1, 2$ sequentially by,

$$
\begin{aligned}
g_i^{(1)[k]} = \arg \min_{g_i^{(1)} \in [G_1]} \sum_{j=1}^{N_2} \sum_{t=1}^{T} \Big\{ & Y_{ij,t} \\
& - \Big(\lambda_{g_i^{(1)}}^{(1)[k]} \sum_{m=1}^{N_1} \frac{a_{im}^{(1)}}{n_{1i}} Y_{mj,(t-1)} + \lambda_{g_j^{(2)[k-1]}}^{(2)[k]} \sum_{m=1}^{N_2} \frac{a_{jm}^{(2)}}{n_{2j}} Y_{im,(t-1)} \Big) \\
& - \alpha_{g_i^{(1)} g_j^{(2)[k-1]}}^{[k]} Y_{ij,(t-1)} - \Big(\mathbf{x}_{it}^{(1)\top} \boldsymbol{\zeta}_{g_i^{(1)}}^{(1)[k]} + \mathbf{x}_{jt}^{(2)\top} \boldsymbol{\zeta}_{g_j^{(2)[k-1]}}^{(2)[k]} \Big) \Big\}^2
\end{aligned}
\tag{A.10}
$$

and

$$
\begin{aligned}
g_j^{(2)[k]} = \arg \min_{g_j^{(2)} \in [G_2]} \sum_{i=1}^{N_1} \sum_{t=1}^{T} \Big\{ & Y_{ij,t} \\
& - \Big(\lambda_{g_i^{(1)[k]}}^{(1)[k]} \sum_{m=1}^{N_1} \frac{a_{im}^{(1)}}{n_{1i}} Y_{mj,(t-1)} + \lambda_{g_j^{(2)}}^{(2)[k]} \sum_{m=1}^{N_2} \frac{a_{jm}^{(2)}}{n_{2j}} Y_{im,(t-1)} \Big) \\
& - \alpha_{g_i^{(1)[k]} g_j^{(2)}}^{[k]} Y_{ij,(t-1)} - \Big(\mathbf{x}_{it}^{(1)\top} \boldsymbol{\zeta}_{g_i^{(1)[k]}}^{(1)[k]} + \mathbf{x}_{jt}^{(2)\top} \boldsymbol{\zeta}_{g_j^{(2)}}^{(2)[k]} \Big) \Big\}^2
\end{aligned}
\tag{A.11}
$$

4: **Output:** Final estimator and memberships: $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}^{[K]}$ and $\widehat{\mathcal{G}} = \{\mathcal{G}_l^{[K]} : l = 1, 2\}$. Here K is the final number of iteration rounds.

for all $i \in [N_1]$ and $t \in [T]$. Let $\boldsymbol{\eta}_1 \in \mathbb{R}^{p_1}$ be a constant vector satisfying $\|\boldsymbol{\eta}_1\| \leq c$, where c is a positive constant. Define $\mathbf{x}_t^{(1)\eta} = (\mathbf{x}_{it}^{(1)\top}\boldsymbol{\eta}_1 : i \in [N_1])^\top \in \mathbb{R}^{N_1}$. Assume $\mathbf{x}^{(1)\eta} = (\mathbf{x}_t^{(1)\eta^\top} : 0 \leq t \leq T)^\top \in \mathbb{R}^{N_1(T+1)}$ satisfies the K-convex concentration property for some constant K according to Definition 1.

B.2 Model Inference

In this subsection, we provide an estimator to the asymptotic covariance when $q = 2$. This procedure can be easily extended to $q > 2$. With the parameter estimator $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\Theta}}_{ij} = (\widehat{\boldsymbol{\theta}}^{(1)\top}_{\widehat{g}_i^{(1)}}, \widehat{\boldsymbol{\theta}}^{(2)\top}_{\widehat{g}_j^{(2)}}, \widehat{\alpha}_{\widehat{g}_i^{(1)}\widehat{g}_j^{(2)}})^\top)$, we first estimate σ^2 as follows

$$\widehat{\sigma}^2 = \frac{1}{N_1N_2T}\sum_{i=1}\sum_{j=1}\sum_{t=1}(Y_{ij,t} - \mathcal{X}_{ij,t}^\top\widehat{\boldsymbol{\Theta}}_{ij})^2, \tag{A.12}$$

where $\mathcal{X}_{ij,t}$ is defined in Assumption 2. Next, we estimate $\mathbf{M}^0$ by $\widehat{\mathbf{M}}$, where $\widehat{\mathbf{M}}$ is obtained by plugging estimated parameters $\widehat{\boldsymbol{\Theta}}$ into the expression in (13). In the following theorem, we show that the covariance estimator is consistent.

Theorem 10. *Suppose Assumption 1–5 and assume that there exists n, such that $c_1 n \leq \min_l N_l \leq \max_l N_l \leq c_2 n$ for some constants $c_1, c_2 > 0$. When $G_1 = G_{1,0}$ $G_2 = G_{2,0}$, assume $\{\log(N_1N_2)\}^2/T \to 0$, then the following holds,*

$$\widehat{\sigma}^2 \to_p \sigma^2, \quad (n^2T)^{-1}\widehat{\sigma}^2\boldsymbol{\eta}^\top(\widehat{\mathbf{M}})^{-1}\boldsymbol{\eta} \to_p (n^2T)^{-1}\sigma^2\boldsymbol{\eta}^\top(\mathbf{M}^0)^{-1}\boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is defined in Theorem 7.

The proof of Theorem 10 can be found in Appendix B.2.1. Theorem 10 indicates that we can obtain a consistent estimator for the asymptotic variance by plugging in the estimators $\widehat{\boldsymbol{\Theta}}$ and the consistent estimator $\widehat{\sigma}^2$. This assures a valid statistical inference procedure. We next present a number of simulation studies to examine the finite sample performances of the model estimation and inference procedures.

B.2.1 Proof of Theorem 10

To prove Theorem 10, we prove the following two statements separately,

$$\widehat{\sigma}^2 \to_p \sigma^2, \tag{A.13}$$

$$(n^2T)^{-1}\widehat{\mathbf{M}}_{nT} \to_p (n^2T)^{-1}\mathbf{M}^0. \tag{A.14}$$

(1) Proof of (A.13)

Note that $\sigma^2 = (N_1N_2T)^{-1}\sum_{i,j,t}E(\varepsilon_{ij,t}^2) = (N_1N_2T)^{-1}E\{\sum_{i,j}Q_{ij}(\boldsymbol{\Theta}_{ij}^0)\} = (N_1N_2T)^{-1}Q^*(\boldsymbol{\Theta}^0)$, and $\widehat{\sigma}^2 = (N_1N_2T)^{-1}\sum_{i,j}Q_{ij}(\widehat{\boldsymbol{\Theta}}_{ij}) = (N_1N_2T)^{-1}Q(\widehat{\boldsymbol{\Theta}})$. To prove (A.13), it suffices to show that $(N_1N_2T)^{-1}|Q(\widehat{\boldsymbol{\Theta}}) - Q^*(\boldsymbol{\Theta}^0)| = o_p(1)$. Further note that

$$\frac{1}{N_1N_2T}|Q(\widehat{\boldsymbol{\Theta}}) - Q^*(\boldsymbol{\Theta}^0)| \leq \frac{1}{N_1N_2T}\left\{|Q(\widehat{\boldsymbol{\Theta}}) - Q^*(\widehat{\boldsymbol{\Theta}})| + |Q^*(\widehat{\boldsymbol{\Theta}}) - Q^*(\boldsymbol{\Theta}^0)|\right\},$$

hence we bound the two terms on the right hand side.

(1.1) Order of $(N_1N_2T)^{-1}|Q(\widehat{\mathbf{\Theta}})-Q^*(\widehat{\mathbf{\Theta}})|$.

By Lemma 17, we have that

$$\sup_{\|\mathbf{\Theta}\|_{\max}\le R}\left|\frac{1}{N_1N_2T}\{Q(\mathbf{\Theta})-Q^*(\mathbf{\Theta})\}\right|=O_p\left\{T^{-1/2}(m+\log(N_1N_2))\right\},$$

where $m=p_1+p_2+3$. Thus when $\{m+\log(N_1N_2)\}/\sqrt{T}\to 0$, we have $(N_1N_2T)^{-1}|Q(\widehat{\mathbf{\Theta}})-Q^*(\widehat{\mathbf{\Theta}})|=o_p(1)$.

(1.2) Order of $(N_1N_2T)^{-1}|Q^*(\widehat{\mathbf{\Theta}})-Q^*(\mathbf{\Theta}^0)|$

By Lemma 23 we have $(N_1N_2T)^{-1}\{Q^*(\widehat{\mathbf{\Theta}})-Q^*(\mathbf{\Theta}^0)\}\le\tau_{\max}d(\widehat{\mathbf{\Theta}},\mathbf{\Theta}^0)$. Since $\tau_{\max}$ is bounded due to Lemma 33 and $d(\widehat{\mathbf{\Theta}},\mathbf{\Theta}^0)=O_p\{T^{-1/2}(m+\log(N_1N_2))\}$ according to Theorem 2, we have that $(N_1N_2T)^{-1}|Q^*(\widehat{\mathbf{\Theta}})-Q^*(\mathbf{\Theta}^0)|=O_p\{T^{-1}(\log(N_1N_2)^2)\}=o_p(1)$ when $(\log N_1N_2)^2/T\to 0$. This condition can be implied by $\{m+\log(N_1N_2)\}/\sqrt{T}\to 0$.

From both (1.1) and (1.2), we have that $(N_1N_2T)^{-1}|Q(\widehat{\mathbf{\Theta}})-Q^*(\mathbf{\Theta}^0)|=o_p(1)$ when $\{m+\log(N_1N_2)\}/\sqrt{T}\to 0$. This implies that $\widehat{\sigma}^2\to_p\sigma^2$ under the same condition. This finishes the proof of (A.13).

(2) Proof of (A.14)

Recall the expression of $\mathbf{M}$ defined in (A.7) when $q=2$, we need to show that the nine elements in $(n^2T)^{-1}\widehat{\mathbf{M}}_{nT}$ are consistent in an analogous manner. Hence, we take the first element, i.e. $(n^2T)^{-1}\widehat{\mathbf{M}}^{(1)}\in\mathbb{R}^{G_1(p_1+1)\times G_1(p_1+1)}$, as an example to show the consistency, where n is defined in Theorem 10. Other terms could be proved similarly.

Recall that $\mathbf{M}^{(1)}=\mathrm{diag}\{\mathbf{M}^{(1)}_{g^{(1)}}:g^{(1)}\in[G_1]\}$ and $\mathbf{M}^{(1)}_{g^{(1)}}=\sum_{t,g^{(2)}}\mathbb{X}^\top_{g^{(1)}g^{(2)},t}\mathbb{X}_{g^{(1)}g^{(2)},t}$, we show that for any $g^{(1)}\in[G_{1,0}]$, it holds that $\big|(n^2T)^{-1}\widehat{\mathbf{M}}^{(1)}_{g^{(1)}}-\lim_{n,T\to\infty}(n^2T)^{-1}E(\widehat{\mathbf{M}}^{(1)}_{g^{(1)}})\big|=o_p(1)$. Recall that

$$\begin{aligned}\mathbb{X}_{g^{(1)}g^{(2)},t}&=\left(\mathrm{vec}(\mathbf{W}_1^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\mathbf{Y}_{t-1}^{(\cdot,\mathcal{R}^{(2)}_{g^{(2)}})}),\mathbf{1}_{N_{2g^{(2)}}}\otimes\mathbf{X}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\right)\\&=\left((\mathbf{I}_{N_{2g^{(2)}}}\otimes\mathbf{W}_1^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)})\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}},\mathbf{1}_{N_{2g^{(2)}}}\otimes\mathbf{X}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\right)\\&\stackrel{\mathrm{def}}{=}\left(\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}},\mathbf{1}_{N_{2g^{(2)}}}\otimes\mathbf{X}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\right)\in\mathbb{R}^{(N_{1g^{(1)}}N_{2g^{(2)}})\times(p_1+1)},\end{aligned}$$

where $\mathbb{W}_{g^{(1)}g^{(2)}}\stackrel{\mathrm{def}}{=}\mathbf{I}_{N_{2g^{(2)}}}\otimes\mathbf{W}_1^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\in\mathbb{R}^{N_{2g^{(2)}}N_{1g^{(1)}}\times N_{2g^{(2)}}N_1}$ and $\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}}\stackrel{\mathrm{def}}{=}\mathrm{vec}(\mathbf{Y}_{t-1}^{(\cdot,\mathcal{R}^{(2)}_{g^{(2)}})})\in\mathbb{R}^{N_{2g^{(2)}}N_1}$. Hence, it is equivalent to prove that

$$\begin{aligned}&(n^2T)^{-1}\sum_{t,g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}}\\&\to_p\lim_{n,T\to\infty}(n^2T)^{-1}\sum_{t,g^{(2)}}E(\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}}),\qquad(A.15)\\&(n^2T)^{-1}\sum_{t,g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbf{1}_{N_{2g^{(2)}}}\otimes\mathbf{X}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\end{aligned}$$

$$\to_p \lim_{n,T\to\infty}(n^2T)^{-1}\sum_{t,g^{(2)}} E(\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbf{1}_{N_{2g^{(2)}}}\otimes\mathbf{X}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}). \tag{A.16}$$

We prove (A.15) for example, and (A.16) could be proved similarly. Similar as the decomposition in (A.6), write $\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}} = \mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},c} + \mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},e}$, where $\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},c} = \sum_{k=0}^{t-1}(\mathbf{B}_0^{\mathcal{R}^{(2)}_{g^{(2)}}})^k\mathbf{c}_{t-k}^{\mathcal{R}^{(2)}_{g^{(2)}}}$, $\mathbf{B}_0^{\mathcal{R}^{(2)}_{g^{(2)}}} = \mathbf{I}_{N_{2g^{(2)}}}\otimes\mathbf{L}_0\mathbf{W}_1 + \mathbf{G}_0^{(\mathcal{R}^{(2)}_{g^{(2)}},\mathcal{R}^{(2)}_{g^{(2)}})\top}\mathbf{W}_2^{(\mathcal{R}^{(2)}_{g^{(2)}},\mathcal{R}^{(2)}_{g^{(2)}})}\otimes\mathbf{I}_{N_1}\in\mathbb{R}^{N_1N_{2g^{(2)}}\times N_{2g^{(2)}}N_1}$, and $\mathbf{c}_t^{\mathcal{R}^{(2)}_{g^{(2)}}}\in\mathbb{R}^{N_1N_{2g^{(2)}}}$ is a similar modification of $\mathbf{c}_t$. $\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},e}$ could be represented analogously. Then the left hand side of (A.15) could be separated as

$$\begin{aligned}
&\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}}\\
&=\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},c\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},c} + 2\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},c\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},e}\\
&+\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},e\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}},e}
\end{aligned}$$

We take the first term as an example. Let $\mathbf{N}_{g^{(1)}g^{(2)}} = \mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\in\mathbb{R}^{N_{2g^{(2)}}N1\times N_{2g^{(2)}}N_1}$, the we have

$$\begin{aligned}
&\frac{1}{n^2T}\sum_{t,g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}}\\
&=\frac{1}{n^2T}\sum_{t,g^{(2)}}^{T,G_2}\sum_{l_1=0}^{t-1}\sum_{l_2=0}^{t-1}\mathbf{c}_{l_1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}(\mathbf{B}_0^{\mathcal{R}^{(2)}_{g^{(2)}}})^{t-l_1\top}\mathbf{N}_{g^{(1)}g^{(2)}}(\mathbf{B}_0^{\mathcal{R}^{(2)}_{g^{(2)}}})^{t-l_2}\mathbf{c}_{l_2}^{\mathcal{R}^{(2)}_{g^{(2)}}}\\
&=\frac{1}{n}\sum_{g^{(2)}}\sum_{l_1=0}^{T-1}\sum_{l_2=0}^{T-1}\mathbf{c}_{l_1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\left\{\frac{1}{nT}\sum_{t=\max(1,l_1,l_2)}^{T}(\mathbf{B}_0^{\mathcal{R}^{(2)}_{g^{(2)}}})^{t-l_1\top}\mathbf{N}_{g^{(1)}g^{(2)}}(\mathbf{B}_0^{\mathcal{R}^{(2)}_{g^{(2)}}})^{t-l_2}\right\}\mathbf{c}_{l_2}^{\mathcal{R}^{(2)}_{g^{(2)}}}\\
&\stackrel{\text{def}}{=}\frac{1}{n}\sum_{g^{(2)}}\sum_{l_1=0}^{T-1}\sum_{l_2=0}^{T-1}\mathbf{c}_{l_1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbf{E}_{l_1,l_2}^{\mathcal{R}^{(2)}_{g^{(2)}}}\mathbf{c}_{l_2}^{\mathcal{R}^{(2)}_{g^{(2)}}}.
\end{aligned}$$

Define $\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}}} = (\mathbf{c}_0^{\mathcal{R}^{(2)}_{g^{(2)}}\top},\cdots,\mathbf{c}_{T-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top})^\top$, and $\mathbf{E}^{\mathcal{R}^{(2)}_{g^{(2)}}} = (\mathbf{E}_{l_1,l_2}^{\mathcal{R}^{(2)}_{g^{(2)}}}, 0\le l_1,l_2\le T-1)$, then we have

$$\begin{aligned}
&\frac{1}{n^2T}\sum_{t,g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}} = \frac{1}{n}\sum_{g^{(2)}}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbf{E}^{\mathcal{R}^{(2)}_{g^{(2)}}}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}}}\\
&=\frac{1}{n}\sum_{g^{(2)}}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}},(1)\top}\mathbf{E}^{\mathcal{R}^{(2)}_{g^{(2)}}}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}},(1)} + \frac{2}{n}\sum_{g^{(2)}}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}},(1)\top}\mathbf{E}^{\mathcal{R}^{(2)}_{g^{(2)}},z}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}},(2)}\\
&+\frac{1}{n}\sum_{g^{(2)}}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}},(2)\top}\mathbf{E}^{\mathcal{R}^{(2)}_{g^{(2)}}}\mathbf{c}^{\mathcal{R}^{(2)}_{g^{(2)}},(2)}.
\end{aligned}$$

Then we could follow the proof of Lemma 21 to show the result. Since

$$n^{-2}\sum_{g^{(2)}}\mathbf{1}^\top_{N_{2g^{(2)}}N_1}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbf{1}_{N_{2g^{(2)}}N_1}=O(1)$$

under Assumption that there exists n that $c_1 n \le \min_l N_l \le max_l N_l \le c_2 n$, we would have

$$P\Big\{\Big|\frac{1}{n^2T}\sum_{t,g^{(2)}}\{\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}}-E(\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}\top}\mathbb{W}^\top_{g^{(1)}g^{(2)}}\mathbb{W}_{g^{(1)}g^{(2)}}\mathbb{Y}_{t-1}^{\mathcal{R}^{(2)}_{g^{(2)}}})\}\Big|\ge u\Big\}$$
$$\le C\exp\Big\{-c\min(Tu^2,\sqrt{T}u)\Big\}.$$

This would lead to the statement (A.15). Then, (A.16) can be shown similarly. Therefore, the consistency in (A.14) could be obtained when $T\to\infty$. Overall, the consistency of the asymptotic covariance estimation in Theorem 10 could be obtained by Step (1) and Step (2).

B.3 Existence of Asymptotic Covariance

We illustrate the asymptotic covariance matrix $\mathbf{M}^0$ in Theorem 7 exists when $q=2$, which can be directly extended to the general q case. For notation simplicity, in this subsection, we use g and h to represent group in the first and second dimension, respectively.

Assumption A.9. *Recall that* $\mathbb{Y}_t=\mathrm{vec}(\mathbf{Y}_t)\in\mathbb{R}^{N_1N_2}$. *Denote* $\mathbb{Y}_{gh,t}=\mathrm{vec}(\mathbf{Y}_t^{(\mathcal{R}^{(1)}_g,\mathcal{R}^{(2)}_h)})$. *For any* $g_1,g_2\in[G_1]$ *and* $h_1,h_2\in[G_2]$, *denote* $\mathbf{\Sigma}^{\mathcal{E}}_{g_1h_1,g_2h_2}=E(\mathbb{Y}^e_{g_1h_1,t}\mathbb{Y}^{e\top}_{g_2h_2,t})$, $\mathbf{\Sigma}^{\mathbf{X}}_{g_1h_1,g_2h_2}=E(\mathbb{Y}^c_{g_1h_1,t}\mathbb{Y}^{c\top}_{g_2h_2,t})$. *Denote* $\mathcal{W}^1_{gh}=(\mathbf{I}_{N_{2h}}\otimes\mathbf{W}_1^{(\mathcal{R}^{(1)}_g,\cdot)})\in\mathbb{R}^{N_{1g}N_{2h}\times N_1N_{2h}}$, *and denote* $\mathcal{W}^2_{gh}=(\mathbf{I}_{N_{1g}}\otimes(\mathbf{W}_2^\top)^{(\cdot,\mathcal{R}^{(2)}_h)})\in\mathbb{R}^{N_{1g}N_{2h}\times N_{1g}N_2}$. *Assume the following limits exist,*

$$\kappa^{11}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^1_{g_1h_1}\mathbf{\Sigma}^{\mathcal{E}}_{\cdot h_1,\cdot h_2}\mathcal{W}^{1\top}_{g_1h_2}),\ \nu^{11}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^1_{g_1h_1}\mathbf{\Sigma}^{\mathbf{X}}_{\cdot h_1,\cdot h_2}\mathcal{W}^{1\top}_{g_1h_2}),$$
$$\kappa^{22}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^2_{g_1h_1}\mathbf{\Sigma}^{\mathcal{E}}_{g_1\cdot,g_2\cdot}\mathcal{W}^{2\top}_{g_1h_1}),\ \nu^{22}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^2_{g_1h_1}\mathbf{\Sigma}^{\mathbf{X}}_{g_1\cdot,g_2\cdot}\mathcal{W}^{2\top}_{g_1h_1}),$$
$$\kappa^{12}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^1_{g_1h_1}\mathbf{\Sigma}^{\mathcal{E}}_{\cdot h_1,g_2\cdot}\mathcal{W}^{2\top}_{g_1h_1}),\ \nu^{12}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^1_{g_1h_1}\mathbf{\Sigma}^{\mathbf{X}}_{\cdot h_1,g_2\cdot}\mathcal{W}^{2\top}_{g_1h_1}),$$
$$\kappa^{1}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^1_{g_1h_1}\mathbf{\Sigma}^{\mathcal{E}}_{\cdot h_1,g_2h_2}),\ \nu^{1}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^1_{g_1h_1}\mathbf{\Sigma}^{\mathbf{X}}_{\cdot h_1,g_2h_2}),$$
$$\kappa^{2}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^2_{g_1h_1}\mathbf{\Sigma}^{\mathcal{E}}_{g_1\cdot,g_2h_2}),\ \nu^{2}_{g_1h_1,g_2h_2}=\lim_{n\to\infty}n^{-2}tr(\mathcal{W}^2_{g_1h_1}\mathbf{\Sigma}^{\mathbf{X}}_{g_1\cdot,g_2h_2}),$$
$$\kappa^{\alpha}_{g_1h_1,g_2,h_2}=\lim_{n\to\infty}n^{-2}tr(\mathbf{\Sigma}^{\mathcal{E}}_{g_1h_1,g_2h_2}),\quad \nu^{\alpha}_{g_1h_1,g_2,h_2}=\lim_{n\to\infty}n^{-2}tr(\mathbf{\Sigma}^{\mathbf{X}}_{g_1h_1,g_2h_2}),$$

where $\mathbf{\Sigma}^{\mathcal{E}}_{\cdot h_1,\cdot h_2},\mathbf{\Sigma}^{\mathbf{X}}_{\cdot h_1,\cdot h_2}\in\mathbb{R}^{N_1N_{2h_1}\times N_1N_{2h_2}}$, $\mathbf{\Sigma}^{\mathcal{E}}_{g_1\cdot,g_2\cdot},\mathbf{\Sigma}^{\mathbf{X}}_{g_1\cdot,g_2\cdot}\in\mathbb{R}^{N_{1g_1}N_2\times N_{1g_2}N_2}$, *and* $\mathbf{\Sigma}^{\mathcal{E}}_{g_1\cdot,g_2h_2}$, $\mathbf{\Sigma}^{\mathbf{X}}_{g_1\cdot,g_2h_2}\in\mathbb{R}^{N_{1g_1}N_2\times N_{1g_2}N_{2h_2}}$, $\mathbf{\Sigma}^{\mathcal{E}}_{g_1h_1,g_2\cdot},\mathbf{\Sigma}^{\mathbf{X}}_{g_1h_1,g_2\cdot}\in\mathbb{R}^{N_{1g_1}N_{2h_2}\times N_{1g_1}N_2}$. *Further assume that* $\mathbf{\Sigma}_{X_1}=E(\mathbf{x}^{(1)}_{it}\mathbf{x}^{(1)\top}_{it})\in\mathbb{R}^{p_1\times p_1}$, $\mathbf{\Sigma}_{X_2}=E(\mathbf{x}^{(2)}_{jt}\mathbf{x}^{(2)\top}_{jt})\in\mathbb{R}^{p_2\times p_2}$ *exist.*

Lemma 11. *Under Assumptions A.9, the asymptotic covariance matrix* $\mathbf{M}^0=\lim_{n,T\to\infty}\mathbf{M}^0_{n,T}$ *can be written as*

$$\mathbf{M}^0=\begin{pmatrix}\mathbf{\Xi}_1 & \mathbf{\Xi}_{12} & \mathbf{\Xi}_{1\alpha}\\ \mathbf{\Xi}_{12}^\top & \mathbf{\Xi}_2 & \mathbf{\Xi}_{2\alpha}\\ \mathbf{\Xi}_{1\alpha}^\top & \mathbf{\Xi}_{2\alpha}^\top & \mathbf{\Xi}_\alpha\end{pmatrix}.$$

where

$$\begin{aligned}
&\mathbf{\Xi}_1 = diag\left(\left(\begin{array}{cc} s_{11,g} & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{\Sigma}_{X_1} \end{array}\right) : g \in [G_1]\right), \quad \mathbf{\Xi}_{2,h} = diag\left(\left(\begin{array}{cc} s_{22,h} & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{\Sigma}_{X_2} \end{array}\right) : h \in [G_2]\right), \\
&\mathbf{\Xi}_{12} = \left(\left(\begin{array}{cc} s_{12,gh} & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{0} \end{array}\right) : g \in [G_1], h \in [G_2]\right), \\
&\mathbf{\Xi}_{1\alpha} = \left((s_{31,gh} I(g = g'), \mathbf{0}^\top) : g \in [G_1], \mathcal{I}_{g'h} \in [G_1 G_2]\right), \\
&\mathbf{\Xi}_{2\alpha} = \left((s_{32,gh} I(h = h'), \mathbf{0}^\top) : h \in [G_2], \mathcal{I}_{gh'} \in [G_1 G_2]\right), \\
&\mathbf{\Xi}_{\alpha} = \left(s_{4,gh} I(g = g', h = h') : \mathcal{I}_{gh} \in [G_1 G_2], \mathcal{I}_{g'h'} \in [G_1 G_2]\right),
\end{aligned} \tag{A.17}$$

In (A.17), $s_{11,g} = \sum_h \nu^{11}_{gh,gh} + \kappa^{11}_{gh,gh}$, $s_{22,h} = \sum_g \nu^{22}_{gh,gh} + \kappa^{22}_{gh,gh}$, $s_{12,gh} = \nu^{12}_{gh,gh} + \kappa^{12}_{gh,gh}$, $s_{31,gh} = \nu^{1}_{gh,gh} + \kappa^{1}_{gh,gh}$, $s_{32,gh} = \nu^{2}_{gh,gh} + \kappa^{2}_{gh,gh}$ *and* $s_{4,gh} = \nu^{\alpha}_{gh,gh} + \kappa^{\alpha}_{gh,gh}$, *with all the constants defined in Assumption A.9.*

Proof Recall the expression of $\mathbf{M}$ when $q = 2$ is defined in (A.7),

$$\mathbf{M} = \left(\begin{array}{ccc} \mathbf{M}^{(1)} & \mathbf{M}^{(12)} & \mathbf{M}^{(1\alpha)} \\ \mathbf{M}^{(12)\top} & \mathbf{M}^{(2)} & \mathbf{M}^{(2\alpha)} \\ \mathbf{M}^{(1\alpha)\top} & \mathbf{M}^{(2\alpha)\top} & \mathbf{M}^{\alpha} \end{array}\right).$$

Due to the similar format of the terms, we derive the expression of

$$\begin{aligned}
&\mathbf{s}_1 \overset{\text{def}}{=} \lim_{n,T\to\infty} (n^2 T)^{-1} E(\mathbf{M}^{(1)}), \quad \mathbf{s}_2 \overset{\text{def}}{=} \lim_{n,T\to\infty} (n^2 T)^{-1} E(\mathbf{M}^{(12)}), \\
&\mathbf{s}_3 \overset{\text{def}}{=} \lim_{n,T\to\infty} (n^2 T)^{-1} E(\mathbf{M}^{(1\alpha)}), \quad \mathbf{s}_4 \overset{\text{def}}{=} \lim_{n,T\to\infty} (n^2 T)^{-1} E(\mathbf{M}^{\alpha})
\end{aligned}$$

in the following four steps.

(1). Expression of $\mathbf{s}_1$.

We prove that $\mathbf{s}_1 = \text{diag}((s_{11,g}, \mathbf{s}_{12,g}^\top; \mathbf{s}_{12,g}, \mathbf{s}_{13}) : g \in [G_1])$ with $s_{11,g} = \sum_h s_{11,gh} = \sum_h \nu^{11}_{gh,gh} + \kappa^{11}_{gh,gh}$, $\mathbf{s}_{12,g} = \sum_h \mathbf{s}_{12,gh} = \mathbf{0}$ and $\mathbf{s}_{13,g} = \mathbf{\Sigma}_{X_1}$, which is defined in Assumption A.9. Recall that $\mathbf{M}^{(1)} = \text{diag}(\mathbf{M}^{(1)}_g : g \in [G_1]) \in \mathbb{R}^{G_1(p_1+1)\times G_1(p_1+1)}$ with $\mathbf{M}^{(1)}_g = \sum_{t,h} \mathbb{X}^{(1)\top}_{gh,t} \mathbb{X}^{(1)}_{gh,t}$, where $\mathbb{X}^{(1)}_{gh,t}$ is defined in (A.8). By (A.8), denote $\mathbb{Y}_{\cdot h,(t-1)} = \text{vec}(\mathbf{Y}^{(\cdot,\mathcal{R}^{(2)}_h)}_{t-1})$, we have

$$\begin{aligned}
s_{11,gh} &= \lim_{n\to\infty} (n^2 T)^{-1} \sum_t E\{\mathbb{Y}^\top_{\cdot h,(t-1)} (\mathbf{I}_{N_{2h}} \otimes \mathbf{W}_1^{(\mathcal{R}_g,\cdot)})^\top (\mathbf{I}_{N_{2h}} \otimes \mathbf{W}_1^{(\mathcal{R}_g,\cdot)}) \mathbb{Y}_{\cdot h,(t-1)}\}, \\
\mathbf{s}_{12,gh} &= \lim_{n,T\to\infty} (n^2 T)^{-1} \sum_t E\{\mathbb{Y}^\top_{\cdot h,(t-1)} (\mathbf{I}_{N_{2h}} \otimes \mathbf{W}_1^{(\mathcal{R}_g,\cdot)})^\top (\mathbf{1}_{N_{2h}} \otimes (\mathbf{X}_t^{(1)})^{(\mathcal{R}_g,\cdot)})\}, \\
\mathbf{s}_{13,gh} &= \lim_{n,T\to\infty} (n^2 T)^{-1} \sum_t E\{(\mathbf{1}_{N_{2h}} \otimes (\mathbf{X}_t^{(1)})^{(\mathcal{R}_g,\cdot)})^\top (\mathbf{1}_{N_{2h}} \otimes (\mathbf{X}_t^{(1)})^{(\mathcal{R}_g,\cdot)})\}.
\end{aligned}$$

Since $\mathbb{Y}_{\cdot h,(t-1)}$ is independent with $\mathbf{X}_t^{(1)}$, we have $s_{12} = \mathbf{0}$. Further note that

$$s_{11,gh} = \lim_{n\to\infty} (n^2 T)^{-1} \sum_t \left[\text{tr}(\mathcal{W}^1_{gh} \mathbf{\Sigma}^{\mathbf{X}}_{\cdot h,\cdot h} \mathcal{W}^{1\top}_{gh}) + \text{tr}(\mathcal{W}^1_{gh} \mathbf{\Sigma}^{\mathcal{E}}_{\cdot h,\cdot h} \mathcal{W}^{1\top}_{gh})\right]$$

$$= \lim_{n\to\infty} n^{-2}\big[\mathrm{tr}(\mathcal{W}^1_{gh}\mathbf{\Sigma}^{\mathbf{X}}_{\cdot h,\cdot h}\mathcal{W}^{1\top}_{gh}) + \mathrm{tr}(\mathcal{W}^1_{gh}\mathbf{\Sigma}^{\mathcal{E}}_{\cdot h,\cdot h}\mathcal{W}^{1\top}_{gh})\big] = \nu^{11}_{gh,gh} + \kappa^{11}_{gh,gh}.$$

Recall that $\mathbf{x}^{(1)}_{it}$s are independent and identically distributed covariates across all $i \in [N_1]$ and $t \in [T]$, denote the covariance as $\mathbf{\Sigma}_{X_1} = E(\mathbf{x}^{(1)}_{it}\mathbf{x}^{(1)\top}_{it}) \in \mathbb{R}^{p_1\times p_1}$. Therefore we have $\mathbf{s}_{13,gh} = \mathbf{\Sigma}_{X_1}$.

(2). Expression of $\mathbf{s}_2$.

Next, we prove that $\mathbf{s}_2 = ((s_{21,gh}, \mathbf{s}_{22,gh}; \mathbf{s}^\top_{22,gh}, \mathbf{s}_{23,gh}) : g \in [G_1], h \in [G_2])$ with $\mathbf{s}_{21,gh} = \nu^{12}_{gh,gh} + \kappa^{12}_{gh,gh}$, $\mathbf{s}_{22,gh} = \mathbf{0}$ and $\mathbf{s}_{23,gh} = \mathbf{0}$. Since $\mathbf{M}^{(12)} = \Big(\mathbf{M}^{(12)}_{gh} : g \in [G_1], h \in [G_2]\Big)$, we derive the expression of $\lim_{n,T\to\infty}(n^2T)^{-1}E(\mathbf{M}^{(12)}_{gh}) = \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t E(\mathbb{X}^{(1)\top}_{gh,t}\mathbb{X}^{(2)}_{gh,t})$. By the definition of $\mathbb{X}^{(1)}_{gh,t}$ and $\mathbb{X}^{(2)}_{gh,t}$ in (A.8) and (A.9), we have

$$\begin{aligned}
s_{21,gh} &= \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t E\{\mathbb{Y}^\top_{\cdot h,(t-1)}(\mathbf{I}_{N_{2h}} \otimes \mathbf{W}_1^{(\mathcal{R}^{(1)}_g,\cdot)})^\top(\mathbf{I}_{N_{1g}} \otimes \mathbf{W}_2^{(\cdot,\mathcal{R}^{(2)}_h)\top})\mathbb{Y}_{g\cdot,(t-1)}\},\\
\mathbf{s}_{22,gh} &= \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t E\{\mathbb{Y}^\top_{\cdot h,(t-1)}(\mathbf{I}_{N_{2h}} \otimes \mathbf{W}_1^{(\mathcal{R}^{(1)}_g,\cdot)})^\top((\mathbf{X}^{(2)}_t)^{(\mathcal{R}^{(2)}_h,\cdot)} \otimes \mathbf{1}_{N_{1g}})\},\\
\mathbf{s}_{23,gh} &= \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t E\{((\mathbf{X}^{(2)}_t)^{(\mathcal{R}^{(2)}_h,\cdot)} \otimes \mathbf{1}_{N_{1g}})^\top(\mathbf{1}_{N_{2h}} \otimes (\mathbf{X}^{(1)}_t)^{(\mathcal{R}^{(1)}_g,\cdot)})\}.
\end{aligned}$$

Since $\mathbb{Y}_{\cdot h,(t-1)}$ is independent with $\mathbf{X}^{(2)}_t$, we have $\mathbf{s}_{22} = \mathbf{0}$, and due to the independence between $\mathbf{X}^{(1)}_t$ and $\mathbf{X}^{(2)}_t$, we have $\mathbf{s}_{23} = \mathbf{0}$. Note that

$$s_{21,gh} = \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t\big[\mathrm{tr}(\mathcal{W}^1_{gh}\mathbf{\Sigma}^{\mathbf{X}}_{\cdot h,g\cdot}\mathcal{W}^{2\top}_{gh}) + \mathrm{tr}(\mathcal{W}^1_{gh}\mathbf{\Sigma}^{\mathcal{E}}_{\cdot h,g\cdot}\mathcal{W}^{2\top}_{gh})\big] = \nu^{12}_{gh,gh} + \kappa^{12}_{gh,gh},$$

we obtain the final expression.

(3). Expression of $\mathbf{s}_3$.

We show that $\mathbf{s}_3 = ((s_{31,gh}I(g = g'), \mathbf{s}_{32,gh}) : g \in [G_1], \mathcal{I}_{g'h} \in [G_1G_2])$ with $\mathcal{I}_{g'h} = (h-1)G_1 + g'$, where $s_{31,gh} = \nu^1_{gh,gh} + \kappa^1_{gh,gh}$ and $\mathbf{s}_{32,gh} = \mathbf{0}$. Since $\mathbf{M}^{(1\alpha)} = (\mathbf{M}^{(1\alpha)}_{g\mathcal{I}_{g'h}} : g \in [G_1], \mathcal{I}_{g'h} \in [G_1G_2])$ and $\mathbf{M}^{(1\alpha)}_{g\mathcal{I}_{g'h}} = \sum_t \mathbb{X}^{(1)\top}_{gh,t}\mathbb{Y}_{gh,(t-1)}I(g = g')$ with $\mathcal{I}_{g'h} = (h-1)G_1 + g'$, we derive the expression of $\lim_{n,T\to\infty}(n^2T)^{-1}\sum_t \mathbb{X}^{(1)\top}_{gh,t}\mathbb{Y}_{gh,(t-1)}$, and we have

$$\begin{aligned}
s_{31,gh} &= \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t E\{\mathbb{Y}^\top_{\cdot h,t}(\mathbf{I}_{N_{2h}} \otimes \mathbf{W}_1^{(\mathcal{R}^{(1)}_g,\cdot)})^\top\mathbb{Y}_{gh,(t-1)}\},\\
\mathbf{s}_{32,gh} &= \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t E\{(\mathbf{1}_{N_{2h}} \otimes (\mathbf{X}^{(1)}_t)^{(\mathcal{R}^{(1)}_g,\cdot)})^\top\mathbb{Y}_{gh,(t-1)}\}.
\end{aligned}$$

Since $\mathbb{Y}_{(t-1)}$ is independent with $\mathbf{X}^{(1)}_t$, we have $\mathbf{s}_{32,gh} = \mathbf{0}$. Note that

$$\begin{aligned}
s_{31,gh} &= \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t E\{\mathbb{Y}^\top_{\cdot h,(t-1)}(\mathbf{I}_{N_{2h}} \otimes \mathbf{W}_1^{(\mathcal{R}^{(1)}_g,\cdot)})^\top\mathbb{Y}_{gh,(t-1)}\}\\
&= \lim_{n,T\to\infty}(n^2T)^{-1}\sum_t\big[\mathrm{tr}(\mathcal{W}^{(1)}_{gh}\mathbf{\Sigma}^{\mathbf{X}}_{\cdot h,gh}) + \mathrm{tr}(\mathcal{W}^{(1)}_{gh}\mathbf{\Sigma}^{\mathcal{E}}_{\cdot h,gh})\big] = \nu^1_{gh,gh} + \kappa^1_{gh,gh}.
\end{aligned}$$

Therefore, we obtain the final expression.

(4). Expression of $\mathbf{s}_4$.

We next show that $\mathbf{s}_4 = (s_{4,gh} I(g = g', h = h') : \mathcal{I}_{gh} \in [G_1 G_2], \mathcal{I}_{g'h'} \in [G_1 G_2])$ with $s_{4,gh} = \nu^{\alpha}_{gh,gh} + \kappa^{\alpha}_{gh,gh}$. Since $\mathbf{M}^{\alpha} = (\mathbf{M}^{\alpha}_{\mathcal{I}_{gh}\mathcal{I}_{g'h'}} : \mathcal{I}_{gh} \in [G_1 G_2], \mathcal{I}_{g'h'} \in [G_1 G_2])$, where $\mathbf{M}^{\alpha}_{\mathcal{I}_{gh}\mathcal{I}_{g'h'}} = \sum_t \mathbb{Y}^{\top}_{gh,(t-1)} \mathbb{Y}_{gh,(t-1)} I(g = g', h = h')$, we derive $s_{4,gh} = \lim_{n,T\to\infty} (n^2 T)^{-1} \sum_t E(\sum_t \mathbb{Y}^{\top}_{gh,(t-1)} \mathbb{Y}_{gh,(t-1)})$.

$$
\begin{aligned}
s_{4,gh} &= \lim_{n,T\to\infty} (n^2 T)^{-1} \sum_t E(\mathbb{Y}^{\top}_{gh,(t-1)} \mathbb{Y}_{gh,(t-1)}) \\
&= \lim_{n,T\to\infty} (n^2 T)^{-1} \sum_t \left[\mathrm{tr}(\mathbf{\Sigma}^{\mathbf{X}}_{gh,gh}) + \mathrm{tr}(\mathbf{\Sigma}^{\mathcal{E}}_{gh,gh}) \right] = \nu^{\alpha}_{gh,gh} + \kappa^{\alpha}_{gh,gh}.
\end{aligned}
$$

Therefore, we obtain the final expression. ■

B.4 Additional Discussion of Weighted Least Squares Estimation with Group-specific Error Variances

B.4.1 Technical Assumption

We need the Assumption A.10 in Appendix B.4.1 instead of the Assumption 3 in Section 4.

Assumption A.10. (Distribution of Noise Term with Group-specific Variance) *Assume $\varepsilon_{ij,t}$ is a zero-mean sub-Gaussian variable, and independently distributed across $i \in [N_1], j \in [N_2]$ and $t \in [T]$ and independent with $\{\mathbf{Y}_s : s \le t-1\}$, $\{\mathbf{X}^{(1)}_s = (\mathbf{x}^{(1)}_{is} : i \in [N_1])^{\top} : s \le t\}$ and $\{\mathbf{X}^{(2)}_s = (\mathbf{x}^{(2)}_{js} : j \in [N_2])^{\top} : s \le t\}$. Assume that var$(\varepsilon_{ij,t}) = \sigma^2_{g_i^{(1)0} g_j^{(2)0}}$, where $g_i^{(1)0}$ and $g_j^{(2)0}$ are true group memberships. It holds $c_1 \le \min_{g^{(1)},g^{(2)}} \sigma^2_{g^{(1)}g^{(2)}} \le \max_{g^{(1)},g^{(2)}} \sigma^2_{g^{(1)}g^{(2)}} \le c_2$, where c_1 and c_2 are two finite positive constants.*

B.4.2 Proof of Theorem 9 (i)

Proof We prove the proposition in three steps. In step (1), we first show that $d(\mathbf{\Theta}, \mathbf{\Theta}^0) = O_p\{T^{-1}(\log(N_1 N_2)^2\}$ holds under the clustered error case. This is a required condition for the Theorem 5. In step (2), we show that when $G_1 \ge G_{1,0}$, $G_2 \ge G_{2,0}$ and $c_{\text{gap}} \gg d(\mathbf{\Theta}, \mathbf{\Theta}^0)$, the three conclusions in Theorem 5 still hold. In the final step (3), we come to the proposition.

Step (1). Order of $d(\mathbf{\Theta}, \mathbf{\Theta}^0)$.

We first show that Lemma 17 and Lemma 23 holds in step (A), and then show that Theorem 2 holds in step (B).

Step (A). Lemma 17 and Lemma 23.

First, we show that (A.66) and (A.67) hold in Lemma 17. To validate (A.66), it suffices to show (A.68)–(A.70) hold. By scrutinizing the three inequality, we only need to take care of (A.68) and (A.69), which are related to the error term. Rewrite the two inequality by

$$
P\Big(\Big| \frac{1}{T} \sum_t \varepsilon^2_{ij,t} - \sigma^2_{g_i^{(1)} \nu^2_{g_j^{(2)}}} \Big| > x/3 \Big) \le 2 \exp\{-c_1 T \min(x^2, x)\}, \tag{A.18}
$$

$$P\Big(\sup_{\|\boldsymbol{\xi}\|_{\max}<2R}\Big|\frac{2}{T}\sum_{t}\varepsilon_{ij,t}\mathcal{X}_{ij,t}^{\top}\boldsymbol{\xi}\Big|>x/3\Big)\leq\exp\Big\{-c_1\min(Tx^2,T^{1/2}x)+c_3m\Big\},\qquad(A.19)$$

where (A.18) could be obtained by Lemma 27. Since $E(\varepsilon_{ij,t})=0$, (A.19) could be proved following the similar procedure to prove (A.69). Hence, we have

$$P\Big(\sup_{\|\boldsymbol{\Theta}_{ij}\|_{\max}<R}\Big|\frac{1}{T}Q_{ij}(\boldsymbol{\Theta}_{ij})-\frac{1}{T}Q_{ij}^{*}(\boldsymbol{\Theta}_{ij})>x\Big|\Big)\leq\exp\Big\{-c_1\min(Tx^2,T^{1/2}x)+c_2m\Big\}$$

still holds for clustered error variance. Likewise, (A.67) could be validated similarly.

Next, we show that Lemma 23 holds. First note that Lemma 33 still holds under Assumption 4 and Assumption 5, and the two assumptions could still be satisfied if the error terms have group-specific variance. Hence, (A.85)–(A.87) in Lemma 23 are obtained when Assumption 2 holds.

Step (B). Theorem 2 holds.

By Step (A) we know that Lemma 17 and Lemma 23 hold, then we could borrow the idea of proof for Theorem 2, and show that $d(\boldsymbol{\Theta},\boldsymbol{\Theta}^0)=O_p\{T^{-1}(\log(N_1N_2)^2\}$.

Step (2). Proof of Theorem 5.

We prove the two statements in Lemma 25 first in Step (A)–(B), and then we prove the statement in Theorem 5 in Step (C).

Step (A). Proof of statement (i) of Lemma 25.

Since the definition of $\mathcal{N}_{\eta}^{(1)}$ and $\mathcal{A}_{\eta}(\boldsymbol{\xi},g^{(2)0},\mathcal{G}_{-1}^{0})$ are the same as the case under homogeneity, hence by the similar techniques in the proof of (i) of Theorem 5, one could keep the conclusion under heterogeneity case.

Step (B). Proof of statement (ii) of Lemma 25.

First, we show some required lemmas still hold. Note that in the clustered error case, the Assumption 2 for convexity and Assumption 6 still maintain. Since we have shown that Lemma 17 and Lemma 23 holds, we next show that Lemma 33 also holds. Because in the clustered error case, Assumption 4 and Assumption 5 are not affected, Lemma 33 could be proved similarly.

Then it comes to the proof for statement (ii). Similar as the proof procedure in Theorem 5, by using Lemma 23 and Assumption 2, one could have for any $g^{(2)}\notin\mathcal{A}_{\eta}(\boldsymbol{\theta},g_j^{(2)0})$,

$$\frac{1}{N_1T}Q_j^{*}(\boldsymbol{\xi}_{g^{(2)}}^{(2)};\boldsymbol{\xi}_{g^{(1)}}^{(1)},\widehat{\mathcal{G}}_1(\boldsymbol{\xi}))-\frac{1}{N_1T}Q_j^{*}(\boldsymbol{\xi}_{g_j^{(2)0}}^{(2)};\boldsymbol{\xi}_{g^{(1)0}}^{(1)0},\mathcal{G}_1^{0})\geq\tau_{\min}(c_{\text{gap}}c_{\pi}/2-\eta),\qquad(A.20)$$

when $N_1\to\infty$. On the other hand, for $\widetilde{g}_j^{(2)}\in\mathcal{A}_{\eta}^{(2)}(\boldsymbol{\xi},g_j^{(2)0},\mathcal{G}_1^{0})$, the following still holds,

$$\frac{1}{N_1T}Q_j^{*}(\boldsymbol{\xi}_{\widetilde{g}_j^{(2)}}^{(2)};\boldsymbol{\xi}_{g^{(1)}}^{(1)},\widehat{\mathcal{G}}_1(\boldsymbol{\xi}))-\frac{1}{N_1T}Q_j^{*}(\boldsymbol{\xi}_{g_j^{(2)0}}^{(2)};\boldsymbol{\xi}_{g^{(1)0}}^{(1)0},\mathcal{G}_1^{0})\leq\tau_{\max}\big\{d(\boldsymbol{\Theta},\boldsymbol{\Theta}^0)+\eta\big\}.\qquad(A.21)$$

Combining (A.20) and (A.21), and by the order of $d(\boldsymbol{\Theta},\boldsymbol{\Theta}^0)$ derived in Step (1), similar conclusion as in the proof of Theorem 5 could be obtained,

$$W_{jg^{(2)}}(\boldsymbol{\xi})\leq 2I\left(\sup_{i,j}\sup_{\|\boldsymbol{\Theta}_{ij}\|_{\max}<R}\Big|\frac{1}{T}Q_{ij}(\boldsymbol{\Theta}_{ij})-\frac{1}{T}Q_{ij}^{*}(\boldsymbol{\Theta}_{ij})\Big|\geq\epsilon_{\eta}/2\right).$$

Further by Step (1), the results could be obtained.

Step (C). Proof of statement (iii) of Lemma 25.

We first show that Lemma 24 and Proposition 4 hold, and then (iii) could be proved. Since Theorem 2 holds, one could use the same procedure to prove Lemma 24. To prove Proposition 4, we need a new assumption for the distribution of noise term, namely Assumption A.10. Together with the previously stated Lemma 17, Lemma 23, Lemma 24 and Theorem 2, Proposition 4 holds under the Assumption 1–2, 4–5 and A.10.

Then it comes to the proof of (iii) of Lemma 25. Similar as the proof of (iii) for Theorem 5, first it could be obtained that

$$\begin{aligned}&\max_{\widetilde{g}^{(2)}\in[G_2]}\min_{g^{(2)}\in[G_{2,0}]}\Big(\|\widehat{\boldsymbol{\theta}}^{c}_{\widetilde{g}^{(2)}}-\boldsymbol{\theta}^{(2)0}_{g^{(2)}}\|^2+\frac{1}{N_1}\sum_i|\widehat{\alpha}_{\widehat{g}_i^{(1)}\widetilde{g}^{(2)}}-\alpha^0_{g_i^{(1)0}g^{(2)}}|^2\Big)\\&=O_p(c_\pi^{-1}T^{-1}(\log(N_1N_2))^2)\end{aligned}$$

by Proposition 4, and due to Lemma 24, it also holds that

$$\begin{aligned}&\max_{g^{(2)}\in[G_2,0]}\min_{\widetilde{g}^{(2)}\in[G_2]}\Big(\|\widehat{\boldsymbol{\theta}}^{(2)}_{\widetilde{g}^{(2)}}-\boldsymbol{\theta}^{(2)0}_{g^{(2)}}\|^2+\frac{1}{N_1}\sum_i|\widehat{\alpha}_{\widehat{g}_I^{(1)}\widetilde{g}^{(2)}}-\alpha^0_{g_I^{(1)0}g^{(2)}}|^2\Big)\\&=O_p(c_\pi^{-1}T^{-1}(\log(N_1N_2))^2).\end{aligned}$$

Then we have $\widehat{\boldsymbol{\xi}}\in\mathcal{N}_\eta^{(2)}$ with $\eta=\tau_{\min}c_{\text{gap}}c_\pi/\{8(\tau_{\min}+\tau_{\max})\}$ under the condition that $c_{\text{gap}}\gg T^{-1}\log(N_1N_2)^2$. Then the statements (i) and (ii) in Lemma 25 hold, and then one could demonstrate Theorem 5 by using the conclusion Lemma 25. Other proof details are consistent with those in the i.i.d. error case.

Step (3). Membership Consistency.

Similarly as the proof for Corollary 6, using the statement in Theorem 5 proved in Step (2), the arguments (34) and (35) hold. ■

B.4.3 Proof of Theorem 9 (ii)

Proof We prove the result by two steps. In step (1), we first show that $|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|=o_p(1)$ when $\log(N_{1g^{(1)}}N_{2g^{(2)}})^2/T\to 0$. Then in step (2), we show that $\sup_{g^{(1)},g^{(2)}}|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|=o_p(1)$ under the condition $\sqrt{T}\gg\log(G_1G_2n^2)+m$.

Step (1). Order of $|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|$.

