Journal of Machine Learning Research 24 (2023) 1-49 Submitted 12/21; Revised 1/23; Published 2/23

Dimension-Grouped Mixed Membership Models for Multivariate Categorical Data

Yuqi Gu YUQI.GU@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Elena A. Erosheva EROSHEVA@UW.EDU
Department of Statistics, School of Social Work,
and the Center for Statistics and the Social Sciences
University of Washington
Seattle, WA 98195, USA

Gongjun Xu GONGJUN@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

David B. Dunson DUNSON@DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708, USA

Editor: Mingyuan Zhou

Abstract

Mixed Membership Models (MMMs) are a popular family of latent structure models for complex multivariate data. Instead of forcing each subject to belong to a single cluster, MMMs incorporate a vector of subject-specific weights characterizing partial membership across clusters. With this flexibility come challenges in uniquely identifying, estimating, and interpreting the parameters. In this article, we propose a new class of *Dimension-Grouped* MMMs (Gro-M^{3}s) for multivariate categorical data, which improve parsimony and interpretability. In Gro-M^{3}s, observed variables are partitioned into groups such that the latent membership is constant for variables within a group but can differ across groups. Traditional latent class models are obtained when all variables are in one group, while traditional MMMs are obtained when each variable is in its own group. The new model corresponds to a novel decomposition of probability tensors. Theoretically, we derive transparent identifiability conditions for both the unknown grouping structure and model parameters in general settings. Methodologically, we propose a Bayesian approach for Dirichlet Gro-M^{3}s to inferring the variable grouping structure and estimating model parameters. Simulation results demonstrate good computational performance and empirically confirm the identifiability results. We illustrate the new methodology through applications to a functional disability survey dataset and a personality test dataset.

Keywords: Bayesian Methods, Grade of Membership Model, Identifiability, Mixed Membership Model, Multivariate Categorical Data, Probabilistic Tensor Decomposition.

1. Introduction

Mixed membership models (MMMs) are a popular family of latent structure models for complex multivariate data. Building on classical latent class and finite mixture models (McLachlan and Peel, 2000), which assign each subject to a single cluster, MMMs include a vector of probability weights characterizing partial membership. MMMs have seen many applications in a wide variety of fields, including social science surveys (Erosheva et al., 2007), topic modeling and text mining (Blei et al., 2003), population genetics and bioinformatics (Pritchard et al., 2000; Saddiki et al., 2015), biological and social networks (Airoldi et al., 2008b), collaborative filtering (Mackey et al., 2010), and data privacy (Manrique-Vallier and Reiter, 2012); see Airoldi et al. (2014) for more examples.

Although MMMs are conceptually appealing and very flexible, with the rich modeling capacity come challenges in identifying, accurately estimating, and interpreting the parameters. MMMs have been popular in many applications, yet key theoretical issues remain understudied. The handbook of Airoldi et al. (2014) emphasized theoretical difficulties of MMMs ranging from non-identifiability to multi-modality of the likelihood. Finite mixture models have related challenges, and the additional complexity of the *individual-level* mixed membership incurs extra difficulties. A particularly important case is MMMs for multivariate categorical data, such as survey response (Woodbury et al., 1978; Erosheva et al., 2007; Manrique-Vallier and Reiter, 2012). In this setting, MMMs provide an attractive alternative to the latent class model of Goodman (1974). However, little is known about what is fundamentally identifiable and learnable from observed data under such models.

Identifiability is a key property of a statistical model, meaning that the model parameters can be uniquely obtained from the observables. An identifiable model is a prerequisite for reproducible statistical inferences and reliable applications. Indeed, interpreting parameters estimated from an unidentifiable model is meaningless, and may lead to misleading conclusions in practice. It is thus important to study the identifiability of MMMs and to provide theoretical support to back up the conceptual appeal. Even better would be to expand the MMM framework to allow variations that aid interpretability and identifiability. With this motivation, and focused on mixed membership modeling of multivariate categorical data, this paper makes the following key contributions.

We propose a new class of models for multivariate categorical data, which retains the same flexibility offered by MMMs, while favoring greater parsimony and interpretability. The key innovation is to allow the p-dimensional latent membership vector to belong to G ($G \ll p$) groups; memberships are the same for different variables within a group but can differ across groups. We deem the new model the *Dimension-Grouped Mixed Membership Model* (Gro-M^3). Gro-M^3 improves interpretability by allowing the potentially high-dimensional observed variables to belong to a small number of meaningful groups. Theoretically, we show that both the continuous model parameters, and the discrete variable grouping structure, can be identified from the data for models in the Gro-M^3 class under transparent conditions on how the variables are grouped. This challenging identifiability issue is addressed by carefully leveraging the dimension-grouping structure to write the model as certain structured tensor products, and then invoking Kruskal's fundamental theorem on the uniqueness of three-way tensor decompositions (Kruskal, 1977; Allman et al., 2009).

To illustrate the methodological usefulness of the proposed class of models, we consider a special case in which each subject's mixed membership proportion vector follows a Dirichlet distribution. This is among the most popular modeling assumptions underlying various MMMs (Blei et al., 2003; Erosheva et al., 2007; Manrique-Vallier and Reiter, 2012; Zhao et al., 2018). For such a Dirichlet Gro-M^3, we employ a Bayesian inference procedure and develop a Metropolis-Hastings-within-Gibbs algorithm for posterior computation. The algorithm has excellent computational performance. Simulation results demonstrate this approach can accurately learn the identifiable quantities of the model, including both the variable-grouping structure and the continuous model parameters. This also empirically confirms the model identifiability result.

The rest of this paper is organized as follows. Section 2 reviews existing mixed membership models, introduces the proposed Gro-M^3, and provides an interesting probabilistic tensor decomposition perspective of the models. Section 3 is devoted to the study of the identifiability of the new model. Section 4 focuses on the Dirichlet distribution induced Gro-M^3 and proposes a Bayesian inference procedure. Section 5 includes simulation studies and Section 6 applies the new model to reanalyze the NLTCS disability survey data. Section 7 provides discussions.