By Assumption A.10, and define the event $\mathcal{O}_{g^{(1)}g^{(2)}}=\{\widehat{\sigma}^2_{g^{(1)}g^{(2)}}\geq\sigma^2_{g^{(1)}g^{(2)}}/2\}=\{\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}\leq 2\sigma^{-2}_{g^{(1)}g^{(2)}}\leq 2c\}$, where c is a positive constant. Then, we have

$$\begin{aligned}P\Big\{|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|\geq t\Big\}&=P\Big\{\Big|\sigma^{-2}_{g^{(1)}g^{(2)}}(\sigma^2_{g^{(1)}g^{(2)}}-\widehat{\sigma}^2_{g^{(1)}g^{(2)}})\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}\Big|\geq t\Big\}\\&\leq P\Big\{\Big|\sigma^{-4}_{g^{(1)}g^{(2)}}(\sigma^2_{g^{(1)}g^{(2)}}-\widehat{\sigma}^2_{g^{(1)}g^{(2)}})\Big|\geq t/2\Big\}+P(\mathcal{O}^c_{g^{(1)}g^{(2)}})\\&\leq P\Big\{\Big|(\sigma^2_{g^{(1)}g^{(2)}}-\widehat{\sigma}^2_{g^{(1)}g^{(2)}})\Big|\geq\frac{t}{2c^2}\Big\}+P(\mathcal{O}^c_{g^{(1)}g^{(2)}}).\end{aligned}$$

Note that for the event $\mathcal{O}^c_{g^{(1)}g^{(2)}}$, we could calculate the probability as

$$\begin{aligned}P(\mathcal{O}^c_{g^{(1)}g^{(2)}}) = P\Big\{\widehat{\sigma}^2_{g^{(1)}g^{(2)}} < \frac{1}{2}\sigma^2_{g^{(1)}g^{(2)}}\Big\} &= P\Big\{\widehat{\sigma}^2_{g^{(1)}g^{(2)}} - \sigma^2_{g^{(1)}g^{(2)}} \le -\frac{1}{2}\sigma^2_{g^{(1)}g^{(2)}}\Big\} \\ &\le P\Big\{|\widehat{\sigma}^2_{g^{(1)}g^{(2)}} - \sigma^2_{g^{(1)}g^{(2)}}| \ge \frac{1}{2}\sigma^2_{g^{(1)}g^{(2)}}\Big\} \\ &\le P\Big\{|\widehat{\sigma}^2_{g^{(1)}g^{(2)}} - \sigma^2_{g^{(1)}g^{(2)}}| \ge \frac{1}{2c}\Big\}.\end{aligned}$$

Hence, we have

$$\begin{aligned}&P\Big\{|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}} - \sigma^{-2}_{g^{(1)}g^{(2)}}| \ge t\Big\} \\ &\le P\Big\{\Big|\sigma^2_{g^{(1)}g^{(2)}} - \widehat{\sigma}^2_{g^{(1)}g^{(2)}}\Big| \ge \frac{t}{2c^2}\Big\} + P\Big\{|\widehat{\sigma}^2_{g^{(1)}g^{(2)}} - \sigma^2_{g^{(1)}g^{(2)}}| \ge \frac{1}{2c}\Big\} \\ &\le P\Big\{\Big|\sigma^2_{g^{(1)}g^{(2)}} - \widehat{\sigma}^2_{g^{(1)}g^{(2)}}\Big| \ge \frac{t}{2c^2}\Big\} + P\Big\{|\widehat{\sigma}^2_{g^{(1)}g^{(2)}} - \sigma^2_{g^{(1)}g^{(2)}}| \ge \frac{t}{2c}\Big\} \\ &\le 2P\Big\{\Big|\sigma^2_{g^{(1)}g^{(2)}} - \widehat{\sigma}^2_{g^{(1)}g^{(2)}}\Big| \ge \frac{t}{C}\Big\}.\end{aligned}$$

when $0 < t < 1$ and $C = \max(2c^2, 2c)$. Note that

$$\begin{aligned}|\widehat{\sigma}^2_{g^{(1)}g^{(2)}} - \sigma^2_{g^{(1)}g^{(2)}}| &= (N_{1g^{(1)}}N_{2g^{(2)}}T)^{-1}\Big|\sum_{i\in\mathcal{R}^{(1)}_{g^{(1)}}}\sum_{j\in\mathcal{R}^{(2)}_{g^{(2)}}}\big\{Q_{ij}(\widehat{\boldsymbol{\Theta}}_{ij}) - Q^*_{ij}(\boldsymbol{\Theta}^0_{ij})\big\}\Big| \\ &\le (N_{1g^{(1)}}N_{2g^{(2)}}T)^{-1}\Big[\Big|\sum_{i\in\mathcal{R}^{(1)}_{g^{(1)}}}\sum_{j\in\mathcal{R}^{(2)}_{g^{(2)}}}\big\{Q_{ij}(\widehat{\boldsymbol{\Theta}}_{ij}) - Q^*_{ij}(\widehat{\boldsymbol{\Theta}}_{ij})\big\}\Big| \\ &\quad + \Big|\sum_{i\in\mathcal{R}^{(1)}_{g^{(1)}}}\sum_{j\in\mathcal{R}^{(2)}_{g^{(2)}}}\big\{Q^*_{ij}(\widehat{\boldsymbol{\Theta}}_{ij}) - Q^*_{ij}(\boldsymbol{\Theta}^0_{ij})\big\}\Big|\Big]\end{aligned}$$

Then we derive the order of the two terms on the right hands separately. First, by Lemma 17, we could directly obtain that

$$\begin{aligned}&(N_{1g^{(1)}}N_{2g^{(2)}}T)^{-1}\Big|\sum_{i\in\mathcal{R}^{(1)}_{g^{(1)}}}\sum_{j\in\mathcal{R}^{(2)}_{g^{(2)}}}\big\{Q_{ij}(\widehat{\boldsymbol{\Theta}}_{ij}) - Q^*_{ij}(\widehat{\boldsymbol{\Theta}}_{ij})\big\}\Big| \\ &\le (N_{1g^{(1)}}N_{2g^{(2)}}T)^{-1}\Big|\sum_{i\in\mathcal{R}^{(1)}_{g^{(1)}}}\sum_{j\in\mathcal{R}^{(2)}_{g^{(2)}}}\sup_{\|\boldsymbol{\Theta}\|_{\max}\le R}\big\{Q_{ij}(\boldsymbol{\Theta}) - Q^*_{ij}(\boldsymbol{\Theta})\big\}\Big| \\ &\le \sup_{i\in\mathcal{R}^{(1)}_{g^{(1)}}, j\in\mathcal{R}^{(2)}_{g^{(2)}}}\sup_{\|\boldsymbol{\Theta}\|_{\max}\le R}\big|T^{-1}\big\{Q_{ij}(\boldsymbol{\Theta}) - Q^*_{ij}(\boldsymbol{\Theta})\big\}\big|,\end{aligned}$$

and

$$\begin{aligned}&P\Big\{\sup_{i\in\mathcal{R}^{(1)}_{g^{(1)}}, j\in\mathcal{R}^{(2)}_{g^{(2)}}}\sup_{\|\boldsymbol{\Theta}\|_{\max}\le R}\big|T^{-1}\big\{Q_{ij}(\boldsymbol{\Theta}) - Q^*_{ij}(\boldsymbol{\Theta})\big\}\big| \ge t\Big\} \\ &\le N_{1g^{(1)}}N_{2g^{(2)}}\exp\big\{-c_1\min(Tt^2, \sqrt{T}t) + c_2 m\big\}.\end{aligned}$$

Then, for the second term, one could borrow the idea of the proof of Lemma 23 to obtain,

$$P\Big\{\frac{1}{N_{1g^{(1)}}N_{2g^{(2)}}T}\big|\sum_{i\in\mathcal{R}^{(1)}_{g^{(1)}}}\sum_{j\in\mathcal{R}^{(2)}_{g^{(2)}}}\{Q^*_{ij}(\widehat{\mathbf{\Theta}}_{ij})-Q^*_{ij}(\mathbf{\Theta}^0_{ij})\}\big|\geq t\Big\}$$
$$\leq N_{1g^{(1)}}N_{2g^{(2)}}\exp\big\{-c_1\min(Tt^2,\sqrt{T}t)+c_2m\big\}.$$

Hence, we have

$$P\Big\{|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|\geq t\Big\}\leq 2P\Big\{\Big|\sigma^2_{g^{(1)}g^{(2)}}-\widehat{\sigma}^2_{g^{(1)}g^{(2)}}\Big|\geq\frac{t}{C}\Big\}$$
$$\leq N_{1g^{(1)}}N_{2g^{(2)}}\exp\big\{-c_1\min(Tt^2,\sqrt{T}t)+c_2m\big\}.$$

Step (2). Order of $\sup_{g^{(1)}g^{(2)}}|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|$**.**

By step (1), we have

$$P\Big\{\sup_{g^{(1)}g^{(2)}}|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|\geq t\Big\}\leq G_1G_2n^2\exp\{-c_1\min(Tt^2,\sqrt{T}t)+c_2m\},$$

where n is defined in the statement of Theorem 9. This implies that $\sup_{g^{(1)}g^{(2)}}|\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}}-\sigma^{-2}_{g^{(1)}g^{(2)}}|=O_p\{(\log(G_1G_2n^2)+m)/\sqrt{T}\}$, and thus the proof is completed. ■

B.4.4 Proof of Theorem 9 (iii)

Proof Note that

$$nT^{1/2}(\widetilde{\boldsymbol{\theta}}^w-\boldsymbol{\theta}^0)=nT^{1/2}\widetilde{\mathbf{M}}^{-1}\breve{\boldsymbol{\delta}}=((n^2T)^{-1}\widetilde{\mathbf{M}})^{-1}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}}\overset{\text{def}}{=}\widetilde{\mathbf{M}}^{-1}_{nT}\mathbf{\Lambda}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}},$$

where $\breve{\boldsymbol{\delta}}=(\breve{\boldsymbol{\delta}}^{(1)\top},\breve{\boldsymbol{\delta}}^{(2)\top},\breve{\boldsymbol{\delta}}^{\alpha\top})^\top$ and

$$\breve{\boldsymbol{\delta}}^{(1)}_{g^{(1)}}=\sum_{t,g^{(2)}}(\widehat{\sigma}_{g^{(1)}g^{(2)}})^{-2}\mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)},t}\mathbb{E}_{g^{(1)}g^{(2)},t},\quad\breve{\boldsymbol{\delta}}^{(1)}=(\breve{\boldsymbol{\delta}}^{(1)\top}_g:g^{(1)}\in[G_1])^\top,$$

$$\breve{\boldsymbol{\delta}}^{(2)}_{g^{(2)}}=\sum_{t,g^{(1)}}(\widehat{\sigma}_{g^{(1)}g^{(2)}})^{-2}\mathbb{X}^{(2)\top}_{g^{(1)}g^{(2)},t}\mathbb{E}_{g^{(1)}g^{(2)},t},\quad\breve{\boldsymbol{\delta}}^{(2)}=(\breve{\boldsymbol{\delta}}^{(2)\top}_{g^{(2)}}:g^{(2)}\in[G_2])^\top,$$

$$\breve{\boldsymbol{\delta}}^{\alpha}_{\mathcal{I}_{g^{(1)}g^{(2)}}}=\sum_t(\widehat{\sigma}_{g^{(1)}g^{(2)}})^{-2}\mathbb{Y}^\top_{g^{(1)}g^{(2)}(t-1)}\mathbb{E}_{g^{(1)}g^{(2)},t},\quad\breve{\boldsymbol{\delta}}^{\alpha}=(\breve{\boldsymbol{\delta}}^{\alpha\top}_{\mathcal{I}_{g^{(1)}g^{(2)}}}:\mathcal{I}_{g^{(1)}g^{(2)}}\in[g^{(1)}g^{(2)}])^\top.$$

Then, we have

$$nT^{1/2}\boldsymbol{\eta}^\top(\widetilde{\boldsymbol{\theta}}^w-\boldsymbol{\theta}^0)=\boldsymbol{\eta}^\top\widetilde{\mathbf{M}}^{-1}_{nT}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}}$$
$$=\boldsymbol{\eta}^\top\{\widetilde{\mathbf{M}}^{-1}_{nT}-(\widetilde{\mathbf{M}}^0_{nT})^{-1}\}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}}+\boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}}$$
$$=\boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}}+\boldsymbol{\eta}^\top\widetilde{\mathbf{M}}^{-1}_{nT}\{\widetilde{\mathbf{M}}^0_{nT}-\widetilde{\mathbf{M}}_{nT}\}(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}}$$

$$= \boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}\widetilde{\boldsymbol{\delta}} + \boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}(\breve{\boldsymbol{\delta}} - \widetilde{\boldsymbol{\delta}})$$
$$+ \boldsymbol{\eta}^\top\widetilde{\mathbf{M}}^{-1}_{nT}\{\widetilde{\mathbf{M}}^0_{nT} - \widetilde{\mathbf{M}}_{nT}\}(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}}.$$

Hence, it suffices to show that

$$\boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}\widetilde{\boldsymbol{\delta}} \to_d N(0, \boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0)^{-1}\boldsymbol{\eta}), \tag{A.22}$$

$$\boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}(\breve{\boldsymbol{\delta}} - \widetilde{\boldsymbol{\delta}}) = o_p(1), \tag{A.23}$$

$$\boldsymbol{\eta}^\top\widetilde{\mathbf{M}}^{-1}_{nT}\{\widetilde{\mathbf{M}}^0_{nT} - \widetilde{\mathbf{M}}_{nT}\}(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}\breve{\boldsymbol{\delta}} = o_p(1). \tag{A.24}$$

While it is obvious that (A.24) holds when $n, T \to \infty$, we prove the (A.22)–(A.23) in the following two steps.

1. Proof of (A.22)

By noting that (A.22) and (A.56) have similar forms, one could borrow the idea from the proof for (A.56) with the new Assumption A.10. Then we could obtain the similar results by using the central limit theorem of the martingale difference array.

2. Proof of (A.23)

Note that $|\boldsymbol{\eta}^\top(\widetilde{\mathbf{M}}^0_{nT})^{-1}(nT^{1/2})^{-1}(\breve{\boldsymbol{\delta}}-\widetilde{\boldsymbol{\delta}})| \le \|(\widetilde{\mathbf{M}}^0_{nT})^{-1}\boldsymbol{\eta}\|\|(nT^{1/2})^{-1}(\breve{\boldsymbol{\delta}}-\widetilde{\boldsymbol{\delta}})\|$, and we have assumed that $\lambda_{\min}(\mathbf{M}^0) > 0$ and $\max_{g^{(1)}g^{(2)}} \sigma^2_{g^{(1)}g^{(2)}} < \infty$, hence it holds that $\|(\widetilde{\mathbf{M}}^0_{nT})^{-1}\boldsymbol{\eta}\| < c < \infty$. Then to derive the order of $\|(nT^{1/2})^{-1}(\breve{\boldsymbol{\delta}}-\widetilde{\boldsymbol{\delta}})\|$, we focus on the order of $(n^2T)^{-1}(\breve{\boldsymbol{\delta}}-\widetilde{\boldsymbol{\delta}})^\top(\breve{\boldsymbol{\delta}}-\widetilde{\boldsymbol{\delta}})$. We take the term $\breve{\boldsymbol{\delta}}^{(1)} \in \mathbb{R}^{G\times(p_1+1)}$ and $\widetilde{\boldsymbol{\delta}}^{(1)} \in \mathbb{R}^{G\times(p_1+1)}$ for example to illustrate. Note that

$$\begin{aligned}
&(n^2T)^{-1}(\breve{\boldsymbol{\delta}}^{(1)} - \widetilde{\boldsymbol{\delta}}^{(1)})^\top(\breve{\boldsymbol{\delta}}^{(1)} - \widetilde{\boldsymbol{\delta}}^{(1)}) \\
&= (n^2T)^{-1}\sum_{g^{(1)},g^{(2)},t}(\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}} - \sigma^{-2}_{g^{(1)}g^{(2)}})^2\mathbb{E}^\top_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)}_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)},t}\mathbb{E}_{g^{(1)}g^{(2)},t} \\
&\le (n^2T)^{-1}\sum_{g^{(1)},g^{(2)},t}\Big\{\sup_{g^{(1)}g^{(2)}}(\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}} - \sigma^{-2}_{g^{(1)}g^{(2)}})^2\Big\}\mathbb{E}^\top_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)}_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)},t}\mathbb{E}_{g^{(1)}g^{(2)},t}.
\end{aligned}$$

By Theorem 9 (ii), one has $\sup_{g^{(1)}g^{(2)}}(\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}} - \sigma^{-2}_{g^{(1)}g^{(2)}})^2 = O_p\{T^{-1}(m + \log(G_1G_2n^2)\}$, hence we next derive the order of $(n^2T)^{-1}\sum_{g^{(1)},g^{(2)},t} E(\mathbb{E}^\top_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)}_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)},t}\mathbb{E}_{g^{(1)}g^{(2)},t})$.

By Assumption A.10 and $\max_{g^{(1)}g^{(2)}}\sigma_{g^{(1)}g^{(2)}} \le c$, we know that with $\boldsymbol{\eta}_1 \in \mathbb{R}^{p+1}$ and $\|\boldsymbol{\eta}_1\| = 1$,

$$\begin{aligned}
&(n^2T)^{-1}\sum_{g^{(1)},g^{(2)},t} E(\mathbb{E}^\top_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)}_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)},t}\mathbb{E}_{g^{(1)}g^{(2)},t}) \\
&= (n^2T)^{-1}\sum_{g^{(1)},g^{(2)},t} E\{\sigma^2_{g^{(1)}g^{(2)}}\mathrm{tr}(\mathbb{X}^{(1)}_{g^{(1)}g^{(2)},t}\mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)},t})\} \\
&\le c^2(n^2T)^{-1}\sum_{g^{(1)},g^{(2)},t} E\{\mathrm{tr}\big(\mathbf{W}_1^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\mathbf{Y}_t^{(\cdot,\mathcal{R}^{(2)}_{g^{(2)}})}\mathbf{Y}_t^{(\cdot,\mathcal{R}^{(2)}_{g^{(2)}})\top}\mathbf{W}_1^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)\top}\big)\}
\end{aligned}$$

$$+ c^2(n^2T)^{-1} \sum_{g^{(1)},g^{(2)},t} E\{\mathrm{vec}(\mathbf{1}_{N_{2g^{(2)}}} \otimes (\mathbf{X}_t^{(1)})^{(\mathcal{R}^{(1)}_{g^{(1)},\cdot})})^\top \mathrm{vec}(\mathbf{1}_{N_{2g^{(2)}}} \otimes (\mathbf{X}_t^{(1)})^{(\mathcal{R}^{(1)}_{g^{(1)},\cdot})})\}$$
$$= c^2(n^2T)^{-1} \sum_t E\{\mathrm{tr}\big(\mathbf{W}_1 \mathbf{Y}_t \mathbf{Y}_t^\top \mathbf{W}_1^\top\big)\} + c^2 N_2 (n^2T)^{-1} \sum_t \sum_{i\in[N_1]} E\|\mathbf{x}_{it}^{(1)}\|^2.$$

Note that

$$\sum_t E\{\mathrm{tr}\big(\mathbf{W}_1 \mathbf{Y}_t \mathbf{Y}_t^\top \mathbf{W}_1^\top\big)\} \le c_1 \sum_t \mathrm{tr}\big(\mathbf{W}_1 \mathbf{1}_{N_1} \mathbf{1}_{N_2}^\top \mathbf{1}_{N_2} \mathbf{1}_{N_1}^\top \mathbf{W}_1^\top\big) = O(n^2T),$$

where $c_1 = E(Y_{ij,t}^2) < \infty$ by Lemma 32, and also note that

$$\sum_t \sum_{i\in[N_1]} E\|\mathbf{x}_{it}^{(1)}\|^2 \le c_K p_1 T N_1 = O(nTs)$$

due to the K-convexity property of $\mathbf{x}_{it}^{(1)}$ and the definition of n, we have that

$$(n^2T)^{-1} \sum_{g^{(1)},g^{(2)},t} (\mathbb{E}^\top_{g^{(1)}g^{(2)},t} \mathbb{X}^{(1)}_{g^{(1)}g^{(2)},t} \mathbb{X}^{(1)\top}_{g^{(1)}g^{(2)},t} \mathbb{E}_{g^{(1)}g^{(2)},t}) = O_p(1).$$

Together with $\sup_{g^{(1)}g^{(2)}} (\widehat{\sigma}^{-2}_{g^{(1)}g^{(2)}} - \sigma^{-2}_{g^{(1)}g^{(2)}})^2 = O_p\{T^{-1}(m + \log(GHn^2)\}$, we know that

$$(n^2T)^{-1} (\breve{\boldsymbol{\delta}}^{(1)} - \widetilde{\boldsymbol{\delta}}^{(1)})^\top (\breve{\boldsymbol{\delta}}^{(1)} - \widetilde{\boldsymbol{\delta}}^{(1)}) = O_p\{(m + \log(G_1G_2n^2))/T\} = o_p(1)$$

when $T \gg (\log(G_1G_2n^2))$ due to that m is a fixed constant.

By step 1–2, we could arrive at the final conclusion.

■

B.4.5 Simulation Results of Weighted Least Squares Estimation

In this section, we conduct the simulation experiment when $G_{1,0} = G_{2,0} = 3$. We set the group-specific error variances as $\widetilde{\boldsymbol{\Sigma}} = (\sigma_{g^{(1)}g^{(2)}} : g^{(1)} \in [G_{1,0}], g^{(2)} \in [G_{2,0}]) \in \mathbb{R}^{3\times 3}$ as

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

Then we set $(N_1, N_2) \in \{(100, 80), (200, 150)\}$ and set $T \in \{20, 40, 80\}$. Next, we repeat the experiment for $R = 300$ times, and show the RMSEs results of $\widetilde{\boldsymbol{\theta}}^w$ in Table A.5. From the table, we first focus on the estimation results under weighted least squares method (marked as "WLS" in the table). We could see that when the error has group-specific variances, our proposed estimation could estimated the true group memberships with high accuracy as the sample size grows. Furthermore, we note that RMSEs of all parameters decrease toward the corresponding oracle results when the sample sizes are large. For the inference results of WLS, when both N_1, N_2, T are large, the CPs are all around 0.95. Then, we compare the

results of WLS and the ordinary method in the main text (marked as "OLS"). One could see that the RMSEs of WLS estimator are always smaller than those of OLS estimator, and the RMSEs of the corresponding oracle estimators have the same pattern. This shows that when the error has group-specific variances, our proposed weighted least squares estimator outperforms the ordinary least squares estimator.

Appendix C. Specific Expressions of M in Equation (13)

$$\mathbf{M}^{(l)}_{g^{(l)}} = \sum_{t,g^{-(l)}} \mathbb{X}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t} \mathbb{X}^{(l)}_{g^{(1)}\cdots g^{(q)},t}, \quad \mathbf{M}^{(l)} = \operatorname{diag}\big(\mathbf{M}^{(l)}_{g^{(l)}} : g^{(l)} \in [G_l]\big) \in \mathbb{R}^{G_l(p_l+1)\times G_l(p_l+1)},$$

$$\mathbf{M}^{(lm)}_{g^{(l)}g^{(m)}} = \sum_{t,g^{-(l,m)}} \mathbb{X}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t} \mathbb{X}^{(m)}_{g^{(1)}\cdots g^{(q)},t},$$

$$\mathbf{M}^{(lm)} = \big(\mathbf{M}^{(lm)}_{g^{(l)}g^{(m)}} : g^{(l)} \in [G_l], g^{(m)} \in [G_m]\big) \in \mathbb{R}^{G_l(p_l+1)\times G_m(p_m+1)},$$

$$\mathbf{M}^{(l\alpha)}_{g^{(l)}\mathcal{I}_{g^{(l)\prime}g^{-(l)}}} = \sum_t \mathbb{X}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t} \mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)} I(g^{(l)} = g^{(l)\prime}),$$

$$\text{with } \mathcal{I}_{g^{(l)\prime}g^{-(l)}} = g^{(l)\prime} + \sum_{m\neq l}((g^{(m)}-1)G_l),$$

$$\mathbf{M}^{(l\alpha)} = \Big(\mathbf{M}^{(l\alpha)}_{g^{(l)}\mathcal{I}_{g^{(l)\prime}g^{-(l)}}} : g^{(l)} \in [G_l], \mathcal{I}_{g^{(l)\prime}g^{-(l)}} \in \Big[\prod_l G_l\Big]\Big) \in \mathbb{R}^{G_l(p_l+1)\times(\prod_l G_l)},$$

$$\mathbf{M}^{\alpha}_{\mathcal{I}_{g^{(1)},\cdots,g^{(q)}},\mathcal{I}_{g^{(1)\prime},\cdots g^{(q)\prime}}} = \sum_t \|\mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)}\|^2 I\{(g^{(1)},\cdots,g^{(q)}) = (g^{(1)\prime},\cdots,g^{(q)\prime})\},$$

$$\mathbf{M}^{\alpha} = \Big(\mathbf{M}^{\alpha}_{\mathcal{I}_{g^{(1)},\cdots,g^{(q)}},\mathcal{I}_{g^{(1)\prime},\cdots g^{(q)\prime}}} : \mathcal{I}_{g^{(1)},\cdots,g^{(q)}} \in \Big[\prod_l G_l\Big], \mathcal{I}_{g^{(1)\prime},\cdots g^{(q)\prime}} \in \Big[\prod_l G_l\Big]\Big)$$

$$\in \mathbb{R}^{(\prod_l G_l)\times(\prod_l G_l)},$$

$$\mathbf{b}^{(l)}_{g^{(l)}} = \sum_{t,g^{-(l)}} \mathbb{X}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t} \mathbb{Y}_{g^{(1)}\cdots g^{(q)},t}, \quad \mathbf{b}^{(l)} = \big(\mathbf{b}^{(l)}_{g^{(l)}} : g^{(l)} \in [G_l]\big) \in \mathbb{R}^{G_l(p_l+1)},$$

$$\mathbf{b}^{\alpha}_{\mathcal{I}_{g^{(1)},\cdots,g^{(q)}}} = \sum_t \mathbb{Y}^{\top}_{g^{(1)}\cdots g^{(q)},(t-1)} \mathbb{Y}_{g^{(1)}\cdots g^{(q)},t},$$

$$\mathbf{b}^{\alpha} = \Big(\mathbf{b}^{\alpha}_{\mathcal{I}_{g^{(1)},\cdots,g^{(q)}}} : \mathcal{I}_{g^{(1)},\cdots,g^{(q)}} \in \Big[\prod_l G_l\Big]\Big) \in \mathbb{R}^{\Pi_l G_l},$$

where $\sum_{t,g^{-(l,m)}}$ is the simplified notation for $\sum_{\{g^{(1)},\cdots,g^{(q)}\}\setminus\{g^{(l)},g^{(m)}\}}$.

Appendix D. GTNAR Algorithm

We summarize the GTNAR algorithm in the following Algorithm A.2.

Appendix E. Local Convergence of the Numerical Algorithm

In this section, we would show that our estimation Algorithm A.2 is convergent, and its solution is indeed a local minimizer of (9). We first show the convergence of the algorithm in (1), followed by the local minimization proof in (2).

Table A.5: RMSEs (×1000) of estimated parameters under scenario 3 ($G_{1,0} = 3, G_{2,0} = 3$) with 300 replications. The performances are evaluated for different sample sizes N_1, N_2 and the time length T. "OLS" means the results estimated using the ordinary least squares method proposed in Section 3, "WLS" means the results using weighted least squares proposed in Appendix B.4. The corresponding CPs are shown in the parenthesis.

N_1	N_2	T	Estimation	$\widehat{\boldsymbol{\lambda}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}^{(2)}$	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\lambda}}^{(1)\text{or}}$	$\widehat{\boldsymbol{\lambda}}^{(2)\text{or}}$	$\widehat{\boldsymbol{\zeta}}^{(1)\text{or}}$	$\widehat{\boldsymbol{\zeta}}^{(2)\text{or}}$	$\widehat{\boldsymbol{\alpha}}^{\text{or}}$	η_1	η_2
100	80	20	OLS	37.34	38.54	128.98	107.75	104.58	22.8	16.12	70.85	72.94	22.17	0.3025	0.0930
				(0.671)	(0.637)	(0.746)	(0.822)	(0.515)	(0.864)	(0.919)	(0.944)	(0.924)	(0.914)		
			WLS	36.64	38.57	99.58	106.13	103.84	15.61	14.92	40.94	69.28	21.87		
				(0.657)	(0.676)	(0.541)	(0.87)	(0.462)	(0.966)	(0.949)	(0.948)	(0.952)	(0.945)		
		40	OLS	24.79	12.47	87.59	52.22	46.10	15.10	11.18	51.11	50.26	14.68	0.1960	0.0077
				(0.71)	(0.919)	(0.774)	(0.93)	(0.696)	(0.883)	(0.93)	(0.945)	(0.936)	(0.936)		
			WLS	24.45	12.02	68.28	49.96	45.45	11.20	10.23	28.77	47.34	14.53		
				(0.673)	(0.928)	(0.576)	(0.951)	(0.635)	(0.952)	(0.947)	(0.947)	(0.954)	(0.955)		
		80	OLS	15.86	8.31	53.54	33.93	22.80	9.39	8.30	39.10	33.94	10.71	0.0877	0.0000
				(0.774)	(0.949)	(0.87)	(0.952)	(0.861)	(0.909)	(0.952)	(0.952)	(0.952)	(0.95)		
			WLS	14.50	7.61	39.13	32.16	22.56	7.24	7.38	22.81	31.87	10.63		
				(0.783)	(0.949)	(0.765)	(0.947)	(0.81)	(0.948)	(0.959)	(0.946)	(0.946)	(0.949)		
200	150	20	OLS	23.50	12.61	87.96	41.33	52.51	11.6	8.55	36.19	38.21	10.66	0.2564	0.0189
				(0.584)	(0.818)	(0.616)	(0.904)	(0.562)	(0.876)	(0.912)	(0.947)	(0.929)	(0.924)		
			WLS	25.15	12.5	79.29	40.03	51.98	8.47	8.09	20.16	36.66	10.63		
				(0.524)	(0.827)	(0.409)	(0.94)	(0.47)	(0.95)	(0.933)	(0.957)	(0.95)	(0.954)		
		40	OLS	12.26	5.48	40.14	26.10	16.32	7.85	5.48	26.85	26.10	7.59	0.1033	0.0000
				(0.706)	(0.928)	(0.84)	(0.929)	(0.79)	(0.867)	(0.93)	(0.945)	(0.928)	(0.933)		
			WLS	11.06	5.09	32.77	25.12	16.07	5.66	4.98	15.58	24.89	7.54		
				(0.946)	(0.942)	(0.673)	(0.949)	(0.74)	(0.95)	(0.94)	(0.952)	(0.949)	(0.95)		
		80	OLS	5.89	4.21	17.95	19.14	5.49	5.80	4.21	17.71	19.14	5.29	0.0078	0.0000
				(0.863)	(0.910)	(0.944)	(0.920)	(0.919)	(0.864)	(0.910)	(0.949)	(0.920)	(0.922)		
			WLS	4.12	3.90	10.83	18.19	5.42	3.99	3.89	10.46	18.17	5.23		
				(0.950)	(0.949)	(0.939)	(0.948)	(0.944)	(0.952)	(0.944)	(0.954)	(0.948)	(0.951)		

Algorithm A.2 Estimation of the GTNAR Model

1: **Input:** $\{\mathcal{Y}_t, \mathbf{X}_t^{(l)}, \mathbf{W}^{(l)}, G_l\}$ for $1 \le l \le q$.

2: Obtain initial group memberships $\mathcal{G}_l^{[0]}$ according to Algorithm A.3 in Appendix F. Let $\{\boldsymbol{\xi}^{[k]}, \mathcal{G}_l^{[k]}\}$ be the estimators and memberships in the kth iteration.

3: Repeat Step 1 and Step 2 for $k = 1, 2, \cdots$ until convergence.

Step 1. Given $\{\mathcal{G}_l^{[k-1]}\}$ for all layers $l \in [q]$, calculate $\boldsymbol{\xi}^{[k-1]} = (\mathbf{M}^{[k-1]})^{-1}\mathbf{b}^{[k-1]}$, where $\mathbf{M}^{[k-1]}$ and $\mathbf{b}^{[k-1]}$ are obtained from (13) with $\mathcal{G}_l^{[k-1]}$s specified.

Step 2. Given $\boldsymbol{\xi}^{[k-1]}$, sequentially update memberships $\mathcal{G}_l^{[k]} = (g_{i_l}^{(l)[k]} : 1 \le i_l \le N_l)^\top$ for $1 \le l \le q$ as follows,

$$
\begin{aligned}
g_{i_l}^{(l)[k]} = \arg \min_{g_{i_l}^{(l)} \in [G_l]} \sum_{i_{-l}} \sum_{t=1}^{T} \Big\{ & Y_{i_1 i_2 \ldots i_q, t} - \Big(\sum_{l'=1}^{l-1} \lambda_{g_{i_{l'}}^{(l')[k]}}^{(l')[k]} \sum_{m=1}^{N_{l'}} \frac{a_{i_{l'}m}^{(l')}}{n_{l' i_{l'}}} Y_{i_1 \ldots i_{l'-1} m i_{l'+1} \ldots i_q, (t-1)} \\
& + \lambda_{g_{i_l}^{(l)}}^{(l)[k]} \sum_{m=1}^{N_l} \frac{a_{i_l m}^{(l)}}{n_{l i_l}} Y_{i_1 \ldots i_{l-1} m i_{l+1} \ldots i_q, (t-1)} + \sum_{l''=l+1}^{q} \lambda_{g_{i_{l''}}^{(l'')[k-1]}}^{(l'')[k]} \sum_{m=1}^{N_{l''}} \frac{a_{i_{l''}m}^{(l'')}}{n_{l'' i_{l''}}} Y_{i_1 \ldots i_{l''-1} m i_{l''+1} \ldots i_q, (t-1)} \Big) \\
& - \alpha_{g_{i_1}^{(1)[k]} \cdots g_{i_{l-1}}^{(l-1)[k]} g_{i_l}^{(l)} g_{i_{l+1}}^{(l+1)[k-1]} \cdots g_{i_q}^{(q)[k-1]}}^{[k]} Y_{i_1 i_2 \ldots i_q, (t-1)} \\
& - \Big(\sum_{l'=1}^{l-1} \mathbf{x}_{i_l' t}^{(l')\top} \boldsymbol{\zeta}_{g_{i_l'}^{(l')[k]}}^{(l')[k]} + \mathbf{x}_{i_l t}^{(l)\top} \boldsymbol{\zeta}_{g_{i_l}^{(l)}}^{(l)[k]} + \sum_{l''=l+1}^{q} \mathbf{x}_{i_l'' t}^{(l'')\top} \boldsymbol{\zeta}_{g_{i_l''}^{(l'')[k-1]}}^{(l'')[k]} \Big) \Big\}^2
\end{aligned} \tag{A.25}
$$

4: **Output:** Final estimator and memberships: $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}^{[K]}$ and $\widehat{\mathcal{G}} = \{\mathcal{G}_l^{[K]} : l \in [q]\}$. Here K is the final number of iteration rounds.

(1) Convergence of the algorithm.

We show that the objective function defined in (9) is monotonically decreasing and hence could obtain the convergence. Denote the estimators in the kth iteration as $(\boldsymbol{\xi}^{(k)}, \mathcal{G}^{(k)})$. Then we have that

$$\begin{aligned} Q(\boldsymbol{\xi}^{(k+1)}, \mathcal{G}^{(k+1)}) &= \min_{\boldsymbol{\xi}} Q(\boldsymbol{\xi}, \mathcal{G}^{(k+1)}) \leq Q(\boldsymbol{\xi}^{(k)}, \mathcal{G}^{(k+1)}) \\ &\overset{①}{\leq} Q(\boldsymbol{\xi}^{(k)}, \{\mathcal{G}_1^{(k)}, \mathcal{G}_2^{(k+1)}, \cdots, \mathcal{G}_q^{(k+1)}\}) \\ &\overset{②}{\leq} \cdots\cdots \\ &\overset{③}{\leq} Q(\boldsymbol{\xi}^{(k)}, \{\mathcal{G}_1^{(k)}, \mathcal{G}_2^{(k)}, \cdots, \mathcal{G}_q^{(k)}\}) = Q(\boldsymbol{\xi}^{(k)}, \mathcal{G}^{(k)}), \end{aligned}$$

where the inequalities ①–③ are derived by

$$\mathcal{G}_l^{(k+1)} = \text{argmin}_{\mathcal{G}_l} Q(\boldsymbol{\xi}^{(k)}, \{\mathcal{G}_1^{(k+1)}, \cdots, \mathcal{G}_{q-1}^{(k+1)}, \mathcal{G}_l, \mathcal{G}_{l+1}^{(k)}, \cdots, \mathcal{G}_q^{(k)}\})$$

in our updating mechanism (14). Then, we have that

$$Q(\boldsymbol{\xi}^{(0)}, \mathcal{G}^{(0)}) \geq Q(\boldsymbol{\xi}^{(1)}, \mathcal{G}^{(1)}) \geq \cdots \geq Q(\boldsymbol{\xi}^{(k^*)}, \mathcal{G}^{(k^*)}) = Q(\boldsymbol{\xi}^{(k^*+1)}, \mathcal{G}^{(k^*+1)})$$

Consequently, the estimators could be obtained by $Q(\boldsymbol{\xi}^{(k^*)}, \mathcal{G}^{(k^*)})$.

(2) Local Minimality.

Next, we prove that the solution $(\boldsymbol{\xi}^{(k^*)}, \mathcal{G}^{(k^*)})$ is the local minimum of the function $\boldsymbol{\xi} \overset{\text{def}}{=} \min_{\mathcal{G}} Q(\boldsymbol{\xi}, \mathcal{G})$. Define $\widetilde{\mathcal{G}} \overset{\text{def}}{=} \text{argmin}_{\mathcal{G}} Q(\boldsymbol{\xi}^{(k^*)} + \boldsymbol{\delta}, \mathcal{G})$, where $\boldsymbol{\delta}$ is a small perturbation. When $\|\boldsymbol{\delta}\|$ is small enough, we have that $\widetilde{\mathcal{G}} = \mathcal{G}^{(k^*)}$ due to the group memberships' discreteness. Then, we have

$$\begin{aligned} &\min_{\mathcal{G}} Q(\boldsymbol{\xi}^{(k^*)} + \boldsymbol{\delta}, \mathcal{G}) = Q(\boldsymbol{\xi}^{(k^*)} + \boldsymbol{\delta}, \widetilde{\mathcal{G}}) \\ &= Q(\boldsymbol{\xi}^{(k^*)} + \boldsymbol{\delta}, \mathcal{G}^{(k^*)}) \geq Q(\boldsymbol{\xi}^{(k^*)}, \mathcal{G}^{(k^*)}) = \min_{\mathcal{G}} Q(\boldsymbol{\xi}^{(k^*)}, \mathcal{G}). \end{aligned}$$

This completes the proof.

Appendix F. Initialization

We provide the initialization procedure for the group memberships $\mathcal{G}_l$ in the Algorithm A.3.

Appendix G. Additional Discussion of Mixed GTNAR Model (24)

G.1 Estimation Procedure

The element-wise model form can be given as

$$Y_{ij,t} = \gamma^{(1)}_{g_i^{(1)}} \gamma^{(2)}_{g_j^{(2)}} \sum_k \sum_m \frac{a^{(1)}_{im}}{n_{1i}} Y_{mk,t-1} \frac{a^{(2)}_{kj}}{n_{2j}}$$

Algorithm A.3 Initialization of the GTNAR Model

1: **Input:** $\{\mathcal{Y}_t, \mathbf{X}_t^{(l)}, \mathbf{W}^{(l)}, G_l\}$ for $1 \le l \le q$.

2: Treat each node as a group for all dimensions $l \in [q]$, estimate $\widehat{\boldsymbol{\theta}}^{(l)}$ and $\widehat{\boldsymbol{\alpha}}$ by (13). Here $\widehat{\boldsymbol{\theta}}^{(l)} \in \mathbb{R}^{N_l(p_l+1)}$ and $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^{N_1 \times \cdots \times N_q}$.

3: Run the k-means clustering for the above estimators for $T_{\text{init}} = 3$ trials. For each trial $t = 1, \cdots, T_{\text{init}}$, try the following two clustering types:

Type 1. (Clustering by time effect)

Step 1. Clustering the nodes in the lth dimension by the mode-l matricization of *self-driven time effect* $\widehat{\boldsymbol{\alpha}}_{(l)} \in \mathbb{R}^{N_l \times (\prod_{m \neq l} N_m)}$. Then one obtain the $\mathcal{G}_l^{1[t]}$ for $1 \le l \le q$.

Step 2. Calculate the loss in the tth trial for type 1 by $Q^{[t]}(\widehat{\boldsymbol{\xi}}, \mathcal{G}^{1[t]})$, where $\mathcal{G}^{1[t]} = \{\mathcal{G}_l^{1[t]} : l \in [q]\}$ is the first type initial group memberships.

Type 2. (Clustering by network and covariate effect)

Step 1. Clustering by *network effects* and *covariate effect* on $\widehat{\boldsymbol{\theta}}^{(l)} = (\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)}} : g^{(l)} \in [G_l])$. Then one obtain the $\mathcal{G}_l^{2[t]}$ for $1 \le l \le q$.

Step 2. Calculate the loss in the tth trial for type 2 by $Q^{[t]}(\widehat{\boldsymbol{\xi}}, \mathcal{G}^{2[t]})$, where $\mathcal{G}^{2[t]} = \{\mathcal{G}_l^{2[t]} : l \in [q]\}$ is the second type initial group memberships.

4: Select the best initial trial $t^* = \text{argmin}_t[\min\{Q^{[t]}(\widehat{\boldsymbol{\xi}}, \mathcal{G}^{1[t]}), Q^{[t]}(\widehat{\boldsymbol{\xi}}, \mathcal{G}^{2[t]})\}]$, and the corresponding initial memberships are denoted as $\mathcal{G}^{[0]}$.

5: **Output:** Best initial memberships: $\mathcal{G}^{[0]}$.

$$+ \Big(\lambda^{(1)}_{g_i^{(1)}} \sum_k \frac{a_{ik}^{(l)}}{n_{1i}} Y_{kj,t-1} + \lambda^{(2)}_{g_j^{(2)}} \sum_k \frac{a_{kj}^{(2)}}{n_{2j}} Y_{ik,t-1}\Big)$$
$$+ \alpha_{g_i^{(1)} g_j^{(2)}} Y_{ij,(t-1)} + \Big(\mathbf{x}_{it}^{(1)\top} \boldsymbol{\zeta}^{(1)}_{g_i^{(1)}} + \mathbf{x}_{jt}^{(2)\top} \boldsymbol{\zeta}^{(2)}_{g_j^{(2)}}\Big) + \varepsilon_{ij,t}.$$

For simplicity, denote the interactive parameters as $\boldsymbol{\psi} = (\boldsymbol{\psi}^{(1)\top}, \boldsymbol{\psi}^{(2)\top})^\top \in \mathbb{R}^{G_1+G_2}$ with $\boldsymbol{\psi}^{(1)} = (\gamma_1^{(1)}, \cdots, \gamma_{G_1}^{(1)})^\top \in \mathbb{R}^{G_1}$ and $\boldsymbol{\psi}^{(2)} = (\gamma_1^{(2)}, \cdots, \gamma_{G_2}^{(2)})^\top \in \mathbb{R}^{G_2}$. Recall that $\boldsymbol{\xi} = (\boldsymbol{\theta}^{(1)\top}, \boldsymbol{\theta}^{(2)\top}, \text{vec}(\boldsymbol{\alpha})^\top)^\top \in \mathbb{R}^{G_1(p_1+1)+G_2(p_2+1)+G_1G_2}$ includes the other group parameters except for the interactive parameters $\boldsymbol{\psi}$. To estimate the unknown parameters, we minimize the following objective function,

$$Q(\boldsymbol{\xi}, \boldsymbol{\psi}, \mathcal{G}) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T} \Big\{ Y_{ij,t} - \gamma^{(1)}_{g_i^{(1)}} \gamma^{(2)}_{g_j^{(2)}} \sum_k \sum_m \frac{a_{im}^{(1)}}{n_{1i}} Y_{mk,t-1} \frac{a_{kj}^{(2)}}{n_{2j}}$$
$$- \Big(\lambda^{(1)}_{g_i^{(1)}} \sum_k \frac{a_{ik}^{(l)}}{n_{1i}} Y_{kj,t-1} + \lambda^{(2)}_{g_j^{(2)}} \sum_k \frac{a_{kj}^{(2)}}{n_{2j}} Y_{ik,t-1}\Big)$$
$$- \alpha_{g_i^{(1)} g_j^{(2)}} Y_{ij,(t-1)} - \Big(\mathbf{x}_{it}^{(1)\top} \boldsymbol{\zeta}^{(1)}_{g_i^{(1)}} + \mathbf{x}_{jt}^{(2)\top} \boldsymbol{\zeta}^{(2)}_{g_j^{(2)}}\Big) \Big\}^2. \quad \text{(A.26)}$$

We utilize the following iterative algorithm.

(1) Update the group parameters $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$.

First, fix the group memberships $g_i^{(1)}$ and $g_j^{(2)}$ and then we update the group parameters. In this parameter updating step, we iteratively update $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ as follows, which can be conducted in a fast speed. To be more specific, given interactive network effects $\boldsymbol{\psi}$, we could calculate

$$\widetilde{\mathbf{Y}}_t = \mathbf{Y}_t - (\boldsymbol{\Gamma}^{(1)}\mathbf{W}^{(1)})\mathbf{Y}_{t-1}(\mathbf{W}^{(2)}\boldsymbol{\Gamma}^{(2)}).$$

Then the parameter $\boldsymbol{\xi}$ can be obtained by $\widehat{\boldsymbol{\xi}} = (\mathbf{M})^{-1}\mathbf{b}$, whose specific expressions can be found in (A.7), with $\mathbb{Y}_{g^{(1)}g^{(2)},t}$ replaced by $\mathrm{vec}\big(\widetilde{\mathbf{Y}}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\mathcal{R}^{(2)}_{g^{(2)}})}\big)$.

Subsequently, given $\boldsymbol{\xi}$, denote $\breve{\mathbf{Y}}_t = \mathbf{Y}_t - \mathbf{L}^{(1)}\mathbf{W}^{(1)}\mathbf{Y}_{t-1} - \mathbf{Y}_{t-1}\mathbf{W}^{(2)}\mathbf{L}^{(2)} - \mathbf{A}\circ\mathbf{Y}_{t-1} - \boldsymbol{\beta}^{(1)}_{X_1,t}\mathbf{1}^\top_{N_2} - \mathbf{1}_{N_1}\boldsymbol{\beta}^{(2)\top}_{X_2,t}$. We need to minimize the following least squares problem,

$$\min_{\boldsymbol{\Gamma}^{(1)},\boldsymbol{\Gamma}^{(2)}} \sum_t \|\breve{\mathbf{Y}}_t - \boldsymbol{\Gamma}^{(1)}\mathbf{W}^{(1)}\mathbf{Y}_{t-1}\mathbf{W}^{(2)}\boldsymbol{\Gamma}^{(2)}\|_F^2.$$

Denote that $\widetilde{\mathbb{X}}_t = \mathbf{W}^{(1)}\mathbf{Y}_{t-1}\mathbf{W}^{(2)}\boldsymbol{\Gamma}^{(2)} \in \mathbb{R}^{N_1\times N_2}$. By the first order condition, we have that

$$\widehat{\gamma}^{(1)}_{g^{(1)}} = \Big(\sum_t \sum_{i\in\mathcal{R}^{(1)}_{g^{(1)}}} \sum_j \widetilde{\mathbb{X}}_{ij,t}\breve{\mathbf{Y}}_{ij,t}\Big) \Big/ \sum_t \Big\|\widetilde{\mathbb{X}}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdot)}\Big\|_F^2, \quad g^{(1)} \in [G_1] \tag{A.27}$$

$$\widehat{\gamma}^{(2)}_{g^{(2)}} = \Big(\sum_t \sum_i \sum_{j\in\mathcal{R}^{(2)}_{g^{(2)}}} \widetilde{\mathbb{X}}_{ij,t}\breve{\mathbf{Y}}_{ij,t}\Big) \Big/ \sum_t \Big\|\widetilde{\mathbb{X}}_t^{(\cdot,\mathcal{R}^{(2)}_{g^{(2)}})}\Big\|_F^2, \quad g^{(2)} \in [G_2], \tag{A.28}$$

where $\mathcal{R}^{(1)}_{g^{(1)}} = \{i \in [N_1] : g_i^{(1)} = g^{(1)}\}$ and $\mathcal{R}^{(2)}_{g^{(2)}} = \{j \in [N_2] : g_j^{(2)} = g^{(2)}\}$. Hence, we can iteratively update $\boldsymbol{\Gamma}^{(1)}$ and $\boldsymbol{\Gamma}^{(2)}$ using the above equations (A.27) and (A.28).

(2) Update the group memberships $\mathcal{G}_1$ and $\mathcal{G}_2$.

Second, we consider given $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$, updating the group memberships $\mathcal{G}_1 = (g_i^{(1)} : i \in [N_1])^\top$ and $\mathcal{G}_2 = (g_j^{(2)} : j \in [N_2])^\top$ iteratively. Given $\mathcal{G}_2$, update

$$\begin{aligned}\widehat{g}_i^{(1)} = \mathrm{argmin}_{g_i^{(1)}\in[G_1]} \sum_{j=1}^{N_2}\sum_t \Big\{ & Y_{ij,t} - \gamma^{(1)}_{g_i^{(1)}}\gamma^{(2)}_{g_j^{(2)}} \sum_k\sum_m \frac{a^{(1)}_{im}}{n_{1i}} Y_{mk,t-1}\frac{a^{(2)}_{kj}}{n_{2j}} \\ & - \Big(\lambda^{(1)}_{g_i^{(1)}} \sum_k \frac{a^{(l)}_{ik}}{n_{1i}} Y_{kj,t-1} + \lambda^{(2)}_{g_j^{(2)}} \sum_k \frac{a^{(2)}_{kj}}{n_{2j}} Y_{ik,t-1}\Big) \\ & - \alpha_{g_i^{(1)}g_j^{(2)}} Y_{ij,(t-1)} - \Big(\mathbf{x}_{it}^{(1)\top}\boldsymbol{\zeta}^{(1)}_{g_i^{(1)}} + \mathbf{x}_{jt}^{(2)\top}\boldsymbol{\zeta}^{(2)}_{g_j^{(2)}}\Big)\Big\}^2. \end{aligned} \tag{A.29}$$

The parallel computation is implement to update $\mathcal{G}_2$ given $\mathcal{G}_1$. To summarize, the iterative algorithm is shown in Algorithm A.4.

To estimate the group numbers for model (25), denoted as $\underline{G} = (G_1, G_2)$, we minimize the information criterion.

$$\mathrm{QIC}(\underline{G}) = \log\{Q(\widehat{\boldsymbol{\xi}}(\underline{G}), \widehat{\boldsymbol{\psi}}(\underline{G}), \widehat{\mathcal{G}}(\underline{G}))\} + \lambda(\underline{G}), \tag{A.30}$$

where $\widehat{\boldsymbol{\xi}}_1(\underline{G})$, $\widehat{\boldsymbol{\psi}}(\underline{G})$ and $\widehat{\mathcal{G}}(\underline{G})$ are the estimated parameters when specifying the group numbers as G_1 and G_2, $Q(\cdot,\cdot,\cdot)$ is the objective function defined in (A.26), and $\lambda(\underline{G})$ is the penalty function. Specifically, we set $\lambda(\underline{G}) = \kappa_1(G_1+G_2)$ with κ_1 being a tuning parameter.

We summarize the estimation algorithm in the following Algorithm A.4.

Algorithm A.4 Estimation of the GTNAR Model with Interactive Term When $q = 2$

1: **Input:** $\{\mathbf{Y}_t, \mathbf{X}_t^{(l)}, \mathbf{W}^{(l)}, G_l\}$ for $1 \le l \le q$.

2: Obtain initial group memberships $\mathcal{G}_l^{[0]}$ according to Algorithm A.3. Let $\{\boldsymbol{\xi}^{[k]}, \mathcal{G}_l^{[k]}\}$ be the estimators and memberships in the kth iteration.

3: Repeat Step 1 and Step 2 for $k = 1, 2, \cdots$ until convergence.

Step 1. Given $\{\mathcal{G}_l^{[k-1]}\}$ for all layers $l \in [q]$

Step (1a). Fix the interactive network effects $\boldsymbol{\Gamma}^{(1)[k-1]}$ and $\boldsymbol{\Gamma}^{(2)[k-1]}$, calculate $\widetilde{\mathbf{Y}}_t = \mathbf{Y}_t - (\boldsymbol{\Gamma}^{(1)[k-1]}\mathbf{W}^{(1)})\mathbf{Y}_{t-1}(\mathbf{W}^{(2)}\boldsymbol{\Gamma}^{(2)[k-1]})$. Update $\boldsymbol{\xi}^{[k]} = (\mathbf{M}^{[k-1]})^{-1}\mathbf{b}^{[k-1]}$..

Step (1b). Fix the estimated parameters $\boldsymbol{\xi}^{[k]}$. Given $\boldsymbol{\Gamma}^{(2)[k-1]}$, obtain $\boldsymbol{\Gamma}^{(1)[k]}$ using (A.27); then given $\boldsymbol{\Gamma}^{(1)[k]}$, obtain $\boldsymbol{\Gamma}^{(2)[k]}$ using (A.28).

Step 2. Given $\gamma^{(1)[k]}_{g^{(1)[k-1]}}$, $\gamma^{(2)[k]}_{g^{(2)[k-1]}}$, $\lambda^{(1)[k]}_{g^{(1)[k-1]}}$, $\lambda^{(2)[k]}_{g^{(2)[k-1]}}$, $\alpha^{[k]}_{g^{(1)[k-1]}g^{(2)[k-1]}}$ and $\boldsymbol{\zeta}^{(1)[k]}_{g^{(1)[k-1]}}$, $\boldsymbol{\zeta}^{(2)[k]}_{g^{(2)[k-1]}}$, update the memberships $\mathcal{G}_1^{[}(k)]$ and $\mathcal{G}_2^{[}(k)]$ sequentially by,

$$
\begin{aligned}
g_i^{(1)[k]} = \operatorname{argmin}_{g_i^{(1)}\in[G_1]} \sum_j \sum_t \Big\{ & Y_{ij,t} - \Big(\gamma^{(1)[k]}_{g_i^{(1)}} \gamma^{(2)[k]}_{g_j^{(2)[k-1]}} \sum_k \sum_m \frac{a^{(1)}_{im}}{n_{1i}} Y_{mk,t-1} \frac{a^{(2)}_{kj}}{n_{2j}}\Big) \\
& - \Big(\lambda^{(1)[k]}_{g_i^{(1)[k]}} \sum_k \frac{a^{(l)}_{ik}}{n_{1i}} Y_{kj,t-1} + \lambda^{(2)[k]}_{g_j^{(2)[k-1]}} \sum_k \frac{a^{(2)}_{kj}}{n_{2j}} Y_{ik,t-1}\Big) \\
& - \alpha^{[k]}_{g_i^{(1)} g_j^{(2)[k-1]}} Y_{ij,(t-1)} - \Big(\mathbf{x}_{it}^{(1)\top} \boldsymbol{\zeta}^{(1)[k]}_{g_i^{(1)}} + \mathbf{x}_{jt}^{(2)\top} \boldsymbol{\zeta}^{(2)[k]}_{g_j^{(2)[k-1]}}\Big)\Big\}^2
\end{aligned}
$$

and

$$
\begin{aligned}
g_j^{(2)[k]} = \operatorname{argmin}_{g_j^{(2)}\in[G_2]} \sum_i \sum_t \Big\{ & Y_{ij,t} - \Big(\gamma^{(1)[k]}_{g_i^{(1)[k]}} \gamma^{(2)[k]}_{g_j^{(2)}} \sum_k \sum_m \frac{a^{(1)}_{im}}{n_{1i}} Y_{mk,t-1} \frac{a^{(2)}_{kj}}{n_{2j}}\Big) \\
& - \Big(\lambda^{(1)[k]}_{g_i^{(1)[k]}} \sum_k \frac{a^{(l)}_{ik}}{n_{1i}} Y_{kj,t-1} + \lambda^{(2)[k]}_{g_j^{(2)}} \sum_k \frac{a^{(2)}_{kj}}{n_{2j}} Y_{ik,t-1}\Big) \\
& - \alpha^{[k]}_{g_i^{(1)[k]} g_j^{(2)}} Y_{ij,(t-1)} - \Big(\mathbf{x}_{it}^{(1)\top} \boldsymbol{\zeta}^{(1)[k]}_{g_i^{(1)[k]}} + \mathbf{x}_{jt}^{(2)\top} \boldsymbol{\zeta}^{(2)[k]}_{g_j^{(2)}}\Big)\Big\}^2
\end{aligned}
$$

4: **Output:** Final estimator and memberships: $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}^{[K]}$, $\widehat{\boldsymbol{\psi}} = \boldsymbol{\psi}^{[K]}$, and $\widehat{\mathcal{G}} = \{\mathcal{G}_l^{[K]} : l = 1, 2\}$. Here K is the final number of iteration rounds.

G.2 Technical Assumptions

Denote

$$\mathcal{X}_{ij,t} = (\sum_{k=1}^{N_1} w_{ik}^{(1)} Y_{kj,t}, \mathbf{x}_{1t}^{(1)\top}, \sum_{k=1}^{N_2} w_{kj}^{(2)} Y_{ik,t}, \mathbf{x}_{jt}^{(2)\top}, Y_{ij,t-1}, \sum_{k}^{N_1}\sum_{m}^{N_2} w_{ik}^{(1)} Y_{km,t-1} w_{mj}^{(2)}) \in \mathbb{R}^{p_1+p_2+4}. \tag{A.31}$$

By using $\mathcal{X}_{ij,t}$, we rewrite the objective function as $Q(\mathbf{\Theta}) = \sum_{i,j,t}(Y_{ij,t} - \mathcal{X}_{ij,t}^\top \mathbf{\Theta}_{ij})^2$. Denote $\widehat{\mathbf{\Theta}}$ as the estimator of $\mathbf{\Theta}$ by minimizing the loss function $Q(\mathbf{\Theta})$. We require the following assumptions.

Assumption A.11. (Parameter Space) *The parameter satisfies that* $\|\mathbf{\Theta}\|_{\max} < \infty$.

Assumption A.12. (Convexity) *Denote* $\mathbf{\Sigma}_{ij} = E(\mathcal{X}_{ij,t}\mathcal{X}_{ij,t}^\top)$, *where* $\mathcal{X}_{ij,t}$ *is defined in* (A.31). *Assume that* $\tau_{\min}^1 \overset{\text{def}}{=} \min_{i,j} \mathbf{\Sigma}_{ij} > 0$ *is a constant.*

Assumption A.13. (Stability) *Suppose* $\mathbb{Y}_0 \overset{\text{def}}{=} \text{vec}(\mathcal{Y}_0) = \mathbf{0}$, *and assume that*

$$\max_{g^{(1)}\in[G_{1,0}], g^{(2)}\in[G_{2,0}]} \left|\lambda_{g^{(1)}}^{(1)0} + \lambda_{g^{(2)}}^{(2)0} + \alpha_{g^{(1)}g^{(2)}}^{0} + \gamma_{g^{(1)}}^{(1)0}\gamma_{g^{(2)}}^{(2)0}\right| \le \kappa_{\max} < 1,$$

where $G_{l,0}$ *is the true number of groups in the* l*th dimension, and* $\kappa_{\max}$ *is a positive constant.*

Assumption A.14. (Group Difference) *For the first dimension, assume*

$$\min_{g_1^{(1)}\neq g_2^{(1)}} \left[\left\|\boldsymbol{\theta}_{g_1^{(1)}}^{(1)0} - \boldsymbol{\theta}_{g_2^{(1)}}^{(1)0}\right\|^2 + \max_{g^{(2)}\in[G_2^0]} \left\{\left|\alpha_{g_1^{(1)}g^{(2)}}^{0} - \alpha_{g_2^{(1)}g^{(2)}}^{0}\right|^2 + \left|\gamma_{g_1^{(1)}}^{(1)0}\gamma_{g^{(2)}}^{(2)0} - \gamma_{g_2^{(1)}}^{(1)0}\gamma_{g^{(2)}}^{(2)0}\right|^2\right\}\right] \ge c_{\text{gap}},$$

where $c_{\text{gap}} \gg T^{-1}(\log(N_1N_2))^2(G_1G_2)$ *as* $T \to \infty$. *Here* G_1, G_2 *are finite. Assume the same holds for the group* $g_1^{(2)} \neq g_2^{(2)}$ *in the second dimension.*

G.3 Numerical Studies

The estimation results in Table A.6 demonstrate excellent finite sample performance: all RMSEs decrease with increasing time length T or increasing sizes N_1, N_2 (e.g., RMSE of $\widehat{\boldsymbol{\lambda}}^{(1)}$ declines from 0.0131 at $T = 40$ to 0.0028 at $T = 80$), while the coverage probabilities in parentheses converge closely to the nominal 95% level at larger T. Furthermore, the group membership errors $\widehat{\eta}_1$ and $\widehat{\eta}_2$ remain near-zero across all configuration with consistently diminishing magnitudes at larger sample sizes, indicating exceptionally accurate group classification even for moderate sample sizes.

G.4 Technical Proofs

G.4.1 Proof of Theorem 8 (i)

By using the $\mathcal{X}_{ij,t}$ defined in (A.31), we define the objective function,

$$Q(\mathbf{\Theta}) = \sum_{i=1}^{N_1}\sum_{j=2}^{N_2}\sum_{t=1}^{T}(Y_{ij,t} - \mathcal{X}_{ij,t}^\top \mathbf{\Theta}_{ij})^2 = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T} Q_{ij}(\mathbf{\Theta}_{ij}).$$

Table A.6: RMSEs of estimated parameters under when $G_{1,0} = G_{2,0} = 3$ with 100 replications. The performances are evaluated for different sample sizes N_1, N_2 and the time length T. The corresponding CPs are shown in the parenthesis.

G_1	G_2	N_1	N_2	T	$\widehat{\boldsymbol{\lambda}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}^{(2)}$	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\Gamma}}_{\text{norm}}$	$\widehat{\eta}_1$	$\widehat{\eta}_2$
3	3	100	80	20	0.0131 (0.90)	0.0119 (0.93)	0.0137 (0.93)	0.0150 (0.92)	0.0369 (0.92)	3.0282	0.0122	0.0086
				40	0.0079 (0.93)	0.0081 (0.95)	0.0089 (0.95)	0.0094 (0.95)	0.0131 (0.95)	1.2326	0.0005	0.0001
				80	0.0055 (0.93)	0.0052 (0.95)	0.0064 (0.95)	0.0067 (0.95)	0.0091 (0.95)	0.5685	0.0000	0.0001
		200	150	20	0.0061 (0.94)	0.0053 (0.93)	0.0070 (0.94)	0.0069 (0.94)	0.0095 (0.94)	2.5897	0.0004	0.0002
				40	0.0040 (0.95)	0.0039 (0.93)	0.0047 (0.95)	0.0048 (0.96)	0.0068 (0..95)	1.0622	0.0000	0.0002
				80	0.0027 (0.95)	0.0028 (0.93)	0.0032 (0.96)	0.0035 (0.96)	0.0047 (0.94)	0.6653	0.0000	0.0001

Given this objective function, we can follow the proof of Theorem 2 using the partition of the new parameter space. To obtain the conclusion, we use the Lemma 15 and Lemma 13 instead of the usage of Lemma 23 and Lemma 19 in the original proof of Theorem 2. Since the rest of the proof remains the same, we omit the details here.

G.4.2 Proof of Theorem 8 (iii)

We first declare some notations in the proof. Denote $\widetilde{\boldsymbol{\gamma}}^{(1)}_{g^{(1)}} = (\gamma^{(1)}_{g^{(1)}}\gamma^{(2)}_{g^{(2)}} : g^{(2)} \in [G_2])^\top \in \mathbb{R}^{G_2}$ and $\widetilde{\boldsymbol{\gamma}}^{(2)}_{g^{(2)}} = (\gamma^{(2)}_{g^{(1)}}\gamma^{(2)}_{g^{(2)}} : g^{(1)} \in [G_1])^\top \in \mathbb{R}^{G_1}$. Denote $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\boldsymbol{\gamma}}^{(1)\top}_{g^{(1)}} : g^{(1)} \in [G_1])^\top \in \mathbb{R}^{G_1G_2}$. Further denote that $\widetilde{\boldsymbol{\alpha}} = (\text{vec}(\boldsymbol{\alpha}_{g^{(1)}.})^\top : g^{(1)} \in [G_1])^\top \in \mathbb{R}^{G_1G_2}$. Recall that $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\theta}^{(1)\top}_{g^{(1)}} : g^{(1)} \in [G_1])^\top$. Denote $\boldsymbol{\xi}^{(1)}_{g^{(1)}} = (\boldsymbol{\theta}^{(1)\top}_{g^{(1)}}, \text{vec}(\boldsymbol{\alpha}_{g^{(1)}.})^\top, \widetilde{\boldsymbol{\gamma}}^{(1)\top}_{g^{(1)}})^\top$ as the parameters of the $g^{(1)}$th group ($g^{(1)} \in [G_1]$). Correspondingly, denote $\boldsymbol{\xi}^{(1)0}_{g^{(1)0}}$ as the true parameters of the true group $g^{(1)0} \in [G_{1,0}]$. Let $\boldsymbol{\xi}^{(1)} = (\boldsymbol{\theta}^{(1)\top}, \widetilde{\boldsymbol{\alpha}}^\top, \widetilde{\boldsymbol{\gamma}}^\top)^\top$. In parallel, we can define $\boldsymbol{\xi}^{(2)}_{g^{(2)}} = (\boldsymbol{\theta}^{(2)\top}_{g^{(2)}}, \widetilde{\boldsymbol{\alpha}}^\top, \widetilde{\boldsymbol{\gamma}}^\top)^\top$ and $\boldsymbol{\xi}^{(2)} = (\boldsymbol{\theta}^{(2)\top}, \widetilde{\boldsymbol{\alpha}}^\top, \widetilde{\boldsymbol{\gamma}}^\top)^\top$. At last, denote $\boldsymbol{\xi} = (\boldsymbol{\theta}^{(1)\top}, \boldsymbol{\theta}^{(2)\top}, \widetilde{\boldsymbol{\alpha}}^\top, \widetilde{\boldsymbol{\gamma}}^\top)^\top$. In the following, we prove the theorem for the first dimension, and the second dimension group membership consistency can be proved similarly. For $g^{(1)} \in [G_1]$ and $g^{(1)0} \in [G_{1,0}]$, define

$$\begin{aligned}\mathcal{L}^{(1)}(\boldsymbol{\xi}^{(1)}_{g^{(1)}}, \boldsymbol{\xi}^{(1)0}_{g^{(1)0}}; \mathcal{G}_2, \mathcal{G}_2^0) &= \big\|\boldsymbol{\theta}^{(1)}_{g^{(1)}} - \boldsymbol{\theta}^{(1)0}_{g^{(1)0}}\big\|^2 \\ &+ \frac{1}{N_2}\sum_j \Big\{\big|\alpha_{g^{(1)}g^{(2)}_j} - \alpha^0_{g^{(1)0}g^{(2)0}_j}\big|^2 + \big|\gamma^{(1)}_{g^{(1)}}\gamma^{(2)}_{g^{(2)}_j} - \gamma^{(1)0}_{g^{(1)0}}\gamma^{(2)0}_{g^{(2)0}_j}\big|^2\Big\}. \end{aligned} \tag{A.32}$$

For notation simplicity, denote $\mathcal{L}^{(1)}_{g^{(1)},g^{(1)0}} \overset{\text{def}}{=} \mathcal{L}^{(1)}(\boldsymbol{\xi}^{(1)}_{g^{(1)}}, \boldsymbol{\xi}^{(1)0}_{g^{(1)0}}; \mathcal{G}_2, \mathcal{G}_2^0)$. Define the distance

$$d^{(1)}_L(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(1)0}; \mathcal{G}_2, \mathcal{G}_2^0) = \max\Big\{\max_{g^{(1)0}\in[G_{1,0}]}\min_{g^{(1)}\in[G_1]}\mathcal{L}^{(1)}_{g^{(1)},g^{(1)0}}, \max_{g^{(1)}\in[G_1]}\min_{g^{(1)0}\in[G_{1,0}]}\mathcal{L}^{(1)}_{g^{(1)},g^{(1)0}}\Big\}.$$

In addition, given parameters $\boldsymbol{\xi}$, denote the estimated group memberships in the second dimension as $\widehat{\mathcal{G}}_2(\boldsymbol{\xi}) = (\widehat{g}_j^{(2)}(\boldsymbol{\xi}) : j \in [N_2])^\top$, define

$$\mathcal{N}_\eta^{(1)} = \{\boldsymbol{\xi} : d_L^{(1)}(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(1)0}; \widehat{\mathcal{G}}_2(\boldsymbol{\xi}), \mathcal{G}_2^0) < \eta\}$$

and denote

$$\mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g^{(1)0}, \mathcal{G}_2^0) = \Big\{ g^{(1)} \in [G_1] : \\ \big\|\boldsymbol{\theta}_{g^{(1)}}^{(1)} - \boldsymbol{\theta}_{g^{(1)0}}^{(1)0}\big\|^2 + \frac{1}{N_2}\sum_j \Big\{ \big|\alpha_{g^{(1)}\widehat{g}_j^{(2)}(\boldsymbol{\xi})} - \alpha^0_{g^{(1)0}g_j^{(2)0}}\big|^2 + \big|\gamma^{(1)}_{g^{(1)}}\gamma^{(2)}_{\widehat{g}_j^{(2)}(\boldsymbol{\xi})} - \gamma^{(1)0}_{g^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}}\big|^2 \Big\} \le \eta \Big\}.$$

Based on the above notations, we first give the following lemma, whose proof can be found in Appendix G.4.3.

Lemma 12. *Under Assumptions 3, 4, 7, A.11–A.14, and assume that $G_1 \ge G_{1,0}$ and $G_2 \ge G_{2,0}$. For $\boldsymbol{\Theta}$ satisfying the condition*

$$d(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = O_p\{T^{-1}(\log(N_1N_2))^2\} = o_p(c_{\text{gap}}) \tag{A.33}$$

as $N_l \to \infty$, $l = 1, 2$, the following conclusions hold,
(i) For all $\boldsymbol{\xi} \in \mathcal{N}_\eta^{(1)}$ with $\eta < c_{\text{gap}}c_\pi/4$, then we have that $\{\mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g^{(1)0}, \mathcal{G}_2^0) : g^{(1)0} \in [G_{1,0}]\}$ is a partition of $[G_1]$.
(ii) For all $\boldsymbol{\xi} \in \mathcal{N}_\eta^{(1)}$ with $\eta \le \tau^1_{\min}c_{\text{gap}}c_\pi/(8(\tau^1_{\min} + \tau^1_{\max}))$, define the event $\mathcal{O} = \{\widehat{g}_i^{(1)}(\boldsymbol{\xi}) \in \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}_2^0) : i \in [N_1]\}$. Then we have $P(\mathcal{O}^c) \le C\exp(-c_1\sqrt{T}c_{\text{gap}} + c_2 m + \log N_1 + \log N_2)$, where C, c_1, c_2 are positive constants, and $m = p_1 + p_2 + 4$.

Next, by using Lemma 12, we first show that $\widehat{\boldsymbol{\xi}} \in \mathcal{N}_\eta^{(1)}$ when $\eta = \tau^1_{\min}c_{\text{gap}}c_\pi/(8(\tau^1_{\min} + \tau^1_{\max}))$ with probability tending to 1, then we use (i) and (ii) in Lemma 12 to obtain the final results.

Step 1. $P\big(\widehat{\boldsymbol{\xi}} \in \mathcal{N}_\eta^{(1)}\big) \to 1$**.**

Using the similar procedure in the Step 1 of proof of Theorem 5, we can show that $d_L^{(1)}(\widehat{\boldsymbol{\xi}}^{(1)}, \boldsymbol{\xi}^{(1)0}; \widehat{\mathcal{G}}_2(\widehat{\boldsymbol{\xi}}), \mathcal{G}_2^0) = O_p\{T^{-1}(\log(N_1N_2))^2\}$. By the condition that

$$c_{\text{gap}} \gg T^{-1}(\log(N_1N_2))^2,$$

we have $\widehat{\boldsymbol{\xi}} \in \mathcal{N}_\eta^{(1)}$ with probability tending to 1.

Step 2. Conclusion of Theorem 8 (iii).

For $\widetilde{g}^{(1)} \in [G_1]$, suppose there are $i_1, i_2 \in \widehat{\mathcal{R}}^{(1)}_{\widetilde{g}^{(1)}}$, then $\widehat{g}^{(1)}_{i_1} = \widehat{g}^{(2)}_{i_2} = \widetilde{g}^{(1)}$. By Lemma 12(ii), we have $\widetilde{g}^{(1)} \in \mathcal{A}_\eta^{(1)}(\widehat{\boldsymbol{\xi}}, g^{(1)0}_{i_1}, \mathcal{G}_2^0)$ and $\widetilde{g}^{(1)} \in \mathcal{A}_\eta^{(1)}(\widehat{\boldsymbol{\xi}}, g^{(1)0}_{i_2}, \mathcal{G}_2^0)$ hold with probability tending to 1. By Lemma 12(i), we know that $\mathcal{A}_\eta^{(1)}(\widehat{\boldsymbol{\xi}}, \cdot, \mathcal{G}_2^0)$ is a partition of $[G_1]$, hence $g^{(1)0}_{i_1} = g^{(1)0}_{i_2}$. Then, by setting $g^{(1)} = g^{(1)0}_{i_1} = g^{(1)0}_{i_2}$, we have $i_1, i_2 \in \mathcal{R}^{(1)0}_{g^{(1)}}$.

G.4.3 Proof of Lemma 12

(1) Proof of (i).

For all $\boldsymbol{\xi} \in \mathcal{N}_\eta^{(1)}$, by the definition of $\mathcal{N}_\eta^{(1)}$ and the group set definition $\mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g^{(1)0}, \mathcal{G}_2^0)$, we have $\cup_{g^{(1)0}=1}^{G_{1,0}} \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g^{(1)0}, \mathcal{G}_2^0) = [G_1]$. Then we prove $\mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g_1^{(1)0}, \mathcal{G}_2^0) \cap \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g_2^{(1)0}, \mathcal{G}_2^0) = \emptyset$ for $g_1^{(1)0}, g_2^{(1)0} \in [G_{1,0}]$ and $g_1^{(1)0} \neq g_2^{(1)0}$. We use the contradiction. Assume there exists $g_3^{(1)} \in [G_1]$ s.t. $g_3^{(1)} \in \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g_1^{(1)0}, \mathcal{G}_2^0) \cap \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g_2^{(1)0}, \mathcal{G}_2^0)$. Then under Assumption A.14, we have

$$
\begin{aligned}
c_{\text{gap}} &\leq \Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}^{(1)0}_{g_2^{(1)0}}\Big\|^2 + \max_{g^{(2)0}\in[G_2^0]} \Big\{\Big|\alpha^0_{g_1^{(1)0}g^{(2)0}} - \alpha^0_{g_2^{(1)0}g^{(2)0}}\Big|^2 + \Big|\gamma^{(1)0}_{g_1^{(1)0}}\gamma^{(2)0}_{g^{(2)0}} - \gamma^{(1)0}_{g_2^{(1)0}}\gamma^{(2)0}_{g^{(2)0}}\Big|^2\Big\} \\
&\leq 2\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 + 2\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 \\
&+ \max_{g^{(2)0}\in[G_{2,0}]} \Big\{\frac{1}{\pi^{(2)}_{g^{(2)0},N_2}N_2}\sum_j \Big|\alpha^0_{g_1^{(1)0}g_j^{(2)0}} - \alpha^0_{g_2^{(1)0}g_j^{(2)0}}\Big|^2 I(g_j^{(2)0} = g^{(2)0})\Big\} \\
&+ \max_{g^{(2)0}\in[G_{2,0}]} \Big\{\frac{1}{\pi^{(2)}_{g^{(2)0},N_2}N_2}\sum_j \Big|\gamma^{(1)0}_{g_1^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}} - \gamma^{(1)0}_{g_2^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}}\Big|^2 I(g_j^{(2)0} = g^{(2)0})\Big\} \\
&\leq 2\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 + 2\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 \\
&+ \max_{g^{(2)0}\in[G_{2,0}]} \Big(\frac{1}{\pi^{(2)}_{g^{(2)0},N_2}N_2}\Big)\sum_j \Big\{\Big|\alpha^0_{g_1^{(1)0}g_j^{(2)0}} - \alpha^0_{g_2^{(1)0}g_j^{(2)0}}\Big|^2 + \Big|\gamma^{(1)0}_{g_1^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}} - \gamma^{(1)0}_{g_2^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}}\Big|^2\Big\} \\
&\leq 2\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 + 2\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 \\
&+ \Big(\frac{1}{\min_{g^{(2)0}\in[G_{2,0}]}\pi^{(2)}_{g^{(2)0},N_2}N_2}\Big)\sum_j \Big\{\Big|\alpha^0_{g_1^{(1)0}g_j^{(2)0}} - \alpha^0_{g_2^{(1)0}g_j^{(2)0}}\Big|^2 \\
&\qquad\qquad + \Big|\gamma^{(1)0}_{g_1^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}} - \gamma^{(1)0}_{g_2^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}}\Big|^2\Big\} \\
&\leq \Big(\frac{2}{\min_{g^{(2)0}\in[G_{2,0}]}\pi^{(2)}_{g^{(2)0},N_2}}\Big)\Big\{\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 + 2\Big\|\boldsymbol{\theta}^{(1)0}_{g_1^{(1)0}} - \boldsymbol{\theta}_{g_3^{(1)}}\Big\|^2 \\
&+ \frac{1}{N_2}\sum_j \Big|\alpha^0_{g_1^{(1)0}g_j^{(2)0}} - \alpha_{g_3^{(1)}\widehat{g}_j^{(2)}(\boldsymbol{\xi})}\Big|^2 + \frac{1}{N_2}\sum_j \Big|\alpha^0_{g_2^{(1)0}g_j^{(2)0}} - \alpha_{g_3^{(1)}\widehat{g}_j^{(2)}(\boldsymbol{\xi})}\Big|^2 \\
&+ \frac{1}{N_2}\sum_j \Big|\gamma^{(1)0}_{g_1^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}} - \gamma^{(1)}_{g_3^{(1)}}\gamma^{(2)}_{\widehat{g}_j^{(2)}(\boldsymbol{\xi})}\Big|^2 + \frac{1}{N_2}\sum_j \Big|\gamma^{(1)0}_{g_2^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}} - \gamma^{(1)}_{g_3^{(1)}}\gamma^{(2)}_{\widehat{g}_j^{(2)}(\boldsymbol{\xi})}\Big|^2\Big\} \\
&\leq \frac{4\eta}{\min_{g^{(2)0}\in[G_{2,0}]}\pi^{(2)}_{g^{(2)0},N_2}},
\end{aligned}
$$

where the last line holds due to $g_3^{(1)} \in \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g_1^{(1)0}, \mathcal{G}_2^0) \cap \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}, g_2^{(1)0}, \mathcal{G}_2^0)$. However, this conclusion contradicts $\eta < c_{\text{gap}}c_\pi/4$ as $N_2 \to \infty$. Therefore, we prove that $\mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}; g_1^{(1)0}, \mathcal{G}_2^0) \cap \mathcal{A}_\eta^{(1)}(\boldsymbol{\xi}; g_2^{(1)0}, \mathcal{G}_2^0) = \emptyset$, which completes the proof of (i).

(2) Proof of (ii).

Denote $\boldsymbol{\xi}^{(1)}_{g_i^{(1)}} = (\boldsymbol{\theta}^{(1)\top}_{g_i^{(1)}}, \mathrm{vec}(\boldsymbol{\alpha}_{g_i^{(1)}\cdot})^\top, \widetilde{\boldsymbol{\gamma}}^{(1)\top}_{g_i^{(1)}})$. Define the loss function

$$\begin{aligned} Q_i(\boldsymbol{\xi}^{(1)}_{g_i^{(1)}}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \mathcal{G}_2) = \sum_j \sum_t \Big\{ & Y_{ij,t} - \gamma^{(1)}_{g_i^{(1)}} \gamma^{(2)}_{g_j^{(2)}} \sum_k \sum_m w^{(1)}_{im} Y_{mk,t-1} w^{(2)}_{kj} \\ & - \lambda^{(1)}_{g_i^{(1)}} \sum_k w^{(1)}_{ik} Y_{kj,t-1} - \lambda^{(2)}_{g_j^{(2)}} \sum_k w_{kj} Y_{ik,t-1} \\ & - \alpha_{g_i^{(1)} g_j^{(2)}} Y_{ij,t-1} - \Big(\mathbf{x}^{(1)\top}_{it} \boldsymbol{\zeta}^{(1)}_{g_i^{(1)}} + \mathbf{x}^{(2)\top}_{jt} \boldsymbol{\zeta}^{(2)}_{g_j^{(2)}} \Big) \Big\}^2. \end{aligned} \tag{A.34}$$

For simplicity, for any given $\boldsymbol{\xi}$ satisfying $\boldsymbol{\xi} \in \mathcal{N}^{(1)}_\eta$, denote $\widehat{g}^{(1)}_i(\boldsymbol{\xi})$ as $\widehat{g}^{(1)}_i$. Hence we have that for any $\widetilde{g}^{(1)} \neq g^{(1)}$,

$$\big\{\widehat{g}^{(1)}_i = g^{(1)}\big\} \subseteq \big\{ Q_i(\boldsymbol{\xi}^{(1)}_{g^{(1)}}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \widehat{\mathcal{G}}_2(\boldsymbol{\xi})) \leq Q_i(\boldsymbol{\xi}^{(1)}_{\widetilde{g}^{(1)}}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \widehat{\mathcal{G}}_2(\boldsymbol{\xi})) \big\}.$$

Hence, for $\widetilde{g}^{(1)}_i \in \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2)$,

$$\begin{aligned} & I\big(\widehat{g}^{(1)}_i \notin \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2)\big) = \sum_{g^{(1)}=1}^{G_1} I\big(g^{(1)} \notin \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2)\big) I\big(\widehat{g}^{(1)}_i = g^{(1)}\big) \\ & \leq \sum_{g^{(1)}=1}^{G_1} I\big(g^{(1)} \notin \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2)\big) I\big\{ Q_i(\boldsymbol{\xi}^{(1)}_{g^{(1)}}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \widehat{\mathcal{G}}_2(\boldsymbol{\xi})) \big\} \overset{\text{def}}{=} \sum_{g^{(1)}=1}^{G_1} W_{ig^{(1)}}(\boldsymbol{\xi}). \end{aligned}$$

On one hand, for all $g^{(1)} \notin \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2)$ and $g^{(1)} \in [G_1]$, there exists a $g^{(1)0}_{i_2} \neq g^{(1)0}_i$, such that $g^{(1)} \in \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g^{(1)0}_{i_2}, \mathcal{G}^0_2)$ due to the conclusion of (i). Then we have

$$\begin{aligned} & \big\|\boldsymbol{\theta}^{(1)}_{g^{(1)}} - \boldsymbol{\theta}^{(1)0}_{g_i^{(1)0}}\big\|^2 + \frac{1}{N_2} \sum_j \Big\{ \big|\alpha_{g^{(1)}\widehat{g}^{(2)}_j(\boldsymbol{\xi})} - \alpha^0_{g_i^{(0)} g_j^{(2)0}}\big|^2 + \big|\gamma^{(1)}_{g^{(1)}} \gamma^{(2)}_{\widehat{g}^{(2)}_j(\boldsymbol{\xi})} - \gamma^{(1)0}_{g_i^{(0)}} \gamma^{(2)0}_{g_j^{(2)0}}\big|^2 \Big\} \\ & \geq \frac{1}{2}\big\|\boldsymbol{\theta}^{(1)0}_{g^{(1)0}_{i_2}} - \boldsymbol{\theta}^{(1)0}_{g_i^{(1)0}}\big\|^2 + \frac{1}{2N_2} \sum_j \Big\{ \big|\alpha^0_{g^{(1)0}_{i_2} g_j^{(2)0}} - \alpha^0_{g_i^{(1)0} g_j^{(2)0}}\big|^2 + \big|\gamma^{(1)0}_{g^{(1)0}_{i_2}} \gamma^{(2)0}_{g_j^{(2)0}} - \gamma^{(1)0}_{g_i^{(1)0}} \gamma^{(2)0}_{g_j^{(2)0}}\big|^2 \Big\} \\ & - \Big[\big\|\boldsymbol{\theta}^{(1)}_{g^{(1)}} - \boldsymbol{\theta}^{(1)0}_{g^{(1)0}_{i_2}}\big\|^2 + \frac{1}{N_2} \sum_j \Big\{ \big|\alpha_{g^{(1)}\widehat{g}^{(2)}_j(\boldsymbol{\xi})} - \alpha^0_{g^{(1)0}_{i_2} g_j^{(2)0}}\big|^2 + \big|\gamma^{(1)}_{g^{(1)}} \gamma^{(2)}_{\widehat{g}^{(2)}_j(\boldsymbol{\xi})} - \gamma^{(1)0}_{g^{(1)0}_{i_2}} \gamma^{(2)0}_{g_j^{(2)0}}\big|^2 \Big\} \Big] \\ & \geq c_{\mathrm{gap}} c_\pi/2 - \eta, \end{aligned}$$

when $N_2 \to \infty$, where c_{gap} and c_π are defined in Assumption A.14 and Assumption 7, respectively. By Lemma 15, it holds for any $g^{(1)} \notin \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2)$,

$$\frac{1}{N_2 T} \Big\{ Q^*_i(\boldsymbol{\xi}^{(1)}_{g^{(1)}}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \widehat{\mathcal{G}}_2(\boldsymbol{\xi})) - Q^*_i(\boldsymbol{\xi}^{(1)}_{g_i^{(1)0}}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \mathcal{G}^0_2) \Big\} \geq \tau^1_{\min}(c_{\mathrm{gap}} c_\pi/2 - \eta). \tag{A.35}$$

On the other hand, for $\widetilde{g}^{(1)}_i \in \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2)$, it holds that

$$\frac{1}{N_2 T} \Big\{ Q^*_i(\boldsymbol{\xi}^{(1)}_{\widetilde{g}^{(1)}_i}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \widehat{\mathcal{G}}_2(\boldsymbol{\xi})) - Q^*_i(\boldsymbol{\xi}^{(1)}_{g_i^{(1)0}}; \boldsymbol{\xi}^{(2)}_{g^{(2)}}, \mathcal{G}^0_2) \Big\}$$

$$\leq \tau^1_{\max}\Big\{\big\|\boldsymbol{\theta}^{(1)}_{\widetilde{g}_i^{(1)}} - \boldsymbol{\theta}^{(1)0}_{g_i^{(1)0}}\big\|^2 + \frac{1}{N_2}\sum_j \big\|\boldsymbol{\theta}^{(1)}_{\widehat{g}_j^{(2)}(\boldsymbol{\xi})} - \boldsymbol{\theta}^{(1)0}_{g_j^{(2)0}}\big\|^2$$
$$+ \frac{1}{N_2}\sum_j \big(\big|\alpha_{\widetilde{g}_i^{(1)}\widehat{g}_j^{(2)}(\boldsymbol{\xi})} - \alpha^0_{g_i^{(1)0}g_j^{(2)0}}\big|^2 + \big|\gamma^{(1)}_{\widetilde{g}_i^{(1)}}\gamma^{(2)}_{\widehat{g}_j^{(2)}(\boldsymbol{\xi})} - \gamma^{(1)0}_{g_i^{(1)0}}\gamma^{(2)0}_{g_j^{(2)0}}\big|^2\big)\Big\}$$
$$\leq \tau^1_{\max}(d(\boldsymbol{\Theta},\boldsymbol{\Theta}^0) + \eta) \leq \tau^1_{\max}\{CT^{-1}(\log N_1 + \log N_2)^2 + \eta\}, \tag{A.36}$$

with probability tending to 1, where the first inequality holds due to Lemma 15, and the last inequality holds due to (A.33). Together with (A.35), we have that with probability tending to 1,

$$\frac{1}{N_2T}Q^*_i(\boldsymbol{\xi}^{(1)}_{g^{(1)}};\boldsymbol{\xi}^{(2)}_{g^{(2)}},\widehat{\mathcal{G}}_2(\boldsymbol{\xi})) - Q^*_i(\boldsymbol{\xi}^{(1)}_{\widetilde{g}_i^{(1)}};\boldsymbol{\xi}^{(2)}_{g^{(2)}},\widehat{\mathcal{G}}_2(\boldsymbol{\xi}))$$
$$\geq \tau^1_{\min}(c_{\text{gap}}c_\pi/2 - \eta) - \tau^1_{\max}\{CT^{-1}(\log N_1 + \log N_2)^2 + \eta\} \stackrel{\text{def}}{=} \epsilon_\eta.$$

Then we apply the similar techniques as in (A.93) and (A.94), to obtain that

$$P\Big\{\sup_i I(\widehat{g}_i^{(1)} \notin \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2))\Big\}$$
$$\leq \sum_{g^{(1)}=1}^{G_1} P\Big\{\sup_i W_{ig^{(1)}}(\boldsymbol{\xi}) > 0\Big\}$$
$$\leq G_1 \exp\Big\{-C_2\min(T\epsilon_\eta^2, \sqrt{T}\epsilon_\eta) + C_3 m + \log N_1 + \log N_2\Big\}.$$

by Lemma 13. Since $c_{\text{gap}} \gg T^{-1}(\log(N_1N_2)^2$, by the condition on η, we have $\eta < \{\tau^1_{\min}c_{\text{gap}}c_\pi/4 - \tau^1_{\max}CT^{-1}(\log(N_1N_2))^2\}/(\tau^1_{\min} + \tau^1_{\max})$. Hence, by the definition of ϵ_η, we have $\epsilon_\eta > \tau^1_{\min}c_{\text{gap}}c_\pi/4$ as $N_2, T \to \infty$. This leads to that

$$P\Big\{\sup_i I(\widehat{g}_i^{(1)} \notin \mathcal{A}^{(1)}_\eta(\boldsymbol{\xi}, g_i^{(1)0}, \mathcal{G}^0_2))\Big\} \leq C\exp\big\{-c_1\sqrt{T}c_{\text{gap}} + c_2 m + \log N_1 + \log N_2\big\},$$

where $m = p_1 + p_2 + 4$.

G.4.4 Technical Lemmas for Interactive Model

Lemma 13. *Under Assumptions 3–4 and A.11–A.13, we have*

$$P\Big\{\sup_{\|\boldsymbol{\Theta}_{ij}\|_{\max}\leq R}\Big|T^{-1}Q_{ij}(\boldsymbol{\Theta}_{ij}) - Q^*_{ij}(\boldsymbol{\Theta}_{ij})\Big| \geq x\Big\}$$
$$\leq C_1\exp\Big\{-C_2\min(Tx^2, \sqrt{T}x) + C_3 m\Big\} \tag{A.37}$$

Further recall that $S_{ij}(\boldsymbol{\Theta}_{ij}) = Q_{ij}(\boldsymbol{\Theta}_{ij}) - Q_{ij}(\boldsymbol{\Theta}^0_{ij})$, we have that

$$P\Big\{\sup_{\boldsymbol{\Theta}_{ij}} T^{-1}\frac{|S_{ij}(\boldsymbol{\Theta}_{ij}) - S^*_{ij}(\boldsymbol{\Theta}_{ij})|}{d_{ij}(\boldsymbol{\Theta}_{ij},\boldsymbol{\Theta}^0_{ij})} > x\Big\} \leq C_1\exp\Big\{-C_2\min(Tx^2,\sqrt{T}x) + C_3 m\Big\}. \tag{A.38}$$

with $m = p_1 + p_2 + 4$, and C_1, C_2, C_3 are positive constants, and

$$d_{ij}(\boldsymbol{\Theta}_{ij},\boldsymbol{\Theta}^0_{ij}) = \|\widehat{\boldsymbol{\theta}}^{(1)}_{\widehat{g}_i^{(1)}} - \boldsymbol{\theta}^{(1)}_{g_i^{(1)}}\|^2 + \|\widehat{\boldsymbol{\theta}}^{(2)}_{\widehat{g}_j^{(2)}} - \boldsymbol{\theta}^{(2)}_{g_j^{(2)}}\|^2$$
$$+ |\widehat{\alpha}_{\widehat{g}_i^{(1)}\widehat{g}_j^{(2)}} - \alpha_{g_i^{(1)}g_j^{(2)}}|^2 + |\widehat{\gamma}^{(1)}_{\widehat{g}_i^{(1)}}\widehat{\gamma}^{(2)}_{\widehat{g}_j^{(2)}} - \gamma^{(1)}_{g_i^{(1)}}\gamma^{(2)}_{g_j^{(2)}}|^2.$$

Proof

1. Proof of (A.37)**.**

Note that

$$T^{-1}Q_{ij}(\mathbf{\Theta}_{ij}) = T^{-1}\sum_t \left(\varepsilon_{ij,t} + \mathcal{X}_{ij,t}^\top \mathbf{\Theta}_{ij}^0 - \mathcal{X}_{ij,t}^\top \mathbf{\Theta}_{ij}\right)^2,$$

where $\mathcal{X}_{ij,t}$ is defined in (A.31). We use the similar procedure in the proof of Lemma 17. The three core steps in the proof of Lemma 17 are (A.68)–(A.70). While the implementation of (A.68) remains the same, the interactive network terms in $\mathcal{X}_{ij,t}$ and $\mathbf{\Theta}_{ij}$ cause changes in using (A.69) and (A.70). We take the change in (A.70) for example.

Specifically, the diagonal elements of $T^{-1}\sum_t \mathcal{X}_{ij,t}\mathcal{X}_{ij,t}^\top$ take the forms $T^{-1}\sum_t \mathbf{w}^\top \mathbb{Y}_t \mathbb{Y}_t^\top \mathbf{w}$, $T^{-1}\sum_t \|\mathbf{x}_{it}^{(1)}\|^2$, and $T^{-1}\sum_t \|\mathbf{x}_{jt}^{(2)}\|^2$. For the additive network term, $\mathbf{w}$ is replaced by $\mathbf{W}_{i\cdot}^{(1)\top} \otimes \mathbf{1}_{N_2} \in \mathbb{R}^{N_1\times N_2}$, while for the interactive network term, $\mathbf{w}$ is replaced by $\mathbf{W}_{i\cdot}^{(1)\top} \otimes \mathbf{W}_{\cdot j}^{(2)} \in \mathbb{R}^{N_1\times N_2}$. Both types of network terms satisfy that $\|\mathbf{w}\|_1 = 1$. Hence, for $T^{-1}\sum_t \mathbf{w}^\top \mathbb{Y}_t \mathbb{Y}_t^\top \mathbf{w}$, we apply Lemma 18 to prove the concentration inequality. For $T^{-1}\sum_t \|\mathbf{x}_{it}^{(1)}\|^2$, and $T^{-1}\sum_t \|\mathbf{x}_{jt}^{(2)}\|^2$, we use Lemma 18 and Assumptions 4.

Then, similar with the Step 2 in the proof of Lemma 17, by taking union upper bound for all $\|\mathbf{\Theta}_{ij}\|_{\max} \le R$, we obtain the final conclusion.

2. Proof of (A.38)**.**

Note that

$$\begin{aligned} S_{ij}(\mathbf{\Theta}_{ij}) - S_{ij}^*(\mathbf{\Theta}_{ij}) = &\left[(\mathbf{\Theta}_{ij} - \mathbf{\Theta}_{ij}^0)^\top \mathcal{X}_{ij,t}\mathcal{X}_{ij,t}^\top (\mathbf{\Theta}_{ij} - \mathbf{\Theta}_{ij}^0)\right] \\ &- E\left[(\mathbf{\Theta}_{ij} - \mathbf{\Theta}_{ij}^0)^\top \mathcal{X}_{ij,t}\mathcal{X}_{ij,t}^\top (\mathbf{\Theta}_{ij} - \mathbf{\Theta}_{ij}^0)\right] \qquad \text{(A.39)} \\ &+ 2\mathcal{X}_{ij,t}^\top (\mathbf{\Theta}_{ij} - \mathbf{\Theta}_{ij}^0)\varepsilon_{ij,t}. \qquad \text{(A.40)} \end{aligned}$$

We borrow the idea to prove Lemma 17 to derive the results. For instance, we can apply (A.66) on (A.39) by using the fact that $\|\mathbf{\Theta}_{ij} - \mathbf{\Theta}_{ij}^0\|/\sqrt{d_{ij}(\mathbf{\Theta}_{ij}, \mathbf{\Theta}_{ij}^0)} \le cR$ for some constant c. Similar techniques can be applied on proof of (A.40). ■

Lemma 14. *Under Assumptions 3–4 and A.11–A.13, for vector $\mathbf{\Delta} \in \mathbb{R}^{p_1+p_2+4}$ in the parameter space of the interactive model* (25)*, we have*

$$\begin{aligned} &P\Big\{\sup_i \sup_{\|\mathbf{\Delta}\|^2\le\omega^2} (N_2T)^{-1}\Big|\sum_j\sum_t \Big\{\mathbf{\Delta}^\top \mathcal{X}_{ij,t}\mathcal{X}_{ij,t}^\top \mathbf{\Delta} - \left[\mathbf{\Delta}^\top E(\mathcal{X}_{ij,t}\mathcal{X}_{ij,t}^\top)\mathbf{\Delta}\right]\Big\}\Big| \ge x\Big\} \\ &\le C_1 \exp\Big\{-C_2\min\Big(Tx^2/\omega^2, \sqrt{T}x/\omega\Big) + C_3 m + \log(N_1N_2)\Big\}, \\ &P\Big\{\sup_i \sup_{\|\mathbf{\Delta}\|^2\le\omega^2} (N_2T)^{-1}\Big|\sum_j\sum_t \mathcal{X}_{ij,t}^\top \mathbf{\Delta}\varepsilon_{ij,t}\Big| \ge x\Big\} \\ &\le C_1 \exp\Big\{-C_2\min\Big(Tx^2/\omega^2, \sqrt{T}x/\omega\Big) + C_3 m + \log(N_1N_2)\Big\}, \end{aligned}$$

where $\mathcal{X}_{ij,t}$ is defined in equation (A.31)*, $m = p_1 + p_2 + 4$, C_1, C_2, C_3 are positive constants.*

Proof The proof is similar with the proof of Lemma 20. The key difference lies in the expression of $\mathcal{X}_{ij,t}$. The technical details can refer to the proof of Lemma 13, which is omitted here. ■

Lemma 15. *Under Assumption A.12, it holds that*

$$\begin{aligned}
&\tau_{\min}^1 d_i(\mathbf{\Theta}_{i\cdot,1}, \mathbf{\Theta}_{i\cdot}^0) \le \frac{1}{N_2 T}\Big\{Q_i^*(\boldsymbol{\xi}_{g_i^{(1)}}^{(1)}; \boldsymbol{\xi}_{g^{(2)}}^{(2)}, \mathcal{G}_2) - Q_i^*(\boldsymbol{\xi}_{g_i^{(1)0}}^{(1)0}; \boldsymbol{\xi}_{g^{(2)0}}^{(2)0}, \mathcal{G}_2^0)\Big\} \\
&\le \tau_{\max}^1 d_i(\mathbf{\Theta}_{i\cdot}, \mathbf{\Theta}_{i\cdot}^0), \\
&\tau_{\min}^1 d(\mathbf{\Theta}, \mathbf{\Theta}^0) \le \frac{1}{N_2 T}\Big\{Q^*(\mathbf{\Theta}) - Q^*(\mathbf{\Theta}^0)\Big\} \le \tau_{\max}^1 d(\mathbf{\Theta}, \mathbf{\Theta}^0)
\end{aligned}$$

where $\tau_{\min}^1$ is defined in Assumption A.12, and $\tau_{\max}^1$ is defined in Lemma 16. Besides, $\mathbf{\Theta}_{i\cdot} = (\mathbf{\Theta}_{ij}^\top : j \in [N_2]) \in \mathbb{R}^{N_2 \times m}$, $\mathbf{\Theta}_{ij} = (\boldsymbol{\theta}_{g_i^{(1)}}^{(1)\top}, \boldsymbol{\theta}_{g_j^{(2)}}^{(2)\top}, \alpha_{g_i^{(1)} g_j^{(2)}}, \gamma_{g_i^{(1)}} \gamma_{g_j^{(2)}})^\top \in \mathbb{R}^m$, and

$$\begin{aligned}
d_i(\mathbf{\Theta}_{i\cdot}, \mathbf{\Theta}_{i\cdot}^0) &= N_2^{-1} \sum_j \|\mathbf{\Theta}_{ij} - \mathbf{\Theta}_{ij}^0\|^2 \\
&= \|\boldsymbol{\theta}_{g_i^{(1)}}^{(1)} - \boldsymbol{\theta}_{g_i^{(1)0}}^{(1)0}\|^2 + N_2^{-1} \sum_j \|\boldsymbol{\theta}_{g_j^{(2)}}^{(2)} - \boldsymbol{\theta}_{g_j^{(2)0}}^{(2)0}\|^2 \\
&\quad + N_2^{-1} \sum_j \Big\{\big|\alpha_{g_i^{(1)} g_j^{(2)}} - \alpha_{g_i^{(1)0} g_j^{(2)0}}^0\big|^2 + \big|\gamma_{g_i^{(1)}}^{(1)} \gamma_{g_j^{(2)}}^{(2)} - \gamma_{g_i^{(1)0}}^{(1)0} \gamma_{g_j^{(2)0}}^{(2)0}\big|^2\Big\}.
\end{aligned}$$

Proof By the definition of loss function $Q_i(\boldsymbol{\xi}_{g_i^{(1)}}^{(1)}; \boldsymbol{\xi}_{g^{(2)}}^{(2)}, \mathcal{G}_2)$ in (A.34), we know $Q_i^*(\boldsymbol{\xi}_{g_i^{(1)}}^{(1)}; \boldsymbol{\xi}_{g^{(2)}}^{(2)}, \mathcal{G}_2) = \sum_j Q_{ij}^*(\mathbf{\Theta}_{ij})$ and $Q_i^*(\boldsymbol{\xi}_{g_i^{(1)0}}^{(1)0}; \boldsymbol{\xi}_{g^{(2)0}}^{(2)0}, \mathcal{G}_2^0) = \sum_j Q_{ij}^*(\mathbf{\Theta}_{ij}^0)$ by following the similar derivation in (A.1). Since $Q_{ij}(\mathbf{\Theta}_{ij}) = \sum_t (\varepsilon_{ij,t} + \mathcal{X}_{ij,t}^\top \mathbf{\Theta}_{ij}^0 - \mathcal{X}_{ij,t}^\top \mathbf{\Theta}_{ij})^2$, where $\mathcal{X}_{ij,t}$ is defined in (A.31), we can use similar techniques in the proof of Lemma 23 to obtain the conclusion. ■

Lemma 16. *Under Assumption 4 and A.13, we have $\tau_{\max}^1 = \max_{i,j} \lambda_{\max}(E(\mathcal{X}_{ij,t} \mathcal{X}_{ij,t}^\top)) < \infty$ with $\mathcal{X}_{ij,t}$ defined in* (A.31)*.*

Proof We use the similar techniques in the proof of Lemma 33. Due to the interactive network term, the different term in $\mathcal{X}_{ij,t}$ is $(\mathbf{W}_{\cdot j}^{(2)} \otimes \mathbf{W}_{i\cdot}^{(1)\top})^\top \mathbb{Y}_t$, with $\|\mathbf{W}_{\cdot j}^{(2)} \otimes \mathbf{W}_{i\cdot}^{(1)\top}\|_1 = 1$. Denote $\mathbf{w} = \mathbf{W}_{\cdot j}^{(2)} \otimes \mathbf{W}_{i\cdot}^{(1)\top}$, this typical term has the property that $\mathrm{var}(\mathbf{w}^\top \mathbb{Y}_t) = \mathbf{w}^\top \mathbf{\Gamma} \mathbf{w} \le \|\mathbf{\Gamma}\|_{\max} |\mathbf{w}^\top \mathbf{1}\mathbf{1}^\top \mathbf{w}| = \|\mathbf{\Gamma}\|_{\max} < c_\Gamma$, where c_Γ is the upper bound conclusion in Lemma 32. Other terms are the same as in the Lemma 33. ■

Appendix H. Discussion about Future Extension

We remark that the GTNAR model can be easily extended to a general-purpose machine learning model by considering nonlinear terms. Take $q = 2$ for example, model (1) can be

extended to

$$Y_{ij,t} = \lambda^{(1)}_{g_i^{(1)}} \sum_{k=1}^{N_1} \frac{a^{(1)}_{ik}}{n_{1i}} Y_{kj,(t-1)} + \lambda^{(2)}_{g_j^{(2)}} \sum_{k=1}^{N_2} \frac{a^{(2)}_{kj}}{n_{2j}} Y_{ik,(t-1)} \tag{A.41}$$

$$+ \alpha_{g_i^{(1)} g_j^{(2)}} Y_{ij,(t-1)} + f_1(\mathbf{x}^{(1)}_{it}) + f_2(\mathbf{x}^{(2)}_{jt}) + \varepsilon_{ij,t}, \tag{A.42}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are non-linear functions. To estimate model (A.42), we could conduct the following (1) and (2) iteratively.

(1) Estimate the linear part. First, given the non-linear part $f(\mathbf{x}^{(1)}_{it})$ and $f(\mathbf{x}^{(2)}_{jt})$, calculate $\widetilde{Y}_{ij,t} = Y_{ij,t} - f_1(\mathbf{x}^{(1)}_{it}) - f_2(\mathbf{x}^{(2)}_{jt})$. Then we implement Algorithm A.2 to estimate the other parameters and group memberships iteratively.

(2) Estimate the non-linear part. Given the parameters $\lambda^{(1)}_{g_i^{(1)}}$, $\lambda^{(2)}_{g_j^{(2)}}$, and $\alpha_{g_i^{(1)} g_j^{(2)}}$, calculate $\breve{Y}_{ij,t} = Y_{ij,t} - \lambda^{(1)}_{g_i^{(1)}} \sum_{k=1}^{N_1} \frac{a^{(1)}_{ik}}{n_{1i}} Y_{kj,(t-1)} - \lambda^{(2)}_{g_j^{(2)}} \sum_{k=1}^{N_2} \frac{a^{(2)}_{kj}}{n_{2j}} Y_{ik,(t-1)} - \alpha_{g_i^{(1)} g_j^{(2)}} Y_{ij,(t-1)}$. Subsequently, one can use deep learning model to fit the function $f_1(\cdot)$ and $f_2(\cdot)$.