2. Dimension-Grouped Mixed Membership Models

2.1 Existing Mixed Membership Models

In this subsection, we briefly review the existing MMM literature to give our proposal appropriate context. Let K be the number of extreme latent profiles. Denote the K-dimensional probability simplex by $\Delta^{K-1} = \{(\pi_1, \ldots, \pi_K) : \pi_k \geq 0 \text{ for all } k,\ \sum_{k=1}^K \pi_k = 1\}$. Each subject i has an individual proportion vector $\boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,K}) \in \Delta^{K-1}$, which indicates the degrees to which subject i is a member of the K extreme profiles. The general mixed membership models summarized in Airoldi et al. (2014) have the following distribution,

$$p\left(\left\{y_{i,1}^{(r)}, \ldots, y_{i,p}^{(r)}\right\}_{r=1}^{R}\right) = \int_{\Delta^{K-1}} \prod_{j=1}^{p} \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \pi_{i,k} f(y_{i,j}^{(r)} \mid \boldsymbol{\lambda}_{j,k})\right) dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i), \tag{1}$$

where $\boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,K})$ follows the distribution $D_{\boldsymbol{\alpha}}$ and is integrated out; the $\boldsymbol{\alpha}$ refers to some generic population parameters depending on the specific model. The hierarchical Bayesian representation for the model in (1) can be written as follows.

$$\begin{aligned} y_{ij}^{(1)}, \ldots, y_{ij}^{(R)} \mid z_{ij} = k &\overset{\text{i.i.d.}}{\sim} \text{Categorical}([d_j];\ \boldsymbol{\lambda}_{j,k}), \quad j \in [p]; \\ z_{i1}, \ldots, z_{ip} \mid \boldsymbol{\pi}_i &\overset{\text{i.i.d.}}{\sim} \text{Categorical}([K];\ \boldsymbol{\pi}_i), \quad i \in [n]; \\ \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n &\overset{\text{i.i.d.}}{\sim} D_{\boldsymbol{\alpha}}. \end{aligned}$$

where "i.i.d." is short for "independent and identically distributed". The number p in (1) is the number of "characteristics", and R is the number of "replications" per characteristic. As shown in (1), for each characteristic j, there are a corresponding set of K conditional distributions indexed by parameter vectors $\{\boldsymbol{\lambda}_{j,k} : k = 1, \ldots, K\}$. Many different mixed

membership models are special cases of the general setup (1). For example, the popular Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Blei, 2012; Anandkumar et al., 2014) for topic modeling takes a document i as a subject, and assumes there is only $p = 1$ distinct characteristic (one single set of K topics which are distributions over the word vocabulary) with $R > 1$ replications (a document i contains R words which are conditionally i.i.d. given $\boldsymbol{\pi}_i$); LDA further specifies $D_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i)$ to be the Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$.

Focusing on MMMs for multivariate categorical data, there are generally many characteristics with $p \gg 1$ and one replication of each characteristic with $R = 1$ in (1). Each variable $y_{i,j} \in \{1, \ldots, d_j\}$ takes one of d_j unordered categories. For each subject i, the observables $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,p})^\top$ are a vector of p categorical variables. MMMs for such data are traditionally called Grade of Membership models (GoMs) (Woodbury et al., 1978). GoMs have been extensively used in applications, such as disability survey data (Erosheva et al., 2007), scholarly publication data (Erosheva et al., 2004), and data disclosure risk and privacy (Manrique-Vallier and Reiter, 2012). GoMs are also useful for psychological measurements where data are Likert scale responses to psychology survey items, and educational assessments where data are students' correct/wrong answers to test questions (e.g. Shang et al., 2021).

In GoMs, the conditional distribution $f(y_{i,j} \mid \boldsymbol{\lambda}_{j,k})$ in (1) can be written as $\mathbb{P}(y_{i,j} \mid \boldsymbol{\lambda}_{j,k}) = \prod_{c_j=1}^{d_j} \lambda_{j,c_j,k}^{\mathbb{I}(y_{i,j}=c_j)}$. Hence, the probability mass function of $\boldsymbol{y}_i$ in a GoM is

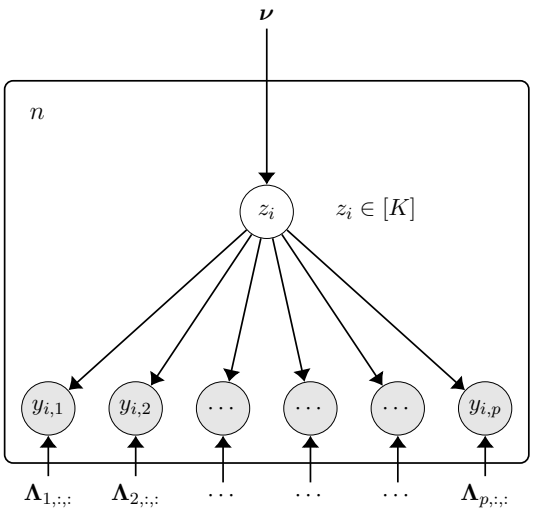
$$p^{\mathrm{GoM}}(y_{i,1}, \ldots, y_{i,p} \mid \boldsymbol{\Lambda}, \boldsymbol{\alpha}) = \int_{\Delta^{K-1}} \prod_{j=1}^{p} \left[\sum_{k=1}^{K} \pi_{i,k} \prod_{c_j=1}^{d_j} \lambda_{j,c_j,k}^{\mathbb{I}(y_{i,j}=c_j)} \right] dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i). \tag{2}$$

The hierarchical Bayesian representation for the model in (2) can be written as follows.