Since the discussion of theoretical guarantees is beyond our scope, we leave it as an interesting future topic.

Appendix I. Proof of Main Theorems

I.1 Proof of Theorem 2

Recall that $d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta})$ is defined in (18). To prove $d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}^0) = O_p(a^2_{NT})$, it suffices to show that

$$P\Big\{ a^{-1}_{NT} \sqrt{d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}^0)} > K \Big\} \to 0 \text{ as } K \to \infty.$$

Denote $\mathbf{\Omega}$ as the parameter space, and partition the space into shells $\mathbf{\Omega}_j = \{\mathbf{\Theta} \in \mathbf{\Omega} : j - 1 < a^{-1}_{NT} \sqrt{d(\mathbf{\Theta}, \mathbf{\Theta}^0)} \le j\}$, where $j \ge 1$ takes over all positive integers. For a positive integer K, if $a^{-1}_{NT} \sqrt{d(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}^0)} \ge K$, then $\widehat{\mathbf{\Theta}} \in \mathbf{\Omega}_j$ with $j \ge K$. Since $\widehat{\mathbf{\Theta}}$ minimizes objective function $Q(\mathbf{\Theta})$, we have $(\prod_l N_l T)^{-1} S(\widehat{\mathbf{\Theta}}) \overset{\text{def}}{=} (\prod_l N_l T)^{-1} Q(\widehat{\mathbf{\Theta}}) - (\prod_l N_l T)^{-1} Q(\mathbf{\Theta}^0) \le 0$. Denote $S^*(\mathbf{\Theta}) = E\{S(\mathbf{\Theta})\}$. Then we have

$$\begin{aligned}
P\Big\{ a^{-1}_{NT} \sqrt{d(\mathbf{\Theta}, \widehat{\mathbf{\Theta}})} > K \Big\} &\le P\Big\{ \inf_{\mathbf{\Theta} \in \bigcup_{j \ge K} \mathbf{\Omega}_j} (\prod_l N_l T)^{-1} S(\mathbf{\Theta}) \le 0 \Big\} \\
&\le \sum_{j \ge K} P\Big\{ \inf_{\mathbf{\Theta} \in \mathbf{\Omega}_j} (\prod_l N_l T)^{-1} S(\mathbf{\Theta}) \le 0 \Big\} \\
&\le \sum_{j \ge K} P\Big\{ \inf_{\mathbf{\Theta} \in \mathbf{\Omega}_j} (\prod_l N_l T)^{-1} \{S(\mathbf{\Theta}) - S^*(\mathbf{\Theta})\} \\
&\quad + \inf_{\mathbf{\Theta} \in \mathbf{\Omega}_j} (\prod_l N_l T)^{-1} S^*(\mathbf{\Theta}) \le 0 \Big\}
\end{aligned}$$

By Lemma 23, we have that $(\prod_l N_l T)^{-1} S^*(\mathbf{\Theta}) = (\prod_l N_l T)^{-1}\{Q^*(\mathbf{\Theta}) - Q^*(\mathbf{\Theta}^0)\} \geq \tau_{\min} d(\mathbf{\Theta}, \mathbf{\Theta}^0)$ with $\tau_{\min}$ defined in Assumption 2. Since $d(\mathbf{\Theta}, \mathbf{\Theta}^0) > (j-1)^2 a_{NT}^2$ for all $\mathbf{\Theta} \in \mathbf{\Omega}_j$, we have

$$
\begin{aligned}
P\Big\{ a_{NT}^{-1} \sqrt{d(\mathbf{\Theta}, \widehat{\mathbf{\Theta}})} > K \Big\} &\leq \sum_{j \geq K} P\Big\{ \inf_{\mathbf{\Theta} \in \mathbf{\Omega}_j} (\prod_l N_l T)^{-1} \{ S(\mathbf{\Theta}) - S^*(\mathbf{\Theta}) \} \leq -\tau_{\min} (j-1)^2 a_{NT}^2 \Big\} \\
&\leq \sum_{j \geq K} P\Big\{ \sup_{\mathbf{\Theta} \in \mathbf{\Omega}_j} (\prod_l N_l T)^{-1} \big| S(\mathbf{\Theta}) - S^*(\mathbf{\Theta}) \big| \geq \tau_{\min} (j-1)^2 a_{NT}^2 \Big\}.
\end{aligned} \tag{A.43}
$$

Denote $S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) = Q_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - Q_{i_1 \cdots i_q}(\mathbf{\Theta}^0_{i_1 \cdots i_q})$, where $Q_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q})$ is defined in (16) in the main text. For simplicity we denote $d_{i_1 \cdots i_q} = d_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}, \mathbf{\Theta}^0_{i_1 \cdots i_q})$. We next show that

$$
\begin{aligned}
&P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} (\prod_l N_l T)^{-1} \big| S(\mathbf{\Theta}) - S^*(\mathbf{\Theta}) \big| \geq x \Big\} \\
&\leq C_1 \exp\Big\{ - C_2 \min(T x^2/\omega^2, \sqrt{T} x/\omega) + C_3 m + \log\big(\prod_l N_l \big) \Big\}.
\end{aligned}
$$

We have

$$
\begin{aligned}
&P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} (\prod_l N_l T)^{-1} \big| S(\mathbf{\Theta}) - S^*(\mathbf{\Theta}) \big| \geq x \Big\} \\
&\leq P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} (\prod_l N_l)^{-1} \sum_l \sum_{i_l} T^{-1} \big| S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - S^*_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) \big| \geq x \Big\} \\
&= P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} \frac{1}{\prod_l N_l} \sum_l \sum_{i_l} \frac{T^{-1} \big| S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - S^*_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) \big|}{\sqrt{d_{i_1 \cdots i_q}}} \sqrt{d_{i_1 \cdots i_q}} \geq x \Big\} \\
&\leq P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} \sqrt{\frac{1}{\prod_l N_l} \sum_l \sum_{i_l} \frac{T^{-2} \big| S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - S^*_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) \big|^2}{d_{i_1 \cdots i_q}}} \sqrt{d_{i_1 \cdots i_q}} \geq x \Big\} \\
&= P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} \frac{1}{\prod_l N_l} \sum_l \sum_{i_l} \frac{T^{-2} \big| S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - S^*_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) \big|^2}{d_{i_1 \cdots i_q}} d_{i_1 \cdots i_q} \geq x^2 \Big\} \\
&\leq P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} \frac{1}{\prod_l N_l} \sum_l \sum_{i_l} \frac{T^{-2} \big| S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - S^*_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) \big|^2}{d_{i_1 \cdots i_q}} \omega^2 \geq x^2 \Big\} \\
&\leq \sum_l \sum_{i_l} P\Big\{ \sup_{d(\mathbf{\Theta}, \mathbf{\Theta}^0) \leq \omega^2} \frac{T^{-2} \big| S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - S^*_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) \big|^2}{d_{i_1 \cdots i_q}} \omega^2 \geq x^2 \Big\} \\
&\leq \sum_l \sum_{i_l} P\Big\{ \sup_{\mathbf{\Theta}_{i_1 \cdots i_q}} \frac{T^{-1} \big| S_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) - S^*_{i_1 \cdots i_q}(\mathbf{\Theta}_{i_1 \cdots i_q}) \big|}{\sqrt{d_{i_1 \cdots i_q}}} \omega \geq x \Big\} \\
&\leq C_1 \exp\Big\{ - C_2 \min(T x^2/\omega^2, \sqrt{T} x/\omega) + C_3 m + \log\big(\prod_l N_l \big) \Big\},
\end{aligned}
$$

where the third inequality holds due to $\sum_i (a_i^2 b_i^2) \le (\sum_i a_i^2)(\sum_i b_i^2)$ and the last inequality holds due to Lemma 19. Here $m = \sum_l (p_l + 1) + 1$ is a fixed constant. Then plugging the last line into (A.43), we have

$$
\begin{aligned}
&P\Big\{a_{NT}^{-1}\sqrt{d(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}})} > K\Big\} \le \sum_{j\ge K} P\Big\{ \sup_{\boldsymbol{\Theta}\in\boldsymbol{\Omega}_j} (\prod_l N_l T)^{-1} \big|S(\boldsymbol{\Theta}) - S^*(\boldsymbol{\Theta})\big| \ge \tau_{\min}(j-1)^2 a_{NT}^2\Big\} \\
&\le \sum_{j\ge K} C_1 \exp\Big\{ - C_2 \min\Big(\frac{T(j-1)^4 a_{NT}^2}{j^2}, \frac{\sqrt{T}(j-1)^2 a_{NT}}{j}\Big) + C_3 m + \log\big(\prod_l N_l\big)\Big\} \\
&\le \sum_{j\ge K} C_1 \exp\Big\{ - C_2 \min\Big(\frac{(j-1)^4 \log^2(\prod_l N_l)}{j^2}, \frac{(j-1)^2 \log(\prod_l N_l)}{j}\Big) + C_3 m + \log\big(\prod_l N_l\big)\Big\} \\
&= \sum_{j\ge K} C_1 \exp\Big\{ - C_2 \frac{(j-1)^2 \log(\prod_l N_l)}{j} + C_3 m + \log\big(\prod_l N_l\big)\Big\} \\
&\le \sum_{j\ge K} C_1 \exp\Big\{ - C_2 (j-1)(1-K^{-1}) \log(\prod_l N_l) + C_3 m + \log\big(\prod_l N_l\big)\Big\} \\
&= \frac{\exp\big\{ - C_2 (K-1)(1-K^{-1}) \log(\prod_l N_l) + C_3 m + \log\big(\prod_l N_l\big)\big\}}{1 - \exp\big\{ - C_2 (1-K^{-1}) \log(\prod_l N_l)\big\}} \\
&\le \frac{\exp(-C_2 (K-1)(1-K^{-1}) - 1) \log(\prod_l N_l) + C_3 m}{1 - \exp(-C_2 \log(\prod_l N_l))} \to 0
\end{aligned} \tag{A.44}
$$

when $K \to \infty$. The third inequality holds by plugging in the expression of $a_{NT}^2 = \log^2(\prod_l N_l) T^{-1}$. The equality holds due to the sum of a geometric sequence. Then we finish the proof.

I.2 Proof of Theorem 3

In the following proof we write $Q(\widehat{\boldsymbol{\xi}}(\underline{G}), \widehat{\mathcal{G}}(\underline{G}))$ as $Q(\widehat{\boldsymbol{\Theta}}(\underline{G}))$ as we defined in (16). We use the notations here to denote $\widehat{\boldsymbol{\Theta}}(\underline{G})$ as the estimate when group numbers $G_1, \cdots, G_q$ are specified. Define $G_{-l} = (G_k : k \ne l)^\top \in \mathbb{R}^{q-1}$ and we use $G_{-l} \ge G_{-l,0}$ to denote that $G_k \ge G_{k,0}$ for $k \ne l$. To establish the consistency property, we consider both the underfitted model (there exists an l with $G_l < G_{l,0}$) and the overfitted model ($\{G_l > G_{l,0}, G_{-l} \ge G_{-l,0}\}$ for any l). In the following we show $\mathrm{QIC}(\underline{G}_0) < \mathrm{QIC}(\underline{G})$ with probability tending to 1 under both circumstances. We write the difference of the two criteria as

$$
\begin{aligned}
\mathrm{QIC}(\underline{G}) - \mathrm{QIC}(\underline{G}_0) &= \log\{Q(\widehat{\boldsymbol{\Theta}}(\underline{G}))\} - \log\{Q(\widehat{\boldsymbol{\Theta}}(\underline{G}_0))\} + \lambda(\underline{G}) - \lambda(\underline{G}_0) \\
&= \log\Big\{1 + \frac{Q(\widehat{\boldsymbol{\Theta}}(\underline{G})) - Q(\widehat{\boldsymbol{\Theta}}(\underline{G}_0))}{Q(\widehat{\boldsymbol{\Theta}}(\underline{G}_0))}\Big\} + \lambda(\underline{G}) - \lambda(\underline{G}_0).
\end{aligned}
$$

1. Overfitted Model. In the following we show that $(\prod_l N_l T)^{-1}\{Q(\widehat{\boldsymbol{\Theta}}(\underline{G}_0)) - Q(\widehat{\boldsymbol{\Theta}}(\underline{G}))\} = (\prod_l N_l T)^{-1}\{[Q(\widehat{\boldsymbol{\Theta}}(\underline{G}_0)) - Q(\boldsymbol{\Theta}^0)] - [Q(\widehat{\boldsymbol{\Theta}}(\underline{G})) - Q(\boldsymbol{\Theta}^0)]\} = (\prod_l N_l T)^{-1}\{S(\widehat{\boldsymbol{\Theta}}(G_0)) - S(\widehat{\boldsymbol{\Theta}}(\underline{G}))\} = O_p(a_{NT})$, where $a_{NT} = T^{-1}(\sum_l \log N_l)^2$. Note that we have

$$
\frac{1}{(\prod_l N_l) T}\big|S(\widehat{\boldsymbol{\Theta}}(\underline{G})) - S(\widehat{\boldsymbol{\Theta}}(\underline{G}_0))\big| \le \frac{1}{(\prod_l N_l) T}\big|S(\widehat{\boldsymbol{\Theta}}(\underline{G}_0)) - S^*(\widehat{\boldsymbol{\Theta}}(\underline{G}_0))\big|
$$

$$+\frac{1}{(\prod_l N_l)T}\big|S^*(\widehat{\mathbf{\Theta}}(\underline{G}_0))-S^*(\widehat{\mathbf{\Theta}}(\underline{G}))\big|+\frac{1}{(\prod_l N_l)T}\big|S(\widehat{\mathbf{\Theta}}(\underline{G}))-S^*(\widehat{\mathbf{\Theta}}(\underline{G}))\big|$$
$$=O_p(T^{-1}(\sum_l \log N_l)^2)+\frac{1}{(\prod_l N_l)T}\big|S^*(\widehat{\mathbf{\Theta}}(\underline{G}_0))-S^*(\widehat{\mathbf{\Theta}}(\underline{G}))\big|$$

by Lemma 19.

Further by Lemma 23 and Theorem 2, it holds that

$$\frac{1}{(\prod_l N_l)T}\big|S^*(\widehat{\mathbf{\Theta}}(\underline{G}_0))-S^*(\widehat{\mathbf{\Theta}}(\underline{G}))\big|$$
$$\le\frac{1}{(\prod_l N_l)T}\big|Q^*(\widehat{\mathbf{\Theta}}(\underline{G}_0))-Q^*(\mathbf{\Theta}^0)\big|+\frac{1}{(\prod_l N_l)T}\big|Q^*(\mathbf{\Theta}^0)-Q^*(\widehat{\mathbf{\Theta}}(\underline{G}))\big|$$
$$\le\tau_{\max}\Big\{d(\widehat{\mathbf{\Theta}}(\underline{G}_0),\mathbf{\Theta}^0)+d(\widehat{\mathbf{\Theta}}(\underline{G}),\mathbf{\Theta}^0)\Big\}=O_p(T^{-1}(\sum_l \log N_l)^2),$$

where $\tau_{\max}$ is defined in Assumption 2. Since m is a fixed constant, $((\prod_l N_l)T)^{-1}\big|Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))-Q(\widehat{\mathbf{\Theta}}(\underline{G}))\big|=O_p\{T^{-1}(\sum_l \log N_l)^2\}$, i.e.,

$$\frac{1}{(\prod_l N_l)T}Q(\widehat{\mathbf{\Theta}}(\underline{G}))=\frac{1}{(\prod_l N_l)T}Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))+O_p\{T^{-1}(\sum_l \log N_l)^2\}.$$

Next we show $((\prod_l N_l)T)^{-1}Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))=\sigma^2+o_p(1)$. Similar to above while replacing $\widehat{\mathbf{\Theta}}(\underline{G})$ by $\mathbf{\Theta}^0$, we have

$$((\prod_l N_l)T)^{-1}Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))=((\prod_l N_l)T)^{-1}Q^*(\mathbf{\Theta}^0)+O_p\{T^{-1}(\sum_l \log N_l)^2\}. \tag{A.45}$$

Here we have $((\prod_l N_l)T)^{-1}Q^*(\mathbf{\Theta}^0)=\sigma^2>0$. This shows that $((\prod_l N_l)T)^{-1}Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))=\sigma^2+o_p(1)$. It implies $\mathrm{QIC}(\underline{G})-\mathrm{QIC}(\underline{G}_0)=O_p\{T^{-1}(\sum_l \log N_l)^2\}+\lambda(\underline{G})-\lambda(\underline{G}_0)$. We can verify that $\lambda(\underline{G})-\lambda(\underline{G}_0)\ge\kappa$ under the overfitting case, where $\kappa=\lambda(\underline{G})/(\sum_l G_l)$. Further note that $\kappa\gg T^{-1}(\sum_l \log N_l)^2$ as we assume, this implies $\mathrm{QIC}(\underline{G}_0)<\mathrm{QIC}(\underline{G})$ with probability tending to 1 under the overfitting case.

2. Underfitted Model. From the specification of $\lambda(\underline{G})$, we have $|\lambda(\underline{G})-\lambda(\underline{G}_0)|=o(c_{\mathrm{gap}}/(\prod_l G_l))$. It suffices to show that $((\prod_l N_l)T)^{-1}\{Q(\widehat{\mathbf{\Theta}}(\underline{G}))-Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))\}\ge C_1 c_{\mathrm{gap}}(c_\pi)^q/(\prod_l G_l)$ for a positive constant C_1 when $N_l\to\infty$. Note that

$$((\prod_l N_l)T)^{-1}\{Q(\widehat{\mathbf{\Theta}}(\underline{G}))-Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))\}$$
$$=((\prod_l N_l)T)^{-1}\{Q(\widehat{\mathbf{\Theta}}(\underline{G}))-Q^*(\mathbf{\Theta}^0)-Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))+Q^*(\mathbf{\Theta}^0)\},$$

and we have already proved that $((\prod_l N_l)T)^{-1}\{Q(\widehat{\mathbf{\Theta}}(\underline{G}_0))-Q^*(\mathbf{\Theta}^0)\}=O_p(T^{-1}(\sum_l \log N_l)^2)=o_p\{c_{\mathrm{gap}}/(\prod_l G_l)\}$ in (A.45). Also note that c_π is a constant according to Assumption 7. Hence it suffices to show $((\prod_l N_l)T)^{-1}\{Q(\widehat{\mathbf{\Theta}}(\underline{G}))-Q^*(\mathbf{\Theta}^0)\}\ge C c_{\mathrm{gap}}(c_\pi)^q/(\prod_l G_l)$ for some positive constant C when $N_l\to\infty$.

Without loss of generality, we consider $G_l < G_{l,0}$. To prove the result, we use two steps. First, we show that by Assumption 6, we have

$$\max_{g^{(l)}\in[G_{l,0}]}\max_{g^{-(l)}\in[G^0_{-l}]}\min_{g^{(l)\prime}\in[G_l]}\Big\{\|\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2+|\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})} - \alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\Big\}\geq c_{\text{gap}}/4, \quad \text{(A.46)}$$

where $\varphi_l(g^{(l)})=\text{argmax}_{g^{(l)\prime}\in[G_l]}\sum_{i_l}I(g^{(l)0}_{i_l}=g^{(l)},\widehat{g}^{(l)}_{i_l}=g^{(l)\prime})$, and $\{g^{-(l)}\in[G^0_{-l}]\}$ denotes $\{g^{(1)}\in[G_{1,0}],\cdots,g^{(l-1)}\in[G_{l-1,0}],g^{(l+1)}\in[G_{l+1,0}],\cdots,g^{(q)}\in[G_{q,0}]\}$. We prove (A.46) by contradiction. Assume (A.46) does not hold. Define $\sigma_l:[G_{l,0}]\to[G_l]$, such that

$$\begin{aligned}\sigma_l(g^{(l)})=\text{argmin}_{g^{(l)\prime}\in[G_l]}\Big\{&\|\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2\\&+\max_{g^{-(l)}\in[G^0_{-l}]}|\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}\\&-\alpha^0_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}|^2\Big\}.\end{aligned}$$

Since $G_{l,0}>G_l$, there exists at least two $g^{(l)}_1,g^{(l)}_2\in[G_{l,0}]$, such that $\sigma_l(g^{(l)}_1)=\sigma_l(g^{(l)}_2)$. Then we have for all $g^{-(l)}\in[G^0_{-l}]$,

$$\begin{aligned}&\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}_1}-\boldsymbol{\theta}^{(l)0}_{g^{(l)}_2}\|^2+|\alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}_1g^{(l+1)}\cdots g^{(q)}}-\alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}_2g^{(l+1)}\cdots g^{(q)}}|^2\\&\leq 2\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}_1}-\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)}_1)}\|^2+2\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}_2}-\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)}_2)}\|^2\\&+2\Big\{|\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})\sigma_l(g^{(l)}_1)\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}-\alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}_1g^{(l+1)}\cdots g^{(q)}}|^2\\&+|\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})\sigma_l(g^{(l)}_2)\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}-\alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}_2g^{(l+1)}\cdots g^{(q)}}|^2\Big\}\\&<c_{\text{gap}}/2+c_{\text{gap}}/2=c_{\text{gap}},\end{aligned}\quad \text{(A.47)}$$

which contradicts Assumption 6. We explain the last inequality as follows. For any $g^{(l)}\in[G_{l,0}],g^{-(l)}\in[G^0_{-l}],g^{(l)\prime}\in[G_l]$, we have

$$\begin{aligned}&\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}}-\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)})}\|^2+|\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})\sigma_l(g^{(l)})\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}-\alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\\&\leq\max_{g^{-(l)}\in[G^0_{-l}]}\Big\{\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}}-\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)})}\|^2\\&+|\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})\sigma_l(g^{(l)})\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}-\alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\Big\}\\&\leq\max_{g^{-(l)}\in[G^0_{-l}]}\Big\{\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}}-\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}}\|^2\\&+|\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}-\alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\Big\}\\&\leq\max_{g^{(l)}\in[G_{l,0}]}\max_{g^{-(l)}\in[G^0_{-l}]}\Big\{\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}}-\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}}\|^2\end{aligned}$$

$$+ |\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})} - \alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\Big\},$$

where the second inequality holds by the definition of $\sigma_l(g^{(l)})$. Since the above inequality holds for any $g^{(l)\prime}$, hence by setting

$$g^{(l)\prime} = \text{argmin}_{g^{(l)\prime}\in[G_l]}\Big\{\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}} - \widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}}\|^2 + |\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})} - \alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\Big\},$$

we get

$$\begin{aligned}&\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}} - \widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)})}\|^2 + |\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})\sigma_l(g^{(l)})\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})} - \alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\\ &\le \max_{g^{(l)}\in[G_{l,0}]}\max_{g^{-(l)}\in[G^0_{-l}]}\min_{g^{(l)\prime}\in[G_l]}\Big\{\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}} - \widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}}\|^2 + |\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})}\\ &\qquad - \alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\Big\} < c_{\text{gap}}/4.\end{aligned}$$

Hence, by the contradiction (A.47), we could obtain that (A.46) holds.

Subsequently, define the mapping $\varphi_l(g^{(l)}) = \text{argmax}_{g^{(l)\prime}\in[G_l]}\sum_{i_l} I(g^{(l)0}_{i_l} = g^{(l)}, \widehat{g}^{(l)}_{i_l} = g^{(l)\prime})$. In addition, let

$$\begin{aligned}&(g^{(1)*},\cdots,g^{(q)*}) = \text{argmax}_{g^{(1)}\in[G_{1,0}],\cdots,g^{(q)}\in[G_{q,0}]}\Big[\min_{g^{(l)\prime}\in[G_l]}\Big\{\|\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2\\ &+ |\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})} - \alpha^0_{g^{(1)}\cdots g^{(l-1)}g^{(l)}g^{(l+1)}\cdots g^{(q)}}|^2\Big\}\Big].\end{aligned} \tag{A.48}$$

In this case, we have

$$\begin{aligned}&d(\widehat{\boldsymbol{\Theta}}(\underline{G}), \boldsymbol{\Theta}^0) = \frac{1}{\prod_l N_l}\sum_{i_1=1}^{N_1}\cdots\sum_{i_q=1}^{N_q}\Big\|\widehat{\boldsymbol{\Theta}}_{i_1\cdots i_q} - \boldsymbol{\Theta}_{i_1\cdots i_q}\Big\|^2\\ &= \sum_l\frac{1}{N_l}\sum_{i_l=1}^{N_l}\sum_{g^{(l)\prime}=1}^{G_l}\sum_{g^{(l)}=1}^{G_{l,0}} I(\widehat{g}^{(l)}_{i_l} = g^{(l)\prime}, g^{(l)0}_{i_l} = g^{(l)})\|\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2\\ &+ \frac{1}{\prod_l N_l}\sum_{i_1,\cdots,i_q}\sum_{g^{(1)\prime}=1}^{G_1}\cdots\sum_{g^{(q)\prime}=1}^{G_q}\sum_{g^{(1)}=1}^{G_{1,0}}\cdots\sum_{g^{(q)}=1}^{G_{q,0}}\\ &I(\widehat{g}^{(1)}_{i_1} = g^{(1)\prime},\cdots,\widehat{g}^{(q)}_{i_q} = g^{(q)\prime}, g^{(1)0}_{i_1} = g^{(1)},\cdots,g^{(q)0}_{i_q} = g^{(q)})|\widehat{\alpha}_{g^{(1)\prime}\cdots g^{(q)\prime}} - \alpha^0_{g^{(1)}\cdots g^{(q)}}|^2\\ &\ge \sum_l\frac{1}{N_l}\sum_{i_l=1}^{N_l} I(\widehat{g}^{(l)}_{i_l} = \varphi_l(g^{(l)*}), g^{(l)0}_{i_l} = g^{(l)*})\|\widehat{\boldsymbol{\theta}}^{(l)}_{\varphi_l(g^{(l)*})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)*}}\|^2\\ &+ \frac{1}{\prod_l N_l}\sum_{i_1,\cdots,i_q} I(\widehat{g}^{(1)}_{i_1} = \varphi_1(g^{(1)*}),\cdots,\widehat{g}^{(q)}_{i_q} = \varphi_q(g^{(q)*}), g^{(1)0}_{i_1} = g^{(1)*},\cdots,g^{(q)0}_{i_q} = g^{(q)*})\\ &|\widehat{\alpha}_{\varphi_1(g^{(1)*})\cdots\varphi(g^{(q)*})} - \alpha^0_{g^{(1)*}\cdots g^{(q)*}}|^2.\end{aligned}$$

By (A.46), we further have

$$
\begin{aligned}
&d(\widehat{\boldsymbol{\Theta}}(\underline{G}), \boldsymbol{\Theta}^0) \\
&\geq \frac{1}{\prod_l N_l} \sum_{i_1,\cdots,i_q} I(\widehat{g}_{i_1}^{(1)} = \varphi_1(g^{(1)*}), \cdots, \widehat{g}_{i_q}^{(q)} = \varphi_q(g^{(q)*}), g_{i_1}^{(1)0} = g^{(1)*}, \cdots, g_{i_q}^{(q)0} = g^{(q)*}) \\
&\quad \|\widehat{\boldsymbol{\theta}}^{(l)}_{\varphi_l(g^{(l)*})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)*}}\|^2 \\
&+ \frac{1}{\prod_l N_l} \sum_{i_1,\cdots,i_q} I(\widehat{g}_{i_1}^{(1)} = \varphi_1(g^{(1)*}), \cdots, \widehat{g}_{i_q}^{(q)} = \varphi_q(g^{(q)*}), g_{i_1}^{(1)0} = g^{(1)*}, \cdots, g_{i_q}^{(q)0} = g^{(q)*}) \\
&|\widehat{\alpha}_{\varphi_1(g^{(1)*})\cdots\varphi(g^{(q)*})} - \alpha^0_{g^{(1)*}\cdots g^{(q)*}}|^2 \\
&\geq \frac{c_{\text{gap}}}{4\prod_l N_l} \sum_{i_1,\cdots,i_q} I(\widehat{g}_{i_1}^{(1)} = \varphi_1(g^{(1)*}), \cdots, \widehat{g}_{i_q}^{(q)} = \varphi_q(g^{(q)*}), g_{i_1}^{(1)0} = g^{(1)*}, \cdots, g_{i_q}^{(q)0} = g^{(q)*}) \\
&= \frac{c_{\text{gap}}}{4\prod_l N_l} \sum_{l=1}^{q} \sum_{i_l=1}^{N_l} I\{\widehat{g}_{i_l}^{(l)} = \varphi_l(g^{(l)*}), g_{i_l}^{(l)0} = g^{(l)*}\}.
\end{aligned}
$$

where the last inequality is due to that

$$
\begin{aligned}
&\|\widehat{\boldsymbol{\theta}}^{(l)}_{\varphi_l(g^{(l)*})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)*}}\|^2 + |\widehat{\alpha}_{\varphi_1(g^{(1)*})\cdots\varphi(g^{(q)*})} - \alpha^0_{g^{(1)*}\cdots g^{(q)*}}|^2 \\
&\geq \min_{g^{(l)\prime}\in[G_l]} \Big\{ \|\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)*}}\|^2 + |\widehat{\alpha}_{\varphi_1(g^{(1)*})\cdots\varphi_{l-1}(g^{(l-1)*})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)*})\cdots\varphi_q(g^{(q)*})} - \alpha^0_{g^{(1)*}\cdots g^{(q)*}}|^2 \Big\} \\
&= \max_{g^{(l)}\in[G_{l,0}]} \max_{g^{-(l)}\in[G^0_{-l}]} \min_{g^{(l)\prime}\in[G_l]} \Big\{ \|\widehat{\boldsymbol{\theta}}^{(l)}_{g^{(l)\prime}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2 \\
&+ |\widehat{\alpha}_{\varphi_1(g^{(1)})\cdots\varphi_{l-1}(g^{(l-1)})g^{(l)\prime}\varphi_{l+1}(g^{(l+1)})\cdots\varphi_q(g^{(q)})} - \alpha^0_{g^{(1)}\cdots g^{(q)}}|^2 \Big\} \geq \frac{c_{\text{gap}}}{4}
\end{aligned}
$$

by the definition (A.48) and (A.46).

By the definition of $\varphi_l(\cdot)$, we have

$$
N_l^{-1} \sum_{i_l} I(\widehat{g}_{i_l}^{(l)} = \varphi_l(g^{(l)*}), g_{i_l}^{(l)0} = g^{(l)*}) \geq N_l^{-1} G_l^{-1} \sum_{i_l} I(g_{i_l}^{(l)0} = g^{(l)*}) = G_l^{-1} \pi^{(l)}_{g^{(l)*}, N_l}
$$

From Assumption 7, we know that there exists $M > 0$, such that for $N_l > M$ it holds $G_l^{-1}\pi^{(l)}_{g^{(l)*},N_l} > c_\pi/2G_l$. This yields $d(\widehat{\boldsymbol{\Theta}}(\underline{G}), \boldsymbol{\Theta}^0) \geq \frac{c_{\text{gap}}(c_\pi)^q}{2^{q+2}\prod_l G_l}$ as $N_l \to \infty$ for all $l \in [q]$ by Assumption 7. Then by Lemmas 19 and 23, we have

$$
\begin{aligned}
&((\prod_l N_l)T)^{-1}\{Q(\widehat{\boldsymbol{\Theta}}(\underline{G})) - Q(\widehat{\boldsymbol{\Theta}}(\underline{G}_0))\} \\
&= O_p(T^{-1}(\sum_l \log N_l)^2) + ((\prod_l N_l)T)^{-1}\Big\{Q^*(\widehat{\boldsymbol{\Theta}}(\underline{G})) - Q^*(\boldsymbol{\Theta}^0)\Big\} \\
&\geq \tau_{\min} d(\widehat{\boldsymbol{\Theta}}(\underline{G}), \boldsymbol{\Theta}^0) + O_p(T^{-1}(\sum_l \log N_l)^2)) \\
&\geq \frac{\tau_{\min} c_{\text{gap}}(c_\pi)^q}{2^{q+2}\prod_l G_l} + O_p(T^{-1/2}(m + \sum_l \log N_l)) \geq \frac{C c_{\text{gap}}(c_\pi)^q}{\prod_l G_l}.
\end{aligned}
$$

with probability tending to 1, where $C = \tau_{\min}/2^{q+2}$ is a positive constant, and the last inequality holds since the condition (19) that $T^{-1}(\sum_l \log N_l)^2 \ll c_{\text{gap}}/(\prod_l G_l)$.

I.3 Proof of Proposition 4

Recall that

$$Q_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}};\boldsymbol{\xi}^{-(l)}_{g^{-(l)}},\mathcal{G}_{-l}) = \sum_{m\neq l}\sum_{i_m=1}^{N_m}\sum_{t=1}^{T}\Big\{Y_{i_1\cdots i_q,t} - \sum_{l=1}^{q}\lambda^{(l)}_{g^{(l)}_{i_l}}\sum_{k=1}^{N_l}w^{(l)}_{i_lk}Y_{i_1\cdots i_{l-1}ki_{l+1}\cdots i_q,(t-1)} - \alpha_{g^{(1)}_{i_1}\cdots g^{(q)}_{i_q}}Y_{i_1\cdots i_q,(t-1)} - \sum_{l=1}^{q}\mathbf{x}^{(l)\top}_{i_lt}\boldsymbol{\zeta}^{(l)}_{g^{(l)}_{i_l}}\Big\}^2,$$

which is defined in (A.1). Define

$$\sigma_l(g^{(l)}) = \mathrm{argmin}_{\widetilde{g}^{(l)}\in[G_l]}\Big\{\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widetilde{g}^{(l)}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2 + \frac{1}{\prod_{m\neq l}N_m}\sum_{i_{-l}}|\widehat{\alpha}_{\widehat{g}^{(1)}_{i_1}\cdots\widehat{g}^{(l-1)}_{i_{l-1}}\widetilde{g}^{(l)}\widehat{g}^{(l+1)}_{i_{l+1}}\cdots\widehat{g}^{(q)}_{i_q}} - \alpha^0_{g^{(1)0}_{i_1}\cdots g^{(l-1)0}_{i_{l-1}}g^{(l)}g^{(l+1)0}_{i_{l+1}}\cdots g^{(q)0}_{i_q}}|^2\Big\}.$$

Let $R_1 = \sup_l\sup_{i_l}\sup_{\|\boldsymbol{\Theta}_{i_1\cdots i_q}\|_{\max}<R}|\frac{1}{T}\{Q_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q}) - Q^*_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q})\}|$ and we have $R_1 = O_p(T^{-1/2}(m+\sum_l\log N_l))$ due to Lemma 17. Therefore again by Lemma 17 and the fact that $\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}} = (\widehat{\boldsymbol{\theta}}^{(l)\top}_{\widehat{g}^{(l)}_{i_l}}, \mathrm{vec}(\widehat{\boldsymbol{\alpha}}_{\cdot,\widehat{g}^{(l)}_{i_l}\cdot})^\top)^\top$ minimize $Q_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}};\boldsymbol{\xi}^{-(l)}_{g^{-(l)}},\mathcal{G}_{-l})$, we have

$$\begin{aligned}&\frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l}) - R_1 \le \frac{1}{(\prod_{m\neq l}N_m)T}Q_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l})\\ &\le \frac{1}{(\prod_{m\neq l}N_m)T}Q_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l}) \le \frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l}) + R_1\end{aligned}$$

for all $i_l\in[N_l]$. Therefore it holds

$$\sup_{i_l}\left\{\frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l}) - \frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l})\right\} \le 2R_1. \tag{A.49}$$

On the other hand by Lemma 23 we have

$$\begin{aligned}0 \le &\sup_{i_l}\Big\{\frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l}) - \frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\boldsymbol{\xi}^{(l)0}_{g^{(l)0}_{i_l}};\boldsymbol{\xi}^{-(l)0}_{g^{-(l)}},\mathcal{G}^0_{-l})\Big\}\\ &\le \tau_{\max}\sup_{i_l}\Big\{\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}\|\widehat{\boldsymbol{\theta}}^{(m)}_{\widehat{g}^{(m)}_{i_m}} - \boldsymbol{\theta}^{(m)0}_{g^{(m)0}_{i_m}}\|^2 + \|\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2\\ &+ \frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m=1}^{N_m}|\widehat{\alpha}_{\widehat{g}^{(1)}_{i_1}\cdots\widehat{g}^{(l-1)}_{i_{l-1}}\sigma_l(g^{(l)0}_{i_l})\widehat{g}^{(l+1)}_{i_{l+1}}\cdots\widehat{g}^{(q)}_{i_q}} - \alpha^0_{g^{(1)0}_{i_1}\cdots g^{(q)0}_{i_q}}|^2\Big\}\\ &= O_p(T^{-1}(\sum_l\log N_l)^2) + O_p(T^{-1}(\sum_l\log N_l)^2) = O_p(T^{-1}(\sum_l\log N_l)^2)\end{aligned} \tag{A.50}$$

where the second last equation is obtained by Theorem 2 and Lemma 24.

Next, we use the order in (A.50) to improve the order of R_1 in (A.49). Denote $\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot}$ as the parameter by replacing the $\widehat{g}^{(l)}_{i_l}$ by $\sigma(g^{(l)0}_{i_l})$ in $\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}$ while keep the other terms the same. By (A.50), there exists a positive constant C, the event $\mathcal{O}_1 = \{\sup_{i_l} d_{i_l}(\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot}, \mathbf{\Theta}^0_{\cdot i_l \cdot}) \leq Ca^2_{NT}\}$ has the probability $P(\mathcal{O}_1) > 1 - \epsilon$ for $\forall \epsilon > 0$, where $a_{NT} = (\sum_l \log N_l)/\sqrt{T}$. To derive the order of R_1, we need to derive the order of distance $d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}, \widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot})$. We next show that $P(a^{-1}_{NT} \sup_{i_l} \sqrt{d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}, \widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot})} > K) \to 0$ as $K \to \infty$, which implies that $d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}, \widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot}) = O_p(a^2_{NT})$. Partition the parameter $\mathbf{\Theta}_{\cdot i_l \cdot}$ into space $\mathbf{\Omega}_j = \{\mathbf{\Theta}_{\cdot i_l \cdot} \in \mathbf{\Omega} : \gamma(j-1) < a^{-1}_{NT}\sqrt{d_{i_l}(\mathbf{\Theta}_{\cdot i_l \cdot}, \widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot})} \leq \gamma j\}$, with $j \geq 1$ taking over all positive integers and $\gamma = \sqrt{1 + C(\tau_{\max} + \tau_{\min})/\tau_{\min}}$ is a positive constant. For a positive integer K, if $a^{-1}_{NT}\sqrt{d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}, \widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot})} > K$, then $\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot} \in \bigcup_{j \geq K} \mathbf{\Omega}_j$. Define

$$S_{i_l}(\mathbf{\Theta}_{\cdot i_l \cdot}) = Q_{i_l}(\boldsymbol{\xi}^{(l)}_{\widehat{g}^{(l)}_{i_l}}; \widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l}) - Q_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})}; \widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l}).$$

By the definition of $\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}$, we know that $S_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}) \leq 0$. Hence, we have

$$\begin{aligned}
&P\Big\{a^{-1}_{NT} \sup_{i_l} \sqrt{d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}, \widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot})} \geq K\Big\} \leq P\Big\{\inf_{i_l} \inf_{\mathbf{\Theta}_{\cdot i_l \cdot} \in \bigcup_{j \geq K} \mathbf{\Omega}_j} (\prod_{m \neq l} N_m T)^{-1} S_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}) \leq 0\Big\} \\
&\leq \sum_{j \geq K} P\Big\{\inf_{i_l}\Big[\inf_{\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot} \in \mathbf{\Omega}_j} (\prod_{m \neq l} N_m T)^{-1}\{S_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}) - S^*_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot})\} \\
&+ \inf_{\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot} \in \mathbf{\Omega}_j} (\prod_{m \neq l} N_m T)^{-1} S^*_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot})\Big] \leq 0\Big\}
\end{aligned}$$

Under event $\mathcal{O}_1$, we know that if $\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot} \in \mathbf{\Omega}_j$,

$$\begin{aligned}
(\prod_{m \neq l} N_m T)^{-1} \inf_{i_l} S^*_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}) &= (\prod_{m \neq l} N_m T)^{-1} \inf_{i_l} \{Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}}; \widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l}) - Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})}; \widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l})\} \\
&= (\prod_{m \neq l} N_m T)^{-1} \inf_{i_l} \Big[\{Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}}; \widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l}) - Q^*_{i_l}(\boldsymbol{\xi}^{(l)0}_{g^{(l)0}_{i_l}}; \boldsymbol{\xi}^{-(l)0}_{g^{-(l)}}, \mathcal{G}^0_{-l})\} \\
&+ \{Q^*_{i_l}(\boldsymbol{\xi}^{(l)0}_{g^{(l)0}_{i_l}}; \boldsymbol{\xi}^{-(l)0}_{g^{-(l)}}, \mathcal{G}^0_{-l}) - Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})}; \widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l})\}\Big] \\
&\overset{①}{\geq} \tau_{\min} \inf_{i_l} d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}, \mathbf{\Theta}^0_{\cdot i_l \cdot}) - \tau_{\max} \sup_{i_l} d_{i_l}(\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot}, \mathbf{\Theta}^0_{\cdot i_l \cdot}) \\
&\overset{②}{\geq} \tau_{\min} \inf_{i_l} d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l \cdot}, \widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot}) - (\tau_{\min} + \tau_{\max}) \sup_{i_l} d_{i_l}(\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l \cdot}, \mathbf{\Theta}^0_{\cdot i_l \cdot}) \\
&\overset{③}{\geq} \tau_{\min}(j-1)^2(1 + C\frac{\tau_{\max} + \tau_{\min}}{\tau_{\min}})a^2_{NT} - C(\tau_{\max} + \tau_{\min})a^2_{NT} \\
&= \tau_{\min} a^2_{NT}(j-1)^2 + C\{(j-1)^2 - 1\}(\tau_{\min} + \tau_{\max})a^2_{NT} \\
&\overset{④}{\geq} \tau_{\min} a^2_{NT}(j-1)^2,
\end{aligned}$$

where the inequality ① holds due to Lemma 23, inequality ② holds due to the triangle inequality, inequality ③ holds under $\mathcal{O}_1$ and the definition of the partition, and inequality ④ holds when $j > 1$. Therefore, we have

$$
\begin{aligned}
&P\Big\{a_{NT}^{-1}\sup_{i_l}\sqrt{d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot},\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l\cdot})}\geq K\Big\}\\
&\leq\sum_{j\geq K}P\Big\{\inf_{i_l}\Big[\inf_{\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot}\in\mathbf{\Omega}_j}(\prod_{m\neq l}N_mT)^{-1}\{S_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})-S^*_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})\}\\
&+\inf_{\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot}\in\mathbf{\Omega}_j}(\prod_{m\neq l}N_mT)^{-1}S^*_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})\Big]\leq 0\Big\}\\
&\leq\sum_{j\geq K}P\Big\{\sup_{i_l}\sup_{\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot}\in\mathbf{\Omega}_j}(\prod_{m\neq l}N_mT)^{-1}|S_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})-S^*_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})|\geq\tau_{\min}a_{NT}^2(j-1)^2\Big\}+P(\mathcal{O}_1^c).
\end{aligned}
\tag{A.51}
$$

Next, we derive the upper bound for the first term in (A.51) under event $\mathcal{O}_1$. Note that

$$
\begin{aligned}
S_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})&=Q_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})-Q_{i_l}(\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l\cdot})\\
&=\sum_{m\neq l}\sum_{i_m}\sum_t\Big\{\Big(Y_{i_1\cdots i_q,t}-\mathcal{X}^\top_{i_1\cdots i_q,t}\widehat{\mathbf{\Theta}}_{i_1\cdots i_q}\Big)^2-\Big(Y_{i_1\cdots i_q,t}-\mathcal{X}^\top_{i_1\cdots i_q,t}\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q}\Big)^2\Big\}\\
&=\sum_{m\neq l}\sum_{i_m}\sum_t\Big\{(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})\\
&-(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})\\
&+2\mathcal{X}^\top_{i_1\cdots i_q,t}(\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})\varepsilon_{i_1\cdots i_q,t}\Big\}
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
&|S_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})-S^*_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot})|\leq\\
&\leq 2\Big|\sum_{m\neq l}\sum_{i_m}\sum_t\Big\{(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})\\
&-\big[(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t})(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})\big]\Big\}\Big| \qquad \text{(A.52)}\\
&+\Big|\sum_{m\neq l}\sum_{i_m}\sum_t\Big\{(\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})\\
&-\big[(\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t})(\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})\big]\Big\}\Big| \qquad \text{(A.53)}\\
&+2\Big|\sum_{m\neq l}\sum_{i_m}\sum_t\mathcal{X}^\top_{i_1\cdots i_q,t}(\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}_{i_1\cdots i_q})\varepsilon_{i_1\cdots i_q,t}\Big| \qquad \text{(A.54)}
\end{aligned}
$$

Due to the definition of parameter $\mathbf{\Theta}_{\cdot i_l\cdot}$'s partition $\mathbf{\Omega}_j$ and under event $\mathcal{O}_1$, the concentration inequalities for (A.52)–(A.54) can be obtained by Lemma 20 that,

$$
\sum_{j\geq K}P\Big\{\sup_{i_l}\sup_{\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot}\in\mathbf{\Omega}_j}(\prod_{m\neq l}N_mT)^{-1}
$$

$$\Big|\sum_{m\neq l}\sum_{i_m}\sum_{t}\Big\{(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})$$
$$-E\big[(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}^0_{i_1\cdots i_q}-\widehat{\mathbf{\Theta}}^{\sigma}_{i_1\cdots i_q})\big]\Big\}\Big|\geq\tau_{\min}a^2_{NT}(j-1)^2\Big\}$$
$$\leq\sum_{j\geq K}C_1\exp\Big\{-C_2\min\Big(\frac{T(j-1)^4a^2_{NT}}{C},\frac{\sqrt{T}(j-1)^2a_{NT}}{\sqrt{C}}\Big)+C_3m+\log(\prod_l N_l)\Big\}$$
$$=\sum_{j\geq K}C_1\exp\Big\{-\widetilde{C}_2(j-1)^2\log(\prod_l N_l)+C_3m+\log(\prod_l N_l)\Big\}$$
$$\leq\sum_{j\geq K}C_1\exp\Big\{-\frac{\widetilde{C}_2(j-1)^2\log(\prod_l N_l)}{j}+C_3m+\log(\prod_l N_l)\Big\}\to 0$$

as $K\to\infty$, where the equality holds as $a_{NT}=\log(\prod_l N_l)/\sqrt{T}$, the last inequality holds when $j\geq 1$, and the limit holds by the same calculation in (A.44). Therefore, by substituting the tail bounds for the first term in (A.51), we obtain that $P\Big\{a^{-1}_{NT}\sup_{i_l}\sqrt{d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot},\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l\cdot})}\geq K\Big\}\to 0$ as $K\to\infty$.

This implies that $\sup_{i_l}d_{i_l}(\widehat{\mathbf{\Theta}}_{\cdot i_l\cdot},\widehat{\mathbf{\Theta}}^{\sigma}_{\cdot i_l\cdot})=O_p(a^2_{NT})$. By using Lemma 23, we have that

$$\sup_{i_l}\left\{\frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l})-\frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l})\right\}$$
$$=O_p(T^{-1}(\sum_l\log N_l)^2).$$

Therefore, together with (A.50), we have

$$\sup_{i_l}\left\{\frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\widehat{\boldsymbol{\xi}}^{(l)}_{\widehat{g}^{(l)}_{i_l}};\widehat{\boldsymbol{\xi}}^{-(l)}_{g^{-(l)}},\widehat{\mathcal{G}}_{-l})-\frac{1}{(\prod_{m\neq l}N_m)T}Q^*_{i_l}(\boldsymbol{\xi}^{(l)0}_{g^{(l)0}_{i_l}};\boldsymbol{\xi}^{-(l)0}_{g^{-(l)}},\mathcal{G}^0_{-l})\right\}$$
$$=O_p(T^{-1}(\sum_l\log N_l)^2).\tag{A.55}$$

By Lemma 23, we get

$$\sup_{i_l}\left\{\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{i_{-l}}|\widehat{\alpha}_{\widehat{g}^{(1)}_{i_1}\cdots\widehat{g}^{(q)}_{i_q}}-\alpha^0_{g^{(1)0}_{i_1}\cdots g^{(q)0}_{i_q}}|^2\right\}$$
$$=O_p\{\tau^{-1}_{\min}T^{-1}(\sum_l\log N_l)^2\}=O_p\{T^{-1}(\sum_l\log N_l)^2\}.$$

I.4 Proof of Theorem 5

The Figure A.6 visualizes the strong group memberships consistency when true group number $G_{1,0}=3$ and estimated group number $\widehat{G}_1=4$.

To prove the result, we use Lemma 25 by setting $\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}$. First we note that $\widehat{\mathbf{\Theta}}$ satisfies (A.88) by Theorem 2 and the condition for c_{gap}. In the following we first show that $\widehat{\boldsymbol{\xi}}\in\mathcal{N}^{(l)}_\eta$

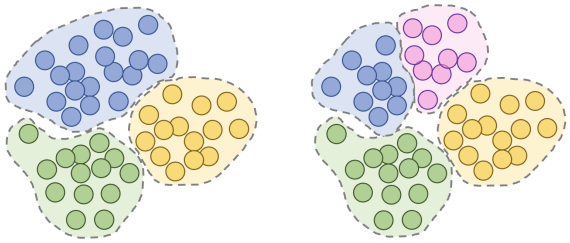


Figure A.6: The group memberships when $G_{1,0}=3$ and $\widehat{G}_1=4$. The units belonging to the blue group are split into two sub groups (marked as blue and pink).

with $\eta = \tau_{\min}c_{\text{gap}}(c_\pi)^{q-1}/\{8(\tau_{\min}+\tau_{\max})\}$ with probability tending to one. Then the conclusions of Lemma 25 hold. Second, we use the conclusion to prove the result. We remark that the notations in the following proof are borrowed in the statement in Lemma 25.

Step 1. Proof of $\widehat{\boldsymbol{\xi}} \in \mathcal{N}_\eta^{(l)}$ with probability tending to one.

First by Lemma 24, we have

$$\max_{g^{(l)}\in[G_{l,0}]}\min_{\widetilde{g}^{(l)}\in[G_l]}\Big(\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widetilde{g}^{(l)}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widetilde{g}^{(l)}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2\Big)$$
$$=O_p(T^{-1}(\sum_l \log N_l)^2).$$

Next we show

$$\max_{\widetilde{g}^{(l)}\in[G_l]}\min_{g^{(l)}\in[G_{l,0}]}\Big(\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widetilde{g}^{(l)}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widetilde{g}^{(l)}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2\Big)$$
$$=O_p(T^{-1}(\sum_l \log N_l)^2).$$

Define the set $\mathcal{O}_{\widetilde{g}^{(l)}}=\{i_l\in[N_l]:\widehat{g}^{(l)}_{i_l}=\widetilde{g}^{(l)}\}$, where $\widehat{g}^{(l)}_{i_l}$ is obtained by (14) by using $\widehat{\boldsymbol{\xi}}$. Therefore for all $i_l\in\mathcal{O}_{\widetilde{g}^{(l)}}$, it holds

$$\min_{g^{(l)}\in[G_{l,0}]}\Big(\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widetilde{g}^{(l)}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widetilde{g}^{(l)}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2\Big)$$
$$\leq\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widetilde{g}^{(l)}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widetilde{g}^{(l)}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)0}_{i_l}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2$$
$$=\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widehat{g}^{(l)}_{i_l}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)0}_{i_l}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2.$$

Then it yields

$$\begin{aligned}
&\max_{\widetilde{g}^{(l)}\in[G_l]}\min_{g^{(l)}\in[G_{l,0}]}\Big(\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widetilde{g}^{(l)}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widetilde{g}^{(l)}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2\Big)\\
&\leq\max_{\widetilde{g}^{(l)}\in[G_l]}\sup_{i_l\in\mathcal{O}_{\widetilde{g}^{(l)}}}\left\{\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widehat{g}^{(l)}_{i_l}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)0}_{i_l}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2\right\}\\
&=\max_{i_l\in[N_l]}\left\{\|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}}-\boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2+\frac{1}{\prod_{m\neq l}N_m}\sum_{m\neq l}\sum_{i_m}|\widehat{\alpha}_{\widehat{g}^{(l)}_{i_l}\widehat{g}^{-(l)}_{\mathbf{i}_{-l}}}(\boldsymbol{\xi})-\alpha^0_{g^{(l)0}_{i_l}g^{-(l)0}_{\mathbf{i}_{-l}}}|^2\right\}\\
&=O_p\big\{T^{-1}(\sum_l\log N_l)^2\big\},
\end{aligned}$$

where the last equation is obtained by Proposition 4. By the condition that $c_{\text{gap}}\gg T^{-1}(\sum_l\log N_l)^2$, and the definition of $\mathcal{N}^{(l)}_\eta$, we can conclude $\widehat{\boldsymbol{\xi}}\in\mathcal{N}^{(l)}_\eta$ with probability tending to one.