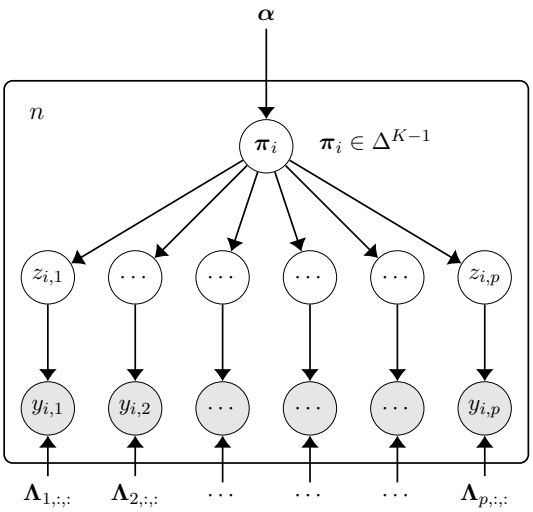
$$\begin{aligned} y_{ij} \mid z_{ij} = k &\overset{\text{i.i.d.}}{\sim} \text{Categorical}([d_j];\ \boldsymbol{\lambda}_{j,k}), \quad j \in [p]; \\ z_{i1}, \ldots, z_{ip} \mid \boldsymbol{\pi}_i &\overset{\text{i.i.d.}}{\sim} \text{Categorical}([K];\ \boldsymbol{\pi}_i), \quad i \in [n]; \\ \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n &\overset{\text{i.i.d.}}{\sim} D_{\boldsymbol{\alpha}}. \end{aligned}$$

See a graphical model representation of the GoM with sample size n in Figure 1(b), where individual latent indicator variables $(z_{i,1}, \ldots, z_{i,p}) \in [K]^p$ are introduced to better describe the data generative process.

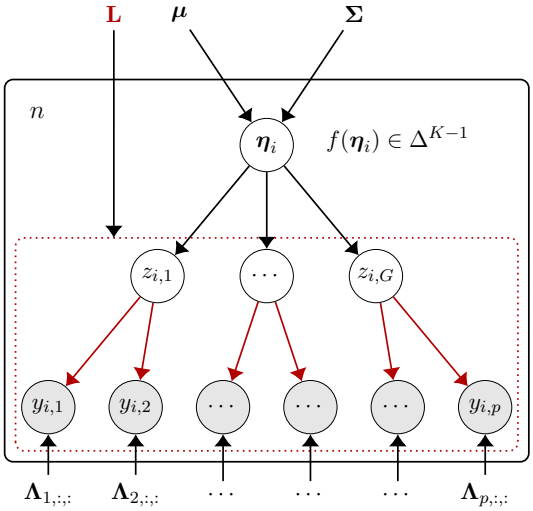
We emphasize that the case with $p > 1$ and $R = 1$ is fundamentally different from the topic models with $p = 1$ and $R > 1$, with the former typically involving many more parameters. This is because the "bag-of-words" assumption in the topic model with $R > 1$ disregards word order in a document and assumes all words in a document are *exchangeable*. In contrast, our mixed-membership model for multivariate categorical data does not assume a subject's responses to the p items in a survey/questionnaire are exchangeable. In other words, given a subject's mixed membership vector $\boldsymbol{\pi}_i$, his/her responses to the p items are independent *but not* identically distributed (because they follow categorical distributions governed by p different sets of parameters $\{\boldsymbol{\lambda}_{j,k} \in \mathbb{R}^d : \ k \in [K\}$ for $j = 1, \ldots, p)$; whereas in a topic model, given a document's latent topic proportion vector $\boldsymbol{\pi}_i$, the p words in the document are independent *and* identically distributed, following the categorical distribution



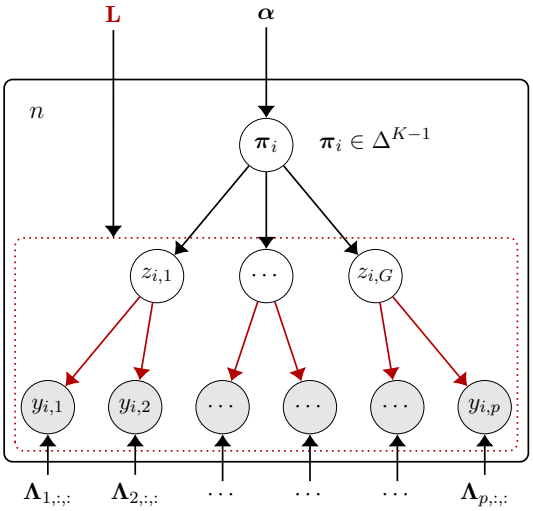
(a) Latent Class Model
(Probabilistic CP Decomposition)



(b) Grade of Membership Model, $z_{i,j} \in [K]$
(Probabilistic Tucker Decomposition)



(c) (New) Gro-M^3, $f(\boldsymbol{\eta}_i)$ logit normal
(Probabilistic Hybrid Decomposition)



(d) (New) Gro-M^3, $\boldsymbol{\pi}_i$ Dirichlet
(Probabilistic Hybrid Decomposition)

Figure 1: Graphical model representations of LCMs in (a), GoMs in (b), and the proposed family of Gro-M^{3}s with two examples in (c), (d). Shaded nodes $\{y_{i,j}\}$ are observed variables, white nodes are latent variables, quantities outside each solid box are population parameters. In (c) and (d), the dotted red box is the key dimension-grouping structure, where the red edges from $\{z_{i,g}\}$ to $\{y_{i,j}\}$ correspond to entries of "1" in the grouping matrix $\mathbf{L}$.

with the same set of parameters $\{\boldsymbol{\lambda}_k \in \mathbb{R}^V : k \in [K]\}$ (here V denotes the vocabulary size). In this sense, the GoM model has greater modeling flexibility than topic models and are more suitable for modeling item response data, where it is inappropriate to assume that the items in the survey/questionnaire are exchangeable or share the same set of parameters.

This fact is made clear also in Figure 1(b), where for each $j \in [p]$ there is a population quantity, the parameter node $\mathbf{\Lambda}_{j,:,:}$ (also denoted by $\mathbf{\Lambda}_j$ for simplicity), that governs its distribution. Thus identifiability is a much greater challenge for GoM models. To our best knowledge, the identifiability issue of the grade-of-membership (GoM) models for item response data considered in Woodbury et al. (1978) and Erosheva et al. (2007) has not been rigorously investigated so far. Motivated by the difficulty of identifying GoM in its original setting due to the large parameter complexity, we next propose a new modeling grouping component to enhance identifiability. Our resulting model still does not make any exchangeability assumption of the items, but rather leverages the variable grouping structure to reduce model complexity.