Step 2. Proof of conclusion of Theorem 5.

Since $\widehat{\boldsymbol{\xi}}\in\mathcal{N}^{(l)}_\eta$, we have that the two conclusions in Lemma 25 holds for $\widehat{\boldsymbol{\xi}}$. Then it should be equivalent to prove that for $i_{l1},i_{l2}\in\widehat{\mathcal{R}}^{(l)}_{\widetilde{g}^{(l)}}$ for some $\widetilde{g}^{(l)}\in[G_l]$, it holds $i_{l1},i_{l2}\in\mathcal{R}^{(l)0}_{g^{(l)}}$ for some $g^{(l)0}\in[G_{l,0}]$. Suppose $i_{l1},i_{l2}\in\widehat{\mathcal{R}}^{(l)}_{\widetilde{g}^{(l)}}$ for some $\widetilde{g}^{(l)}\in[G_l]$, then $\widehat{g}^{(l)}_{i_{l1}}=\widehat{g}^{(l)}_{i_{l2}}=\widetilde{g}^{(l)}$ in this case. By conclusion (ii) in Lemma 25, we have $\widetilde{g}^{(l)}\in\mathcal{A}^{(l)}_\eta(\widehat{\boldsymbol{\xi}},g^{(l)0}_{i_{l1}},\mathcal{G}^0_2)$ and $\widetilde{g}^{(l)}\in\mathcal{A}^{(l)}_\eta(\widehat{\boldsymbol{\xi}},g^{(l)0}_{i_{l2}},\mathcal{G}^0_2)$ with probability tending to 1. Next, by conclusion (i) in Lemma 25, it holds $g^{(l)0}_{i_{l1}}=g^{(l)0}_{i_{l2}}$ since $\mathcal{A}^{(l)}_\eta(\widehat{\boldsymbol{\xi}},\cdot,\mathcal{G}^0_2)$ is a partition of $[G_l]$. By defining $g^{(l)}\overset{\text{def}}{=}g^{(l)0}_{i_{l1}}=g^{(l)0}_{i_{l2}}\in[G_{l,0}]$, we have $i_{l1},i_{l2}\in\mathcal{R}^{(l)0}_{g^{(l)}}$. Then the conclusion we finish the proof.

I.5 Proof of Corollary 6

By the Step 1 in the proof of Theorem 5, we have $\widehat{\boldsymbol{\xi}}\in\mathcal{N}^{(l)}_\eta$ with probability tending to one. For $G_l=G_{l,0}$, since $\{\mathcal{A}^{(l)}_\eta(\widehat{\boldsymbol{\xi}},g^{(l)},\mathcal{G}^0_2),g^{(l)}\in[G_{l,0}]\}$ is a partition of $[G_l]$, then we can conclude that $\mathcal{A}^{(l)}_\eta(\widehat{\boldsymbol{\xi}},g^{(l)},\mathcal{G}^0_2)$ contains only one element in $[G_l]$. Consequently, we can define a permutation by $\mathcal{A}^{(l)}_\eta(\widehat{\boldsymbol{\xi}},\cdot,\mathcal{G}^0_2):[G_{l,0}]\to[G_l]$. By Theorem 5, this implies that $\lim_{\min\{N_1,\cdots,N_q,T\}\to\infty}P\left(\widehat{\mathcal{G}}_l=\mathcal{G}^0_l\right)\to1$. Next, given that the probability of $\{\widehat{\mathcal{G}}_1=\mathcal{G}^0_1,\cdots,\widehat{\mathcal{G}}_q=\mathcal{G}^0_q\}$ goes to 1, (22) can be directly obtained.

I.6 Proof of Theorem 7

By Corollary 6, we have $\widehat{\boldsymbol{\xi}}=\widehat{\boldsymbol{\xi}}^{\text{or}}$ with probability tending to 1. Then it suffices to treat $\widehat{\boldsymbol{\xi}}$ as $\widehat{\boldsymbol{\xi}}^{\text{or}}$ equivalently. Recall that $\mathcal{E}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdots,\mathcal{R}^{(q)}_{g^{(q)}})}\in\mathbb{R}^{N_{1g^{(1)}}\times\cdots\times N_{qg^{(q)}}}$ is defined as the subset of the tensor $\mathcal{E}_t$ with each dimension selected by the index sets $\mathcal{R}^{(l)}_{g^{(l)}}$. By the vectorization of tensor defined in the main text, denote $\mathbb{Y}_{g^{(1)}\cdots g^{(q)},t}=\text{vec}\big(\mathcal{Y}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdots,\mathcal{R}^{(q)}_{g^{(q)}})}\big)$ and $\mathbb{E}_{g^{(1)}\cdots g^{(q)},t}=$

$\mathrm{vec}\big(\mathcal{E}_t^{(\mathcal{R}^{(1)}_{g^{(1)}},\cdots,\mathcal{R}^{(q)}_{g^{(q)}})}\big)$, then we have

$$n^{q/2}T^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^0)=n^{q/2}T^{1/2}\mathbf{M}^{-1}\boldsymbol{\delta}=(n^{-q}T^{-1}\mathbf{M})^{-1}(n^{-q/2}T^{-1/2}\boldsymbol{\delta}),$$

where $\boldsymbol{\delta}=(\boldsymbol{\delta}^{(1)\top},\cdots,\boldsymbol{\delta}^{(q)\top},\boldsymbol{\delta}^{\alpha\top})^\top$ and

$$\boldsymbol{\delta}^{(l)}_{g^{(l)}}=\sum_{t,g^{-(l)}}\mathbb{X}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t}\mathbb{E}_{g^{(1)}\cdots g^{(q)},t},\quad \boldsymbol{\delta}^{(l)}=(\boldsymbol{\delta}^{(l)\top}_{g^{(l)}}:g^{(l)}\in[G_l])^\top\in\mathbb{R}^{G_l(p_l+1)},$$

$$\boldsymbol{\delta}^{\alpha}_{\mathcal{I}_{g^{(1)},\cdots,g^{(q)}}}=\sum_t\mathbb{Y}^\top_{g^{(1)}\cdots g^{(q)},(t-1)}\mathbb{E}_{g^{(1)}\cdots g^{(q)},t},\quad \boldsymbol{\delta}^{\alpha}=(\boldsymbol{\delta}^{\alpha\top}_{\mathcal{I}_{g^{(1)},\cdots,g^{(q)}}}:\mathcal{I}_{g^{(1)},\cdots,g^{(q)}}\in\Big[\prod_l G_l\Big])^\top.$$

Here

$$\mathbb{X}^{(l)}_{g^{(1)}\cdots g^{(q)},t}=\Big(\mathrm{vec}\big\{(\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)})_{g^{(1)}\cdots g^{(q)}}\big\},\\ \mathbf{1}_{N_{1g^{(1)}}}\otimes\cdots\otimes(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)}},\cdot)}\otimes\cdots\otimes\mathbf{1}_{N_{qg^{(q)}}}\Big)\in\mathbb{R}^{(N_{1g^{(1)}}\cdots N_{qg^{(q)}})\times(p_l+1)}$$

is defined in (10). Recall that $\mathbf{M}_{nT}=n^{-q}T^{-1}\mathbf{M}$ and then we have

$$\begin{aligned}&n^{q/2}T^{1/2}\boldsymbol{\eta}^\top(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^0)=\boldsymbol{\eta}^\top\mathbf{M}^{-1}_{nT}(n^{-q/2}T^{-1/2}\boldsymbol{\delta})\\&=n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top(\mathbf{M}^0_{nT})^{-1}\boldsymbol{\delta}+n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top\left\{\mathbf{M}^{-1}_{nT}-(\mathbf{M}^0_{nT})^{-1}\right\}\boldsymbol{\delta}\\&=n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top(\mathbf{M}^0_{nT})^{-1}\boldsymbol{\delta}+n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top\mathbf{M}^{-1}_{nT}\left(\mathbf{M}^0_{nT}-\mathbf{M}_{nT}\right)(\mathbf{M}^0_{nT})^{-1}\boldsymbol{\delta}.\end{aligned}$$

Since we have $n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top\mathbf{M}^{-1}_{nT}\left(\mathbf{M}^0_{nT}-\mathbf{M}_{nT}\right)(\mathbf{M}^0_{nT})^{-1}\boldsymbol{\delta}=o_p(1)$, then it suffices to show

$$n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top(\mathbf{M}^0_{nT})^{-1}\boldsymbol{\delta}\to_d N\{0,\sigma^2\boldsymbol{\eta}^\top(\mathbf{M}^0)^{-1}\boldsymbol{\eta}\}.\tag{A.56}$$

Next, we provide the proof of (A.56).

Define $\widetilde{\boldsymbol{\eta}}\overset{\text{def}}{=}(\mathbf{M}^0_{nT})^{-1}\boldsymbol{\eta}=(\widetilde{\boldsymbol{\eta}}^{(1)\top}_1,\cdots,\widetilde{\boldsymbol{\eta}}^{(1)\top}_{G_1},\cdots,\widetilde{\boldsymbol{\eta}}^{(q)\top}_1,\cdots,\widetilde{\boldsymbol{\eta}}^{(q)\top}_{G_q},\widetilde{\boldsymbol{\eta}}^{\alpha}_{1\cdots1},\cdots,\widetilde{\boldsymbol{\eta}}^{\alpha}_{G_1\cdots G_q})^\top$, where $\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)}}\in\mathbb{R}^{p_l+1}$ and $\widetilde{\boldsymbol{\eta}}^{\alpha}_{g^{(1)}\cdots g^{(q)}}\in\mathbb{R}$. Then we have

$$\begin{aligned}n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top(\mathbf{M}^0_{nT})^{-1}\boldsymbol{\delta}&=\frac{1}{\sqrt{n^qT}}\sum_l\sum_{g^{(l)}}\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)}}\sum_{t,g^{-(l)}}\mathbb{X}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t}\mathbb{E}_{g^{(1)}\cdots g^{(q)},t}\\&\quad+\frac{1}{\sqrt{n^qT}}\sum_l\sum_{g^{(l)}=1}^{G_l}\widetilde{\boldsymbol{\eta}}^{\alpha}_{g^{(1)}\cdots g^{(q)}}\sum_t\mathbb{Y}^\top_{g^{(1)}\cdots g^{(q)},(t-1)}\mathbb{E}_{g^{(1)}\cdots g^{(q)},t}\\&\overset{\text{def}}{=}\sum_t\sum_{g^{(l)}}\mathcal{Z}^\top_{g^{(1)}\cdots g^{(q)},t}\mathbb{E}_{g^{(1)}\cdots g^{(q)},t}\overset{\text{def}}{=}\sum_t\mathcal{Z}^\top_t\mathbb{E}_t.\end{aligned}$$

Define $\mathcal{F}_t$ as the sigma field generated by $\{\varepsilon_{i_1\cdots i_q,s}:i_l\in[N_l],l\in[q],s\le t\}$. We have $\{\mathcal{Z}^\top_t\mathbb{E}_t,\mathcal{F}_t\}$ as a martingale difference array. By the central limit theorem of Hall and Heyde (2014), it suffices to verify for any $u>0$

$$\sum_{t=1}^T E\left\{(\mathcal{Z}^\top_t\mathbb{E}_t)^2I(|\mathcal{Z}^\top_t\mathbb{E}_t|>u)|\mathcal{F}_{t-1}\right\}\to_p0\tag{A.57}$$

$$\sum_{t=1}^{T} E\left\{(\mathcal{Z}_t^\top \mathbb{E}_t)^2 | \mathcal{F}_{t-1}\right\} \to_p \boldsymbol{\eta}^\top (\mathbf{M}^0)^{-1} \mathbf{M}_b^0 (\mathbf{M}^0)^{-1} \boldsymbol{\eta}. \tag{A.58}$$

where $\mathbf{M}_b^0 \stackrel{\text{def}}{=} \lim_{\min(n,T)\to\infty} n^{-q}T^{-1}\text{var}(\boldsymbol{\delta}) = \sigma^2 \lim_{\min(n,T)\to\infty} \mathbf{M}_{nT}^0 = \sigma^2 \mathbf{M}^0$, and $(\mathbf{M}^0)^{-1} = (\mathbf{M}^0)^{-1}\mathbf{M}_b^0(\mathbf{M}^0)^{-1}$ by the condition that $\varepsilon_{i_1\cdots i_q,t}$ is i.i.d over $i_1, \cdots, i_q, t$ with $\text{var}(\varepsilon_{i_1\cdots i_q,t}) = \sigma^2$ and it is also independent of $\{\mathcal{Y}_{s-1}, \mathbf{X}_s^{(1)}, \cdots, \mathbf{X}_s^{(q)} : s \le t\}$, where $\mathbf{X}_s^{(l)} = (\mathbf{x}_{i_l,s}^{(l)} : i_l \in [N_l])^\top$.

(1) Proof of (A.57)

We have

$$\sum_{t=1}^{T} E\left\{(\mathcal{Z}_t^\top \mathbb{E}_t)^2 I(|\mathcal{Z}_t^\top \mathbb{E}_t| > u) | \mathcal{F}_{t-1}\right\} \le u^{-2} \sum_t E\left\{(\mathcal{Z}_t^\top \mathbb{E}_t)^4 | \mathcal{F}_{t-1}\right\} \le cu^{-2} \sum_t (\mathcal{Z}_t^\top \mathcal{Z}_t)^2,$$

where c is the fourth moment of $\varepsilon_{i_1\cdots i_q,t}$, which is a finite constant by Assumption 3. Then it suffices to show $E\{\sum_t (\mathcal{Z}_t^\top \mathcal{Z}_t)^2\} \to 0$. Note that we have

$$\begin{aligned}\mathcal{Z}_{g^{(1)}\cdots g^{(q)},t} &= \sum_l \frac{1}{\sqrt{n^q T}} \mathbb{X}^{(l)}_{g^{(1)}\cdots g^{(q)},t} \widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)}} + \frac{1}{\sqrt{n^q T}} \mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)} \widetilde{\boldsymbol{\eta}}^{\alpha}_{g^{(1)}\cdots g^{(q)}} \\ &\stackrel{\text{def}}{=} \sum_l \mathcal{Z}^{(l)}_{g^{(1)}\cdots g^{(q)},t} + \mathcal{Z}^{\alpha}_{g^{(1)}\cdots g^{(q)},t}.\end{aligned}$$

Then it suffices to show

$$E\{\sum_t (\sum_l \sum_{g^{(l)}} \mathcal{Z}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t} \mathcal{Z}^{(l)}_{g^{(1)}\cdots g^{(q)},t})^2\} \to 0, \tag{A.59}$$

$$E\{\sum_t (\sum_l \sum_{g^{(l)}} \mathcal{Z}^{\alpha\top}_{g^{(1)}\cdots g^{(q)},t} \mathcal{Z}^{\alpha}_{g^{(1)}\cdots g^{(q)},t})^2\} \to 0 \tag{A.60}$$

due to Cauchy-Schwarz inequality. We first show (A.59) in Step (1.1), and then show (A.60) in Step (1.2).

(1.1) Proof of $E\{\sum_t (\sum_l \sum_{g^{(l)}} \mathcal{Z}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t} \mathcal{Z}^{(l)}_{g^{(1)}\cdots g^{(q)},t})^2\} \to 0$

Write $\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)}} = (\widetilde{\eta}^{(l)}_{g^{(l)},1}, \widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)})^\top$. By the definition of $\mathcal{Y}_{t-1} \times_l \mathbf{W}^{(l)}$ in the main text, we have

$$\begin{aligned}&\mathcal{Y}_{t-1} \times_l \mathbf{W}^{(l)(\mathcal{R}_{g^{(l)}},\cdot)} \\ &= (\sum_{i_l=1}^{N_l} Y_{i_1,\cdots,i_l,\cdots,i_q,(t-1)} (a^{(l)}_{s,i_l}/n_{ls}))_{i_1\in[N_1],\cdots,s\in\mathcal{R}^{(l)}_{g^{(l)}},\cdots,i_q\in[N_q]} \in \mathbb{R}^{N_1\times\cdots\times N_{lg^{(l)}}\times N_q}.\end{aligned}$$

We have

$$E\left\{\sum_t (\sum_l \sum_{g^{(l)}} \mathcal{Z}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t} \mathcal{Z}^{(l)}_{g^{(1)}\cdots g^{(q)},t})^2\right\} \le c\frac{1}{n^{2q}T} E\left\{\left(\sum_{g^{(1)},\cdots,g^{(q)}} (\widetilde{\eta}^{(l)}_{g^{(l)},1})^2 \|\mathbb{X}^{(l)\dagger}_{g^{(1)}\cdots g^{(q)},t}\|^2\right)^2\right\}$$

$$+ c\frac{1}{n^{2q}T}E\left\{\left(\sum_{g^{(1)},\cdots,g^{(q)}}(\prod_{m\neq l}N_{mg^{(m)}})\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}\mathbb{X}^{(l)\dagger\dagger\top}_{g^{(1)}\cdots g^{(q)},t}\mathbb{X}^{(l)\dagger\dagger}_{g^{(1)}\cdots g^{(q)},t}\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\right)^2\right\}$$
$$= c\frac{1}{n^{2q}T}E\left\{\left(\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}^\top\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}\right)^2\right\}$$
$$+ c\frac{1}{n^{2q}T}E\left\{\left(\sum_{g^{(l)}}(\prod_{m\neq l}N_m)\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)},\cdot})^\top}(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\right)^2\right\},$$

where c is a constant. By Lemma 22 and Cauchy-Schwarz inequality, it holds

$$E\left\{\left(\sum_{g}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}^\top\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}\right)^2\right\}$$
$$\leq c_4\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\mathrm{tr}\Big(\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}(\mathbf{1}^\top_{N_1}\cdots\mathbf{1}^\top_{N_{l-1}}\mathbf{1}_{N_l}\mathbf{1}^\top_{N_{l+1}}\cdots\mathbf{1}^\top_{N_q})$$
$$(\mathbf{1}_{N_q}\cdots\mathbf{1}_{N_{l+1}}\mathbf{1}^\top_{N_l}\mathbf{1}_{N_{l-1}}\cdots\mathbf{1}_{N_1})\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})^\top}\Big)^2\leq c_4\left(\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2c_2n^q\right)^2\leq cn^{2q}$$

where $c_4=E(Y^4_{i_1\cdots i_q,t-1})<\infty$ and c is a constant. Here the last second inequality holds because $N_l\leq c_2n$ and c_2 is a positive constant. Next, we have

$$\frac{1}{n^{2q}T}E\left\{\left(\sum_{g^{(l)}}(\prod_{m\neq l}N_m)\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)},\cdot})^\top}(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\right)^2\right\}$$
$$=\frac{1}{n^{2q}T}(\prod_{m\neq l}N_m)^2E\left(\sum_{g^{(l)}}\sum_{i_l\in\mathcal{R}^{(l)}_{g^{(l)}}}(\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}\mathbf{x}^{(l)}_{i_lt})^2\right)^2\leq\frac{c}{n^2T}E\left(\sum_{g^{(l)}}\sum_{i_l\in\mathcal{R}^{(l)}_{g^{(l)}}}(\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}\mathbf{x}^{(l)}_{i_lt})^2\right)^2.$$

By the Cauchy's inequality we have

$$E\left\{(\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)}_1,(-1)}\mathbf{x}^{(l)}_{i_{l1}t})^2(\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)}_2,(-1)}\mathbf{x}^{(l)}_{i_{l2}t})^2\right\}\leq\left[E\left\{\big(\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)}_1,(-1)}\mathbf{x}^{(l)}_{i_{l1}t}\big)^4\right\}E\left\{\big(\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)}_2,(-1)}\mathbf{x}^{(l)}_{i_{l2}t}\big)^4\right\}\right]^{1/2}.$$

We also have

$$E\left\{\big(\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}\mathbf{x}^{(l)}_{i_lt}\big)^4\right\}=\Big\|\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\Big\|^4E\left\{\Big(\widetilde{\boldsymbol{\eta}}^{(l)\dagger\top}_{g^{(l)},(-1)}\mathbf{x}^{(l)}_{i_lt}\Big)^4\right\}$$
$$=4\Big\|\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\Big\|^4\int_0^\infty t^3P\left\{\big|\langle\widetilde{\boldsymbol{\eta}}^{(l)\dagger}_{g^{(l)},(-1)},\mathbf{x}^{(l)}_{i_lt}\rangle\big|>t\right\}dt\leq c_K\|\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\Big\|^4$$

where $\widetilde{\boldsymbol{\eta}}^{(l)\dagger}_{g^{(l)},(-1)}=\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}/\|\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\|$, and c_K is a constant related to K. Here the last inequality is obtained by noting that $\langle\widetilde{\boldsymbol{\eta}}^{(l)\dagger}_{g^{(l)},(-1)},\mathbf{x}^{(l)}_{i_lt}\rangle$ is a 1-Lipschitz convex function of $\mathbf{x}^{(l)}_{i_lt}$

with mean zero and $\mathbf{x}_{i_l t}^{(l)}$ is K-convex defined in Definition 1. Consequently we have

$$\frac{c}{n^2 T} E\left(\sum_{g_1^{(l)}, g_2^{(l)}} \sum_{i_{l1} \in \mathcal{R}_{g_1^{(l)}}^{(l)}} \sum_{i_{l2} \in \mathcal{R}_{g_2^{(l)}}^{(l)}} \big(\widetilde{\boldsymbol{\eta}}_{g_1^{(l)},(-1)}^{(l)\top} \mathbf{x}_{i_{l1} t}^{(l)}\big)^2 \big(\widetilde{\boldsymbol{\eta}}_{g_2^{(l)},(-1)}^{(l)\top} \mathbf{x}_{i_{l2} t}^{(l)}\big)^2 \right)$$
$$\leq \frac{C}{n^2 T} \left(\sum_{g_1^{(l)}, g_2^{(l)}} c_K \|\widetilde{\boldsymbol{\eta}}_{g_1^{(l)},(-1)}^{(l)}\|^2 \|\widetilde{\boldsymbol{\eta}}_{g_2^{(l)},(-1)}^{(l)}\|^2 n^2 \right) = O(T^{-1}) = o(1).$$

Consequently, it yields $E\left\{\sum_t (\sum_l \sum_{g^{(l)}} \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}^{(l)\top} \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}^{(l)})^2\right\} = O(T^{-1}) = o(1)$.

(1.2) Proof of $E\{\sum_t (\sum_l \sum_{g^{(l)}} \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}^{\alpha\top} \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}^{\alpha})^2\} \to 0$

Next, we have

$$E\left\{ \sum_t (\sum_l \sum_{g^{(l)}} \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}^{\alpha\top} \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}^{\alpha})^2 \right\}$$
$$= \frac{1}{n^{2q} T} E\left\{ \sum_l \sum_{g^{(l)}} (\widetilde{\boldsymbol{\eta}}_{g^{(1)}\cdots g^{(q)}}^{\alpha})^2 \mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)}^{\top} \mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)} \right\}^2$$
$$\leq \frac{1}{n^{2q} T} c \left\{ \sum_l \sum_{g^{(l)}} (\widetilde{\boldsymbol{\eta}}_{g^{(1)}\cdots g^{(q)}}^{\alpha})^2 n^q \right\}^2 = O(T^{-1}) = o(1),$$

where the inequality is obtained by Lemma 22.

(2) Proof of (A.58)

We have

$$\sum_{t=1}^{T} E\left\{ (\mathcal{Z}_t^\top \mathbb{E}_t)^2 | \mathcal{F}_{t-1} \right\} = \sigma^2 \sum_t \mathcal{Z}_t^\top \mathcal{Z}_t = \sigma^2 \sum_t \sum_l \sum_{g^{(l)}} \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}^\top \mathcal{Z}_{g^{(1)}\cdots g^{(q)},t}$$
$$= \frac{\sigma^2}{n^q T} \sum_t \sum_l \sum_{g^{(1)},\cdots,g^{(q)}} \widetilde{\boldsymbol{\eta}}_{g^{(l)}}^{(l)\top} \mathbb{X}_{g^{(1)}\cdots g^{(q)},t}^{(l)\top} \mathbb{X}_{g^{(1)}\cdots g^{(q)},t}^{(l)} \widetilde{\boldsymbol{\eta}}_{g^{(l)}}^{(l)}$$
$$+ \frac{\sigma^2}{n^q T} \sum_t \sum_{g^{(1)},\cdots,g^{(q)}} (\widetilde{\boldsymbol{\eta}}_{g^{(1)},\cdots,g^{(q)}}^{\alpha})^2 \mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)}^{\top} \mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)}.$$

In the following we prove that

$$\Big| \frac{1}{n^q T} \sum_t \sum_{g^{(1)},\cdots,g^{(q)}} \widetilde{\boldsymbol{\eta}}_{g^{(l)}}^{(l)\top} \mathbb{X}_{g^{(1)}\cdots g^{(q)},t}^{(l)\top} \mathbb{X}_{g^{(1)}\cdots g^{(q)},t}^{(l)} \widetilde{\boldsymbol{\eta}}_{g^{(l)}}^{(l)}$$
$$- \lim_{n\to\infty} \frac{1}{n^q} \sum_{g^{(1)},\cdots,g^{(q)}} \widetilde{\boldsymbol{\eta}}_{g^{(l)}}^{(l)\top} E(\mathbb{X}_{g^{(1)}\cdots g^{(q)},t}^{(l)\top} \mathbb{X}_{g^{(1)}\cdots g^{(q)},t}^{(l)}) \widetilde{\boldsymbol{\eta}}_{g^{(l)}}^{(l)} \Big| \to_p 0, \tag{A.61}$$

$$\Big|\frac{1}{n^qT}\sum_t\sum_{g^{(1)},\cdots,g^{(q)}}(\widetilde{\boldsymbol{\eta}}^{\alpha}_{g^{(1)},\cdots,g^{(q)}})^2\mathbb{Y}^{\top}_{g^{(1)}\cdots g^{(q)},(t-1)}\mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)}$$
$$-\lim_{n\to\infty}\frac{1}{n^q}\sum_{g^{(1)},\cdots,g^{(q)}}(\widetilde{\boldsymbol{\eta}}^{\alpha}_{g^{(1)},\cdots,g^{(q)}})^2E(\mathbb{Y}^{\top}_{g^{(1)}\cdots g^{(q)},(t-1)}\mathbb{Y}_{g^{(1)}\cdots g^{(q)},(t-1)})\Big|\to_p 0. \tag{A.62}$$

We first show (A.61) in the following. Note that by the step (1.1), we have

$$\frac{1}{n^qT}\sum_t\sum_{g^{(1)},\cdots,g^{(q)}}\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)}}\mathbb{X}^{(l)\top}_{g^{(1)}\cdots g^{(q)},t}\mathbb{X}^{(l)}_{g^{(1)}\cdots g^{(q)},t}\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)}}$$
$$=\frac{1}{n^qT}\sum_t\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}^{\top}\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}$$
$$+\frac{1}{n^qT}\sum_t\sum_{g^{(l)}}(\prod_{m\neq l}N_m)\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)},\cdot})^{\top}}(\mathbf{X}^{(l)}_t)^{(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}. \tag{A.63}$$

Note that

$$\frac{1}{n^qT}\sum_t\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}^{\top}\mathrm{vec}\big\{\mathcal{Y}_{t-1}\times_l\mathbf{W}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}\big\}$$
$$=\frac{1}{n^qT}\sum_t\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\left((\mathbf{I}_{N_1}\otimes\cdots\otimes\mathbf{I}_{N_{l-1}}\otimes\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\otimes\mathbf{I}_{N_{l+1}}\cdots\otimes\mathbf{I}_{N_q})\mathbb{Y}_t\right)^{\top}$$
$$\left((\mathbf{I}_{N_1}\otimes\cdots\otimes\mathbf{I}_{N_{l-1}}\otimes\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\otimes\mathbf{I}_{N_{l+1}}\cdots\otimes\mathbf{I}_{N_q})\mathbb{Y}_t\right).$$

Further note that

$$\Big|\frac{1}{n^qT}\sum_t\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\left((\mathbf{I}_{N_1}\otimes\cdots\otimes\mathbf{I}_{N_{l-1}}\otimes\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\otimes\mathbf{I}_{N_{l+1}}\cdots\otimes\mathbf{I}_{N_q})\mathbb{Y}_t\right)^{\top}$$
$$\left((\mathbf{I}_{N_1}\otimes\cdots\otimes\mathbf{I}_{N_{l-1}}\otimes\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\otimes\mathbf{I}_{N_{l+1}}\cdots\otimes\mathbf{I}_{N_q})\mathbb{Y}_t\right)$$
$$-\frac{1}{n^q}\sum_{g^{(l)}}(\widetilde{\eta}^{(l)}_{g^{(l)},1})^2\mathrm{tr}\left\{\big(\mathbf{I}_{N_q}\otimes\cdots\otimes\mathbf{I}_{N_{l-1}}\otimes\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})^{\top}}\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})}\otimes\mathbf{I}_{N_{l+1}}\otimes\cdots\otimes\mathbf{I}_{N_q}\big)\boldsymbol{\Gamma}\right\}\Big|$$
$$=o_p(1) \tag{A.64}$$

by Lemma 21, where $\widetilde{\mathbf{W}}$ in Lemma 21 is set to be $(\prod_{m\neq l}N_m)^{-1/2}\big(\mathbf{I}_{N_1}\otimes\cdots\mathbf{I}_{N_{l-1}}\otimes(\widetilde{\eta}^{(l)}_{g^{(l)},1}\mathbf{W}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})},1\le g^{(l)}\le G_l)^{\top}\otimes\mathbf{I}_{N_{l+1}}\otimes\cdots\otimes\mathbf{I}_{N_q}\big)\in\mathbb{R}^{(\prod_l N_l)\times(\prod_l N_l)}$ and we can verify that $n^{-1}\mathbf{1}^{\top}_{\prod_l N_l}\widetilde{\mathbf{W}}^{\top}\widetilde{\mathbf{W}}\mathbf{1}_{\prod_l N_l}$ is bounded under when $N_l=N_{lg^{(l)}}=O(n)$ for all $l\in[q]$.

Next define $\mathbf{x}^{(l)\dagger}_t\overset{\text{def}}{=}((\mathbf{X}^{(l)(\mathcal{R}_1,\cdot)}_t\widetilde{\boldsymbol{\eta}}^{(l)\dagger}_{1,(-1)})^{\top},(\mathbf{X}^{(l)(\mathcal{R}_2,\cdot)}_t\widetilde{\boldsymbol{\eta}}^{(l)\dagger}_{2,(-1)})^{\top},\cdots,(\mathbf{X}^{(l)(\mathcal{R}_{G_l},\cdot)}_t\widetilde{\boldsymbol{\eta}}^{(l)\dagger}_{G_l,(-1)})^{\top})^{\top}$, where $\widetilde{\boldsymbol{\eta}}^{(l)\dagger}_{g^{(l)},(-1)}=\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}/\|\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\|$. Then we have

$$\frac{1}{nT}\sum_t\left(\sum_{g^{(l)}}\widetilde{\boldsymbol{\eta}}^{(l)\top}_{g^{(l)},(-1)}\mathbf{X}^{(l)(\mathcal{R}^{(l)}_{g^{(l)},\cdot})^{\top}}_t\mathbf{X}^{(l)(\mathcal{R}_{g^{(l)},\cdot})}_t\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\right)=\frac{1}{nT}\sum_t\mathbf{x}^{(l)\dagger\top}_t\mathbf{A}^{(l)}\mathbf{x}^{(l)\dagger}_t \tag{A.65}$$

where $\mathbf{A}^{(l)} = \mathrm{diag}\{\|\widetilde{\boldsymbol{\eta}}^{(l)}_{g^{(l)},(-1)}\|^2 \mathbf{1}_{N_{lg^{(l)}}} : g^{(l)} \in [G_l]\}$. Note that $\mathbf{x}_t^\dagger$ satisfies K-convex concentration property defined in Definition 1. Then by using Lemma 27, we have

$$\big|(nT)^{-1}\sum_t \mathbf{x}_t^{(l)\dagger\top}\mathbf{A}^{(l)}\mathbf{x}_t^{(l)\dagger} - n^{-1}E(\mathbf{x}_t^{(l)\dagger\top}\mathbf{A}^{(l)}\mathbf{x}_t^{(l)\dagger})\big| = o_p(1).$$

Together with (A.64) and (A.63), and by taking the limit, we have proved (A.61).

Next, (A.62) can be similarly proved by using the same technique as (A.64) and Lemma 21. Together with the fact that $n^{-q/2}T^{-1/2}\boldsymbol{\eta}^\top\mathbf{M}_{nT}^{-1}\left(\mathbf{M}_{nT}^0 - \mathbf{M}_{nT}\right)(\mathbf{M}_{nT}^0)^{-1}\boldsymbol{\delta} = o_p(1)$, we reach the conclusion.

Appendix J. Technical Lemmas

Lemma 17. *Suppose Assumptions 1–5 hold. Then we have*

$$\begin{aligned}&P\Big(\sup_{\|\boldsymbol{\Theta}_{i_1\cdots i_q}\|_{\max}<R}\Big|\frac{1}{T}Q_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q}) - \frac{1}{T}Q^*_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q})\Big| > x\Big)\\ &\le \exp\Big\{-c_1\min(Tx^2, T^{1/2}x) + c_2 m\Big\},\end{aligned} \tag{A.66}$$

where $m = \sum_l(p_l+1)+1$, and c_1, c_2 are positive constants, In addition, we have

$$\begin{aligned}&\sup_{\|\boldsymbol{\Theta}\|_{\max}<R}\Big|\frac{1}{(\prod_l N_l)T}\Big\{Q(\boldsymbol{\Theta}) - Q^*(\boldsymbol{\Theta})\Big\}\Big|\\ &\le \sup_{i_1,\cdots,i_q}\sup_{\|\boldsymbol{\Theta}_{i_1\cdots i_q}\|_{\max}<R}\Big|\frac{1}{T}\Big\{Q_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q}) - \frac{1}{T}Q^*_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q})\Big\}\Big|\\ &= O_p\Big(T^{-1/2}(m + \sum_l \log N_l)\Big).\end{aligned} \tag{A.67}$$

Proof Note that we have

$$\begin{aligned}\frac{1}{T}Q_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q}) &= \frac{1}{T}\sum_{t=1}^T\big(\varepsilon_{i_1\cdots i_q,t} + \mathcal{X}^\top_{i_1\cdots i_q,t}\boldsymbol{\Theta}^0_{i_1\cdots i_q} - \mathcal{X}^\top_{i_1\cdots i_q,t}\boldsymbol{\Theta}_{i_1\cdots i_q}\big)^2\\ &= \frac{1}{T}\sum_{t=1}^T\Big\{\varepsilon^2_{i_1\cdots i_q,t} + 2\varepsilon_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\boldsymbol{\Theta}^0_{i_1\cdots i_q} - \boldsymbol{\Theta}_{i_1\cdots i_q})\\ &\quad + (\boldsymbol{\Theta}_{i_1\cdots i_q} - \boldsymbol{\Theta}^0_{i_1\cdots i_q})^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\boldsymbol{\Theta}_{i_1\cdots i_q} - \boldsymbol{\Theta}^0_{i_1\cdots i_q})\Big\}.\end{aligned}$$

Recall that $\boldsymbol{\Sigma}_{i_1\cdots i_q} = E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t})$. It is sufficient to show

$$P\Big(\Big|\frac{1}{T}\sum_t \varepsilon^2_{i_1\cdots i_q,t} - \sigma^2\Big| > x/3\Big) \le 2\exp\{-c_1 T\min(x^2, x)\} \tag{A.68}$$

$$P\Big(\sup_{\|\mathbf{a}\|_{\max}<2R}\Big|\frac{2}{T}\sum_t \varepsilon_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}\mathbf{a}\Big| > x/3\Big) \le \exp\Big\{-c_1\min(Tx^2, T^{1/2}x) + c_3 m\Big\}, \tag{A.69}$$

$$P\Big(\sup_{\|\mathbf{a}\|_{\max}<2R} \Big| \frac{1}{T}\sum_t \mathbf{a}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1^*\cdots i_q,t}^\top \mathbf{a} - \mathbf{a}^\top \mathbf{\Sigma}_{i_1\cdots i_q}\mathbf{a} \Big| > x/3 \Big)$$
$$\le \exp\Big\{ - c_1 \min(Tx^2, T^{1/2}x) + c_3 m \Big\} \tag{A.70}$$

where $m = \sum_l (p_l+1)+1$ as defined in Theorem 2, and c_1, c_3 are positive constants. First by Lemma 27, (A.68) holds directly. We next prove (A.70) in two steps. First we establish the upper bound for a fixed $\mathbf{a}$, then we establish the upper bound for all $\mathbf{a}$ satisfying $\|\mathbf{a}\|_{\max} < 2R$. This yields (A.66) and (A.67) holds subsequently.

Step 1. We first show

$$P\Big(\Big| \frac{1}{T}\sum_t \mathbf{a}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1^*\cdots i_q,t}^\top \mathbf{a} - \mathbf{a}^\top \mathbf{\Sigma}_{i_1\cdots i_q}\mathbf{a} \Big| > x/3 \Big) \le c_0 \exp\Big(- c_3 \min(Tx^2, T^{1/2}x) \Big)$$

for any $\mathbf{a}$, where c_3 and c_0 are positive constants. For the sake of similarity, we only verify the concentration inequality for diagonal elements of $T^{-1}\sum_t \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1^*\cdots i_q,t}^\top$. Note that the diagonal elements of $T^{-1}\sum_t \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1^*\cdots i_q,t}^\top$ takes the form $T^{-1}\sum_t \mathbf{w}^\top \mathbb{Y}_t \mathbb{Y}_t^\top \mathbf{w}$ (with $\mathbf{w} \in \mathbb{R}^{\prod_l N_l}$ and $\|\mathbf{w}\|_1 = 1$), or $T^{-1}\sum_t (\mathbf{x}_{i_l j,t}^{(l)})^2$, where $\mathbf{x}_{i_l j,t}^{(l)}$ is the jth element of the p_l-length vector $\mathbf{x}_{i_l,t}^{(l)}$. For $T^{-1}\sum_t \mathbf{x}_{i_l,t}^{(l)\top}\mathbf{x}_{i_l,t}^{(l)}$. We use Lemma 27 and Assumption 4, and note that $c_0\exp\{-c_3 T\min(x^2,x)\} \le c_0\exp\{-c_3\min(Tx^2, T^{1/2}x)\}$. This yields the concentration inequality. For the $T^{-1}\sum_t \mathbf{w}^\top \mathbb{Y}_t\mathbb{Y}_t^\top \mathbf{w}$, we use Lemma 18 to obtain the result.

Step 2. Here we follow Lemma F.2 of Basu et al. (2015) to prove the result. First we define $\mathcal{M} = \{\mathbf{v} \in \mathbb{R}^m : \|\mathbf{v}\|_{\max} \le 2R\}$, where $m = \sum_l (p_l+1)+1$. Let $\mathcal{A} = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{|\mathcal{A}|}\}$ be a $R/5$-net of $\mathcal{M}$, where $|\mathcal{A}| \le 10^m$. Hence for any $\mathbf{v} \in \mathcal{M}$, there exists some $\boldsymbol{u}_i \in \mathcal{A}$ such that $\|\Delta\mathbf{v}\|_{\max} \le R/5$, where $\Delta\mathbf{v} = \mathbf{v} - \boldsymbol{u}_i$. Define $\mathbf{M} = |T^{-1}\sum_t \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1^*\cdots i_q,t}^\top - \mathbf{\Sigma}_{i_1\cdots i_q}|_e$.

$$\tau \overset{\text{def}}{=} \sup_{\mathbf{v}\in\mathcal{M}} |\mathbf{v}^\top \mathbf{M}\mathbf{v}| \le \max_{1\le k\le|\mathcal{A}|} |\boldsymbol{u}_k^\top \mathbf{M}\boldsymbol{u}_k| + 2\sup_{\mathbf{v}\in\mathcal{M}} |\max_{1\le k\le|\mathcal{A}|} \Delta\mathbf{v}^\top \mathbf{M}\boldsymbol{u}_k| + \sup_{\mathbf{v}\in\mathcal{M}} |\Delta\mathbf{v}^\top\mathbf{M}\Delta\mathbf{v}|$$
$$\le 2\max_{1\le k\le|\mathcal{A}|} |\boldsymbol{u}_k^\top \mathbf{M}\boldsymbol{u}_k| + 2\sup_{\mathbf{v}\in\mathcal{M}} |\Delta\mathbf{v}^\top\mathbf{M}\Delta\mathbf{v}| \le 2\max_{1\le k\le|\mathcal{A}|} |\boldsymbol{u}_k^\top \mathbf{M}\boldsymbol{u}_k| + \tau/50,$$

where the last step is because $10\Delta\mathbf{v} \in \mathcal{M}$. Thus $\tau \le (100/49)\max_{1\le k\le|\mathcal{A}|} |\boldsymbol{u}_k^\top\mathbf{M}\boldsymbol{u}_k|$. This leads to

$$\sup_{\|\mathbf{a}\|_{\max}\le 2R} P\Big(\Big| \frac{1}{T}\sum_t \mathbf{a}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1^*\cdots i_q,t}^\top \mathbf{a} - \mathbf{a}^\top \mathbf{\Sigma}_{i_1\cdots i_q}\mathbf{a} \Big| > x/3 \Big)$$
$$= P\Big\{ \sup_{\mathbf{v}\in\mathcal{M}} |\mathbf{v}^\top\mathbf{M}\mathbf{v}| > \frac{x}{3} \Big\} \le P\Big\{ (100/49) \max_{1\le k\le|\mathcal{A}|} |\boldsymbol{u}_k^\top\mathbf{M}\boldsymbol{u}_k| > x/3 \Big\}$$
$$\le c_0|\mathcal{A}| \exp\Big(- c_4 \min(T(49x/100)^2, T^{1/2}49x/100) \Big) \le \exp\Big\{ - c_1 \min(Tx^2, T^{1/2}x) + c_3 m \Big\},$$

where c_3, c_4 are positive constants. (A.69) can be proved similarly hence we skip the details. ■

Lemma 18. *Under Assumption 4, then we have*

$$P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t\mathbb{Y}_t^\top \mathbf{w} - \mathbf{w}^\top \mathbf{\Gamma}\mathbf{w}\Big| > u\Big) \le 2(q^2+q+1)\exp\Big(-C_1\min(Tu^2, T^{1/2}u)\Big), \tag{A.71}$$

where $\mathbf{w} \in \mathbb{R}^{\prod_l N_l}$ *is element-wisely non-negative and* $\|\mathbf{w}\|_1 = 1$ *and* $\mathbf{\Gamma} = cov(\mathbb{Y}_t)$*, and* C_1, C_2, C_3 *are positive constants.*

Proof By (A.6), we have

$$\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t\mathbb{Y}_t^\top \mathbf{w} = \frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^c\mathbb{Y}_t^{c\top} \mathbf{w} + \frac{2}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^c\mathbb{Y}_t^{e\top} \mathbf{w} + \frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^e\mathbb{Y}_t^{e\top} \mathbf{w}.$$

Specifically, let $\mathbf{\Gamma}^c = \mathrm{cov}(\mathbb{Y}_t^c)$ and $\mathbf{\Gamma}^e = \mathrm{cov}(\mathbb{Y}_t^e)$. It suffices to show

$$P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^c\mathbb{Y}_t^{c\top} \mathbf{w} - \mathbf{w}^\top \mathbf{\Gamma}^c\mathbf{w}\Big| > u/3\Big) \le 2q^2\exp\{-c_1\min(Tu^2, T^{1/2}u)\} \tag{A.72}$$

$$P\Big(\Big|\frac{2}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^c\mathbb{Y}_t^{e\top} \mathbf{w}\Big| > u/3\Big) \le 2q\exp\{-c_2\min(Tu^2, T^{1/2}u)\} \tag{A.73}$$

$$P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^e\mathbb{Y}_t^{e\top} \mathbf{w} - \mathbf{w}^\top \mathbf{\Gamma}^e\mathbf{w}\Big| > u/3\Big) \le 2\exp\Big(-c_3\min\Big(Tu^2, T^{1/2}u\Big)\Big). \tag{A.74}$$

where c_1, c_2, c_3 are positive constants. In summary we obtain the final conclusion.

1. Proof of (A.72)

Let $\mathbf{A} = \mathbf{w}\mathbf{w}^\top \in \mathbb{R}^{(\prod_l N_l)\times(\prod_l N_l)}$. By (A.6),

$$\begin{aligned}
&\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^c\mathbb{Y}_t^{c\top} \mathbf{w} = \frac{1}{T}\sum_{t=1}^T \mathbb{Y}_t^{c\top}\mathbf{A}\mathbb{Y}_t^c = \frac{1}{T}\sum_{t=1}^T\sum_{l_1=0}^t\sum_{s_2=0}^t \mathbf{c}_{l_1}^\top(\mathbf{B}_0^{t-l_1})^\top\mathbf{A}\mathbf{B}_0^{t-s_2}\mathbf{c}_{s_2} \\
&= \sum_{l_1=0}^T\sum_{s_2=0}^T \mathbf{c}_{l_1}^\top\Big\{\frac{1}{T}\sum_{t=\max\{1,l_1,s_2\}}^T (\mathbf{B}_0^{t-l_1})^\top\mathbf{A}\mathbf{B}_0^{t-s_2}\Big\}\mathbf{c}_{s_2} \overset{\text{def}}{=} \sum_{l_1=0}^T\sum_{s_2=0}^T \mathbf{c}_{l_1}^\top\mathbf{D}_{l_1s_2}\mathbf{c}_{s_2}.
\end{aligned} \tag{A.75}$$

For any $l \ge 0$, we define $\mathbf{c} = (\mathbf{c}_0^\top, \mathbf{c}_1^\top, \cdots, \mathbf{c}_T^\top)^\top$, and

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{00} & \mathbf{D}_{01} & \cdots & \mathbf{D}_{0T} \\ \mathbf{D}_{10} & \mathbf{D}_{11} & \cdots & \mathbf{D}_{1T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_{T0} & \mathbf{D}_{T1} & \cdots & \mathbf{D}_{TT} \end{pmatrix}. \tag{A.76}$$

Then we have $T^{-1}\sum_{t=1}^T \mathbb{Y}_t^{c\top}\mathbf{A}\mathbb{Y}_t^c = \mathbf{c}^\top\mathbf{D}\mathbf{c} = \sum_{l=1}^q \mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(l)} + 2\sum_{l\neq m}\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(m)}$, where $\mathbf{c}^{(l)} = (\mathbf{c}_0^{(l)\top}, \cdots, \mathbf{c}_T^{(l)\top})$. Obviously,

$$P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^c\mathbb{Y}_t^{c\top} \mathbf{w} - \mathbf{w}^\top \mathbf{\Gamma}^c\mathbf{w}\Big| > u/3\Big)$$

$$= P\Big(\Big|\sum_{l=1}^{q}\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(l)} + 2\sum_{l\neq m}\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(m)} - E(\sum_{l=1}^{q}\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(l)} + 2\sum_{l\neq m}\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(m)})\Big| > u/3\Big)$$
$$\leq \sum_{l} P\{|\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(l)} - E(\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(l)})| > 2u/3q(q+1)\}$$
$$+ \sum_{l\neq m} P\{|\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(m)} - E(\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(m)})| > 2u/3q(q+1)\}.$$

Recall that $\mathbf{c}_t^{(l)}$ is defined as $\text{vec}(\mathbf{1}_{N_1}\circ\cdots\circ\boldsymbol{\beta}^0_{X_l,t}\circ\cdots\mathbf{1}_{N_q})\in\mathbb{R}^{\prod_l N_l}$ in the notations, and denote $\mathbf{c}^{(l)} = (\mathbf{c}_t^{(l)\top} : 0\leq t\leq T)^\top \in \mathbb{R}^{(\prod_l N_l)(T+1)}$. Also denote the $C_u u = 2u/3q(q+1)$. Due to the bounded assumption for parameters in Assumption 1, we treat $\mathbf{c}_t^{(l)}$ as $\widetilde{\mathbf{x}}_t^{(l)\eta}$ in Lemma 29, which indicates that

$$P\{|\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(l)} - E(\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(l)})| > C_u u\}$$
$$= P\Big\{\Big|\sum_{l_1,s_2=0}^{T}\{\mathbf{c}_{l_1}^{(l)\top}\mathbf{D}_{l_1 s_2}\mathbf{c}_{s_2}^{(l)} - E(\mathbf{c}_{l_1}^{(l)\top}\mathbf{D}_{l_1 s_2}\mathbf{c}_{s_2}^{(l)})\}\Big| > C_u u\Big\}$$
$$\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{(C_u u)^2}{\|\mathcal{D}^{(l)}\|_F^2}, \frac{C_u u}{\|\mathcal{D}^{(l)}\|}\Big)\Big) \leq 2\exp\{-c_1\min(u^2T, uT^{1/2})\},$$

where $\mathcal{D}^{(l)}$ is given in Lemma 29 and the last step is due to Lemma 30. Similar treatment leads to

$$P\{|\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(m)} - E(\mathbf{c}^{(l)\top}\mathbf{D}\mathbf{c}^{(m)})| > C_u u\} \leq 2\exp\{-c_1\min(u^2T, uT^{1/2})\}$$

Combining the above results, we get

$$P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top\mathbb{Y}_t^c\mathbb{Y}_t^{c\top}\mathbf{w} - \mathbf{w}^\top\boldsymbol{\Gamma}^c\mathbf{w}\Big| > u/3\Big) \leq 2q^2\exp\{-c_1\min(Tu^2, T^{1/2}u)\},$$

where c_1 is a constant.

2. Proof of (A.73)

We can obtain that $\frac{1}{T}\sum_t \mathbf{w}^\top\mathbb{Y}_t^c\mathbb{Y}_t^{e\top}\mathbf{w} = \sum_{l_1=0}^{T}\sum_{s_2=0}^{T}\mathbf{c}_{l_1}^\top\mathbf{D}_{l_1 s_2}\mathbb{E}_{s_2} = \mathbf{c}^\top\mathbf{D}\mathbb{E}$, where $\mathbb{E} = (\mathbb{E}_0^\top,\ldots,\mathbb{E}_t^\top)^\top \in \mathbb{R}^{(\prod_l N_l)(T+1)}$. Obviously,

$$P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top\mathbb{Y}_t^c\mathbb{Y}_t^{e\top}\mathbf{w}\Big| > u/6\Big) = P\Big(\Big|\sum_l \mathbf{c}^{(l)\top}\mathbf{D}\mathbb{E}\Big| > u/6\Big) \leq \sum_l P\{|\mathbf{c}^{(l)\top}\mathbf{D}\mathbb{E}| > u/(6q)\}.$$

Treating $\mathbf{c}_t^{(l)}$, which is defined as $\text{vec}(\mathbf{1}_{N_1}\circ\cdots\circ\boldsymbol{\beta}^0_{X_l,t}\circ\cdots\circ\mathbf{1}_{N_q})$, as $\widetilde{\mathbf{x}}_t^{(l)\eta}$ in Lemma 29, we get

$$P\{|\mathbf{c}^{(l)\top}\mathbf{D}\mathbb{E}| > u/(6q)\} = P\{|\sum_{l_1,s_2=0}^{T}|\mathbf{c}_{l_1}^{(l)\top}\mathbf{D}_{l_1 s_2}\mathbb{E}_{s_2}| > u/(6q)\}$$
$$\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{u^2}{\|\widetilde{\mathcal{D}}^{(l)}\|_F^2}, \frac{u}{\|\widetilde{\mathcal{D}}^{(l)}\|}\Big)\Big) \leq 2\exp\{-c_2\min(Tu^2, T^{1/2}u)\},$$

where $\widetilde{\mathcal{D}}^{(l)}$ is given in Lemma 29 and the last step is due to Lemma 30. Therefore, we get

$$P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{e\top}\mathbf{w}\Big| > u/6\Big) \le 2q\exp\{-c_2\min(Tu^2, T^{1/2}u)\},$$

where c_2 is a positive constant.

3. Proof of (A.74)

Similar to the expression of (A.75), we can obtain that

$$\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top}\mathbf{w} = \sum_{l_1=0}^{T}\sum_{s_2=0}^{T} \mathbb{E}_{l_1}^\top \mathbf{D}_{l_1 s_2}\mathbb{E}_{s_2} = \mathbb{E}^\top \mathbf{D}\mathbb{E},$$

where $\mathbb{E} = (\mathbb{E}_0^\top, \ldots, \mathbb{E}_T^\top)^\top \in \mathbb{R}^{\prod_l N_l}$. Since $\mathbb{E}$ follows the K-convex concentration property, following Lemma 27, we have

$$\begin{aligned} &P\Big(\Big|\frac{1}{T}\sum_t \mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top}\mathbf{w} - \mathbf{w}^\top \mathbf{\Gamma}^e \mathbf{w}\Big| > u/3\Big) = P\Big(\Big|\mathbb{E}^\top\mathbf{D}\mathbb{E} - E(\mathbb{E}^\top\mathbf{D}\mathbb{E})\Big| > u/3\Big) \\ &\le 2\exp\Big(-\frac{1}{C}\min\Big(\frac{u^2/9}{K^4\|\mathbf{D}\|_F^2}, \frac{u/3}{K^2\|\mathbf{D}\|}\Big)\Big) \le 2\exp\Big(-c_3\min\Big(Tu^2, T^{1/2}u\Big)\Big), \end{aligned}$$

where c_3 is a finite constant and the last step is due to (A.100) of Lemma 30.

■

Lemma 19. *Under Assumptions 1–5, denote* $d_{i_1\cdots i_q} \overset{\text{def}}{=} d_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}, \mathbf{\Theta}^0_{i_1\cdots i_q})$. *Recall that* $S_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) = Q_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) - Q_{i_1\cdots i_q}(\mathbf{\Theta}^0_{i_1\cdots i_q})$. *we have*

$$\begin{aligned} &P\Big\{\sup_{\mathbf{\Theta}_{i_1\cdots i_q}} T^{-1}\frac{|S_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) - S^*_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q})|}{\sqrt{d_{i_1\cdots i_q}}} > x\Big\} \\ &\le C_1\exp\Big\{-C_2\min(Tx^2, \sqrt{T}x) + C_3 m\Big\}, \end{aligned} \tag{A.77}$$

where $m = \sum_l (p_l+1)+1$. *Furthermore, we have*

$$\max_l \sup_{i_l} \sup_{d_{i_1\cdots i_q}\le\omega^2} T^{-1}|S_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) - S^*_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q})| = O_p\big\{\omega(\sum_l \log N_l)/\sqrt{T}\big\} \tag{A.78}$$

$$\sup_{d(\mathbf{\Theta},\mathbf{\Theta}^0)\le\omega^2} (\prod_l N_l T)^{-1}|S(\mathbf{\Theta}) - S^*(\mathbf{\Theta})| = O_p\big\{\omega(\sum_l \log N_l)/\sqrt{T}\big\}. \tag{A.79}$$

Proof

1. Proof of (A.77)**.**

We first write

$$\begin{aligned} &S_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) - S^*_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) \\ &= \Big[(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}) \end{aligned}$$

$$- E\{(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})\}\Big] \tag{A.80}$$

$$+ 2\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})\varepsilon_{i_1\cdots i_q,t}. \tag{A.81}$$

We use similar techniques in the proof of Lemma 17 to derive the concentration inequality. For example, we can apply (A.70) on term (A.80) by noticing that $\|\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}\|/\sqrt{d_{i_1\cdots i_q}} \leq cR$ for some constant c. Similarly, (A.81) can be shown by applying the proof for (A.69). Hence, we can obtain the conclusion under the same Assumptions of Lemma 17.

2. Proof of (A.78) **and** (A.79)**.**

For simplicity, denote $d_{i_1\cdots i_q} \overset{\text{def}}{=} d_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}, \mathbf{\Theta}^0_{i_1\cdots i_q})$, and $S_{i_1\cdots i_q} \overset{\text{def}}{=} S_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q})$. Note that we have

$$\begin{aligned}
&P\Big\{\sup_{d_{i_1\cdots i_q}\leq\omega^2}|S_{i_1\cdots i_q} - S^*_{i_1\cdots i_q}| > x\Big\} \\
&= P\Big\{\sup_{d_{i_1\cdots i_q}\leq\omega^2}|S_{i_1\cdots i_q} - S^*_{i_1\cdots i_q}|/\omega > x/\omega\Big\} \\
&\leq P\Big\{\sup_{d_{i_1\cdots i_q}\leq\omega^2}|S_{i_1\cdots i_q} - S^*_{i_1\cdots i_q}|/\sqrt{d_{i_1\cdots i_q}} > x/\omega\Big\} \\
&\leq P\Big\{\sup_{\mathbf{\Theta}_{i_1\cdots i_q}}|S_{i_1\cdots i_q} - S^*_{i_1\cdots i_q}|/\sqrt{d_{i_1\cdots i_q}} > x/\omega\Big\} \\
&\leq C_1\exp\big\{-C_2\min(Tx^2/\omega^2, \sqrt{T}x/\omega) + C_3 m\big\},
\end{aligned}$$

where the last line uses (A.77). By taking the union bound, we have

$$\max_l \sup_{i_l}\sup_{d_{i_1\cdots i_q}\leq\omega^2}|S_{i_1\cdots i_q} - S^*_{i_1\cdots i_q}| = O_p\big\{\omega(\sum_l \log N_l)/\sqrt{T}\big\}.$$

Using similar procedure, we can obtain (A.79) holds.

■

Lemma 20. *Under Assumptions 1–5, for vector $\mathbf{\Delta} \in \mathbb{R}^{\sum_l(p_l+1)+1}$ in the parameter space, we have*

$$\begin{aligned}
&P\Big\{\sup_{i_l}\sup_{\|\mathbf{\Delta}\|^2\leq\omega^2}(\prod_{m\neq l}N_m T)^{-1}\Big|\sum_{m\neq l}\sum_{i_m}\sum_t \\
&\qquad\qquad\Big\{\mathbf{\Delta}^\top\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}\mathbf{\Delta} - \big[\mathbf{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t})\mathbf{\Delta}\big]\Big\}\Big| \geq x\Big\} \\
&\leq C_1\exp\Big\{-C_2\min\Big(Tx^2/\omega^2, \sqrt{T}x/\omega\Big) + C_3 m + \sum_l \log N_l\Big\},
\end{aligned} \tag{A.82}$$

$$\begin{aligned}
&P\Big\{\sup_{i_l}\sup_{\|\mathbf{\Delta}\|^2\leq\omega^2}(\prod_{m\neq l}N_m T)^{-1}\Big|\sum_{m\neq l}\sum_{i_m}\sum_t \mathcal{X}^\top_{i_1\cdots i_q,t}\mathbf{\Delta}\varepsilon_{i_1\cdots i_q,t}\Big| \geq x\Big\} \\
&\leq C_1\exp\Big\{-C_2\min\Big(Tx^2/\omega^2, \sqrt{T}x/\omega\Big) + C_3 m + \sum_l \log N_l\Big\},
\end{aligned} \tag{A.83}$$

where $\mathcal{X}_{i_1\cdots i_q,t}$ is defined in equation (17)*, $m = \sum_l(p_l+1)+1$, and C_1, C_2, C_3 are positive constants.*

Proof We first derive (A.82). Note that

$$
\begin{aligned}
&P\Big\{ \sup_{i_l} \sup_{\|\boldsymbol{\Delta}\|^2\le\omega^2} \frac{1}{\prod_{m\neq l} N_m T}\Big| \sum_{m\neq l}\sum_{i_m}\sum_{t} \Big\{ \boldsymbol{\Delta}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top \boldsymbol{\Delta} \\
&\quad - \big[\boldsymbol{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top)\boldsymbol{\Delta}\big]\Big\}\Big| \ge x\Big\} \\
&\le P\Big\{ \sup_{i_l} \sup_{\|\boldsymbol{\Delta}\|^2\le\omega^2} \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l}\sum_{i_m} \\
&\quad \frac{T^{-1}\Big| \sum_t \boldsymbol{\Delta}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top \boldsymbol{\Delta} - \big[\boldsymbol{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top)\boldsymbol{\Delta}\big]\Big|}{\|\boldsymbol{\Delta}\|}\|\boldsymbol{\Delta}\| \ge x\Big\} \\
&\le P\Big\{ \sup_{i_l} \sup_{\|\boldsymbol{\Delta}\|^2\le\omega^2} \\
&\quad \sqrt{\sum_{m\neq l}\sum_{i_m} \frac{T^{-2}\Big| \sum_t \boldsymbol{\Delta}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top \boldsymbol{\Delta} - \big[\boldsymbol{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top)\boldsymbol{\Delta}\big]\Big|^2}{\|\boldsymbol{\Delta}\|^2}}\|\boldsymbol{\Delta}\| \ge x\Big\} \\
&= P\Big\{ \sup_{i_l} \sup_{\|\boldsymbol{\Delta}\|^2\le\omega^2} \sum_{m\neq l}\sum_{i_m} \\
&\quad \frac{T^{-2}\Big| \sum_t \boldsymbol{\Delta}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top \boldsymbol{\Delta} - \big[\boldsymbol{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top)\boldsymbol{\Delta}\big]\Big|^2}{\|\boldsymbol{\Delta}\|^2}\|\boldsymbol{\Delta}\|^2 \ge x^2\Big\} \\
&\le P\Big\{ \sup_{i_l} \sup_{\|\boldsymbol{\Delta}\|^2\le\omega^2} \sum_{m\neq l}\sum_{i_m} \\
&\quad \frac{T^{-2}\Big| \sum_t \boldsymbol{\Delta}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top \boldsymbol{\Delta} - \big[\boldsymbol{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top)\boldsymbol{\Delta}\big]\Big|^2}{\|\boldsymbol{\Delta}\|^2}\omega^2 \ge x^2\Big\} \\
&\le \sum_l \sum_{i_l=1}^{N_l} P\Big\{ \sup_{\|\boldsymbol{\Delta}\|^2\le\omega^2} \\
&\quad \frac{T^{-2}\Big| \sum_t \boldsymbol{\Delta}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top \boldsymbol{\Delta} - \big[\boldsymbol{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top)\boldsymbol{\Delta}\big]\Big|^2}{\|\boldsymbol{\Delta}\|^2}\omega^2 \ge x^2\Big\} \\
&\le \sum_l \sum_{i_l=1}^{N_l} P\Big\{ \sup_{\boldsymbol{\Delta}} \frac{T^{-1}\Big| \sum_t \boldsymbol{\Delta}^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top \boldsymbol{\Delta} - \big[\boldsymbol{\Delta}^\top E(\mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}_{i_1\cdots i_q,t}^\top)\boldsymbol{\Delta}\big]\Big|}{\|\boldsymbol{\Delta}\|}\omega \ge x\Big\} \\
&\le C_1 \exp\Big\{ - C_2 \min(Tx^2/\omega^2, \sqrt{T}x/\omega) + C_3 m + \log(\prod_l N_l)\Big\},
\end{aligned}
$$

where $m = \sum_l (p_l + 1) + 1$, the third line holds due to $\sum_i (a_i^2 b_i^2) \le (\sum_i a_i^2)(\sum_i b_i^2)$. The last line is obtained by applying (A.70) and using that $\|\boldsymbol{\Delta}\|_{\max}/\|\boldsymbol{\Delta}\| \le cR$. The proof of (A.83) can be finished in a similar scheme, while applying (A.69). ■

Lemma 21. *Under Assumption 4 and 5, let* $\mathbf{M} = \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{W}} \in \mathbb{R}^{(\prod_l N_l)\times(\prod_l N_l)}$ *be a symmetric matrix. Here* $\widetilde{\mathbf{W}} \in \mathbb{R}^{m\times(\prod_l N_l)}$ *is elementwisely positive and assume* $n^{-1}\mathbf{1}_{\prod_l N_l}^\top \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{W}} \mathbf{1}_{\prod_l N_l} = O(1)$*, where* n *is the order of* N_l*,* $l \in [q]$*. Then we have*

$$P\Big(\Big|\frac{1}{nT}\sum_t \mathbb{Y}_t^\top \mathbf{M}\mathbb{Y}_t - n^{-1}tr(\mathbf{\Gamma M})\Big| > u\Big) \le 2(q^2+q+1)\exp\Big(-c\min(Tu^2, T^{1/2}u)\Big),$$

and we have $n^{-1}tr(\mathbf{M\Gamma}) \le c$ *for a positive constant* c*.*

Proof By (A.6), we have $\frac{1}{T}\sum_t \mathbb{Y}_t^\top \mathbf{M}\mathbb{Y}_t = \frac{1}{T}\sum_t \mathbb{Y}_t^{c\top}\mathbf{M}\mathbb{Y}_t^c + \frac{2}{T}\sum_t \mathbb{Y}_t^{e\top}\mathbf{M}\mathbb{Y}_t^c + \frac{1}{T}\sum_t \mathbb{Y}_t^{e\top}\mathbf{M}\mathbb{Y}_t^e$. The proof follows the proof of Lemma 18. The difference is that we need to replace the $\mathbf{A}$ matrix with $\mathbf{M}/n$. Since the procedure is the same we omit the details here.

Furthermore, we have $n^{-1}\mathrm{tr}(\mathbf{M\Gamma}) \le n^{-1}\mathrm{tr}(\mathbf{M}\mathbf{1}_{\prod_l N_l}\mathbf{1}_{\prod_l N_l}^\top c_\Gamma) \le c_\Gamma \mathbf{1}_{\prod_l N_l}^\top \mathbf{M}\mathbf{1}_{\prod_l N_l}/n = O(1)$, where the first inequality is obtained by Lemma 32. ■

Lemma 22. *Under Assumption 4, we have* $\max_l \max_{i_l} E(Y_{i_1\cdots i_q,t}^4) \le c$*, where* c *is a positive constant.*

Proof Let $\mathbf{w} = \mathbf{e}_{i_q}^{(N_q)} \otimes \cdots \otimes \mathbf{e}_{i_1}^{(N_1)} \in \mathbb{R}^{\prod_l N_l}$, where $\mathbf{e}_{i_l}^{(N_l)} \in \mathbb{R}^{N_l}$ is a vector whose i_lth element being equal to 1 while others being equal to 0. Then we have $Y_{i_1\cdots i_q,t} = \mathbf{w}^\top \mathbb{Y}_t$. By (A.6), we have

$$\begin{aligned} Y_{i_1\cdots i_q,t}^2 &= \mathbf{w}^\top \mathbb{Y}_t \mathbb{Y}_t^\top \mathbf{w} = \mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{c\top}\mathbf{w} + 2\mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{e\top}\mathbf{w} + \mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top}\mathbf{w} \\ &\le 2\mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{c\top}\mathbf{w} + 2\mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top}\mathbf{w}. \end{aligned}$$

Thus, $Y_{i_1\cdots i_q,t}^4 = (\mathbf{w}^\top \mathbb{Y}_t \mathbb{Y}_t^\top \mathbf{w})^2 \le 4(\mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{c\top}\mathbf{w})^2 + 4(\mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top}\mathbf{w})^2$. It suffices to investigate the above two terms respectively.

1. Proof of $E\{(\mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{c\top}\mathbf{w})^2\} < \infty$

First, let $\mathbf{A} = \mathbf{w}\mathbf{w}^\top \in \mathbb{R}^{\prod_l N_l \times \prod_l N_l}$, then by (A.6) and the definition in Lemma 30,

$$\begin{aligned} &\mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{c\top}\mathbf{w} = \mathbb{Y}_t^{c\top}\mathbf{A}\mathbb{Y}_t^c = \sum_{s_1=0}^t \sum_{s_2=0}^t \mathbf{c}_{s_1}^\top (\mathbf{B}_0^{t-s_1})^\top \mathbf{A}\mathbf{B}_0^{t-s_2}\mathbf{c}_{s_2} \\ &= \sum_{s_1=0}^t \sum_{s_2=0}^t \mathbf{c}_{s_1}^\top \mathbf{L}_{t,s_1s_2}\mathbf{c}_{s_2} = \sum_{s_1=0}^t \sum_{s_2=0}^t (\sum_l \mathbf{c}_{s_1}^{(l)\top})\mathbf{L}_{t,s_1s_2}(\sum_l \mathbf{c}_{s_2}^{(l)}) \\ &\le 2\sum_l \sum_{s_1=0}^t \sum_{s_2=0}^t \mathbf{c}_{s_1}^{(l)\top}\mathbf{L}_{t,s_1s_2}\mathbf{c}_{s_2}^{(l)} = 2\sum_l \sum_{s_1=0}^t \sum_{s_2=0}^t \boldsymbol{\beta}_{X_l,s_1}^{(l)0\top}\mathcal{L}_{t,s_1s_2}^{(l)}\boldsymbol{\beta}_{X_l,s_2}^{(l)0}. \end{aligned} \tag{A.84}$$

where $\mathcal{L}_{t,s_1s_2}^{(l)} = (\mathbf{1}_{N_1}^\top \otimes \cdots \otimes \mathbf{1}_{N_{l-1}}^\top \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}}^\top \otimes \cdots \otimes \mathbf{1}_{N_q}^\top)\mathbf{L}_{t,s_1s_2}(\mathbf{1}_{N_1} \otimes \cdots \otimes \mathbf{1}_{N_{l-1}} \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}} \otimes \cdots \otimes \mathbf{1}_{N_q}) \in \mathbb{R}^{N_l \times N_l}$ as defined in Lemma 30.

Further, let $\mathcal{L}_t^{(l)} = (\widetilde{\mathcal{L}}_{t,s_1s_2}^{(l)} : 0 \le s_1, s_2 \le T)$, where $\widetilde{\mathcal{L}}_{t,s_1s_2}^{(l)} = \mathbf{0}$ for $s_1 > t$ or $s_2 > t$, and $\widetilde{\mathcal{L}}_{t,s_1s_2}^{(l)} = \mathcal{L}_{t,s_1s_2}^{(l)}$ otherwise, as defined in Lemma 30. Let $\boldsymbol{\beta}_{X_l}^{(l)0} = (\boldsymbol{\beta}_{X_l,s}^{(l)0\top} : 0 \le s \le T)^\top$. Then we have $\sum_{s_1=0}^t \sum_{s_2=0}^t \boldsymbol{\beta}_{X_l,s_1}^{(l)0\top}\mathcal{L}_{t,s_1s_2}^{(l)}\boldsymbol{\beta}_{X_l,s_2}^{(l)0} = \boldsymbol{\beta}_{X_l}^{(l)0\top}\mathcal{L}_t^{(l)}\boldsymbol{\beta}_{X_l}^{(l)0}$, which

implies that $(\mathbf{w}^\top \mathbb{Y}_t^c \mathbb{Y}_t^{c\top} \mathbf{w})^2 \le 4\sum_l (\boldsymbol{\beta}_{X_l}^{(l)0\top} \mathcal{L}_t^{(l)} \boldsymbol{\beta}_{X_l}^{(l)0})^2$. Let $\mathcal{L}_t^{(l)} = \sum_{i_l} \lambda_{i_l} \mathbf{u}_{i_l} \mathbf{u}_{i_l}^\top$ be the eigen-decomposition of $\mathcal{L}_t^{(l)}$, where λ_{i_l} and $\mathbf{u}_{i_l}$ are the i_lth eigenvalue and eigenvector respectively. One can verify that $\mathcal{L}_t^{(l)}$ is semi-definite, then we have all $\lambda_{i_l} \ge 0$. Then we have

$$(\boldsymbol{\beta}_{X_l}^{(l)0\top} \mathcal{L}_t^{(l)} \boldsymbol{\beta}_{X_l}^{(l)0})^2 = \sum_{i_{l1}, i_{l2}} \lambda_{i_{l1}} \lambda_{i_{l2}} \left(\boldsymbol{\beta}_{X_l}^{(l)0\top} \mathbf{u}_{i_{l1}} \mathbf{u}_{i_{l1}}^\top \boldsymbol{\beta}_{X_l}^{(l)0}\right) \left(\boldsymbol{\beta}_{X_l}^{(l)0\top} \mathbf{u}_{i_{l2}} \mathbf{u}_{i_{l2}}^\top \boldsymbol{\beta}_{X_l}^{(l)0}\right).$$

Note that

$$\begin{aligned}
&E\left\{\left(\boldsymbol{\beta}_{X_l}^{(l)0\top} \mathbf{u}_{i_{l1}} \mathbf{u}_{i_{l1}}^\top \boldsymbol{\beta}_{X_l}^{(l)0}\right) \left(\boldsymbol{\beta}_{X_l}^{(l)0\top} \mathbf{u}_{i_{l2}} \mathbf{u}_{i_{l2}}^\top \boldsymbol{\beta}_{X_l}^{(l)0}\right)\right\} = E\left(\langle \mathbf{u}_{i_{l1}}, \boldsymbol{\beta}_{X_l}^{(l)0}\rangle^2 \langle \mathbf{u}_{i_{l2}}, \boldsymbol{\beta}_{X_l}^{(l)0}\rangle^2\right) \\
&\le (1/2) E\left(\langle \mathbf{u}_{i_{l1}}, \boldsymbol{\beta}_{X_l}^{(l)0}\rangle^4\right) + (1/2) E\left(\langle \mathbf{u}_{i_{l2}}, \boldsymbol{\beta}_{X_l}^{(l)0}\rangle^4\right) \\
&= 2\int_0^\infty t^3 P\left\{|\langle \mathbf{u}_{i_{l1}}, \boldsymbol{\beta}_{X_l}^{(l)0}\rangle| > t\right\} dt + 2\int_0^\infty t^3 P\left\{|\langle \mathbf{u}_{i_{l2}}, \boldsymbol{\beta}_{X_l}^{(l)0}\rangle| > t\right\} dt \\
&\le 8\int_0^\infty t^3 \exp(-t^2/K^2) dt \stackrel{\text{def}}{=} c_K,
\end{aligned}$$

where c_K is a constant related to K, and it is the constant defined in Definition 1. Here the last inequality is obtained by noting that $\langle \mathbf{u}_{i_l}, \boldsymbol{\beta}_{X_l}^{(l)0}\rangle$ is a 1-Lipschitz convex function of vector $\mathbf{x}$ with mean zero and $\mathbf{x}$ is K-convex defined in Definition 1. Therefore we have

$$E\left\{(\boldsymbol{\beta}_{X_l}^{(l)0\top} \mathcal{L}_t^{(l)} \boldsymbol{\beta}_{X_l}^{(l)0})^2\right\} \le c_K \sum_{i_{l1}, i_{l2}} \lambda_{i_{l1}} \lambda_{i_{l2}} = c_K \mathrm{tr}(\mathcal{L}_t^{(l)})^2 < \infty,$$

where the last inequality is obtained by (A.102) of Lemma 30.

2. Proof of $E\{(\mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top} \mathbf{w})^2\} < \infty$

By the definition in Lemma 30, we have

$$\mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top} \mathbf{w} = \mathbb{Y}_t^{e\top} \mathbf{A} \mathbb{Y}_t^e = \sum_{s_1=0}^t \sum_{s_2=0}^t \mathbb{E}_{s_1}^\top (\mathbf{B}_0^{t-s_1})^\top \mathbf{A} \mathbf{B}_0^{t-s_2} \mathbb{E}_{s_2} = \sum_{s_1=0}^t \sum_{s_2=0}^t \mathbb{E}_{s_1}^\top \mathbf{L}_{t,s_1 s_2} \mathbb{E}_{s_2}$$

Define $\widetilde{\mathbf{L}}_t = (\widetilde{\mathbf{L}}_{t,s_1 s_2} : 0 \le s_1, s_2 \le T)$, where $\widetilde{\mathbf{L}}_{t,s_1 s_2} = \mathbf{0}$ for $s_1 > t$ or $s_2 > t$, and $\widetilde{\mathbf{L}}_{t,s_1 s_2} = \mathbf{L}_{t,s_1 s_2}$ otherwise. Then we have $\mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top} \mathbf{w} = \mathbb{E}^\top \widetilde{\mathbf{L}}_t \mathbb{E}$, where $\mathbb{E} = (\mathbb{E}_0^\top, \cdots, \mathbb{E}_t^\top)^\top$. Similar to the proof given in Step 1, we can verify $E\{(\mathbf{w}^\top \mathbb{Y}_t^e \mathbb{Y}_t^{e\top} \mathbf{w})^2\} = E\{(\mathbb{E}^\top \widetilde{\mathbf{L}}_t \mathbb{E})^2\} \le c\mathrm{tr}(\widetilde{\mathbf{L}}_t)^2 \le c\mathrm{tr}(\mathbf{L}_t)^2 < \infty$, where the second last inequality used the definition of $\mathbf{L}_t$ in Lemma 30 and the last inequality is obtained by (A.102) of Lemma 30. ■

Lemma 23. *Under Assumption 2, it holds that*

$$\tau_{\min} \|\boldsymbol{\Theta}_{i_1 \cdots i_q} - \boldsymbol{\Theta}^0_{i_1 \cdots i_q}\|^2 \le \frac{1}{T}\Big\{Q^*_{i_1 \cdots i_q}(\boldsymbol{\Theta}_{i_1 \cdots i_q}) - Q^*_{i_1 \cdots i_q}(\boldsymbol{\Theta}^0_{i_1 \cdots i_q})\Big\} \le \tau_{\max} \|\boldsymbol{\Theta}_{i_1 \cdots i_q} - \boldsymbol{\Theta}^0_{i_1 \cdots i_q}\|^2 \tag{A.85}$$

$$\tau_{\min} d(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) \le \frac{1}{(\prod_l N_l) T}\Big\{Q^*(\boldsymbol{\Theta}) - Q^*(\boldsymbol{\Theta}^0)\Big\} \le \tau_{\max} d(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0). \tag{A.86}$$

$$\tau_{\min} d_{i_l}(\mathbf{\Theta}_{\cdot i_l \cdot}, \mathbf{\Theta}^0_{\cdot i_l \cdot}) \le \frac{1}{(\prod_{m\neq l} N_l)T}\Big\{Q^*_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}_{i_l}}; \boldsymbol{\xi}^{-(l)}_{g^{-(l)}}, \mathcal{G}_{-l}) - Q^*_{i_l}(\boldsymbol{\xi}^{(l)0}_{g^{(l)0}_{i_l}}; \boldsymbol{\xi}^{-(l)0}_{g^{-(l)}}, \mathcal{G}^0_{-l}))\Big\}$$
$$\le \tau_{\max} d_{i_l}(\mathbf{\Theta}_{\cdot i_l \cdot}, \mathbf{\Theta}^0_{\cdot i_l \cdot}), \tag{A.87}$$

where $\tau_{\min}$ is defined in Assumption 2, $\tau_{\max}$ is defined in notation Section A, and $d_{i_l}(\cdot,\cdot)$ is defined in (A.2).

Proof We have

$$\begin{aligned} Q(\mathbf{\Theta}) &= \sum_{i_1,\cdots,i_q} Q_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) = \sum_l \sum_{i_l=1}^{N_l}\sum_{t=1}^{T} \big(\varepsilon_{i_1\cdots i_q,t} + \mathcal{X}^\top_{i_1\cdots i_q,t}\mathbf{\Theta}^0_{i_1\cdots i_q} - \mathcal{X}^\top_{i_1\cdots i_q,t}\mathbf{\Theta}_{i_1\cdots i_q}\big)^2 \\ &= \sum_l \sum_{i_l=1}^{N_l}\sum_{t=1}^{T}\Big\{\varepsilon^2_{i_1\cdots i_q,t} + 2\varepsilon_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}^0_{i_1\cdots i_q} - \mathbf{\Theta}_{i_1\cdots i_q}) \\ &+ (\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathcal{X}_{i_1\cdots i_q,t}\mathcal{X}^\top_{i_1\cdots i_q,t}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})\Big\}. \end{aligned}$$

We have $\frac{1}{T}\Big\{Q^*_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q}) - Q^*_{i_1\cdots i_q}(\mathbf{\Theta}^0_{i_1\cdots i_q})\Big\} = (\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathbf{\Sigma}_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})$. By using Assumption 2 and Lemma 33, (A.85) can be obtained.

Next, it holds

$$\begin{aligned} &\frac{1}{(\prod_l N_l)T}\Big\{Q^*(\mathbf{\Theta}) - Q^*(\mathbf{\Theta}^0)\Big\} \\ &= \frac{1}{(\prod_l N_l)T}\sum_l \sum_{i=i_l}^{N_l}\sum_{t=1}^{T}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathbf{\Sigma}_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}) \\ &= \frac{1}{\prod_l N_l}\sum_l \sum_{i_l=1}^{N_l}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathbf{\Sigma}_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}). \end{aligned}$$

We have

$$\begin{aligned} &\frac{1}{\prod_l N_l}\sum_l \sum_{i_l=1}^{N_l}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathbf{\Sigma}_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}) \\ &\ge \frac{\tau_{\min}}{\prod_l N_l}\sum_l \sum_{i_l=1}^{N_l}\Big\|\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}\Big\|^2 = \tau_{\min} d(\mathbf{\Theta}, \mathbf{\Theta}^0), \\ &\frac{1}{\prod_l N_l}\sum_l \sum_{i_l=1}^{N_l}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q})^\top \mathbf{\Sigma}_{i_1\cdots i_q}(\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}) \\ &\le \frac{\tau_{\max}}{\prod_l N_l}\sum_l \sum_{i_l=1}^{N_l}\Big\|\mathbf{\Theta}_{i_1\cdots i_q} - \mathbf{\Theta}^0_{i_1\cdots i_q}\Big\|^2 = \tau_{\max} d(\mathbf{\Theta}, \mathbf{\Theta}^0) \end{aligned}$$

by using Assumption 2 and Lemma 33. This proves (A.86).

Similar arguments as that of the proof for (A.86) at a fixed i_l leads to (A.87). ■

Lemma 24. *Suppose $G_l \geq G_{l,0}$ for all $l \in [q]$, where $G_{l,0}$ is the true number of groups. In addition, assume Assumptions 1–5 hold. Define*

$$\sigma_l(g^{(l)}) = argmin_{\widetilde{g}^{(l)} \in [G_l]} \|\widehat{\boldsymbol{\theta}}^{(l)}_{\widetilde{g}^{(l)}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2 + \frac{1}{\prod_{m \neq l} N_m} \sum_{m \neq l} \sum_{i_m} \Big| \widehat{\alpha}_{\widetilde{g}^{(l)} \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}} - \alpha^0_{g^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} \Big|^2.$$

Then we have

$$\begin{aligned}
&\max_{g^{(l)} \in [G_{l,0}]} \Big\{ \|\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2 + \frac{1}{\prod_{m \neq l} N_l} \sum_{m \neq l} \sum_{i_m} \Big| \widehat{\alpha}_{\sigma_l(g^{(l)}) \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}} - \alpha^0_{g^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} \Big|^2 \Big\} \\
&\leq \max_{g^{(l)} \in [G_{l,0}]} \frac{N_l}{N_{l g^{(l)}}} d(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^0) = O_p\Big(T^{-1} (\sum_l \log N_l)^2 \Big).
\end{aligned}$$

Proof We have

$$\begin{aligned}
&\|\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2 + \frac{1}{\prod_{m \neq l} N_l} \sum_{m \neq l} \sum_{i_m} \Big| \widehat{\alpha}_{\sigma_l(g^{(l)}) \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}} - \alpha^0_{g^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} \Big|^2 \\
&= \frac{1}{N_{l g^{(l)}}} \sum_{i_l} I(g^{(l)0}_{i_l} = g^{(l)}) \Big\{ \|\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma_l(g^{(l)0}_{i_l})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2 \\
&+ \frac{1}{\prod_{m \neq l} N_l} \sum_{m \neq l} \sum_{i_m} \Big| \widehat{\alpha}_{\sigma_l(g^{(l)0}_{i_l}) \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}} - \alpha^0_{g^{(l)0}_{i_l} g^{-(l)0}_{\mathbf{i}_{-l}}} \Big|^2 \Big\} \\
&\leq \frac{1}{N_{l g^{(l)}}} \sum_{i_l} I(g^{(l)0}_{i_l} = g^{(l)}) \Big\{ \|\widehat{\boldsymbol{\theta}}^{(l)}_{\widehat{g}^{(l)}_{i_l}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)0}_{i_l}}\|^2 + \frac{1}{\prod_{m \neq l} N_l} \sum_{m \neq l} \sum_{i_m} \Big| \widehat{\alpha}_{\widehat{g}^{(l)}_{i_l} \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}} - \alpha^0_{g^{(l)0}_{i_l} g^{-(l)0}_{\mathbf{i}_{-l}}} \Big|^2 \Big\} \\
&\leq \frac{N_l}{N_{l g^{(l)}}} d(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^0).
\end{aligned}$$

Then by taking $\max_{g^{(l)} \in [G_{l,0}]}$ of both sides, we have

$$\begin{aligned}
&\max_{g^{(l)} \in [G_{l,0}]} \left(\|\widehat{\boldsymbol{\theta}}^{(l)}_{\sigma(g^{(l)})} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}}\|^2 + \frac{1}{\prod_{m \neq l} N_l} \sum_{m \neq l} \sum_{i_m} |\widehat{\alpha}_{\sigma(g^{(l)}) \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}} - \alpha^0_{g^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}}|^2 \right) \\
&\leq \frac{N_l}{\min_{g^{(l)}} N_{l g}} d(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^0) = O_p\Big(T^{-1} (\sum_l \log N_l)^2 \Big)
\end{aligned}$$

by Theorem 2.

■

Lemma 25. *Under Assumptions 1–7, and suppose $G_l \geq G_{l,0}$ for all $l \in [q]$. For any $\boldsymbol{\xi}$, let $\widehat{g}^{(l)}_{i_l}(\boldsymbol{\xi})$ denote the membership obtained with (14). Define $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_{i_1 \cdots i_q} : i_l \in [N_l])$ with $\boldsymbol{\Theta}_{i_1 \cdots i_q} = (\boldsymbol{\theta}^{(1)\top}_{\widehat{g}^{(1)}_{i_1}(\boldsymbol{\xi})}, \cdots, \boldsymbol{\theta}^{(q)\top}_{\widehat{g}^{(q)}_{i_q}(\boldsymbol{\xi})}, \alpha_{\widehat{g}^{(1)}_{i_1}(\boldsymbol{\xi}) \cdots \widehat{g}^{(q)}_{i_q}(\boldsymbol{\xi})})^\top$. Assume we have*

$$d(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = O_p(T^{-1} (\sum_l \log N_l)^2) = o_p(c_{\text{gap}}), \tag{A.88}$$

as $\min_l N_l \to \infty$. *Let* $\boldsymbol{\xi}^{(l)} = (\boldsymbol{\theta}^{(l)\top}, (\mathrm{vec}(\boldsymbol{\alpha}_{\cdot g^{(l)}\cdot})^\top : g^{(l)} \in [G_l]))^\top = \big((\boldsymbol{\theta}^{(l)}_{g^{(l)}} : g^{(l)} \in [G_l])^\top, (\mathrm{vec}(\boldsymbol{\alpha}_{\cdot g^{(l)}\cdot})^\top : g^{(l)} \in [G_l])\big)^\top$. *Define*

$$\begin{aligned}
&\mathcal{M}^{(l)}(\boldsymbol{\xi}^{(l)}_{g^{(l)}}, \boldsymbol{\xi}^{(l)0}_{g^{(l)0}}; \mathcal{G}_{-l}, \mathcal{G}^0_{-l}) = \|\boldsymbol{\theta}^{(l)}_{g^{(l)}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)0}}\|^2 \\
&+ \frac{1}{\prod_{m\neq l} N_l} \sum_{m\neq l} \sum_{i_m=1}^{N_m} \big|\alpha_{g^{(1)}_{i_l} \cdots g^{(l-1)}_{i_{l-1}} g^{(l)} g^{(l+1)}_{i+1} \cdots g^{(q)}_{i_q}} - \alpha^0_{g^{(1)0}_{i_1} \cdots g^{(l-1)0}_{i_{l-1}} g^{(l)0} g^{(l+1)0}_{i+1} \cdots g^{(q)0}_{i_q}}\big|^2 \\
&\stackrel{\mathrm{def}}{=} \|\boldsymbol{\theta}^{(l)}_{g^{(l)}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)0}}\|^2 + \frac{1}{\prod_{m\neq l} N_l} \sum_{m\neq l} \sum_{i_m=1}^{N_m} |\alpha_{g^{(l)} g^{-(l)}_{\mathbf{i}_{-l}}} - \alpha_{g^{(l)0} g^{-(l)0}_{\mathbf{i}_{-l}}}|^2
\end{aligned}$$

with $\mathbf{i}_{-l} = (i_m : m \neq l)^\top$ *and* $g^{-(l)}_{\mathbf{i}_{-l}} = (g^{(m)}_{i_m} : m \neq l)^\top \in \mathbb{R}^{q-1}$. *Further denote*

$$\begin{aligned}
d^{(l)}_S(\boldsymbol{\xi}^{(l)}, \boldsymbol{\xi}^{(l)0}; \mathcal{G}_{-l}, \mathcal{G}^0_{-l}) = \max\Big\{ &\max_{g^{(l)0}\in[G_{l,0}]} \min_{g^{(l)}\in[G_l]} \Big(\mathcal{M}^{(l)}(\boldsymbol{\xi}^{(l)}_{g^{(l)}}, \boldsymbol{\xi}^{(l)0}_{g^{(l)0}}; \mathcal{G}_{-l}, \mathcal{G}^0_{-l})\Big), \\
&\max_{g^{(l)}\in[G_l]} \min_{g^{(l)0}\in[G_{l,0}]} \Big(\mathcal{M}^{(l)}(\boldsymbol{\xi}^{(l)}_{g^{(l)}}, \boldsymbol{\xi}^{(l)0}_{g^{(l)0}}; \mathcal{G}_{-l}, \mathcal{G}^0_{-l})\Big)\Big\}, \quad \text{(A.89)}
\end{aligned}$$

where $\mathcal{G}_{-l} = \{\mathcal{G}_m : m \neq l\} = \{(g^{(m)}_{i_m} : 1 \leq i_m \leq N_m)^\top : m \neq l\}$. *In addition, define* $\mathcal{N}^{(l)}_\eta = \{\boldsymbol{\xi} : d^{(l)}_S(\boldsymbol{\xi}^{(l)}, \boldsymbol{\xi}^{(l)0}; \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi}), \mathcal{G}^0_{-l}) < \eta\}$ *given* $\boldsymbol{\xi}$. *Correspondingly, denote*

$$\begin{aligned}
\mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)0}, \mathcal{G}^0_{-l}) = \Big\{ &g^{(l)} \in [G_l] : \|\boldsymbol{\theta}^{(l)}_{g^{(l)}} - \boldsymbol{\theta}^{(l)0}_{g^{(l)0}}\|^2 \\
&+ \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l} \sum_{i_m} |\alpha_{g^{(l)} \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}(\boldsymbol{\xi})} - \alpha^0_{g^{(l)0} g^{-(l)0}_{\mathbf{i}_{-l}}}| \leq \eta \Big\}. \quad \text{(A.90)}
\end{aligned}$$

Then the following conclusions hold:
(i) For all $\boldsymbol{\xi} \in \mathcal{N}^{(l)}_\eta$ *with* $\eta < (c_\pi)^{q-1} c_{\mathrm{gap}}/4$, *we have* $\{\mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)0}, \mathcal{G}^0_{-l}),\ g^{(l)0} \in [G_{l,0}]\}$ *is a partition of* $[G_l]$;
(ii) Define the event $\boldsymbol{\Omega} = \{\widehat{g}^{(l)}_{i_l}(\boldsymbol{\xi}) \in \mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)0}_{i_l}, \mathcal{G}^0_{-l}),\ \forall i_l \in [N_l]\}$, *where* $\boldsymbol{\xi}$ *satisfying that* $\boldsymbol{\xi} \in \mathcal{N}^{(l)}_\eta$, *and* $\eta \leq \tau_{\min} c_{\mathrm{gap}} (c_\pi)^{q-1}/\{8(\tau_{\min} + \tau_{\max})\}$. *Here,* $\tau_{\max}$ *is given in the notation Section A. Then we have* $P(\boldsymbol{\Omega}^c) \leq C \exp\Big(-c_1 T^{1/2} c_{\mathrm{gap}} + c_2 m + \sum_l \log N_l\Big)$, *where* C, c_1, c_2 *are positive constants.*

Proof 1. Proof of (i)

By the definition of $\mathcal{N}^{(l)}_\eta$ and $\mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)}, \mathcal{G}^0_{-l})$, we have $\cup_{g^{(l)}=1}^{G_{l,0}} \mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)}, \mathcal{G}^0_{-l}) = [G_l]$. Then it remains to show that $\mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)}, \mathcal{G}^0_{-l})$ is a partition of $[G_l]$. That is $\mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)}_1, \mathcal{G}^0_{-l}) \cap \mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)}_2, \mathcal{G}^0_{-l}) = \emptyset$ for any $g^{(l)}_1 \neq g^{(l)}_2$.

We prove by contradiction. Suppose there exists $g^{(l)}_{12} \in [G_l]$ so that $g^{(l)}_{12} \in \mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)}_1, \mathcal{G}^0_{-l}) \cap \mathcal{A}^{(l)}_\eta(\boldsymbol{\xi}, g^{(l)}_2, \mathcal{G}^0_{-l})$ for $g^{(l)}_1, g^{(l)}_2 \in [G_{l,0}]$ and $g^{(l)}_1 \neq g^{(l)}_2$. Denote $g^{-(l)} = (g^{(m)} : m \neq l)$, and correspondingly denote the event $\{g^{-(l)0}_{\mathbf{i}_{-l}} = g^{-(l)}\}$ as $\{g^{(m)0}_{i_m} = g^{(m)} : m \neq l\}$. In this case we have

$$c_{\mathrm{gap}} \leq \{\|\boldsymbol{\theta}^{(l)0}_{g^{(l)}_1} - \boldsymbol{\theta}^{(l)0}_{g^{(l)}_2}\|^2 + \max_{g^{-(l)}\in[G^0_{-l}]} |\alpha^0_{g^{(l)}_1 g^{-(l)}} - \alpha^0_{g^{(l)}_2 g^{-(l)}}|^2\}$$

$$
\begin{aligned}
&\leq 2\|\boldsymbol{\theta}^{(l)0}_{g_1^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 + 2\|\boldsymbol{\theta}^{(l)0}_{g_2^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 \\
&+ \max_{g^{-(l)}\in[G^0_{-l}]} \left(\frac{1}{\prod_{m\neq l}(\pi^{(m)}_{g^{(m)},N_m} N_m)} \sum_{m\neq l}\sum_{i_m} \left\{ |\alpha^0_{g_1^{(l)} g^{-(l)}} - \alpha^0_{g_2^{(l)} g^{-(l)}}|^2 I(g^{-(l)0}_{\mathbf{i}_{-l}} = g^{-(l)}) \right\} \right) \\
&= 2\|\boldsymbol{\theta}^{(l)0}_{g_1^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 + 2\|\boldsymbol{\theta}^{(l)0}_{g_2^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 \\
&+ \max_{g^{-(l)}\in[G^0_{-l}]} \left(\frac{1}{\prod_{m\neq l}(\pi^{(m)}_{g^{(m)},N_m} N_m)} \sum_{m\neq l}\sum_{i_m} \left\{ |\alpha^0_{g_1^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} - \alpha^0_{g_2^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}}|^2 I(g^{-(l)0}_{\mathbf{i}_{-l}} = g^{-(l)}) \right\} \right) \\
&\leq 2\|\boldsymbol{\theta}^{(l)0}_{g_1^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 + 2\|\boldsymbol{\theta}^{(l)0}_{g_2^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 \\
&+ \max_{g^{-(l)}\in[G^0_{-l}]} \left(\frac{1}{\prod_{m\neq l}(\pi^{(m)}_{g^{(m)},N_m} N_m)} \sum_{m\neq l}\sum_{i_m} |\alpha^0_{g_1^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} - \alpha^0_{g_2^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}}|^2 \right) \\
&= 2\|\boldsymbol{\theta}^{(l)0}_{g_1^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 + 2\|\boldsymbol{\theta}^{(l)0}_{g_2^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 \\
&+ \max_{g^{-(l)}\in[G^0_{-l}]} \left(\frac{1}{\prod_{m\neq l}(\pi^{(m)}_{g^{(m)},N_m} N_m)} \right) \sum_{m\neq l}\sum_{i_m} |\alpha^0_{g_1^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} - \alpha^0_{g_2^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}}|^2 \\
&\leq 2\|\boldsymbol{\theta}^{(l)0}_{g_1^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 + 2\|\boldsymbol{\theta}^{(l)0}_{g_2^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 \\
&+ \frac{1}{\min_{g^{-(l)}\in[G^0_{-l}]}(\prod_{m\neq l} \pi^{(m)}_{g^{(m)},N_m} N_m)} \sum_{m\neq l}\sum_{i_m} |\alpha^0_{g_1^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} - \alpha^0_{g_2^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}}|^2 \\
&\leq \frac{2}{\min_{g^{-(l)}\in[G^0_{-l}]}(\prod_{m\neq l} \pi^{(m)}_{g^{(m)},N_m})} \times \Big(\|\boldsymbol{\theta}^{(l)0}_{g_1^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 + \|\boldsymbol{\theta}^{(l)0}_{g_2^{(l)}} - \boldsymbol{\theta}^{(l)}_{g_{12}^{(l)}}\|^2 \\
&+ \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l}\sum_{i_m} |\alpha^0_{g_1^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} - \alpha_{g_{12}^{(l)} \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}(\boldsymbol{\xi})}|^2 \\
&+ \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l}\sum_{i_m} |\alpha^0_{g_2^{(l)} g^{-(l)0}_{\mathbf{i}_{-l}}} - \alpha_{g_{12}^{(l)} \widehat{g}^{-(l)}_{\mathbf{i}_{-l}}(\boldsymbol{\xi})}|^2 \Big) \\
&\leq \frac{4\eta}{\min_{g^{-(l)}\in[G^0_{-l}]}(\prod_{m\neq l} \pi^{(m)}_{g^{(m)},N_m})}.
\end{aligned}
$$

The last line contradicts the definition of η that $\eta < c_{\text{gap}}(c_\pi)^{q-1}/4$ as $N_l \to \infty$. Therefore, there does not exist a $g_{12}^{(l)} \in [G_l]$ such that $g_{12}^{(l)} \in \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_1^{(l)}, \mathcal{G}^0_{-l}) \cap \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_2^{(l)}, \mathcal{G}^0_{-l})$ for any $g_1^{(l)}, g_2^{(l)} \in [G_{l,0}]$ and $g_1^{(l)} \neq g_2^{(l)}$, which suggests that $\mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_1^{(l)}, \mathcal{G}^0_{-l}) \cap \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_2^{(l)}, \mathcal{G}^0_{-l}) = \emptyset$ for any $g_1^{(l)} \neq g_2^{(l)}$. This completes the proof of part (i).

2. Proof of (ii)

For the notation simplicity, we use $\widehat{g}_{i_l}^{(l)}$ to replace $\widehat{g}_{i_l}^{(l)}(\boldsymbol{\xi})$, where $\boldsymbol{\xi}$ satisfies that $\boldsymbol{\xi} \in \mathcal{N}_\eta^{(l)}$ and also (A.88). By the definition of $\widehat{g}_{i_l}^{(l)}$, we have

$$
I(\widehat{g}_{i_l}^{(l)} = g^{(l)}) \leq I\left(Q_{i_l}(\boldsymbol{\xi}^{(l)}_{g^{(l)}}; \boldsymbol{\xi}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) < Q_{i_l}(\boldsymbol{\xi}^{(l)}_{\widetilde{g}^{(l)}}; \boldsymbol{\xi}^{-(l)}_{g^{-(l)}}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) \right)
$$

for any $\widetilde{g}^{(l)} \neq g^{(l)}$. Therefore, for $\widetilde{g}_{i_l}^{(l)} \in \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)$, we have

$$
\begin{aligned}
&I\left(\widehat{g}_{i_l}^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)\right) = \sum_{g^{(l)}=1}^{G_l} I\left(g^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)\right) I(\widehat{g}_{i_l}^{(l)} = g^{(l)}) \\
&\leq \sum_{g^{(l)}=1}^{G_l} I\left(g^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)\right) I\left(Q_{i_l}(\boldsymbol{\xi}_{g^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) < Q_{i_l}(\boldsymbol{\xi}_{\widetilde{g}_{i_l}^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi}))\right) \\
&\stackrel{\text{def}}{=} \sum_{g^{(l)}=1}^{G_l} W_{i_l g^{(l)}}(\boldsymbol{\xi}).
\end{aligned}
$$

For all $g^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)$ and $g^{(l)} \in [G_l]$, due to (i), there exists a $g_{j_l}^{(l)0} \neq g_{i_l}^{(l)0}$, such that $g^{(l)} \in \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{j_l}^{(l)0}, \mathcal{G}_{-l}^0)$. Then we have

$$
\begin{aligned}
&\left\|\boldsymbol{\theta}_{g_{i_l}^{(l)0}}^{(l)0} - \boldsymbol{\theta}_{g^{(l)}}^{(l)}\right\|^2 + \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l} \sum_{i_m} |\alpha_{g^{(l)}\widehat{g}_{\mathbf{i}_{-l}}^{-(l)}(\boldsymbol{\xi})} - \alpha^0_{g_{i_l}^{(l)0} g_{\mathbf{i}_{-l}}^{-(l)0}}|^2 \\
&\geq \frac{1}{2}\left\|\boldsymbol{\theta}_{g_{i_l}^{(l)0}}^{(l)0} - \boldsymbol{\theta}_{g_{j_l}^{(l)0}}^{(l)0}\right\|^2 + \frac{1}{2\prod_{m\neq l} N_m} \sum_{m\neq l} \sum_{i_m} |\alpha^0_{g_{j_l}^{(l)0} g_{\mathbf{i}_{-l}}^{-(l)0}} - \alpha^0_{g_{i_l}^{(l)0} g_{\mathbf{i}_{-l}}^{-(l)0}}|^2 \\
&- \left\{\left\|\boldsymbol{\theta}_{g_{j_l}^{(l)0}}^{(l)0} - \boldsymbol{\theta}_{g^{(l)}}^{(l)}\right\|^2 + \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l} \sum_{i_m} |\alpha_{g^{(l)}\widehat{g}_{\mathbf{i}_{-l}}^{-(l)}(\boldsymbol{\xi})} - \alpha^0_{g_{j_l}^{(l)0} g_{\mathbf{i}_{-l}}^{-(l)0}}|^2\right\} \geq c_{\text{gap}}(c_\pi)^{q-1}/2 - \eta
\end{aligned}
$$

by Assumption 6 when $N_l \to \infty$. By Lemma 23, it holds for any $g^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)$,

$$
\frac{1}{(\prod_{m\neq l} N_m)T}\{Q_{i_l}^*(\boldsymbol{\xi}_{g^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) - Q_{i_l}^*(\boldsymbol{\xi}_{g_{i_l}^{(l)0}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)0}, \mathcal{G}_{-l}^0)\} \geq \tau_{\min}(c_{\text{gap}}(c_\pi)^{q-1}/2 - \eta). \tag{A.91}
$$

On the other hand, for any $\widetilde{g}_{i_l}^{(l)} \in \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)$, it holds

$$
\begin{aligned}
&\frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}^*(\boldsymbol{\xi}_{\widetilde{g}_{i_l}^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) - \frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}^*(\boldsymbol{\xi}_{g_{i_l}^{(l)0}}^{(l)0}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)0}, \mathcal{G}_{-l}^0) \\
&\leq \tau_{\max}\left\{\frac{1}{\prod_m N_m} \sum_{m\neq l} \sum_{i_m} \left\|\boldsymbol{\theta}_{\widehat{g}_{i_m}^{(m)}(\boldsymbol{\xi})}^{(m)} - \boldsymbol{\theta}_{g_{i_m}^{(m)0}}^{(m)0}\right\|^2 + \left\|\boldsymbol{\theta}_{g_{i_l}^{(l)0}}^{(l)0} - \boldsymbol{\theta}_{\widetilde{g}_{i_l}^{(l)}}^{(l)}\right\|^2\right. \\
&\left. + \frac{1}{\prod_{m\neq l} N_m} \sum_{m\neq l} \sum_{i_m} |\widehat{\alpha}_{\widetilde{g}_{i_l}^{(l)}\widehat{g}_{\mathbf{i}_{-l}}^{-(l)}(\boldsymbol{\xi})} - \alpha^0_{g_{i_l}^{(l)0} g_{\mathbf{i}_{-l}}^{-(l)0}}|^2\right\} \\
&\leq \tau_{\max}\{d(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) + \eta\} \leq \tau_{\max}\{CT^{-1}(\sum_l \log N_l)^2 + \eta\}.
\end{aligned} \tag{A.92}
$$

in probability tending to 1 by (A.88), where C is a positive constant. Combining (A.91) and (A.92), we have

$$
\frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}^*(\boldsymbol{\xi}_{g^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) - \frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}^*(\boldsymbol{\xi}_{\widetilde{g}_{i_l}^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi}))
$$

$$\geq \tau_{\min}\{c_{\text{gap}}(c_\pi)^{q-1}/2-\eta\}-\tau_{\max}\{\eta+CT^{-1}(\sum_l \log N_l)^2\} \stackrel{\text{def}}{=} \epsilon_\eta$$

when $N_l \to \infty$. Note that $\tau_{\min}$ and $\tau_{\max}$ are both bounded positive constants due to Assumption 2 and Lemma 33. This leads to

$$\begin{aligned}
&W_{i_l g^{(l)}}(\mathbf{x}) \\
&= I\left(g^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)\right) I\left(Q_{i_l}(\boldsymbol{\xi}_{g^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) < Q_{i_l}(\boldsymbol{\xi}_{\widetilde{g}_{i_l}^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi}))\right) \\
&\leq I\left(g^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)\right) \\
&I\Big(\frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}(\boldsymbol{\xi}_{\widetilde{g}_{i_l}^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) - \frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}(\boldsymbol{\xi}_{g^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) \\
&+ \frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}^*(\boldsymbol{\xi}_{g^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) - \frac{1}{(\prod_{m\neq l} N_m)T} Q_{i_l}^*(\boldsymbol{\xi}_{\widetilde{g}_{i_l}^{(l)}}^{(l)}; \boldsymbol{\xi}_{g^{-(l)}}^{-(l)}, \widehat{\mathcal{G}}_{-l}(\boldsymbol{\xi})) \geq \epsilon_\eta\Big) \\
&\leq 2I\left(\sup_{i_1,\cdots,i_q} \sup_{\|\boldsymbol{\Theta}_{i_1,\cdots i_q}\|_{\max}<R} \Big|\frac{1}{T} Q_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q}) - \frac{1}{T} Q_{i_1\cdots i_q}^*(\boldsymbol{\Theta}_{i_1\cdots i_q})\Big| \geq \epsilon_\eta/2\right). \qquad \text{(A.93)}
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
&P\left\{\sup_{1\leq i_l\leq N_l} I\left(\widehat{g}_{i_l}^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)\right) > 0\right\} \leq \sum_{g^{(l)}=1}^{G_l} P\left\{\sup_{1\leq i_l\leq N_l} W_{i_l g^{(l)}}(\boldsymbol{\xi}) > 0\right\} \\
&\leq \sum_{g^{(l)}=1}^{G_l} P\left\{\sup_{i_1,\cdots,i_q} \sup_{\|\boldsymbol{\Theta}_{i_1,\cdots i_q}\|_{\max}<R} \Big|\frac{1}{T} Q_{i_1\cdots i_q}(\boldsymbol{\Theta}_{i_1\cdots i_q}) - \frac{1}{T} Q_{i_1\cdots i_q}^*(\boldsymbol{\Theta}_{i_1\cdots i_q})\Big| \geq \epsilon_\eta/2\right\} \\
&\leq G_l(\prod_l N_l)\exp\Big\{-c_1' \min(T\epsilon_\eta^2, T^{1/2}\epsilon_\eta) + c_2 m\Big\} \\
&= G_l \exp\Big\{-c_1' \min(T\epsilon_\eta^2, T^{1/2}\epsilon_\eta) + c_2 m + \sum_l \log N_l\Big\} \qquad \text{(A.94)}
\end{aligned}$$

by using (A.66) of Lemma 17. Finally, noting that $c_{\text{gap}} \gg T^{-1}(\sum_l \log N_l)^2$ and we have $\eta < \{\tau_{\min} c_{\text{gap}}(c_\pi)^{q-1}/4 - \tau_{\max} C T^{-1}(\sum_l \log N_l)^2\}/(\tau_{\min}+\tau_{\max})$, then we can obtain $\epsilon_\eta > \tau_{\min} c_{\text{gap}}(c_\pi)^{q-1}/4$ as $\min_l N_l, T \to \infty$. This results in

$$P\left\{\sup_{1\leq i_l\leq N_l} I\left(\widehat{g}_{i_l}^{(l)} \notin \mathcal{A}_\eta^{(l)}(\boldsymbol{\xi}, g_{i_l}^{(l)0}, \mathcal{G}_{-l}^0)\right) > 0\right\} \leq C\exp\Big\{-c_1 T^{1/2} c_{\text{gap}} + c_2 m + \sum_l \log N_l\Big\}.$$

This finishes the proof of (ii).