2.2 New Modeling Component: the Variable Grouping Structure

Generalizing Grade of Membership models for multivariate categorical data, we propose a new structure that groups the p observed variables in the following sense: any subject's latent membership is the same for variables within a group but can differ across groups. To represent the key structure of how the p variables are partitioned into G groups, we introduce a notation of the *grouping matrix* $\mathbf{L} = (\ell_{j,g})$. The $\mathbf{L}$ is a $p \times G$ matrix with binary entries, with rows indexed by the p variables and columns by the G groups. Each row j of $\mathbf{L}$ has exactly one entry of "1" indicating group assignment. In particular,

$$\mathbf{L} = (\ell_{j,g})_{p\times G}, \qquad \ell_{j,g} = \begin{cases} 1, & \text{if the } j\text{th variable belongs to the } g\text{th group;} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Our key specification is the following generative process in the form of a hierarchical Bayesian representation,

$$\begin{aligned} \text{Gro-M}^3\text{:} \qquad \{y_{i,j}\}_{\ell_{j,g}=1} \mid z_{i,g} = k &\overset{\text{ind.}}{\sim} \text{Categorical}\left([d_j];\, \left(\lambda_{j,1,k}, \cdots, \lambda_{j,d_j,k}\right)\right), \quad g \in [G]; \\ z_{i,1}, \ldots, z_{i,G} \mid \boldsymbol{\pi}_i &\overset{\text{i.i.d}}{\sim} \text{Categorical}([K];\, \boldsymbol{\pi}_i); \\ \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n &\overset{\text{i.i.d.}}{\sim} D_{\boldsymbol{\alpha}}. \end{aligned} \tag{4}$$

where "ind." is short for "independent", meaning that conditional on $z_{i,g} = k$, subject i's observed responses to items in group g are independently generated. Hence, given the population parameters $(\mathbf{L}, \mathbf{\Lambda}, \boldsymbol{\alpha})$, the probability distribution of $\boldsymbol{y}_i$ can be written as

$$p^{\text{Gro-M}^3}\left(y_{i,1}, \ldots, y_{i,p} \mid \mathbf{L}, \mathbf{\Lambda}, \boldsymbol{\alpha}\right) = \int_{\Delta^{K-1}} \prod_{g=1}^{G} \left[\sum_{k=1}^{K} \pi_{i,k} \prod_{j:\, \ell_{j,g}=1} \prod_{c_j=1}^{d_j} \lambda_{j,c_j,k}^{\mathbb{I}(y_{i,j}=c_j)} \right] dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i).$$

For a sample with n subjects, assume the observed responses $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent and identically distributed according to the above model.

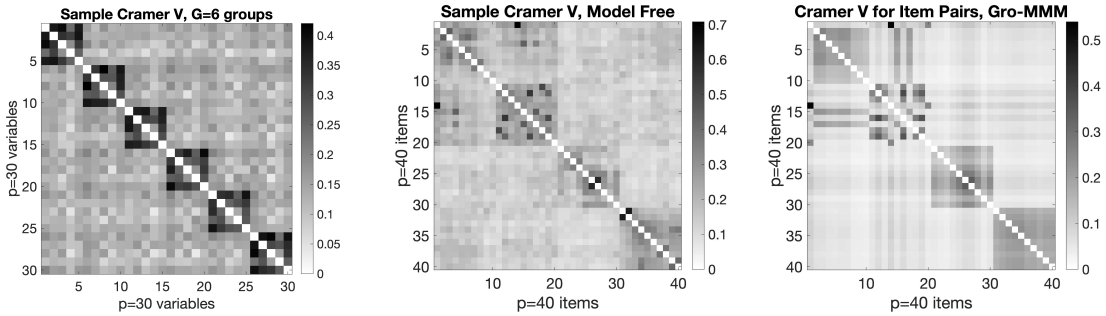
We visualize the proposed model as a probabilistic graphical model to highlight connections to and differences from existing latent structure models for multivariate categorical data. In Figure 1, we show the graphical model representations of two popular latent structure models for multivariate categorical data in (a) and (b), and for the newly proposed Gro-M^3 in (c) and (d). The $\mathbf{\Lambda}_j$ for $j \in [p]$ denotes a $d_j \times K$ matrix with entries $\lambda_{j,c_j,k}$.

Each column of $\mathbf{\Lambda}_j$ characterizes a conditional probability distribution of variable y_j given a particular extreme latent profile. We emphasize that the variable grouping is done at the level of the latent allocation variables z, and that the $\mathbf{\Lambda}_j$ parameters are still free without constraints just as they are in traditional LCMs or GoMs. From the visualizations in Figure 1 we can also easily distinguish our proposed model from another popular family of methods, the co-clustering methods (Dhillon et al., 2003; Govaert and Nadif, 2013). Co-clustering usually refers to simultaneously clustering the subjects and clustering the variables, where subjects within a cluster exhibit similar behaviors and variables within a cluster also share similar characteristics. Our Gro-M^3, however, does not restrict the p variables to have similar characteristics within groups, but rather allows them to have entirely free parameters $\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_p$. The "dimension-grouping" happens at the latent level by constraining the latent allocations behind the p variables to be grouped into G statuses. Such groupings give rise to a novel probabilistic hybrid tensor decomposition visualized in Figure 1(c)–(d); see the next Section 2.3 for details.

Other than facilitating model identifiability (see Section 3), our dimension-grouping modeling assumption is also motivated by real-world applications. In general, our new model Gro-M^3 with the variable grouping component can apply to any multivariate categorical data to simultaneously model individual mixed membership and capture variable similarity. For example, Gro-M^3 can be applied to survey/questionnaire response data in social sciences, where it is not only of interest to model subjects' partial membership to several extreme latent profiles, but also of interest to identify blocks of items which share similar measurement goals within each block. We next provide numerical evidence to demonstrate the merit of the variable grouping modeling component. For a dataset simulated from Gro-M^3 (in the setting as the later Table 2) and also the real-world IPIP personality test dataset (analyzed in the later Section 6), we calculate the sample Cramer's V between item pairs. Cramer's V is a classical measure of association between two categorical variables, which gives a value between 0 and 1, with larger values indicating stronger association. Figure 2 presents the plots of the sample Cramer's V matrix for the simulated data and the real IPIP data. This figure shows that the pairwise item dependence for the Gro-M^3-simulated data looks quite similar to the real-world personality test data. Indeed, after fitting the Gro-M^3 to this IPIP personality test dataset, the estimated model-based Cramer's V shown in Figure 2(c) nicely and more clearly recovers the item block structure. We conjecture that many real world datasets in other applied domains exhibit similar grouped dependence.