■

Appendix K. General Technical Lemmas

Lemma 26. *(Hanson-Wright Inequality) Define the ϕ_2-norm of variable X as*

$$\|X\|_{\phi_2} = \inf\Big\{t>0 : E\exp\Big(\frac{X^2}{t^2}\Big) \leq 2\Big\}.$$

Let $X_1, \cdots, X_n$ be independent variables satisfying $E(X_i) = 0$, $E(X_i^2) = \sigma_i^2$, and $\|X_i\|_{\phi_2} \leq M < \infty$ and $\mathbf{A}$ be a symmetric $n \times n$ matrix. Define $\mathbf{x} = (X_1, \cdots, X_n)^\top \in \mathbb{R}^n$. For any $t > 0$, we have

$$P\Big(|\mathbf{x}^\top \mathbf{A}\mathbf{x} - tr(\mathbf{A}\mathbf{\Sigma}_x)| > t\Big) \leq 2\exp\Big\{ - c\min\Big(\frac{t^2}{M^4\|\mathbf{A}\|_F^2}, \frac{t}{M^2\sigma_1(\mathbf{A})}\Big)\Big\},$$

where $\mathbf{\Sigma}_x = diag\{\sigma_1^2, \cdots, \sigma_n^2\}$, $\sigma_1(\mathbf{A})$ is the maximum singular value of $\mathbf{A}$, $\|\mathbf{A}\|_F$ denotes the Frobenius norm of matrix $\mathbf{A}$, and $c > 0$ is a constant.

Proof The proof is given in Hanson and Wright (1971). ■

Lemma 27. *let $\mathbf{x}$ be a mean zero random vector in $\mathbb{R}^n$ satisfying K-convex concentration property according to Definition 1. Then for any $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $t > 0$, it holds*

$$P\Big(\Big|\mathbf{x}^\top \mathbf{A}\mathbf{x} - E(\mathbf{x}^\top \mathbf{A}\mathbf{x})\Big| \geq t\Big) \leq 2\exp\Big(- \frac{1}{C}\min\Big(\frac{t^2}{K^4\|\mathbf{A}\|_F^2}, \frac{t}{K^2\|\mathbf{A}\|}\Big)\Big).$$

Proof The proof can be found in Theorem 2.5 of Adamczak (2015). ■

Lemma 28. *If $\mathbf{x} \in \mathbb{R}^n$ satisfies the K-convex concentration property according to Definition 1 with $E(\mathbf{x}) = \mathbf{0}$, then we have $\|cov(\mathbf{x})\| \leq 2K^2$.*

Proof For any unit vector $\boldsymbol{u} \in \mathbb{R}^n$ with $\|\boldsymbol{u}\| = 1$, $\langle \boldsymbol{u}, \mathbf{x}\rangle$ is a 1-Lipschitz convex function of $\mathbf{x}$. Then we have

$$\boldsymbol{u}^\top \mathrm{cov}(\mathbf{x})\boldsymbol{u} = E\langle \boldsymbol{u}, \mathbf{x}\rangle^2 = 2\int_0^\infty tP(|\langle \boldsymbol{u}, \mathbf{x}\rangle| > t)dt \leq 4\int_0^\infty t\exp(-t^2/K^2)dt = 2K^2.$$

This implies $\|\mathrm{cov}(\mathbf{x})\| \leq 2K^2$.

■

Lemma 29. *Let $\{\mathbf{D}_{s_1s_2} \in \mathbb{R}^{(\prod_l N_l)\times(\prod_l N_l)} : s_1, s_2 \in \{0, \ldots, T\}$ be a sequence of matrices. Define the following three terms, $\mathcal{D}^{(l)}_{s_1s_2} = (\mathbf{1}_{N_1}^\top \otimes \cdots \otimes \mathbf{1}_{N_{l-1}}^\top \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}}^\top \otimes \cdots \otimes \mathbf{1}_{N_q}^\top)\mathbf{D}_{s_1s_2}(\mathbf{1}_{N_1} \otimes \cdots \otimes \mathbf{1}_{N_{l-1}} \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}} \otimes \cdots \otimes \mathbf{1}_{N_q}) \in \mathbb{R}^{N_l\times N_l}$, $\mathcal{D}^{(lm)}_{s_1s_2} = (\mathbf{1}_{N_1}^\top \otimes \cdots \otimes \mathbf{1}_{N_{l-1}}^\top \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}}^\top \otimes \cdots \otimes \mathbf{1}_{N_q}^\top)\mathbf{D}_{s_1s_2}(\mathbf{1}_{N_1} \otimes \cdots \otimes \mathbf{1}_{N_{m-1}} \otimes \mathbf{I}_{N_m} \otimes \mathbf{1}_{N_{m+1}} \otimes \cdots \otimes \mathbf{1}_{N_q}) \in \mathbb{R}^{N_l\times N_m}$, and $\widetilde{\mathcal{D}}^{(l)}_{s_1s_2} = (\mathbf{1}_{N_1}^\top \otimes \cdots \otimes \mathbf{1}_{N_{l-1}}^\top \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}}^\top \otimes \cdots \otimes \mathbf{1}_{N_q}^\top)\mathbf{D}_{s_1s_2} \in \mathbb{R}^{N_l\times(\prod_l N_l)}$. Correspondingly, let $\mathcal{D}^{(l)} = (\mathcal{D}^{(l)}_{s_1s_2} : s_1, s_2 \in \{0, \ldots, T\}) \in \mathbb{R}^{(N_l(T+1))\times(N_l(T+1))}$, let $\mathcal{D}^{(lm)} = (\mathcal{D}^{(lm)}_{s_1s_2} : s_1, s_2 \in \{0, \ldots, T\}) \in \mathbb{R}^{(N_l(T+1))\times(N_m(T+1))}$, and let $\widetilde{\mathcal{D}}^{(l)} = (\widetilde{\mathcal{D}}^{(l)}_{s_1s_2} : s_1, s_2 \in \{0, \cdots, T\}) \in \mathbb{R}^{N_l(T+1)\times(\prod_l N_l)(T+1)}$. Define $\widetilde{\mathbf{x}}_t^{(l)\eta} = \mathbf{1}_{N_1} \circ \cdots \circ \mathbf{x}_t^{(l)\eta} \circ \cdots \circ \mathbf{1}_{N_q} \in \mathbb{R}^{\prod_l N_l}$, where $\mathbf{x}_t^{(l)\eta}$ is defined in Assumption 4. Further note that $\mathbf{x}^{(l)\eta} = (\mathbf{x}_t^{(l)\eta} : 0 \leq t \leq T) \in \mathbb{R}^{N_l(T+1)}$. Then we have*

$$P\Big\{\Big|\sum_{s_1,s_2=0}^{T} \widetilde{\mathbf{x}}_t^{(l)\eta\top}\mathbf{D}_{s_1s_2}\widetilde{\mathbf{x}}_t^{(l)\eta} - \sum_{s_1,s_2=0}^{T} E(\widetilde{\mathbf{x}}_t^{(l)\eta\top}\mathbf{D}_{s_1s_2}\widetilde{\mathbf{x}}_t^{(l)\eta})\Big| \geq u\Big\}$$

$$\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{u^2}{\|\mathcal{D}^{(l)}\|_F^2},\frac{u}{\|\mathcal{D}^{(l)}\|}\Big)\Big), \tag{A.95}$$

$$P\Big\{\Big|\sum_{s_1,s_2=0}^{T}\widetilde{\mathbf{x}}_{s_1}^{(l)\eta\top}\mathbf{D}_{s_1s_2}\widetilde{\mathbf{x}}_{s_2}^{(m)\eta}\Big|>u\Big\}\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{u^2}{\|\mathcal{D}^{(lm)}\|_F^2},\frac{u}{\|\mathcal{D}^{(lm)}\|}\Big)\Big), \tag{A.96}$$

$$P\Big\{\Big|\sum_{s_1,s_2=0}^{T}\widetilde{\mathbf{x}}_{s_1}^{(l)\eta\top}\mathbf{D}_{s_1s_2}\mathbb{E}_{s_2}\Big|>u\Big\}\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{u^2}{\|\widetilde{\mathcal{D}}^{(l)}\|_F^2},\frac{u}{\|\widetilde{\mathcal{D}}^{(l)}\|}\Big)\Big). \tag{A.97}$$

Proof 1. Proof of (A.95)

Note that we have $\sum_{s_1,s_2=0}^{T}\widetilde{\mathbf{x}}_{s_1}^{(l)\eta\top}\mathbf{D}_{s_1s_2}\widetilde{\mathbf{x}}_{s_2}^{(l)\eta}=\sum_{s_1,s_2=0}^{T}\mathbf{x}_{s_1}^{(l)\eta\top}\mathcal{D}_{s_1s_2}^{(l)}\mathbf{x}_{s_2}^{(l)\eta}=\mathbf{x}^{(l)\eta\top}\mathcal{D}^{(l)}\mathbf{x}^{(l)\eta}$. Here $\mathbf{x}^{(l)\eta}$ satisfies the K'-convex concentration property defined in Definition 1 by Assumption 4 for a constant K'. Therefore by Lemma 27 the concentration inequality in (A.95) is directly obtained.

2. Proof of (A.96)

Note that $\sum_{s_1,s_2=0}^{T}\mathbf{x}_{s_1}^{(l)\eta\top}\mathbf{D}_{s_1s_2}\mathbf{x}_{s_2}^{(m)\eta}=\sum_{s_1,s_2=0}^{T}\mathbf{x}_{s_1}^{(l)\eta\top}\mathcal{D}_{s_1s_2}^{(lm)}\mathbf{x}_{s_2}^{(m)\eta}=\mathbf{x}^{(l)\eta\top}\mathcal{D}^{(lm)}\mathbf{x}^{(m)\eta}$. Define $\boldsymbol{h}^{\eta}=(\mathbf{x}^{(l)\eta\top},\mathbf{x}^{(m)\eta\top})^{\top}$, and $\mathcal{D}=(\mathbf{0},\mathcal{D}^{(lm)};\mathcal{D}^{(lm)\top},\mathbf{0})$. Then we have $\mathbf{x}^{(l)\eta\top}\mathcal{D}^{(lm)}\mathbf{x}^{(m)\eta}=\boldsymbol{h}^{\eta\top}\mathcal{D}\boldsymbol{h}^{\eta}/2$. By Lemma 27 and Assumption 4 we obtain that

$$\begin{aligned}P\Big(\Big|\mathbf{x}^{(l)\eta\top}\mathcal{D}^{(lm)}\mathbf{x}^{(m)\eta}\Big|>u\Big)&\leq 2\exp\Big(-\frac{1}{C_1}\min\Big(\frac{u^2}{\|\mathcal{D}\|_F^2},\frac{u}{\|\mathcal{D}\|}\Big)\Big)\\&\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{u^2}{\|\mathcal{D}^{(lm)}\|_F^2},\frac{u}{\|\mathcal{D}^{(lm)}\|}\Big)\Big).\end{aligned}$$

3. Proof of (A.97)

Denote $\mathbb{E}=(\mathbb{E}_t^{\top}:0\leq t\leq T)^{\top}\in\mathbb{R}^{(\prod_l N_l)(T+1)}$. Note that $\sum_{s_1,s_2=0}^{T}\widetilde{\mathbf{x}}_{s_1}^{(l)\eta\top}\mathbf{D}_{s_1s_2}\mathbb{E}_{s_2}=\mathbf{x}^{(l)\eta\top}\widetilde{\mathcal{D}}^{(l)}\mathbb{E}$, then we follow the proof of (A.96) and obtain the result.

■

Lemma 30. *Let* $\mathbf{D}_{s_1s_2}=1/T\sum_{t=\max\{1,s_1,s_2\}}^{T}(\mathbf{B}_0^{t-s_1})^{\top}\mathbf{w}^{(l)}\mathbf{w}^{(l)\top}\mathbf{B}_0^{t-s_2}\stackrel{\text{def}}{=}1/T\sum_{t=\max\{1,s_1,s_2\}}^{T}\mathbf{L}_{t,s_1s_2}^{(l)}$, *where* $\mathbf{w}^{(l)}=\mathbf{e}_{k_1}^{(N_1)}\otimes\cdots\otimes\mathbf{w}_{i_l\cdot}^{(l)}\otimes\cdots\otimes\mathbf{e}_{k_q}^{(N_q)}\in\mathbb{R}^{\prod_l N_l}$ *for any* $l\in[q]$. *Here* $\mathbf{w}_{li_l\cdot}=(w_{li_lj}:j\in[N_l])^{\top}\in\mathbb{R}^{N_l}$, $w_{li_lj}\geq 0$ *and* $\|\mathbf{w}_{li_l\cdot}\|_1=1$. *In addition, let* $\mathcal{D}^{(l)}$, $\mathcal{D}^{(lm)}$ *and* $\widetilde{\mathcal{D}}^{(l)}$ *be defined as in Lemma 29. Furthermore, let* $\mathbf{L}_t^{(l)}=(\mathbf{L}_{t,s_1s_2}^{(l)}:0\leq s_1,s_2\leq t)$ *and* $\mathcal{L}_{t,s_1s_2}^{(l)}=(\mathbf{1}_{N_1}^{\top}\otimes\cdots\otimes\mathbf{1}_{N_{l-1}}^{\top}\otimes\mathbf{I}_{N_l}\otimes\mathbf{1}_{N_{l+1}}^{\top}\otimes\cdots\otimes\mathbf{1}_{N_q}^{\top})\mathbf{L}_{t,s_1s_2}^{(l)}(\mathbf{1}_{N_1}\otimes\cdots\otimes\mathbf{1}_{N_{l-1}}\otimes\mathbf{I}_{N_l}\otimes\mathbf{1}_{N_{l+1}}\otimes\cdots\otimes\mathbf{1}_{N_q})\in\mathbb{R}^{N_l\times N_l}$. *Define* $\mathcal{L}_t^{(l)}=(\widetilde{\mathcal{L}}_{t,s_1s_2}^{(l)}:0\leq s_1,s_2\leq T)$ *for* $l\in[q]$, *where* $\widetilde{\mathcal{L}}_{t,s_1s_2}^{(l)}=\mathbf{0}$ *for* $s_1>t$ *or* $s_2>t$, *and* $\widetilde{\mathcal{L}}_{t,s_1s_2}^{(l)}=\mathcal{L}_{t,s_1s_2}^{(l)}$ *otherwise. Then we have for any* $l\in[q]$,

$$\|\mathcal{D}^{(l)}\|_F^2\leq c_1T^{-1},\quad\|\mathcal{D}^{(l)}\|\leq c_1^{1/2}T^{-1/2}, \tag{A.98}$$

$$\|\mathcal{D}^{(lm)}\|_F^2\leq c_1T^{-1},\quad\|\mathcal{D}^{(lm)}\|\leq c_1^{1/2}T^{-1/2}, \tag{A.99}$$

$$\|\mathbf{D}\|_F^2\leq c_2T^{-1},\quad\|\mathbf{D}\|\leq c_2^{1/2}T^{-1/2}, \tag{A.100}$$

$$\|\widetilde{\mathcal{D}}^{(l)}\|_F^2\leq c_2T^{-1},\quad\|\widetilde{\mathcal{D}}^{(l)}\|\leq c_2^{1/2}T^{-1/2}, \tag{A.101}$$

$$\max_l tr(\mathcal{L}_t^{(l)}) < \infty, \quad tr(\mathbf{L}_t^{(l)}) < \infty, \tag{A.102}$$

where $\mathbf{D}$ *is defined in* (A.76)*, and* c_1*,* c_2 *are two positive constants.*

Proof 1. Proof of (A.98)–(A.99)

Note that by Lemma 31 we have

$$\begin{aligned}
\|\mathcal{D}^{(l)}\|_F^2 &= \sum_{s_1,s_2=0}^{T} \|\mathcal{D}_{s_1s_2}^{(l)}\|_F^2 \\
&= \sum_{s_1,s_2=0}^{T} \Big\| (\mathbf{1}_{N_1}^\top \otimes \cdots \otimes \mathbf{1}_{N_{l-1}}^\top \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}}^\top \otimes \cdots \otimes \mathbf{1}_{N_q}^\top)\mathbf{D}_{s_1s_2} \\
&\quad (\mathbf{1}_{N_1} \otimes \cdots \otimes \mathbf{1}_{N_{l-1}} \otimes \mathbf{I}_{N_l} \otimes \mathbf{1}_{N_{l+1}} \otimes \cdots \otimes \mathbf{1}_{N_q}) \Big\|_F^2 \\
&\le \sum_{s_1,s_2=0}^{T} \Big(\mathbf{1}_{\prod_l N_l}^\top |\mathbf{D}_{s_1s_2}|_e \mathbf{1}_{\prod_l N_l} \Big)^2 \\
&\le \frac{1}{T^2} \sum_{s_1,s_2=0}^{T} \Big(\sum_{t=\max\{1,s_1,s_2\}}^{T} \kappa_{\max}^{2t-s_1-s_2} \mathbf{1}_{\prod_l N_l}^\top |\mathbf{w}^{(l)}|_e |\mathbf{w}^{(l)\top}|_e \mathbf{1}_{\prod_l N_l} \Big)^2 \\
&= \frac{1}{T^2} \sum_{s_1,s_2=0}^{T} \Big(\sum_{t=\max\{1,s_1,s_2\}}^{T} \kappa_{\max}^{2t-s_1-s_2} \Big)^2 \le \frac{1}{T^2} \sum_{s_1,s_2=0}^{T} \Big(\frac{1}{1-\kappa_{\max}^2} \kappa_{\max}^{|s_1-s_2|} \Big)^2 \le c_1 T^{-1}
\end{aligned}$$

where c_1 is a finite constant. Furthermore, noting that $\|\mathcal{D}^{(l)}\| \le \|\mathcal{D}^{(l)}\|_F \le c_1^{1/2} T^{-1/2}$, we have the upper bound for $\|\mathcal{D}^{(l)}\|$. Subsequently, the upper bounds for $\|\mathcal{D}^{(lm)}\|_F^2$ and $\|\mathcal{D}^{(lm)}\|$ can be established using the same technique.

2. Proof of (A.100)

First, for any two matrices $\mathbf{M}_1, \mathbf{M}_2$,

$$|\mathbf{M}_1\mathbf{M}_2^\top|_e \preccurlyeq \max_{i,j} |\mathbf{M}_{1i\cdot}\mathbf{M}_{2j\cdot}^\top| \mathbf{1}\mathbf{1}^\top \le \max_{i,j} \|\mathbf{M}_{1i\cdot}\| \|\mathbf{M}_{2j\cdot}\| \mathbf{1}\mathbf{1}^\top.$$

When $\|\mathbf{M}_1\|_{\max} \le 1$ and $\|\mathbf{M}_2\|_{\max} \le 1$, we further have $\|\mathbf{M}_{1i\cdot}\| \|\mathbf{M}_{2j\cdot}\| \le \|\mathbf{M}_1\|_\infty \|\mathbf{M}_2\|_\infty$ Hence $|\mathbf{M}_1\mathbf{M}_2^\top|_e \preccurlyeq \|\mathbf{M}_1\|_\infty \|\mathbf{M}_2\|_\infty \mathbf{1}\mathbf{1}^\top$, and therefore

$$\mathbf{w}^{(l)\top}\mathbf{M}_1\mathbf{M}_2^\top\mathbf{w}^{(l)} \le \|\mathbf{M}_1\|_\infty \|\mathbf{M}_2\|_\infty \mathbf{w}^{(l)\top}\mathbf{1}\mathbf{1}^\top\mathbf{w}^{(l)} = \|\mathbf{M}_1\|_\infty \|\mathbf{M}_2\|_\infty. \tag{A.103}$$

Note that Lemma 31 implies $\|\mathbf{B}_0^k\|_\infty \le \kappa_{\max}^n < 1$ for any $k > 0$, hence

$$\begin{aligned}
\|\mathbf{D}\|_F^2 &= \sum_{s_1,s_2=0}^{T} \|\mathbf{D}_{s_1s_2}\|_F^2 \\
&= \frac{1}{T^2} \sum_{s_1,s_2=0}^{T} \sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T} \mathbf{w}^{(l)\top}\mathbf{B}_0^{t_1-s_2}(\mathbf{B}_0^{t_2-s_2})^\top\mathbf{w}^{(l)}\mathbf{w}^{(l)\top}\mathbf{B}_0^{t_2-s_1}(\mathbf{B}_0^{t_1-s_1})^\top\mathbf{w}^{(l)} \\
&\le \frac{1}{T^2} \sum_{s_1,s_2=0}^{T} \sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T} \Big(\|\mathbf{B}_0^{t_1-s_2}\|_\infty \|\mathbf{B}_0^{t_2-s_2}\|_\infty \|\mathbf{B}_0^{t_2-s_1}\|_\infty \|\mathbf{B}_0^{t_1-s_1}\|_\infty \Big)
\end{aligned}$$

$$\leq \frac{1}{T^2}\sum_{s_1,s_2=0}^{T}\sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T}\kappa_{\max}^{2t_1+2t_2-2s_1-2s_2}$$
$$\leq \frac{1}{T^2(1-\kappa_{\max}^2)^2}\sum_{s_1,s_2=0}^{T}\kappa_{\max}^{2|s_1-s_2|} \leq c_2 T^{-1}$$

where c_2 is a finite constant and we used Lemma 31 in the second inequality.

3. Proof of (A.101)

We have

$$\|\widetilde{\mathcal{D}}^{(l)}\|_F^2 = \sum_{s_1,s_2=0}^{T}\|\widetilde{\mathcal{D}}^{(l)}_{s_1s_2}\|_F^2 = \sum_{s_1,s_2=0}^{T}\Big\|(\mathbf{1}_{N_1}^\top\otimes\cdots\otimes\mathbf{1}_{N_{l-1}}^\top\otimes\mathbf{I}_{N_l}\otimes\mathbf{1}_{N_{l+1}}^\top\otimes\cdots\otimes\mathbf{1}_{N_q}^\top)\mathbf{D}_{s_1s_2}\Big\|_F^2$$
$$= \frac{1}{T^2}\sum_{s_1,s_2=0}^{T}\sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T}\mathbf{w}^{(l)\top}\mathbf{B}_0^{t_1-s_2}\mathbf{B}_0^{t_2-s_2\top}\mathbf{w}^{(l)}\mathbf{w}^{(l)\top}\mathbf{B}_0^{t_2-s_1}$$
$$(\mathbf{1}_{N_1}\mathbf{1}_{N_1}^\top\otimes\cdots\otimes\mathbf{1}_{N_{l-1}}\mathbf{1}_{N_{l-1}}^\top\otimes\mathbf{I}_{N_l}\otimes\mathbf{1}_{N_{l+1}}\mathbf{1}_{N_{l+1}}^\top\otimes\cdots\otimes\mathbf{1}_{N_q}\mathbf{1}_{N_q}^\top)(\mathbf{B}_0^{t_1-s_1})^\top\mathbf{w}^{(l)}$$
$$\leq \frac{1}{T^2}\sum_{s_1,s_2=0}^{T}\sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T}\|\mathbf{B}_0^{t_1-s_2}\|_\infty\|\mathbf{B}_0^{t_2-s_2\top}\|_\infty$$
$$\mathbf{w}^{(l)\top}|\mathbf{B}_0^{t_2-s_1}|_e(\mathbf{1}_{\prod_l N_l}\mathbf{1}_{\prod_l N_l}^\top)|\mathbf{B}_0^{t_1-s_1}|_e^\top\mathbf{w}^{(l)}$$
$$\leq \frac{1}{T^2}\sum_{s_1,s_2=0}^{T}\sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T}\|\mathbf{B}_0^{t_1-s_2}\|_\infty\|\mathbf{B}_0^{t_2-s_2}\|_\infty\|\mathbf{B}_0^{t_2-s_1}\|_\infty\|\mathbf{B}_0^{t_1-s_1}\|_\infty\mathbf{w}^{(l)\top}\mathbf{1}\mathbf{1}^\top\mathbf{w}^{(l)}$$
$$= \frac{1}{T^2}\sum_{s_1,s_2=0}^{T}\sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T}\|\mathbf{B}_0^{t_1-s_2}\|_\infty\|\mathbf{B}_0^{t_2-s_2}\|_\infty\|\mathbf{B}_0^{t_2-s_1}\|_\infty\|\mathbf{B}_0^{t_1-s_1}\|_\infty$$
$$\leq \frac{1}{T^2}\sum_{s_1,s_2=0}^{T}\sum_{t_1,t_2=\max\{1,s_1,s_2\}}^{T}\kappa_{\max}^{2t_1+2t_2-2s_1-2s_2}$$
$$\leq \frac{1}{T^2(1-\kappa_{\max}^2)^2}\sum_{s_1,s_2=0}^{T}\kappa_{\max}^{2|s_1-s_2|} \leq c_2 T^{-1}$$

In addition, it further implies that $\|\widetilde{\mathcal{D}}^{(l)}\| \leq c_2^{1/2}T^{-1/2}$.

4. Proof of (A.102)

Note that by Lemma 31 we have

$$\mathrm{tr}(\mathcal{L}_t^{(l)}) = \sum_{s=0}^{t}\mathrm{tr}(\mathcal{L}_{t,ss}^{(l)})$$
$$= \sum_{s=0}^{t}\mathrm{tr}\Big((\mathbf{1}_{N_1}^\top\otimes\cdots\otimes\mathbf{1}_{N_{l-1}}^\top\otimes\mathbf{I}_{N_l}\otimes\cdots\otimes\mathbf{1}_{N_q}^\top)\mathbf{L}_{t,ss}^{(l)}(\mathbf{1}_{N_1}\otimes\cdots\otimes\mathbf{1}_{N_{l-1}}\otimes\mathbf{I}_{N_l}\otimes\cdots\otimes\mathbf{1}_{N_q})\Big)$$
$$\leq \sum_{s=0}^{t}\Big(\mathbf{1}_{\prod_l N_l}^\top|\mathbf{L}_{t,ss}^{(l)}|_e\mathbf{1}_{\prod_l N_l}\Big) \leq \sum_{s=0}^{t}\Big(\kappa_{\max}^{2t-2s}\mathbf{1}_{\prod_l N_l}^\top|\mathbf{w}^{(l)}|_e|\mathbf{w}^{(l)\top}|_e\mathbf{1}_{\prod_l N_l}\Big) \leq \frac{1}{1-\kappa_{\max}^2} < \infty.$$

Consequently the first inequality of (A.102) holds. Next we note

$$\begin{aligned}\mathrm{tr}(\mathbf{L}_t^{(l)}) &\le \sum_{s=0}^{t}(\mathbf{1}_{\prod_l N_l}^\top|\mathbf{L}_{t,ss}^{(l)}|_e\mathbf{1}_{\prod_l N_l}) \\ &\le \sum_{s=0}^{t}(\kappa_{\max}^{2t-2s}\mathbf{1}_{\prod_l N_l}^\top|\mathbf{w}^{(l)}|_e|\mathbf{w}^{(l)\top}|_e\mathbf{1}_{\prod_l N_l}) \le 1/(1-\kappa_{\max}^2) < \infty,\end{aligned}$$

then the second inequality holds. ■

Lemma 31. *Under Assumption 5, we have* $\|\mathbf{B}_0^n\|_\infty \le \kappa_{\max}^n$ *and* $|\mathbf{B}_0|_e^n\mathbf{1}_{\prod_l N_l} \preccurlyeq \kappa_{\max}^n\mathbf{1}_{\prod_l N_l}$.

Proof Note that we have

$$|\mathbf{B}_0|_e\mathbf{1}_{\prod_l N_l} \preccurlyeq \sum_l \mathbf{1}_{N_1}\otimes\cdots\mathbf{1}_{N_{l-1}}\otimes\mathbf{L}_{l,0}\mathbf{1}_{N_l}\otimes\mathbf{1}_{N_{l+1}}\otimes\cdots\otimes\mathbf{1}_{N_q} + |\mathrm{vec}(\mathcal{A}_0)|_e \preccurlyeq \kappa_{\max}\mathbf{1}_{\prod_l N_l}$$

by Assumption 5. As a result, we have $\|\mathbf{B}_0^n\|_\infty \le \||\mathbf{B}_0|_e^n\|_\infty = \||\mathbf{B}_0|_e^n\mathbf{1}_{\prod_l N_l}\|_{\max} \le \kappa_{\max}^n$. ■

Lemma 32. *Under Assumptions 4 and 5, we have* $\|\mathbf{\Gamma}\|_{\max} \le c_\Gamma$*, where* $c_\Gamma > 0$ *is a constant.*

Proof Define $\boldsymbol{h}_t = \mathbf{c}_t + \mathbb{E}_t$ and then $\boldsymbol{h}_t$ follows K'-convex concentration property by Assumption 4 for some constant K'. In addition, let $\mathbb{H}_t = (\boldsymbol{h}_t^\top, \boldsymbol{h}_{t-1}^\top, \cdots, \boldsymbol{h}_0^\top)^\top$, $\mathbb{B} = (\mathbf{I}_{\prod_l N_l}, \mathbf{B}_0, \mathbf{B}_0^2, \cdots, \mathbf{B}_0^t)$. Note that we have

$$\mathbf{\Gamma} = \mathrm{cov}(\mathbb{Y}_t) = \sum_{k_1,k_2=0}^{t}\mathbf{B}_0^{k_1}\mathrm{cov}\Big\{\boldsymbol{h}_{t-k_1}, \boldsymbol{h}_{t-k_2}\Big\}\mathbf{B}_0^{k_2\top} = \mathbb{B}\mathrm{cov}(\mathbb{H}_t)\mathbb{B}^\top.$$

Then we have $\|\mathbf{\Gamma}\|_{\max} = \max_{i,j}|\mathbf{e}_i^\top\mathbf{\Gamma}\mathbf{e}_j| = \max_i|\mathbf{e}_i^\top\mathbf{\Gamma}\mathbf{e}_i| \le \|\mathrm{cov}(\mathbb{H}_t)\|\max_i|\mathbf{e}_i^\top\mathbb{B}\mathbb{B}^\top\mathbf{e}_i|$, where $\mathbf{e}_i \in \mathbb{R}^{\prod_l N_l}$ is a vector with its ith element bing equal to 1 while others being 0. By Lemma 28, it holds $\|\mathrm{cov}(\mathbb{H}_t)\|\max_i|\mathbf{e}_i^\top\mathbb{B}\mathbb{B}^\top\mathbf{e}_i| \le 2K^{*2}|\mathbf{e}_i^\top\mathbb{B}\mathbb{B}^\top\mathbf{e}_i|$ for some constant K^*. Next, we have $\mathbf{e}_i^\top\mathbb{B}\mathbb{B}^\top\mathbf{e}_i = \sum_{k=0}^t\mathbf{e}_i^\top\mathbf{B}_0^k\mathbf{B}_0^{k\top}\mathbf{e}_i \le \sum_{k=0}^t\|\mathbf{B}_0^k\|_\infty^2 \le \sum_{k=0}^t\kappa_{\max}^{2k} \le 1/(1-\kappa_{\max}^2) < \infty$ by Lemma 31 and Assumption 5. Thus the result holds with $c_\Gamma = 1/(1-\kappa_{\max}^2)$. ■

Lemma 33. *Under Assumptions 4 and 5, we have* $\tau_{\max} = \max_{i_1,\cdots,i_q}\lambda_{\max}(\mathbf{\Sigma}_{i_1\cdots i_q}) < \infty$.

Proof Recall that we have defined

$$\begin{aligned}\mathcal{X}_{i_1\cdots i_q,t} \overset{\mathrm{def}}{=} \big(&\sum_{k=1}^{N_1} w_{i_1k}^{(1)}Y_{ki_2\cdots i_q,(t-1)}, \mathbf{x}_{i_1t}^{(1)\top}, \cdots, \\ &\sum_{k=1}^{N_q} w_{i_qk}^{(q)}Y_{i_1\cdots i_{(q-1)}k,(t-1)}, \mathbf{x}_{i_qt}^{(q)\top}, Y_{i_1\cdots i_q,(t-1)}\big)^\top \in \mathbb{R}^{\sum_l(p_l+1)+1}.\end{aligned}$$

Note that there are $(2q+1)$ elements in $\mathcal{X}_{i_1\cdots i_q,t}$, denote the mth element as $\mathcal{X}^m_{i_1\cdots i_q,t}, 1 \le m \le (2q+1)$. Specifically, $\mathcal{X}^m_{i_1\cdots i_q,t} = \sum_{k=1}^{N_1} w^{(l)}_{i_l k} Y_{i_1\cdots i_{l-1} k i_{l+1}\cdots i_q,(t-1)}$ when m takes $\{1,3,\cdots,2q-1\}$, $\mathcal{X}^m_{i_1\cdots i_q,t} = \mathbf{x}^{(l)\top}_{i_l t}$ when m takes $\{2,4,\cdots,2q\}$, and $\mathcal{X}^{2q+1}_{i_1\cdots i_q,t} = Y_{i_1\cdots i_q,(t-1)}$. We write $\boldsymbol{\eta} = (\eta_1,\eta_2,\eta_3,\eta_4,\cdots,\eta_{2q+1})^\top$ for any vector $\boldsymbol{\eta} \in \mathbb{R}^{\sum_l (p_l+1)+1}$. Here the even elements, i.e., $\eta_2,\eta_4,\cdots,\eta_{2q}$ are vectors with dimension $p_2,p_4,\cdots,p_{2q}$, while other elements are scalar. Then we have

$$\begin{aligned}\lambda_{\max}(\boldsymbol{\Sigma}_{i_1\cdots i_q}) &= \sup_{\|\boldsymbol{\eta}\|=1} \sum_{k_1,k_2=1}^{2q+1} \eta_{k_1}^\top \mathrm{cov}(\mathcal{X}^{k_1}_{i_1\cdots i_q,t}, \mathcal{X}^{k_2}_{i_1\cdots i_q,t})\eta_{k_2} \\ &\le (2q+1) \sup_{\|\boldsymbol{\eta}\|=1} \sum_{k=1}^{2q+1} \eta_k^\top \mathrm{var}(\mathcal{X}^k_{i_1\cdots i_q,t})\eta_k.\end{aligned}$$

Note that $\mathcal{X}^k_{i_1\cdots i_q,t}$ has two typical forms. The first can be written as $\mathbf{w}^\top \mathbb{Y}_t$ $(k=1,3,5,\cdots)$ by taking $\mathbf{w}$ as a vector with non-negative elements and $\|\mathbf{w}\|_1 = 1$. The second is $\mathbf{x}^{(l)}_{i_l t}$, which represents the covariate information. The first form yields $\eta_k \mathrm{var}(\mathbf{w}^\top \mathbb{Y}_t)\eta_k = \eta_k^2 \mathbf{w}^\top \boldsymbol{\Gamma} \mathbf{w} \le \|\boldsymbol{\Gamma}\|_{\max} |\mathbf{w}^\top \mathbf{1}\mathbf{1}^\top \mathbf{w}| = \|\boldsymbol{\Gamma}\|_{\max} < c_\Gamma$, where c_Γ is defined in Lemma 32. The second form yields $\eta_k^\top \mathrm{var}(\mathbf{x}^{(l)}_{i_l t})\eta_k \le \mathrm{var}(\mathbf{x}^{(l)\eta}_t) < 2K^2$ due to Assumption 4 and Lemma 28. Thus, $\tau_{\max} \le (2q+1)((q+1)c_\Gamma + qK^2) < \infty$. ■

Appendix L. Additional Simulation Studies

L.1 Group Number Selection Consistency

To examine the finite sample performance of the group selection consistency, we conduct experiment when $q=2$. The data generating mechanism is the same as in Section 6.1 in the main text. Specifically, we define

$$\varrho(G_1) = R^{-1}\sum_{r=1}^{R} I(\widehat{G}_1^{(r)} = G_1), \quad \varrho(G_2) = R^{-1}\sum_{r=1}^{R} I(\widehat{G}_2^{(r)} = G_2),$$

where $\widehat{G}_1^{(r)}$ and $\widehat{G}_2^{(r)}$ are estimated group numbers in the rth replicate. Hence, $\varrho(G_1)$ and $\varrho(G_2)$ evaluate the proportion of the correctly group numbers for G_1 and G_2. The results are shown in Table A.7, from which one could see that as the sample size increases, the correct group numbers selection percentage is closer to 1, which can illustrate the group numbers selection consistency in Theorem 3 from the finite sample experiment.

L.2 Random Initialization

In the Step 2 of Algorithm A.3 in Appendix F, we use the k-means clustering for initialization of the group memberships. Alternatively, we also consider using random initialization in Step 2, and the estimation results are shown in Table A.8. We take $q=2$ as an example, and set $N_1 = 100, N_2 = 80$. The time lengths vary in $\{40,80,150,200\}$. All the evaluation metrics are the same as Section 6.1 in the main text. The results demonstrate that

Table A.7: The proportion of selected group numbers G_1 and G_2 in $R = 500$ replicates under different settings.

N_1	N_2	T	G_1	G_2	Scenario 1 (SBM)		Scenario 2 (Power-Law)	
					$\varrho(G_1)$	$\varrho(G_2)$	$\varrho(G_1)$	$\varrho(G_2)$
100	80	20	2	2	0.966	0.966	0.838	0.838
			3	3	0.034	0.034	0.162	0.162
			4	4	0.000	0.000	0.000	0.000
		40	2	2	0.134	0.134	0.000	0.000
			3	3	0.866	0.866	1.000	1.000
			4	4	0.000	0.000	0.000	0.000
200	150	20	2	2	0.756	0.756	0.010	0.010
			3	3	0.244	0.244	0.990	0.990
			4	4	0.000	0.000	0.000	0.000
		40	2	2	0.016	0.016	0.000	0.000
			3	3	0.984	0.984	1.000	1.000
			4	4	0.000	0.000	0.000	0.000

k-means initialization yields superior parameter estimation accuracy compared to random initialization under limited temporal lengths. Notably, as the temporal length increases, the RMSE gap between the two initialization methods narrows significantly, i.e., random initialization based procedure converges to the similar performance as the k-means initialization based procedure. This convergence pattern is consistently observed in mis-classification rates (the last two columns in Table A.8). Critically, k-means initialization maintains robust performance regardless of sample size, while random initialization achieves comparable finite-sample performance when sufficient data is available.

L.3 Robustness Evaluation

To show the robustness of our proposed model, we add several simulations in this section, including model mis-specification and heavy-tail noise settings.

L.3.1 Model Mis-specification

In this subsection, we consider two model-specification settings. In both settings, we set $(N_1, N_2) \in \{(100, 80), (200, 150)\}$, and the training time length $T \in \{20, 40, 80, 120\}$.

1. Mis-specified network structures.

First, we consider the case that the network weighting matrices $\{\mathbf{W}^{(l)}\}$ are misspecified when $q = 2$. First, we generate the "correct" networks $\mathbf{A}^{(l)}$ by setting $P(a_{ij,l} = 1) = 1/N_l$ and set $a_{ii} = 0$ by convention. Then, we simulate mis-specified networks by setting $P(a'_{ij,l} = 1) = 1/(10N_l)$ for those $a'_{ij,l} = 0,\ i \neq j$. Consequently, we observe more links than the correctly specified networks. Denote the mis-specified networks by $\mathbf{A}^{(1)}_{\text{mis}}$ and $\mathbf{A}^{(2)}_{\text{mis}}$. Correspondingly, denote the row-normalized weighting matrices as $\mathbf{W}^{(1)}_{\text{mis}}$ and $\mathbf{W}^{(2)}_{\text{mis}}$. Our data is generated by model (1) with the "correct" weighting matrices $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, while the estimation is conducted using mis-specified weighting matrices $\mathbf{W}^{(1)}_{\text{mis}}$ and $\mathbf{W}^{(2)}_{\text{mis}}$.

2. Non-linear covariates.

Table A.8: RMSEs of estimated parameters and mis-clustering rates when $G_{1,0} = G_{2,0} = 3$ with 300 replicates. The Initialization methods include random and k-means initialization.

G_1	G_2	N_1	N_2	Initialization	T	$\widehat{\boldsymbol{\lambda}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}^{(2)}$	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\eta}_1$	$\widehat{\eta}_2$
3	3	100	80	K-means	40	0.0100	0.0071	0.0140	0.0094	0.0153	0.0313	0.0001
					80	0.0054	0.0046	0.0068	0.0067	0.0092	0.0031	0.0001
					150	0.0040	0.0038	0.0047	0.0060	0.0068	0.0000	0.0001
					200	0.0034	0.0032	0.0041	0.0043	0.0057	0.0000	0.0001
				Random	40	0.0157	0.0135	0.0207	0.0135	0.0356	0.0673	0.0124
					80	0.0089	0.0105	0.0112	0.0093	0.0208	0.0296	0.0044
					150	0.0050	0.0039	0.0067	0.0053	0.0084	0.0117	0.0013
					200	0.0043	0.0040	0.0054	0.0054	0.0083	0.0078	0.0036

Table A.9: ReMSPEs of the mis-specified model and the true model in two scenarios.

N_1	N_2	T	**Changing Networks**		**Nonlinear Covariates**	
			$\text{ReMSPE}_{\text{mis}}$	$\text{ReMSPE}_{\text{true}}$	$\text{ReMSPE}_{\text{mis}}$	$\text{ReMSPE}_{\text{true}}$
100	80	20	0.7097	0.7082	0.9190	0.8924
		40	0.6845	0.6833	0.8758	0.8497
		80	0.6445	0.6435	0.8589	0.8339
		120	0.6290	0.6282	0.8514	0.8260
200	150	20	0.7309	0.7300	0.9257	0.8985
		40	0.6926	0.6921	0.8792	0.8531
		80	0.6741	0.6736	0.8560	0.8305
		120	0.6415	0.6410	0.8489	0.8236

Second, we consider the case when covariate structure is mis-specified. We generate data using the following model,

$$Y_{ij,t} = \lambda^{(1)}_{g_i^{(1)}} \sum_{k=1}^{N_1} \frac{a^{(1)}_{ik}}{n_{1i}} Y_{kj,(t-1)} + \lambda^{(2)}_{g_j^{(2)}} \sum_{k=1}^{N_2} \frac{a^{(2)}_{kj}}{n_{2j}} Y_{ik,(t-1)} + \alpha_{g_i^{(1)} g_j^{(2)}} Y_{ij,(t-1)} + f(\mathbf{x}^{(1)}_{it})^\top \boldsymbol{\zeta}^{(1)}_{g_i^{(1)}} + f(\mathbf{x}^{(2)}_{jt})^\top \boldsymbol{\zeta}^{(2)}_{g_j^{(2)}} + \varepsilon_{ij,t}, \tag{A.104}$$

where $f(\mathbf{v}) = (f(v_1), \cdots, f(v_{p_l}))$ with $f(v_m) = 0.1v_m^3 + \sin(0.1v_m\pi)$. Then we estimate the model using our proposed algorithm A.1, mis-specifying the true covariates $f(\mathbf{x}^{(1)}_{it})$ and $f(\mathbf{x}^{(2)}_{jt})$ with $\mathbf{x}^{(1)}_{it}$ and $\mathbf{x}^{(2)}_{jt}$.

The data generating scheme for other parts of the model is the same as Section 6.1 in the main text. To evaluate the performance of the two mis-specification settings, we calculate the out-of-sample mean square prediction error (MSPE) for the fitted response $\widehat{Y}_{ij,t}$, respectively. Specifically, we set the subsequent $T_{test} = T/2$ samples after the training set as the testing set. Denote the predicted response for the testing set as $\widehat{Y}^{\text{te}}_{ij,t}$. Correspondingly, calculate the MSPE as

$$\text{MSPE} = (N_1 N_2 T_{test})^{-1} \sum_{t=T_{train}+1}^{T_{test}} \sum_{i,j} (\widehat{Y}^{\text{te}}_{ij,t} - Y_{ij,t})^2.$$

Then we calculate the relative mean square prediction error as $\text{ReMSPE} = \text{MSPE}/\text{MSPE}_0$, where the baseline $\text{MSPE}_0 = (N_1 N_2 T_{train})^{-1} \sum_{t=1}^{T_{train}} (Y_{ij,t} - \mu_{ij,train})^2$, and $\mu_{ij,train} = T_{train}^{-1} \sum_{t=1}^{T_{train}} Y_{ij,t}$. For comparison, we calculate the ReMSPEs for the true models, denoted as $\text{ReMSPE}_{\text{true}}$. Specifically, we use $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ as the true networks in the first scenario, and use $f(\mathbf{x}^{(1)}_{it})$, $f(\mathbf{x}^{(2)}_{jt})$ as the true covariates in the second scenario. The results are shown in Table A.9. Across both scenarios, ReMSPEs for misspecified models exhibit a pronounced decreasing trend as temporal length increases. Notably, prediction errors under misspecification are close to those of true models, demonstrating the robust performance of the proposed GTNAR framework against model deviations.

L.3.2 Heavy-tail Random Noise

In this subsection, we generate heavy-tailed data to validate the robustness of our proposed methodology. The experimental design specifies $q = 2$ with $G_{1,0} = G_{2,0} = 3$ underlying groups, with network structures generated from Stochastic Block Model (SBM) networks as in Section 6.1. We conduct $R = 300$ replicates for a reliable evaluation. For each replicate, the idiosyncratic noise $\varepsilon_{i_1 i_2,t}$ follows an independent and identical t-distribution with 5 degrees of freedom (i.e., $\varepsilon_{i_1 i_2,t} \overset{\text{i.i.d.}}{\sim} t(5)$). All other simulation parameters align with Section 6.1 of the main text. The results are shown in Table A.10. As both the sample sizes (N_1, N_2) and the time length (T) increase, the RMSEs for all parameters decrease significantly. Our proposed estimators closely approach their oracle counterparts, particularly at larger sample sizes. Notably, the group membership errors $(\widehat{\eta}_1, \widehat{\eta}_2)$ converge rapidly to zero, reaching near-zero values at $T = 40$ for $N_1 \geq 200$, which confirms the high accuracy of the proposed methodology. This demonstrates the GTNAR method's robustness in maintaining finite sample performance despite heavy-tailed noise setting.

L.4 Computational Cost

As we comment in Remark 5, the memberships update equation (A.25) can be computed in parallel for each inner-layer subject (i_l in the lth layer). Therefore, we implement the algorithm by a multi-core computational scheme. For each layer, the group memberships estimation is conducted in R version 4.4.3, running on a Apple M3 platform. Parallel processing implements 4 CPU cores using the doParallelSNOW package (version 1.0.17). Total available hardware resources included 8 physical cores supporting 8 concurrent threads. We report the average computational cost of 10 replicated experiments in Figure A.7. The results demonstrate the computational efficiency of the proposed methodology under varying network sizes and time lengths. The initialization stage (orange dashed line) exhibits near-constant time complexity with stable execution times (minimal fluctuation as T increases), indicating negligible scaling overhead in initialization. The iterative estimation phase (blue solid line) shows linear scaling with time length T. Notably, even at the maximum network size configuration ($N_1 = 300, N_2 = 250$), total computational costs remain no more than 20 seconds across all time horizons, which highlights the GTNAR method's suitability for large-scale applications despite increasing data dimensionality.

L.5 Numerical Convergence Analysis

In this subsection, we take $q = 2$ for example to evaluate the numerical convergence rate of our proposed algorithm. Specifically, denote the loss function in the kth iteration as $Q^{[k]}(\widehat{\mathbf{\Theta}}^{[k]})$, where $\widehat{\mathbf{\Theta}}^{[k]}$ is the estimators in the kth iteration. The convergence criterion is set as $|Q^{[k+1]}(\widehat{\mathbf{\Theta}}^{[k+1]}) - Q^{[k]}(\widehat{\mathbf{\Theta}}^{[k]})| \leq 10^{-6}$. We first show the steps of iteration required for algorithm convergence and the corresponding total loss defined in equation (8) in each iteration step in Figure A.8. Specifically, we set the network sizes $(N_1, N_2) \in \{(100, 80), (200, 150)\}$ and the time lengths $T \in \{20, 40\}$. We repeat the experiments for 50 replicates in each sample size setting. For a baseline comparison, we calculate the oracle loss by using the oracle estimators, which are obtained by setting the group memberships as true values. Figure A.8 demonstrates the highly efficient convergence of our proposed iterative algorithm

Table A.10: RMSEs of estimated parameters when $G_{1,0} = G_{2,0} = 3$ with noise distribution being $t(5)$. The experiments are conducted with 300 replications. The performances are evaluated for different sample sizes N_1, N_2 and the time length T. Results of the oracle scenario (given true group memberships $\mathcal{G}_1^0, \mathcal{G}_2^0$) are provided. The corresponding CPs are shown in the parenthesis.

N_1	N_2	T	$\widehat{\boldsymbol{\lambda}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}^{(2)}$	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\lambda}}^{(1)\text{or}}$	$\widehat{\boldsymbol{\lambda}}^{(2)\text{or}}$	$\widehat{\boldsymbol{\zeta}}^{(1)\text{or}}$	$\widehat{\boldsymbol{\zeta}}^{(2)\text{or}}$	$\widehat{\boldsymbol{\alpha}}^{\text{or}}$	$\widehat{\eta}_1$	$\widehat{\eta}_2$
100	80	20	0.0212	0.0243	0.0582	0.0325	0.0716	0.0127	0.0123	0.0167	0.0169	0.0203	0.2332	0.0569
			(0.7378)	(0.8078)	(0.5885)	(0.8322)	(0.6544)	(0.9400)	(0.9500)	(0.9500)	(0.9422)	(0.9467)		
		40	0.0113	0.0077	0.0223	0.0118	0.0189	0.0087	0.0077	0.0116	0.0118	0.0139	0.0642	0
			(0.8656)	(0.9489)	(0.8185)	(0.9463)	(0.8726)	(0.9478)	(0.9511)	(0.9541)	(0.9470)	(0.9507)		
200	150	20	0.0109	0.0065	0.0248	0.0088	0.0178	0.0067	0.0064	0.0085	0.0088	0.0102	0.0829	0.0001
			(0.7933)	(0.9333)	(0.7430)	(0.9430)	(0.8189)	(0.9389)	(0.9356)	(0.9511)	(0.9444)	(0.9448)		
		40	0.0054	0.0043	0.0074	0.0060	0.0075	0.0049	0.0043	0.0060	0.0060	0.0069	0.0068	0.0001
			(0.9133)	(0.9356)	(0.9256)	(0.9563)	(0.9381)	(0.9367)	(0.9333)	(0.9485)	(0.9563)	(0.9519)		
300	250	20	0.0053	0.0039	0.0085	0.0053	0.0077	0.0042	0.0039	0.0055	0.0053	0.0064	0.0153	0
			(0.9133)	(0.9356)	(0.9256)	(0.9563)	(0.9381)	(0.9367)	(0.9333)	(0.9485)	(0.9563)	(0.9519)		
		40	0.0027	0.0027	0.0039	0.0038	0.0044	0.0027	0.0027	0.0039	0.0038	0.0044	0	0
			(0.9600)	(0.9467)	(0.9478)	(0.9437)	(0.9463)	(0.9600)	(0.9467)	(0.9478)	(0.9437)	(0.9467)		

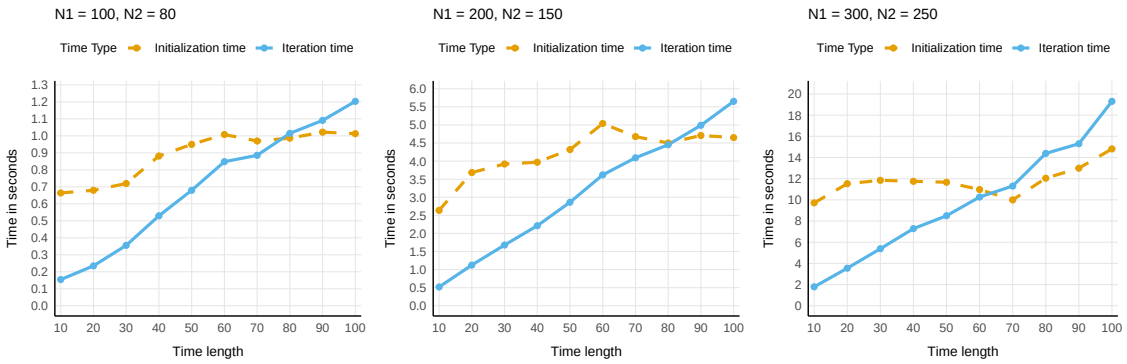


Figure A.7: Computational costs (in second) of the initialization (orange dashed line) and iterative estimation (blue solid line) stage under different sample sizes and time lengths.

to the oracle loss across diverse sample size configurations. Notably, either larger network sizes ($N_1 = 200, N_2 = 150$) or longer time periods ($T = 40$) enhance convergence speed.