2.3 Probabilistic Tensor Decomposition Perspective

The Gro-M^3 class has interesting connections to popular tensor decompositions. For a subject i, the observed vector $\boldsymbol{y}_i$ resides in a contingency table with $\prod_{j=1}^{p} d_j$ cells. Since the MMMs for multivariate categorical data (both traditional GoM and the newly proposed Gro-M^3) induce a probability of $\boldsymbol{y}_i$ being in each of these cells, such probabilities $\{p\,(y_{i,1} = c_1, \ldots, y_{i,p} = c_p \mid -)\,;\, c_j \in [d_j] \text{ for each } j \in [p]\}$ can be arranged as a p-way $d_1 \times d_2 \times \cdots \times d_p$ array. This array is a tensor with p modes and we denote it by $\mathbf{P}$; Kolda and Bader (2009) provided a review of tensors. The tensor $\mathbf{P}$ has all the entries nonnegative and they sum up to one, so we call it a *probability tensor*. We next describe in



(a) Simulated data. (b) IPIP data, sample CRV. (c) IPIP data, Gro-M^3 CRV.

Figure 2: (a): Sample Cramer's V (abbreviated as CRV) for a simulated dataset. (b): Sample Cramer's V for the IPIP data. (c) Gro-M^3 based Cramer's V for the IPIP data.

detail the tensor decomposition perspective of our model; such a perspective will turn out to be useful in the study of identifiability.

The probability mass function of $\boldsymbol{y}_i$ under the traditional GoM model can be written as follows by exchanging the order of product and summation,

$$p^{\mathrm{GoM}}\left(y_{i,1}=c_1,\ldots,y_{i,p}=c_p \mid \boldsymbol{\Lambda},\boldsymbol{\alpha}\right)=\int_{\Delta^{K-1}}\prod_{j=1}^{p}\left[\sum_{k=1}^{K}\pi_{i,k}\lambda_{j,c_j,k}\right]dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i)$$

$$=\sum_{k_1=1}^{K}\cdots\sum_{k_p=1}^{K}\prod_{j=1}^{p}\lambda_{j,c_j,k_j}\underbrace{\int_{\Delta^{K-1}}\pi_{i,k_1}\cdots\pi_{i,k_p}dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i)}_{=:\,\phi^{\mathrm{GoM}}_{k_1,\ldots,k_p}}. \tag{5}$$

Then $\boldsymbol{\Phi}^{\mathrm{GoM}}:=\left(\phi^{\mathrm{GoM}}_{k_1,\ldots,k_p};\ k_j\in[K]\right)$ forms a tensor with p modes, and each mode has dimension K. Further, this tensor $\boldsymbol{\Phi}$ is a probability tensor, because $\phi_{k_1,\ldots,k_p}\geq 0$ and it is not hard to see that its entries sum up to one. Viewed from a tensor decomposition perspective, this is the popular Tucker decomposition (Tucker, 1966); more specifically this is the nonnegative and probabilistic version of the Tucker decomposition. The $\boldsymbol{\Phi}^{\mathrm{GoM}}$ represents the Tucker tensor core, and the product of $\{\lambda_{j,c_j,k}\}$ form the Tucker tensor arms.

It is useful to compare our modeling assumption to that of the the Latent Class Model (LCM; Goodman, 1974), which follows the graphical model shown in Figure 1(a). The LCM is essentially a finite mixture model assuming each subject i belongs to a single cluster. The distribution of $\boldsymbol{y}_i$ under an LCM is

$$p^{\mathrm{LC}}\left(y_{i,1}=c_1,\ldots,y_{i,p}=c_p \mid \boldsymbol{\Lambda},\boldsymbol{\nu}\right)=\sum_{k=1}^{K}\mathbb{P}(z_i=k)\prod_{j=1}^{p}\mathbb{P}(y_{i,j}\mid z_i=k)=\sum_{k=1}^{K}\nu_k\prod_{j=1}^{p}\lambda_{j,c_j,k}. \tag{6}$$

Based on the above definition, each subject i has a single variable $z_i\in[K]$ indicating which latent class it belongs to, rather than a mixed membership proportion vector $\boldsymbol{\pi}_i$. Denoting $\boldsymbol{\nu}^{\mathrm{LC}}=(\nu_k;\ k\in[K])$, then (6) corresponds to the popular CP decomposition of tensors (Hitchcock, 1927), where the CP rank is at most K.

Finally, consider our proposed Gro-M^3,

$$p^{\text{Gro-M}^3}(y_{i,1},\ldots,y_{i,p} \mid \mathbf{L},\mathbf{\Lambda},\boldsymbol{\alpha}) = \int_{\Delta^{K-1}} \prod_{g=1}^{G} \left[\sum_{k=1}^{K} \pi_{i,k} \prod_{j:\,\ell_{j,g}=1} f(y_{i,j} \mid \lambda_{j,c_j,k}) \right] dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i)$$

$$= \sum_{k_1=1}^{K} \cdots \sum_{k_G=1}^{K} \prod_{g=1}^{G} \prod_{j:\,\ell_{j,g}=1} f(y_{i,j} \mid \lambda_{j,c_j,k_g}) \underbrace{\int_{\Delta^{K-1}} \pi_{i,k_1} \cdots \pi_{i,k_G} dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i)}_{=:\,\phi^{\text{Gro-M}^3}_{k_1,\ldots,k_G}}, \tag{7}$$

where $f(y_{i,j}|\lambda_{j,c_j,k})$ generally denotes the conditional distribution of variable $y_{i,j}$ given parameter $\lambda_{j,c_j,k}$. In our Gro-M^3, $\lambda_{j,c_j,k}$ specifically refer to the categorical distribution parameters for the random variable $y_{i,j}$; that is, $\lambda_{j,c_j,k} = \mathbb{P}(y_{i,j} = c_j \mid z_{i,j} = k)$ denotes the probability of responding c_j to item j given that the subject's realization of the latent profile for item j is the kth extreme latent profile. In this case, $\mathbf{\Phi}^{\text{Gro-M}^3} := \left(\phi^{\text{Gro-M}^3}_{k_1,\ldots,k_G};\ k_g \in [K]\right)$ forms a tensor with G modes, and each mode has dimension K. There still is $\sum_{k_1=1}^{K} \cdots \sum_{k_G=1}^{K} \phi^{\text{Gro-M}^3}_{k_1,\ldots,k_G} = 1$. This reduces the size of the core tensor in the classical Tucker decomposition because $G < p$. The Gro-M^3 incorporates aspects of both the CP and Tucker decompositions, providing a *probabilistic hybrid decomposition* of probability tensors. The CP is obtained when all variables are in the same group, while the Tucker is obtained when each variable is in its own group; see Figure 1 for a clear illustration of this fact.

Gro-M^3 is conceptually related to the collapsed Tucker decomposition (c-Tucker) of Johndrow et al. (2017), though they did not model mixed memberships, used a very different model for the core tensor $\mathbf{\Phi}$, and did not consider identifiability. Nonetheless and interestingly, our identifiability results can be applied to establish identifiability of c-Tucker decomposition (see Remark 7 in Section 4). Another work related to our dimension-grouping assumption is Anandkumar et al. (2015), which considered the case of overcomplete topic modeling with the number of topics exceeding the vocabulary size. For such scenarios, the authors proposed a "persistent topic model" which assumes the latent topic assignment persists locally through multiple words, and established identifiability. Our dimension-grouped mixed membership assumption is similar in spirit to this topic persistence assumption. However, the setting we consider here for general multivariate categorical data has the multi-characteristic single-replication nature ($p > 1$ and $R = 1$); as mentioned before, this is fundamentally different from topic models with a single characteristic and multiple replications ($p = 1$ and $R > 1$).

3. Identifiability of Dimension-Grouped MMMs

Identifiability is an important property of a statistical model, generally meaning that model parameters can be uniquely recovered from the observables. Identifiability serves as a fundamental prerequisite for valid statistical estimation and inference. The study of identifiability, however, can be challenging for complicated models and especially latent variable models, including the Gro-M^{3}s considered here. In subsections 3.1 and 3.2, we propose easily checkable and practically useful identifiability conditions for Gro-M^{3}s by carefully inspecting the inherent algebraic structures. Specifically, we will exploit the variable group-

ings to write the model as certain highly structured mixed tensor products, and then prove identifiability by invoking Kruskal's theorem on the uniqueness of tensor decompositions (Kruskal, 1977). We point out that such proof procedures share a similar spirit to those in Allman et al. (2009), but the complicated structure of the new Gro-M^{3}s requires some special care to handle. We provide a high-level summary of our proof approach. First, we write the probability mass function of the observed p-dimensional multivariate categorical vector as a probabilistic tensor with p modes. Second, we unfold this tensor into a G-way tensor with each mode corresponding to a variable group. Third, we further concatenate the transformed tensor and leverage Kruskal's Theorem on the uniqueness of three-way tensor decomposition to establish the identifiability of the model parameters under our proposed Gro-M^3. Our theoretical developments provide a solid foundation for performing estimation of the latent quantities and drawing valid conclusions from data.

3.1 Strict Identifiability Conditions

For LDA and closely related topic models, there is a rich literature investigating identifiability under different assumptions (Anandkumar et al., 2012; Arora et al., 2012; Nguyen, 2015; Wang, 2019). Typically, when there is only one characteristic ($p = 1$), $R \geq 2$ is necessary for identifiability; see Example 2 in Wang (2019). However, there has been limited consideration of identifiability of mixed membership models with multiple characteristics and one replication, i.e., $p > 1$ and $R = 1$. GoM models belong to this category, as does the proposed Gro-M^{3}s, with GoM being a special case of Gro-M^{3}s.

We consider the general setup in (1), where $\mathbf{\Phi}$ denotes the G-mode tensor core induced by any distribution $D(\boldsymbol{\pi}_i)$ on the probability simplex Δ^{K-1}. The following definition formally defines the concept of strict identifiability for the proposed model.

Definition 1 (Strict Identifiability of Gro-M^{3}s) *A parameter space $\mathbf{\Theta}$ of a Gro-M^3 is said to be strictly identifiable, if for any valid set of parameters $(\mathbf{L}, \mathbf{\Lambda}, \mathbf{\Phi}) \in \mathbf{\Theta}$, the following equations hold if and only if $(\mathbf{L}, \mathbf{\Lambda}, \mathbf{\Phi})$ and the alternative $(\overline{\mathbf{L}}, \overline{\mathbf{\Lambda}}, \overline{\mathbf{\Phi}})$ are identical up to permutations of the K extreme latent profiles and permutations of the G variable groups,*

$$\mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \mathbf{L}, \mathbf{\Lambda}, \mathbf{\Phi}) = \mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \overline{\mathbf{L}}, \overline{\mathbf{\Lambda}}, \overline{\mathbf{\Phi}}), \quad \forall \boldsymbol{c} \in \times_{j=1}^{p}[d_j]. \tag{8}$$

Definition 1 gives the strongest possible notion of identifiability of the considered population quantities $(\mathbf{L}, \mathbf{\Lambda}, \mathbf{\Phi})$ in the model. In particular, the strict identifiability notion in Definition 1 requires identification of *both* the continuous parameters $\mathbf{\Lambda}$ and $\mathbf{\Phi}$, *and* the discrete latent grouping structure of variables in $\mathbf{L}$. The following theorem proposes sufficient conditions for the strict identifiability of Gro-M^{3}s.

Theorem 2 *Under a Gro-M^3, the following two identifiability conclusions hold.*

(a) Suppose each column of $\mathbf{L}$ contains at least three entries of "1"s, and the corresponding conditional probability table $\mathbf{\Lambda}_j = (\lambda_{j,c_j,k})_{d_j \times K}$ for each of these three j has full column rank. Then the $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ are strictly identifiable.