We further show the average and the maximum number of iterations. Fix network sizes as $N_1 = 100, N_2 = 80$, we show the results as the time length T grows from 10 to 100; and fix the time length $T = 20$, we increase the network sizes throughout $(N_1, N_2) \in \{(30, 20), (50, 40), (80, 60), (100, 80), (120, 100), (160, 120), (200, 150), (250, 180), (300, 250)\}$. The results are shown in Figure A.9. The average iteration number demonstrates consistent monotonic decline as sample sizes increase, which highlights the algorithm's intrinsic efficiency. While the maximum iteration number exhibits fluctuations at minimal network sizes, it undergoes rapid stabilization followed by progressive decline as sample sizes grow.

L.6 Simulation Results when $q = 3$

In this subsection, we conduct experiment when $q = 3$ to show the finite sample performance of our proposed method. We set the true group numbers in the three dimensions as $G_{1,0} = 3, G_{2,0} = 3, G_{3,0} = 2$. The true parameters are provided in Table A.11. We consider the networks generated from stochastic block models. The network sizes are set as $(N_1, N_2, N_3) \in \{(20, 20, 20), (30, 30, 30)\}$, and the time length is set to be $T \in \{10, 40, 80\}$. The network and the data generation schemes are the same as those in Section 6.1 in the main text. The experiments are repeated for $R = 300$ times.

We first evaluate estimation accuracy under correctly specified group numbers, with comprehensive results presented in Table A.12. We observe three key patterns. First, the mis-classification rates decline rapidly to zero as sample sizes increase, confirming precise recovery of group memberships. Second, the RMSEs decrease monotonically with increasing network sizes or time length, demonstrating consistent estimation efficiency. Third, the coverage probabilities converge to the nominal 95% confidence level under larger samples, validating asymptotic normality of the estimators. Collectively, these findings meet theoretical results for statistical consistency.

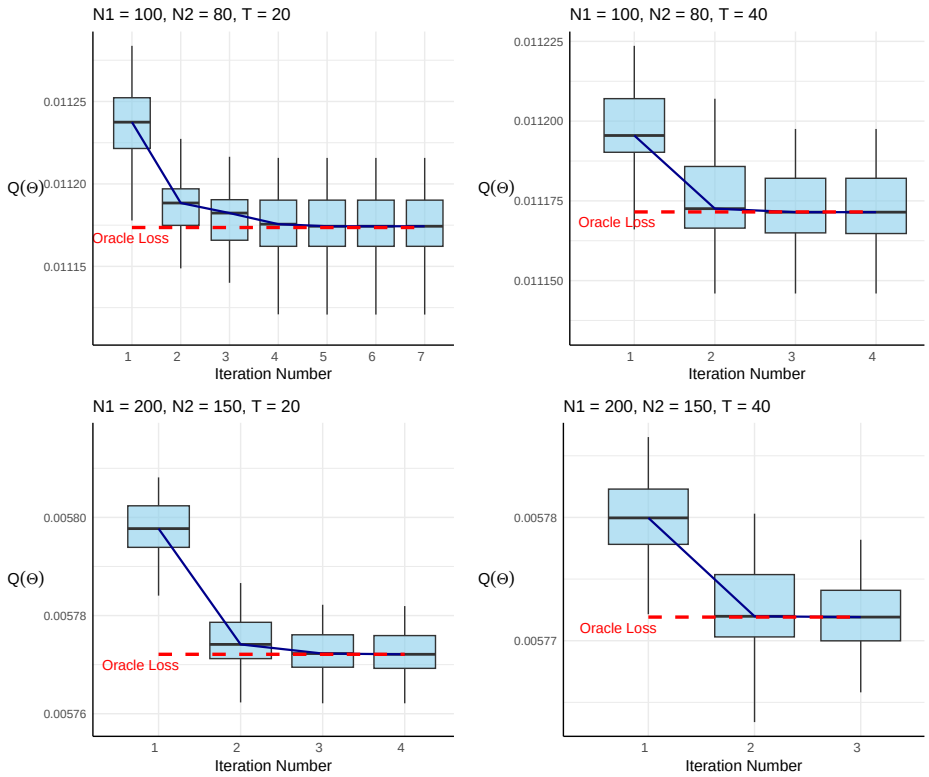


Figure A.8: Total loss in each iteration step and the number of iterations required for algorithm convergence in different network sizes and time lengths. Each box shows the total loss of $R = 50$ replicates. The red line shows the oracle loss.

Table A.11: True parameters when $G_{1,0} = 3, G_{2,0} = 3, G_{3,0} = 2$.

$G_{1,0} = 3$	$G_{2,0} = 3$	$G_{3,0} = 2$
$\lambda^{(1)}_{g^{(1)}}$ (-0.1, 0.2, 0.3)	$\lambda^{(2)}_{g^{(2)}}$ (0.15, 0.2, 0.4)	$\lambda^{(3)}_{g^{(3)}}$ (-0.2, 0.25)
$\boldsymbol{\zeta}^{(1)}_{g^{(1)}}$ $\begin{pmatrix} 0.2 & 0.25 & -0.3 \\ 0.15 & 0.35 & -0.35 \\ 0.24 & 0.30 & -0.32 \end{pmatrix}$	$\boldsymbol{\delta}^{(2)}_{g^{(2)}}$ $\begin{pmatrix} 0.25 & -0.3 & 0.35 \\ 0.2 & -0.25 & 0.32 \\ 0.1 & -0.2 & 0.2 \end{pmatrix}$	$\boldsymbol{\delta}^{(3)}_{g^{(3)}}$ $\begin{pmatrix} -0.1 & 0.2 & 0.4 \\ 0.3 & 0.1 & -0.32 \end{pmatrix}$
$\boldsymbol{\alpha}_{\cdot\cdot 1}$ $\begin{pmatrix} 0.2 & 0.25 & -0.3 \\ 0.15 & 0.35 & -0.35 \\ 0.24 & 0.30 & -0.32 \end{pmatrix}$	$\boldsymbol{\alpha}_{\cdot\cdot 2}$ $\begin{pmatrix} 0.2 & 0.25 & -0.3 \\ 0.15 & 0.35 & -0.35 \\ 0.24 & 0.30 & -0.32 \end{pmatrix}$	

Table A.12: RMSEs of estimated parameters when $G_{1,0} = G_{2,0} = 3$ and $G_{3,0} = 2$ with 300 replications. The performances are evaluated for different sample sizes N_1, N_2, N_3 and the time length T. The corresponding CPs are shown in the parenthesis.

N_1	N_2	N_3	T	$\widehat{\boldsymbol{\lambda}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}^{(2)}$	$\widehat{\boldsymbol{\lambda}}^{(3)}$	$\widehat{\boldsymbol{\zeta}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}^{(3)}$	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\eta}_1$	$\widehat{\eta}_2$	$\widehat{\eta}_3$
20	20	20	10	0.0165 (0.9256)	0.0141 (0.9411)	0.0097 (0.9367)	0.0207 (0.9226)	0.0204 (0.9315)	0.0128 (0.9272)	0.0666 (0.9222)	0.0083	0	0
			40	0.0066 (0.9356)	0.0067 (0.9378)	0.0044 (0.9533)	0.0093 (0.9411)	0.0096 (0.9452)	0.0059 (0.9372)	0.0263 (0.9422)	0	0	0
			80	0.0025 (0.9311)	0.0024 (0.9467)	0.0024 (0.9383)	0.0065 (0.9459)	0.0065 (0.9519)	0.0042 (0.9478)	0.0158 (0.9478)	0	0	0
30	30	30	10	0.0105 (0.9422)	0.0082 (0.9400)	0.0057 (0.9383)	0.0130 (0.9196)	0.0106 (0.9370)	0.0068 (0.9378)	0.0435 (0.9269)	0.0078	0	0
			40	0.0034 (0.9489)	0.0033 (0.9567)	0.0023 (0.9583)	0.0050 (0.9526)	0.0049 (0.9556)	0.0033 (0.9411)	0.0133 (0.9459)	0	0	0
			80	0.0013 (0.9478)	0.0013 (0.9422)	0.0012 (0.9650)	0.0034 (0.9489)	0.0037 (0.9507)	0.0024 (0.9461)	0.0084 (0.9463)	0	0	0

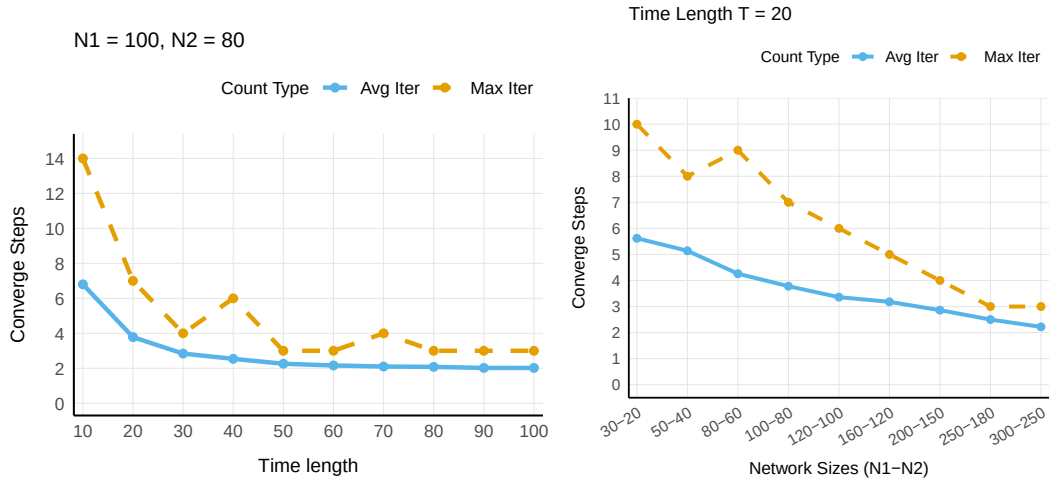


Figure A.9: Average and maximum number of iterations required for convergence under different sample sizes throughout 50 replicates. The orange dashed line shows the maximum number, while the blue solid lines shows the average one.

Next, we consider the setting of unspecified group numbers. We implement the QIC introduced in Section 3.3 in the main text, with κ set as the same with Section 6.1. The results are shown in Table A.13. When the group numbers are under-specified (i.e., $G_l = 2,\ l = 1, 2, 3$), the node-wise RMSEs are large and do not decrease as the sample sizes increase. On the contrary, when the group numbers are correctly- or over-specified, the node-wise RMSEs show a decreasing pattern as the theoretical results show. These findings are consistent with those for two dimensional tensor in Section 6.2 of the main text.

L.7 Comparison with Existing Methods

L.7.1 Comparing with Estimation Using A Series of Matrices Models

To enhance the difference between our proposed GTNAR model and a series of matrix models, we conduct an experiment when $q = 3$ to show the fundamental role of the intrinsic inner-tensor data information. Specifically, we generate data using the same setting as in Appendix L.6. The total time length is set to be $T_{train} + T_{test}$, and we use the first T_{train} as training set, while leaving the following T_{test} as the testing set. By implementing Algorithm A.2, we obtain the in-sample and out-of-sample predicted values $\widehat{\mathcal{Y}}_t^{\text{GTNAR}}$. For comparison, we divide the data by the third dimension into N_3 parallel data slices. Each slice contains a matrix-valued time series. By algorithm A.1, we can estimate the two-dimensional parameters for each series slice. Take the i_3 slice for example, we can calculate the in-sample fitted matrix $\widehat{\mathbf{Y}}_{\cdot\cdot i_3,t}, (t \in [T_{train}])$ and out-of-sample prediction values $\widehat{\mathbf{Y}}_{\cdot\cdot i_3,t}, (t = T_{train}+1, \cdots T_{train}+T_{test})$ for each slice. Then we aggregate the fitted and predicted values in all N_3 slices, obtaining the in-sample and out-of-sample predicted tensor $\widehat{\mathcal{Y}}_t^{\text{MatSlice}}$. We call this method by "Matrix Slices" method. Using the same metrics as in Appendix L.3.1, we calculate ReMSPE for both methods regarding in-sample and out-of-

Table A.13: Simulation results with pre-specified group numbers as well as the QIC selection group numbers $\widehat{G}_l,\ l=1,2,3$. The true group numbers are set as $G_{1,0}=G_{2,0}=3$ and $G_{3,0}=2$. The node-wise RMSEs of different estimators are denoted as $\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(l)}, \boldsymbol{\zeta}_{\text{all}}^{(l)}, \boldsymbol{\alpha}_{\text{all}}$ for $l=1,2,3$.

N_1	N_2	N_3	T	G_1	G_2	G_3	$\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(1)}$	$\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(2)}$	$\widehat{\boldsymbol{\lambda}}_{\text{all}}^{(3)}$	$\widehat{\boldsymbol{\zeta}}_{\text{all}}^{(1)}$	$\widehat{\boldsymbol{\zeta}}_{\text{all}}^{(2)}$	$\widehat{\boldsymbol{\zeta}}_{\text{all}}^{(3)}$	$\widehat{\boldsymbol{\alpha}}_{\text{all}}$	$\widehat{\xi}_1$	$\widehat{\xi}_2$	$\widehat{\xi}_3$
				Oracle			0.0037	0.0038	0.0031	0.0052	0.0046	0.0040	0.0058	-	-	-
				2	2	2	0.0559	0.0805	0.0551	0.0361	0.0468	0.0051	0.1400	0.2500	0.2080	0
20	20	20	40	3	3	2	0.0037	0.0038	0.0031	0.0052	0.0046	0.0040	0.0174	0.1403	0.0142	0
				4	4	4	0.0047	0.0048	0.0047	0.0061	0.0060	0.0063	0.0123	0	0	0
				$\widehat{G}_1$	$\widehat{G}_2$	$\widehat{G}_3$	0.0037	0.0038	0.0031	0.0052	0.0046	0.0040	0.0058	-	-	-
				Oracle			0.0014	0.0017	0.0017	0.0034	0.0034	0.0026	0.0037	-	-	-
				2	2	2	0.1140	0.0571	0.0580	0.0343	0.0213	0.0041	0.1447	0.2449	0.1959	0
20	20	20	80	3	3	2	0.0037	0.0028	0.0028	0.0040	0.0038	0.0026	0.0065	0	0	0
				4	4	4	0.0040	0.0033	0.0038	0.0046	0.0042	0.0043	0.0099	0	0	0
				$\widehat{G}_1$	$\widehat{G}_2$	$\widehat{G}_3$	0.0014	0.0017	0.0017	0.0034	0.0034	0.0026	0.0037	-	-	-

Table A.14: In-sample and out-of-sample ReMSPE of GTNAR and the matrix slices estimation methods.

N_1	N_2	N_3	T	**GTNAR**		**Matrix Slices**	
				$\text{ReMSPE}_{\text{tr}}$	$\text{ReMSPE}_{\text{te}}$	$\text{ReMSPE}_{\text{tr}}$	$\text{ReMSPE}_{\text{te}}$
20	20	20	10	0.6739	0.6549	0.8836	1.0319
			40	0.5882	0.5871	0.7692	0.8003
			80	0.0234	0.0049	2.4769	4.3342
30	30	30	10	0.6261	0.6173	0.8789	1.0281
			40	0.4710	0.4660	0.9787	1.7442
			80	0.0066	0.0006	4.8092	2.9876

sample performance. The experiments are both repeated for $R = 100$ times. The results are shown in Table A.14. GTNAR exhibits consistent improvement in both in-sample and out-of-sample prediction accuracy as sample size increases. In contrast, Matrix Slices shows no systematic improvement trend, with error magnitudes actually increasing substantially at larger T. Compared with the Matrix Slices method, our proposed GTNAR shows a superior prediction performance. These results show that the GTNAR fundamentally differs from simple matrix slicing approaches. The error gaps reveals critical information loss inherent in slice-based approximations when the inherent data is generated from the tensor-valued model.

L.7.2 Comparing with Other Methods

In this section, we conduct the prediction and compare the accuracy with a number of competing methods. Denote T_{train} and T_{test} as the time length for training set and the testing set. We generate the simulation data when $q = 2$, and the generation scheme is the same as that in Section 6.1 in the main text. We set the sample sizes as $(N_1, N_2, T_{train}) \in \{(30, 20, 20), (50, 40, 30), (80, 60, 40), (100, 80, 50), (120, 100, 60), (160, 120, 70), (200, 150, 80), (250, 180, 90)\}$. For each sample size setting, we set the following $T_{test} = 0.5T_{train}$ as the test set. To evaluate the prediction accuracy, we repeat the experiments for each setting for $R = 50$ times, and we compute average relative mean square prediction error (ReMSPE). Specifically, first calculate mean square prediction error as

$$\text{MSPE}^r = (N_1 N_2 T_{test})^{-1} \sum_{t=t_{test,0}}^{t_{test,0}+T_{test}} \sum_{i,j} (\widehat{Y}^r_{ij,t} - Y^r_{ij,t})^2,$$

where $\widehat{Y}^r_{ij,t}$ is the predicted response in the rth replicate, and $t_{test,0} = T_{train} + 1$ is the starting time point. Then, calculate

$$\text{MSPE}^r_0 = (N_1 N_2 T_{test})^{-1} \sum_{t=t_{test,0}}^{t_{test,0}+T_{test}} \sum_{i,j} (Y^r_{ij,t} - \widehat{\mu}^r_{ij,train})^2,$$

where $\widehat{\mu}^r_{ij,train} = T^{-1}_{trian} \sum_{t=1}^{T_{train}} Y^r_{ij,t}$ as the mean response in the training set. Subsequently, we calculate the average ReMSPE as $\text{ReMSPE} = R^{-1} \sum_r \text{MSPE}^r / \text{MSPE}^r_0$.

Specifically, we compare the prediction accuracy estimated by the proposed algorithm with that estimated by the sparse VAR method (sVAR, Nicholson et al. (2020)), the multilinear tensor regression model (BiTR, Hoff (2015)), the group NAR model (GNAR, Zhu et al. (2023)), and the deep learning method for time series, LSTM (Hochreiter and Schmidhuber, 1997). For the sVAR model, we first vectorize the matrix response $\mathbf{Y}_t$ to be in the vector form as $\text{vec}(\mathbf{Y}_t)$. Subsequently, we apply the sVAR method in the bigtime package to $\{\text{vec}(\mathbf{Y}_t)\}$ to obtain the model estimation and prediction result. For BiTR model, we use $\mathbf{Y}_t$ as the response variable, and set the explanatory variables as the concatenation by the row network term $\mathbf{W}^{(1)}\mathbf{Y}_{t-1}$, the and column network term $\mathbf{Y}_{t-1}\mathbf{W}^{(2)}$, the lag term $\mathbf{Y}_{t-1}$ and the covariates $\mathbf{X}_t^{(1)}$, $\mathbf{X}_t^{(2)}$. We apply the alternating least squares algorithm proposed by Hoff (2015) to estimate the model. For the GNAR model proposed by Zhu et al. (2023), we note that it only involves a single network and can only be applied for the vector time series data. Therefore we make prediction with their model form by involving one network matrix a time. To involve the user network (i.e., row network), we apply the GNAR model in the following form

$$Y_{ij,t} = \sum_{k=1}^{N_1} \beta_{g_i^{(1)} g_k^{(1)}} w_{1,ik} Y_{kj,t-1} + \nu_{g_i^{(1)}} Y_{ij,t-1} + \mathbf{x}_i^{(1)\top} \boldsymbol{\zeta}_{g_i^{(1)}} + \varepsilon_{ij,t}, \tag{A.105}$$

and estimate the model parameters and make prediction for each $j \in [N_2]$. The model prediction performance is denoted as GNAR-R in Table A.18. Similarly, to incorporate the spatial network (i.e., column network), we estimate the following GNAR model as

$$Y_{ij,t} = \sum_{k=1}^{N_2} \beta_{g_k^{(2)} g_j^{(2)}} w_{2,kj} Y_{ik,t-1} + \nu_{g_j^{(2)}} Y_{ij,t-1} + \mathbf{x}_j^{(2)\top} \boldsymbol{\zeta}_{g_j^{(2)}} + \varepsilon_{ij,t},$$

for each $i \in [N_1]$, whose prediction performance is denoted as GNAR-C in Figure A.10. For the above two models, we use the open source code provided by Zhu et al. (2023) to obtain the result. Lastly, we apply the LSTM method (Hochreiter and Schmidhuber, 1997) to our data $\{(\mathbf{x}_{it}^{(1)} \in \mathbb{R}^{p_1}, \mathbf{x}_{jt}^{(2)} \in \mathbb{R}^{p_2}, Y_{ij,t}) : i \in [N_1], j \in [N_2], t \in [T]\}$. During training, we set the learning rate as 0.05, the dimension of hidden layers as 64, and set the number of iteration as 100. The momentum coefficient is set as 0.5.

We remark that due to the suboptimal estimation accuracy of BiTR, which exhibits orders of magnitude differences compared to others, we present the ReMSPEs of GTNAR, GNAR-R, GNAR-C, sVAR and LSTM in Figure A.10 for visual clarity. The results for BiTR are provided in Table A.15 for detailed examination. Based on the simulation results, when the data is generated using the GTNAR model, the prediction accuracy of all alternative methods proves suboptimal. Even the LSTM model, which is recognized for its strong fitting capabilities by its non-linear structure, fails to achieve satisfactory prediction accuracy. Furthermore, in Appendix M.3, we conduct rolling-window prediction comparisons across six methods using the Yelp dataset, which also demonstrate the superiority of our proposed approach.

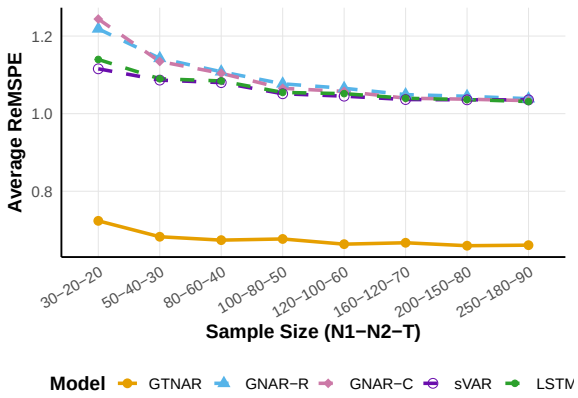


Figure A.10: Average ReMSPEs of GTNAR, GNAR-R, GNAR-C, sVAR and LSTM under different sample sizes.

Table A.15: Average ReMSPEs of GTNAR and BiTR under different sample sizes.

N_1	N_2	T_{train}	**GTNAR**	**BiTR**
30	20	20	0.7237	94.8849
50	30	30	0.6827	160.6119
80	60	40	0.6740	219.2674
100	80	50	0.6770	138.1804
120	100	60	0.6636	105.4973
160	120	70	0.6672	139.3628
200	150	80	0.6596	101.1558
250	180	90	0.6610	28.5175

Appendix M. Additional Yelp Data Analysis

M.1 Additional Results for Yelp Data Analysis

We visualize the relationship between these user-related covariates and the response variable in Figure A.11. The plot reveals that users who receive more tags for their reviews tend to be motivated to contribute more reviews. Notably, VIP users in Scottsdale and Toronto tend to write more reviews, whereas VIP users in Charlotte exhibit comparatively less activity. The estimation results for city Scottsdale and Toronto are provided in Table A.16.

M.2 Ablation Experiment

In this subsection, we use ablation experiments to show the necessity of each component in our GTNAR model. We implement both the GTNAR model (1) and the Mixed GTNAR model (25) to analyze the Yelp dataset. The evaluation metric is the similar as that in Appendix L.7.2 yet under a rolling window. Specifically, first calculate mean square prediction error as

$$\text{MSPE} = (N_1 N_2 T_{test})^{-1} \sum_{t=t_{test,0}}^{t_{test,0}+T_{test}} \sum_{i,j} (\widehat{Y}_{ij,t} - Y_{ij,t})^2,$$

Then, calculate

$$\text{MSPE}_0 = (N_1 N_2 T_{test})^{-1} \sum_{t=t_{test,0}}^{t_{test,0}+T_{test}} \sum_{i,j} (Y_{ij,t} - \widehat{\mu}_{ij,train})^2,$$

where $\widehat{\mu}_{ij,train} = T_{trian}^{-1} \sum_{t=t_{test,0}-T_{train}}^{t_{test,0}-1} Y_{ij,t}$ as the mean response in the training set. Subsequently, we calculate the ReMSPE as $\text{ReMSPE} = \text{MSPE}/\text{MSPE}_0$. Recall the GTNAR model

$$\mathbf{Y}_t = \underbrace{(\mathbf{L}^{(1)}\mathbf{W}^{(1)})\mathbf{Y}_{t-1} + \mathbf{Y}_{t-1}(\mathbf{W}^{(2)}\mathbf{L}^{(2)})}_{\text{network terms}} + \underbrace{\mathbf{A} \odot \mathbf{Y}_{t-1}}_{\text{momentum term}} + \underbrace{\boldsymbol{\beta}_{X_1,t}^{(1)} \mathbf{1}_{N_2}^\top + \mathbf{1}_{N_1} \boldsymbol{\beta}_{X_2,t}^{(2)\top}}_{\text{covariates terms}} + \mathbf{E}_t, \tag{A.106}$$

For the ablation experiment, we remove the network terms, the lag term, and the covariates terms from GTNAR (A.106) and perform predictions, respectively. For performance evaluation, we predict the out-of-sample review number in the same rolling window settings, and calculate the ReMSPE as defined in Appendix L.7.2. The prediction results are shown in Table A.17, from which we can see that both GTNAR and Mixed GTNAR consistently achieve the highest prediction accuracy across most of the five evaluation testing time periods. This performance superiority is particularly evident in Charlotte, Las Vegas, and Scottsdale, where these models outperform all ablated versions at every starting time point. The ablation study reveals performance deterioration when each component is removed, validating the contribution of each model component (network terms, momentum term, and covariates terms) to GTNAR's robust predictive capability.

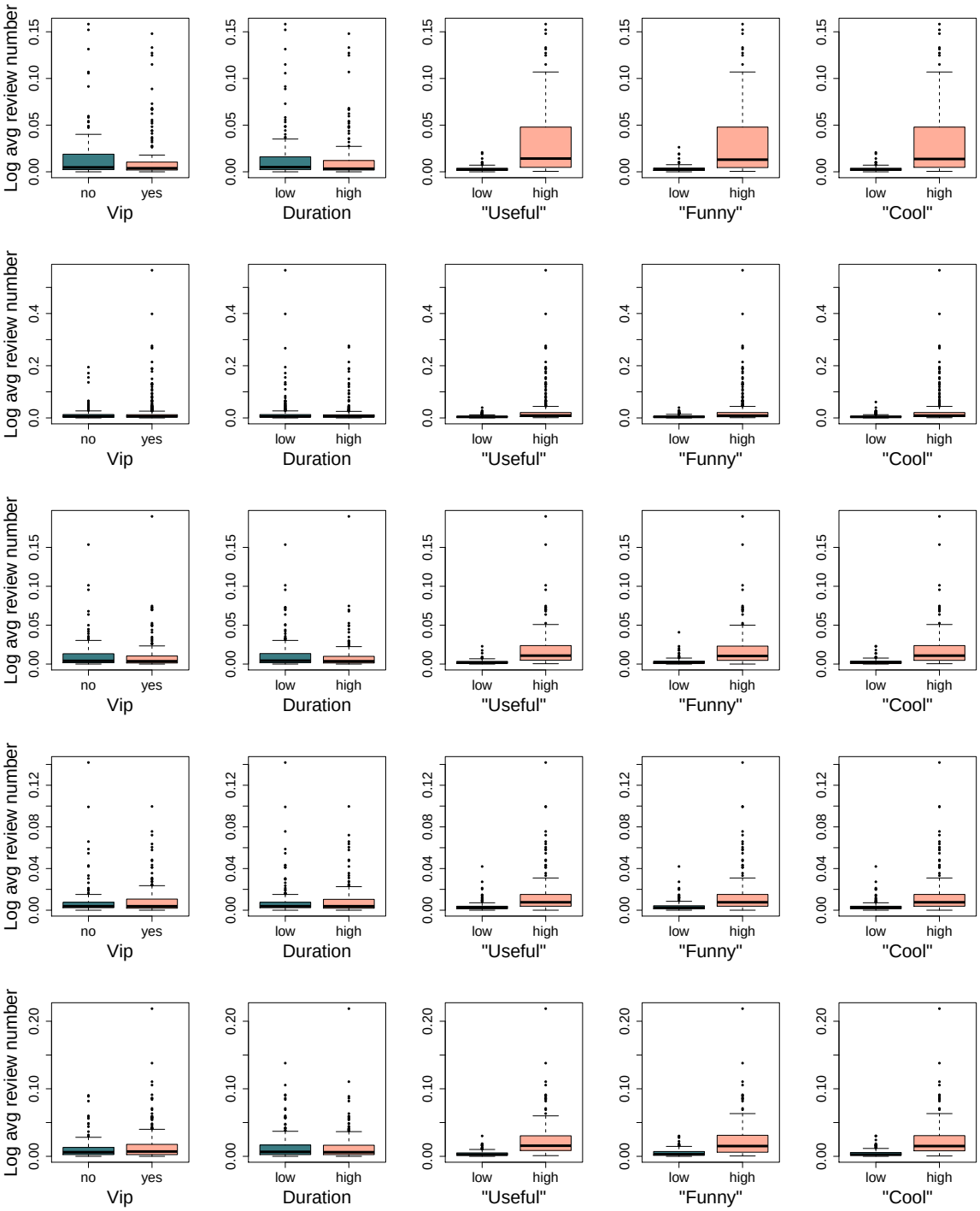


Figure A.11: Boxplots for response variable with regards to the users' covariates in Charlotte, Las Vegas, Phoenix, Scottsdale, and Toronto (displayed in each line). Specifically, the y-axis is the log-transformed average review number given by each user throughout the time span. The x-axis shows the corresponding high or low level for each covariate, separated by the median for continuous variables.

Table A.16: Estimation results for Scottsdale and Toronto. The p-values are shown in the parenthesis.

Parameters	Scottsdale ($N_1 = 391,\ N_2 = 60$)				Toronto ($N_1 = 462,\ N_2 = 56$)			
	$\lambda^{(1)}_{g^{(1)}}$		$\lambda^{(2)}_{g^{(2)}}$		$\lambda^{(1)}_{g^{(1)}}$		$\lambda^{(2)}_{g^{(2)}}$	
	0.067	0.011	0.069	0.239	0.013	0.037	0.305	0.107
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
	$\boldsymbol{\zeta}^{(1)}_{g^{(1)}}$		$\boldsymbol{\zeta}^{(2)}_{g^{(2)}}$		$\boldsymbol{\zeta}^{(1)}_{g^{(1)}}$		$\boldsymbol{\zeta}^{(2)}_{g^{(2)}}$	
Intercept	-0.002	0.002	-0.001	0.003	-0.009	0.009	0.016	-0.001
	(0.175)	(<0.001)	(<0.001)	(<0.001)	(0.001)	(<0.001)	(<0.001)	(0.274)
$\zeta^{(1)}_{\text{dur}}$ / $\zeta^{(2)}_{\text{star}}$	0.009	-0.003	0.002	0.004	10^{-4}	-0.006	0.003	0.003
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.453)	(<0.001)	(0.209)	(<0.001)
$\zeta^{(1)}_{\text{vip}}$ / $\zeta^{(2)}_{\text{num}}$	0.002	0.001	0.003	0.002	0.001	0.001	0.003	0.005
	(0.037)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.155)	(0.009)	(<0.001)
$\zeta^{(1)}_{\text{use}}$	0.227	0.023			0.032	0.069		
	(<0.001)	(<0.001)			(<0.001)	(<0.001)		
$\zeta^{(1)}_{\text{fun}}$	-0.360	-0.025			-0.002	0.013		
	(<0.001)	(<0.001)			(0.401)	(<0.001)		
$\zeta^{(1)}_{\text{cool}}$	0.176	0.028			0.002	0.041		
	(<0.001)	(<0.001)			(0.483)	(<0.001)		
	$\boldsymbol{\alpha}^\top \in \mathbb{R}^{G_1 \times G_2}$				$\boldsymbol{\alpha}^\top \in \mathbb{R}^{G_1 \times G_2}$			
	0.071		0.023		0.055		0.290	
	(<0.001)		(<0.001)		(<0.001)		(<0.001)	
	0.235		0.051		0.033		0.096	
	(<0.001)		(<0.001)		(<0.001)		(<0.001)	

Table A.17: ReMSPEs of the ablation study in five cities under the rolling window setting.

		Starting time point				
		2010-Q1	2010-Q2	2010-Q3	2010-Q4	2011-Q1
Charlotte	**GTNAR**	0.8386	0.8132	**0.8361**	**0.8646**	**0.8951**
	Mixed GTNAR	**0.8228**	**0.7725**	0.8382	0.9019	0.9629
	no network terms	0.8421	0.8140	0.8392	0.8689	0.8979
	no momentum term	0.8717	0.8199	0.8629	0.8824	0.9624
	no covariates terms	0.8539	0.8297	0.8565	0.8772	0.9131
Las Vegas	**GTNAR**	0.8336	0.7811	0.8549	0.9102	**0.9601**
	Mixed GTNAR	**0.8228**	**0.7725**	**0.8382**	**0.9019**	0.9629
	no network terms	0.8373	0.8021	0.8519	0.9237	0.9784
	no momentum term	0.8238	0.7913	0.8520	0.9145	0.9729
	no covariates terms	0.8119	0.7918	0.8395	0.9070	0.9691
Phoenix	**GTNAR**	1.1625	**1.0784**	0.9419	**0.8730**	0.8581
	Mixed GTNAR	1.1406	1.0888	0.9571	0.8822	0.8582
	no network terms	**1.1048**	1.0793	**0.9393**	0.8677	0.8616
	no momentum term	1.1426	1.0859	0.9424	0.8731	**0.8554**
	no covariates terms	1.1310	1.1023	0.9542	0.8877	0.8664
Scottsdale	**GTNAR**	**0.8725**	**0.8588**	**0.8535**	**0.8757**	**0.8415**
	Mixed GTNAR	0.8986	0.8757	0.8629	0.8859	0.8581
	no network terms	0.8761	0.8629	0.8587	0.8771	0.8441
	no momentum term	0.8992	0.8699	0.8545	0.9057	0.8931
	no covariates terms	0.8831	0.8685	0.8636	0.8854	0.8492
Toronto	**GTNAR**	1.5779	**1.4670**	**1.3785**	**1.2241**	**1.0669**
	Mixed GTNAR	1.5829	1.5279	1.3961	1.2313	1.0773
	no network terms	**1.5751**	1.4678	1.3863	1.2334	1.0681
	no momentum term	1.5783	1.4787	1.4079	1.2362	1.0766
	no covariates terms	1.5995	1.4942	1.4019	1.2513	1.0877

M.3 Prediction Accuracy of Comparing Methods

In this section, we conduct the prediction for the Yelp dataset and compare the accuracy with a number of competing methods. To illustrate the prediction performance of our proposed GTNAR model, we take a rolling window prediction setting. Specifically, we set the training length as $T_{train} = 25$ from the first data point, and set $T_{test} = 5$. Then we use a sliding window approach to evaluate the prediction performance by calculating the Root Mean Square Prediction Error (RMSPE) with moving one quarter in each step, which is calculated by

$$\text{RMSPE} = \Big\{(N_1 N_2 T_{test})^{-1} \sum_{t=t_{test,0}}^{t_{test,0}+T_{test}} \sum_{i,j} (\widehat{Y}_{ij,t} - Y_{ij,t})^2 \Big\}^{1/2},$$

where $\widehat{Y}_{ij,t}$ is the predicted response from user i to the district j in the tth quarter, and $t_{test,0}$ is the starting time point. The settings for the competing methods, namely, sVAR, BiTR, GNAR-R, GNAR-C and LSTM are the same as those in Appendix L.7.2.

The prediction results are detailed in Table A.18. In comparison to the sVAR, BiTR, GNAR-R and GNAR-C, the GTNAR method demonstrates superior and more stable prediction accuracy across various staring points in all five cities. In comparison to the LSTM model, we find that the prediction performances are comparable. In particular, for Charlotte, Las Vegas, Scottsdale and Toronto, the GTNAR method outperforms the LSTM with higher prediction accuracy. For Phoenix, the prediction error of the LSTM method in the first start time point is slightly smaller than that of the GTNAR method. We would like to remark that other than our relative robust prediction performance, our modeling approach provides a clearer model interpretation compared with complex nonlinear machine learning methods.

Appendix N. Addtional Trading Data Application

N.1 Data Description

In this section, we embark on an analysis of the monthly import and export volumes of goods among countries/regions, spanning the period from 1993 to 2022, comprising a total of $T = 360$ months. The source for our multilateral trading data is the IMF-DOTS database (IMF, 2017). To enhance the stationarity of the time series, we define the response variable, denoted by $Y_{ij,t}$, as the first-order difference in export volume, which is referred to as the "increased export volume (IEV)" in the subsequent discussions. IEV quantifies the change in export volume from the ith country/region to the jth country/region between the tth month and the preceding $(t-1)$th month.

Covariates. In our analysis, we consider three essential covariates for each country or region. Firstly, the annual gross domestic product ($x^{(1)}_{it,\text{gdp}}/x^{(2)}_{jt,\text{gdp}}$) serves as a measure of the overall size of the economy. Secondly, we incorporate the difference of log-annual average exchange rate to US dollars ($x^{(1)}_{it,\text{exc}}/x^{(2)}_{jt,\text{exc}}$) in consecutive years, which influences the actual costs of exporting goods. The influence from exchange rate toward trading pattern is underscored in prior research (Hooper and Kohlhagen, 1978; Auboin and Ruta, 2013). Lastly, we include the proportion of the population with internet access ($x^{(1)}_{it,\text{net}}/x^{(2)}_{jt,\text{net}}$),

Table A.18: Prediction RMSPE in five cities under the rolling window setting.

City	Model	Start Time Point				
		2010-Q1	2010-Q2	2010-Q3	2010-Q4	2021-Q1
Charlotte	**GTNAR**	0.0848	**0.0841**	**0.0845**	**0.0848**	**0.0854**
	sVAR	0.1145	0.1067	0.0999	0.1048	0.1061
	BiTR	**0.0560**	0.2329	0.7017	0.5040	0.6093
	GNAR-R	0.0964	0.1128	0.1227	0.0923	0.0949
	GNAR-C	2.6539	0.1023	0.2522	0.1080	0.0953
	LSTM	0.0866	0.0851	0.0863	0.0860	0.0871
LasVegas	**GTNAR**	**0.0786**	**0.0760**	**0.0790**	**0.0814**	**0.0829**
	sVAR	0.1118	0.1344	0.1300	0.1273	0.1442
	BiTR	1.5677	1.5457	1.6350	1.6465	1.2869
	GNAR-R	0.0862	0.1124	7.0599	0.6934	0.0849
	GNAR-C	43.9337	3.4841	37.4009	0.0943	0.2487
	LSTM	0.0789	0.0782	0.0790	0.0816	0.0837
Phoenix	**GTNAR**	0.0791	**0.0762**	**0.0716**	**0.0690**	**0.0682**
	sVAR	0.0877	0.0881	0.0861	0.0930	0.0881
	BiTR	0.6108	1.1864	0.2036	0.2322	0.2437
	GNAR-R	0.0801	0.0796	0.0792	0.0736	0.0733
	GNAR-C	1.3056	0.0795	0.1082	0.0745	0.1303
	LSTM	**0.0782**	0.0790	0.0738	0.0700	0.0717
Scottsdale	**GTNAR**	**0.0617**	**0.0614**	**0.0611**	**0.0619**	0.0600
	sVAR	0.0772	0.0720	0.0715	0.0766	0.0730
	BiTR	0.7890	2.0271	0.8091	0.8288	**0.0446**
	GNAR-R	0.0655	0.0639	0.0680	0.0638	0.0639
	GNAR-C	0.1423	0.0778	0.0656	0.0713	0.1265
	LSTM	0.0632	0.0624	0.0616	0.0624	0.0606
Toronto	**GTNAR**	**0.0989**	**0.0972**	**0.0958**	**0.0915**	**0.0865**
	sVAR	0.1117	0.1106	0.1105	0.1067	0.1083
	BiTR	1.3064	0.9224	0.6444	1.2483	0.9182
	GNAR-R	0.1013	0.1006	0.1013	0.1022	0.0951
	GNAR-C	5.1932	3.0433	6.2586	18.4431	0.4216
	LSTM	0.0999	0.0985	0.0969	0.0932	0.0877

a metric that captures the effects of the rapid development of e-commerce and its effect on cross-regional commodity circulation, as discussed in studies such as those by Freund and Weinhold (2004) and Terzi (2011). This covariate data has been sourced from both the World Bank Database[1] and the IMF-DOTS database, encompassing the years 1992 to 2021. Figure A.12 depicts the accumulated response for export volume in the year 2022 versus the three covariates of export countries in 2021, which suggests that a lower exchange rate, higher GDP, or a higher proportion of internet users in 2021 are associated with potentially higher export volumes in 2022.

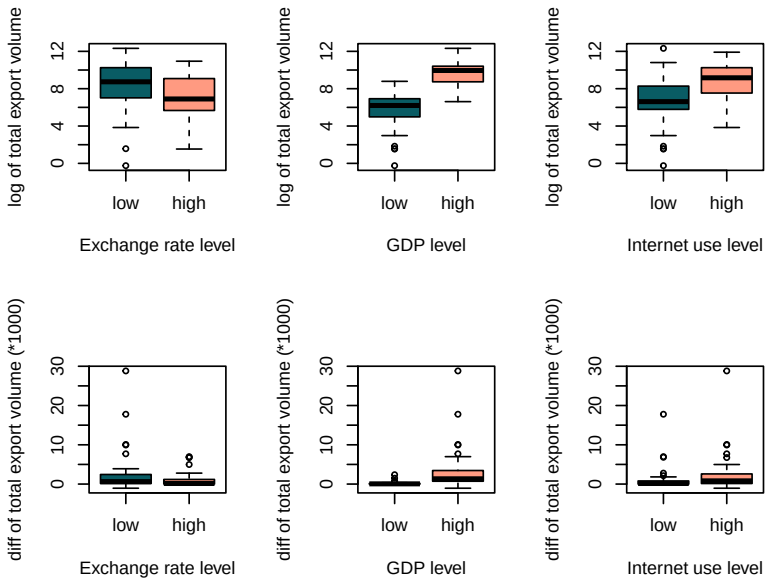


Figure A.12: Top panels: the relationship between the total export volume in 2022 and the levels of three covariates in 2021 (the categorization into "low" and "high" levels is determined by the medians). Bottom panels: the relationship between the first-order difference in exports between 2022 and 2021 and the levels of the same three covariates in 2021.

Spatial Networks. The countries under study are interconnected through various international relationships. In this analysis, our focus is on the spatial network spillover effects, a phenomenon well-documented in the literature (Redding and Venables, 2004; Yin et al., 2020). After filtering out countries/regions with missing data, we have a total of $N_1 = N_2 = 82$ observed units. To construct the adjacency matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, we reference neighboring countries and territories for each country and region, as listed on Wikipedia[22]. In these adjacency matrices, an element $a_{ij}^{(1)} = 1$ ($a_{ij}^{(2)} = 1$) signifies the presence of at least one land border between the two regions. Consequently, our response variable, $Y_{ij,t}$, is indexed by two networks, the exporter network $\mathbf{A}^{(1)}$ and the importer network $\mathbf{A}^{(2)}$. In particular, we note that although these two networks share the same network nodes and

2. https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_number_of_land_borders

topology, they exert different influences on the response variable, accounting for export and import spatial spillover effects on IEV, respectively.

N.2 Estimating Results

We next implement the GMNAR model to investigate the dynamic trading patterns around the world. The QIC suggests two groups for the exporter network ($\mathbf{A}^{(1)}$) nodes and three groups for the importer network ($\mathbf{A}^{(2)}$) nodes. The group patterns are color-coded in Figure A.13 and estimated model parameters are summarized in Table A.19. We summarize the estimated results as follows.

Self-momentum effects. Table A.19 reveals that the self-momentum effects $\widehat{\boldsymbol{\alpha}}$ are consistently negative for all cases, with the exception of $\widehat{\alpha}_{22}$. One possible explanation for the negative auto-correlation is that the demand for exported goods from specific countries within the importer network tends to remain relatively stable over the years. Consequently, a significant increase in imports in the previous year may indicate a reduction in the importation of these goods in the following year. The sole exception is $\widehat{\alpha}_{22} > 0$, representing the self-momentum of exported goods from group 2 of the exporter network (i.e., mainland China and Brazil) to group 2 of the importer network (i.e., Canada, mainland China and Tunisia). This implies that trade between these countries has consistently grown at an accelerated speed from 1992 to 2021. For instance, mainland China's exports to the Canada have increased at annualized rates of 10.8% and 14.7%, respectively, while Brazil's exports to mainland China and Canada have grown at rates of 17.6% and 8.41%[3]. Such observations align well with the estimation by the GTNAR method.

Spatial exporter network effects. Table A.19 indicates that the export network effects ($\widehat{\lambda}^{(1)}_{g^{(1)}}$'s) are negative for both groups, but the second group, consisting of Brazil and mainland China, exhibits a substantially larger magnitude. The negative exporter network effects imply that a country's export is negatively influenced when neighboring countries experience significant increases in their exports, particularly pronounced in the case of large developing countries like Brazil and mainland China. This observation suggests that countries in neighboring regions tend to be competitors rather than collaborators in the export market.

Spatial importer network effects. In Table A.19, we also observe that the import network effects $\widehat{\lambda}_{g^{(2)}}$ are positive in the first two groups (i.e., $\widehat{\lambda}^{(2)}_1 = 0.042$ and $\widehat{\lambda}^{(2)}_2 = 0.032$), but negative in the third group (i.e., $\widehat{\lambda}^{(2)}_3 = -0.005$). The positive values of $\widehat{\lambda}_{g^{(2)}}$ for the first two groups indicate that if neighboring countries import more goods, it positively influences their own imports. This implies a positive spillover effect of the import market for regions within the first two groups (such as mainland China, the United States and Canada). Conversely, the third group predominantly comprises developing countries in South America and Africa, which are shown to be less affected by their neighbors' imports.

Covariates effects. In Table A.19, we can observe distinct effects on IEV (International Export Volume) resulting from the three covariates. Firstly, it is notable that for the second group of exporters, $\widehat{\zeta}^{(1)}_{\text{exc},2} = -1.710$, signifying a significant negative correlation. This suggests that higher export costs lead to reduced IEV in Brazil and mainland China.

3. https://oec.world/en/profile/bilateral-country

Conversely, for the second group of importers, $\widehat{\zeta}^{(2)}_{\text{exc},2} = 0.035$, which is a positive value and aligns with expectations, as higher exchange rates generally correspond to increased import volumes. Secondly, both $\widehat{\zeta}^{(1)}_{\text{gdp}}$ and $\widehat{\zeta}^{(2)}_{\text{gdp}}$ display positive effects on IEV across all countries. This aligns with existing research indicating that domestic economic development bolsters international trade (Bernard and Jensen, 1999, 2004; Berman and Héricourt, 2010). Thirdly, we can observe that $\widehat{\zeta}^{(1)}_{\text{net},1} = 0.228$, suggesting that a higher level of Internet access in most exporting countries leads to increased IEV. However, it's worth noting that countries in the second export group (Brazil and mainland China), as well as the second importer group (mainland China and Canada), exhibit negative coefficients concerning Internet population percentage. This discrepancy may be attributed to the rapid growth of internet users in China, which has outpaced the growth in export volume from 1993 to 2022.

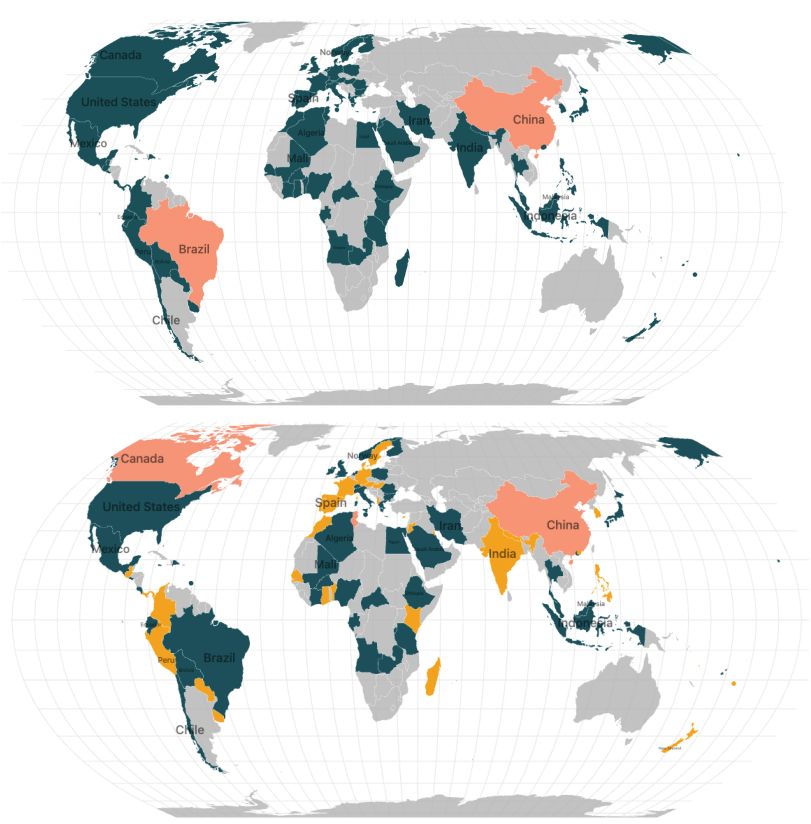


Figure A.13: The group patterns of export (the top) and import (the bottom) markets. The Group 1 and 2 in the export market are marked with dark green and pink. The Group 1, 2, 3 in the import market are marked with dark green, pink and orange, respectively. The gray color refers to the regions excluded in the model dataset.

Table A.19: Estimation results for trading data analysis. The p-values are shown in the parenthesis.

Parameters	$\lambda^{(1)}_{g^{(1)}}$		$\lambda^{(2)}_{g^{(2)}}$		
	-0.003	-0.512	0.042	0.032	-0.005
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
	$\boldsymbol{\zeta}^{(1)}_{g^{(1)}}$		$\boldsymbol{\zeta}^{(2)}_{g^{(2)}}$		
Intercept	-0.011	2.072	-0.369	2.674	-0.244
	(0.960)	(0.002)	(0.112)	(<0.001)	(0.329)
$\boldsymbol{\zeta}^{(1)}_{\text{exc}}/\boldsymbol{\zeta}^{(2)}_{\text{exc}}$	-0.120	-1.710	-0.104	0.035	-1.399
	(0.484)	(0.036)	(0.549)	(<0.001)	(0.027)
$\boldsymbol{\zeta}^{(1)}_{\text{gdp}}/\boldsymbol{\zeta}^{(2)}_{\text{gdp}}$	0.693	2.544	1.046	1.015	1.681
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
$\boldsymbol{\zeta}^{(1)}_{\text{net}}/\boldsymbol{\zeta}^{(2)}_{\text{net}}$	0.228	-2.962	0.029	-1.314	0.206
	(0.022)	(0.010)	(0.838)	(0.001)	(0.186)
	$\boldsymbol{\alpha}^\top \in \mathbb{R}^{G_1 \times G_2}$				
	-0.431		-0.494		
	(<0.001)		(<0.001)		
	-0.287		0.021		
	(<0.001)		(<0.001)		
	-0.315		-0.311		
	(<0.001)		(<0.001)		