(b) In addition to the conditions in (a), if $\mathbf{\Lambda}$ satisfies that for each $j \in [p]$, not all the column vectors of $\mathbf{\Lambda}_j$ are identical, then $\mathbf{L}$ is also identifiable.

Example 1 *Denote by* $\mathbf{I}_G$ *a* $G \times G$ *identity matrix. Suppose* $p = 3G$ *and the matrix* $\mathbf{L}$ *takes the following form,*

$$\mathbf{L} = (\mathbf{I}_G \ \mathbf{I}_G \ \mathbf{I}_G)^\top. \tag{9}$$

Also suppose for each $j \in \{1, \ldots, 3G\}$*, the* $\mathbf{\Lambda}_j$ *of size* $d_j \times K$ *has full column rank* K*. Then the conditions in Theorem 2 hold, so* $\mathbf{\Lambda}$*,* $\mathbf{L}$ *and* $\mathbf{\Phi}$ *are identifiable. Theorem 2 implies that if* $\mathbf{L}$ *contains any additional row vectors other than those in* (9) *the model is still identifiable.*

Theorem 2 requires that each of the G variable groups contains at least three variables, and that for each of these $3G$ variables, the corresponding conditional probability table $\mathbf{\Lambda}_j$ has linearly independent columns. Theorem 2 guarantees not only the continuous parameters are identifiable, but also the discrete variable grouping structure summarized by $\mathbf{L}$ is identifiable. This is important practically as typically the appropriate variable grouping structure is unknown, and hence needs to be inferred from the data.

The conditions in Theorem 2 essentially requires at least $3G$ conditional probability tables, each being a matrix of size $d_j \times K$, to have full column rank. This implicitly requires $d_j \geq K$. Tan and Mukherjee (2017) proposed a moment-based estimation approach for traditional mixed membership models and briefly discussed the identifiability issue, also assuming $d_j \geq K$ with some full-rank requirements. However, the cases where the number of categories d_j's are small but the number of extreme latent profiles K is much larger can arise in applications; for example, the disability survey data analyzed in Erosheva et al. (2007) and Manrique-Vallier (2014) have binary responses with $d_1 = \cdots = d_p = 2$ while the considered K ranges from 2 to 10. Our next theoretical result establishes weaker conditions for identifiability that accommodates $d_j < K$, by taking advantage of the dimension-grouping property of our proposed model class.

Before stating the theorem, we first introduce two useful notions of matrix products. Denote by $\bigotimes$ the Kronecker product of matrices and by $\bigodot$ the Khatri-Rao product. Consider two matrices $\mathbf{A} = (a_{i,j}) \in \mathbb{R}^{m \times r}$, $\mathbf{B} = (b_{i,j}) \in \mathbb{R}^{s \times t}$; and another two matrices $\mathbf{C} = (c_{i,j}) = (\boldsymbol{c}_{:,1} \mid \cdots \mid \boldsymbol{c}_{:,k}) \in \mathbb{R}^{n \times k}$, $\mathbf{D} = (d_{i,j}) = (\boldsymbol{d}_{:,1} \mid \cdots \mid \boldsymbol{d}_{:,k}) \in \mathbb{R}^{\ell \times k}$, then there are $\mathbf{A} \bigotimes \mathbf{B} \in \mathbb{R}^{ms \times rt}$ and $\mathbf{C} \bigodot \mathbf{D} \in \mathbb{R}^{n\ell \times k}$ with

$$\mathbf{A} \bigotimes \mathbf{B} = \begin{pmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,r}\mathbf{B} \\ \vdots & \vdots & \vdots \\ a_{m,1}\mathbf{B} & \cdots & a_{m,r}\mathbf{B} \end{pmatrix}, \qquad \mathbf{C} \bigodot \mathbf{D} = \left(\boldsymbol{c}_{:,1} \bigotimes \boldsymbol{d}_{:,1} \mid \cdots \mid \boldsymbol{c}_{:,k} \bigotimes \boldsymbol{d}_{:,k}\right).$$

The above definitions show the Khatri-Rao product is the column-wise Kronecker product. The Khatri-Rao product of matrices plays an important role in the technical definition of the proposed dimension-grouped MMM. The following Theorem 3 exploits the grouping structure in $\mathbf{L}$ to relax the identifiability conditions in the previous Theorem 2.

Theorem 3 *Denote by* $\mathcal{A}_g = \{j \in [p] : \ell_{j,g} = 1\}$ *the set of variables that belong to group* g*. Suppose each* $\mathcal{A}_g$ *can be partitioned into three sets* $\mathcal{A}_g = \cup_{m=1}^3 \mathcal{A}_{g,m}$*, and for each* $g \in [G]$ *and* $m \in \{1, 2, 3\}$ *the matrix* $\widetilde{\mathbf{\Lambda}}_{g,m}$ *defined below has full column rank* K*.*

$$\widetilde{\mathbf{\Lambda}}_{g,m} := \bigodot_{j \in \mathcal{A}_{g,m}} \mathbf{\Lambda}_j. \tag{10}$$

Also suppose for each $j \in [p]$, not all the column vectors of $\mathbf{\Lambda}_j$ are identical. Then the model parameters $\mathbf{L}$, $\mathbf{\Lambda}$, and $\mathbf{\Phi}$ are strictly identifiable.

Compared to Theorem 2, Theorem 3 relaxes the identifiability conditions by lifting the full-rank requirement on the individual matrices $\mathbf{\Lambda}_j$'s. Rather, as long as the Khatri-Rao product of several different $\mathbf{\Lambda}_j$'s have full column rank as specified in (10), identifiability can be guaranteed. Recall that the Khatri-Rao product of two matrices $\mathbf{\Lambda}_{j_1}$ of size $d_{j_1} \times K$ and $\mathbf{\Lambda}_{j_2}$ of size $d_{j_2} \times K$ has size $(d_{j_1} d_{j_2}) \times K$. So intuitively, requiring $\mathbf{\Lambda}_{j_1} \bigodot \mathbf{\Lambda}_{j_2}$ to have full column rank K is weaker than requiring each of $\mathbf{\Lambda}_{j_1}$ and $\mathbf{\Lambda}_{j_2}$ to have full column rank K. The following Example 2 formalizes this intuition.

Example 2 *Consider $d_1 = d_2 = 2$, $K = 3$ with the following conditional probability tables*

$$\mathbf{\Lambda}_1 = \begin{pmatrix} a_1 & a_2 & a_3 \\ 1-a_1 & 1-a_2 & 1-a_3 \end{pmatrix}; \quad \mathbf{\Lambda}_2 = \begin{pmatrix} b_1 & b_2 & b_3 \\ 1-b_1 & 1-b_2 & 1-b_3 \end{pmatrix}.$$

Suppose variables $j = 1, 2$ belong to the same group, e.g., $\ell_{1,:} = \ell_{2,:}$. Then since $K > d_1 = d_2$, both $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ can not have full column rank K. However, if we consider their Khatri-Rao product, it has size 4×3 in the following form

$$\mathbf{\Lambda}_1 \bigodot \mathbf{\Lambda}_2 = \begin{pmatrix} a_1 b_1 & a_2 b_2 & a_3 b_3 \\ a_1(1-b_1) & a_2(1-b_2) & a_3(1-b_3) \\ (1-a_1) b_1 & (1-a_2) b_2 & (1-a_3) b_3 \\ (1-a_1)(1-b_1) & (1-a_2)(1-b_2) & (1-a_3)(1-b_3) \end{pmatrix}.$$

Indeed, $\mathbf{\Lambda}_1 \bigodot \mathbf{\Lambda}_2$ has full column rank for "generic" parameters $\boldsymbol{\theta} := (a_1, a_2, a_3, b_1, b_2, b_3)$; precisely speaking, for $\boldsymbol{\theta}$ varying almost everywhere in the parameter space $[0,1]^6$ (the 6-dimensional hypercube), the subset of $\boldsymbol{\theta}$ that renders $\mathbf{\Lambda}_1 \bigodot \mathbf{\Lambda}_2$ rank-deficient has Lebesgue measure zero in $\mathbb{R}^6$. To see this, let $\boldsymbol{x} = (x_1, x_2, x_3)^\top \in \mathbb{R}^3$ such that $(\mathbf{\Lambda}_1 \bigodot \mathbf{\Lambda}_2)\boldsymbol{x} = 0$, then

$$\begin{cases} a_1 b_1 x_1 + a_2 b_2 x_2 + a_3 b_3 x_3 = 0; \\ a_1(1-b_1) x_1 + a_2(1-b_2) x_2 + a_3(1-b_3) x_3 = 0; \\ (1-a_1) b_1 x_1 + (1-a_2) b_2 x_2 + (1-a_3) b_3 x_3 = 0; \\ (1-a_1)(1-b_1) x_1 + (1-a_2)(1-b_2) x_2 + (1-a_3)(1-b_3) x_3 = 0; \end{cases}$$

$$\overset{\text{invertible transform}}{\Longleftrightarrow} \quad \begin{cases} a_1 b_1 x_1 + a_2 b_2 x_2 + a_3 b_3 x_3 = 0; \\ a_1 x_1 + a_2 x_2 + a_3 x_3 = 0; \\ b_1 x_1 + b_2 x_2 + b_3 x_3 = 0; \\ x_1 + x_2 + x_3 = 0. \end{cases}$$

Based on the last four equations above, one can use basic algebra to obtain the following set of equations about (x_1, x_2, x_3),

$$\begin{pmatrix} b_1 - b_3 & b_3 - b_2 \\ a_1 - a_3 & a_3 - a_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_2 - b_1 & b_1 - b_3 \\ a_2 - a_1 & a_1 - a_3 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This implies as long as the following inequalities hold, there must be $x_1 = x_2 = x_3 = 0$*,*

$$\begin{cases} (b_1 - b_3)(a_3 - a_2) - (a_1 - a_3)(b_3 - b_2) \neq 0; \\ (b_2 - b_1)(a_1 - a_3) - (a_2 - a_1)(b_1 - b_3) \neq 0 \end{cases} \tag{11}$$

Now note that the subset of the parameter space $\{(a_1, a_2, a_3, b_1, b_2, b_3) \in [0,1]^6 : \textit{Eq. (11) holds}\}$ *is a Lebesgue measure zero subset of* $[0,1]^6$*. This means for such "generic" parameters varying almost everywhere in the parameter space* $[0,1]^6$*, the* $(\mathbf{\Lambda}_1 \bigodot \mathbf{\Lambda}_2)\boldsymbol{x} = \mathbf{0}$ *implies* $\boldsymbol{x} = \mathbf{0}$ *which means* $\mathbf{\Lambda}_1 \bigodot \mathbf{\Lambda}_2$ *has full column rank* $K = 3$*.*

Example 2 shows that the Khatri-Rao product of two matrices seems to have full rank under fairly mild conditions. This indicates that the conditions in Theorem 3 are much weaker than those in Theorem 2 by imposing the full-rankness requirement only on a certain Khatri-Rao product of the $\mathbf{\Lambda}_j$-matrices, instead of on individual $\mathbf{\Lambda}_j$s. To be more concrete, the next Example 3 illustrates Theorem 3, as a counterpart of Example 1.

Example 3 *Consider the following grouping matrix* $\mathbf{L}$ *with* $G = 3$ *and* $p = 6G = 18$*,*

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 \\ \mathbf{L}_1 \\ \mathbf{L}_1 \end{pmatrix}, \quad \textit{where} \quad \mathbf{L}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{12}$$

Then $\mathbf{L}$ *contains six copies of the identity matrix* $\mathbf{I}_G$ *after a row permutation. Thanks to greater variable grouping compared to the previous Example 1, we can use Theorem 3 (instead of Theorem 2) to establish identifiability. Specifically, consider binary responses with* $d_1 = \cdots = d_{18} =: d = 2$ *and* $K = 3$ *extreme latent profiles. For* $g = 1$*, define sets* $\mathcal{A}_{g,1} = \{1, 2\}$*,* $\mathcal{A}_{g,2} = \{7, 8\}$*,* $\mathcal{A}_{g,3} = \{13, 14\}$*; for* $g = 2$*, define sets* $\mathcal{A}_{g,1} = \{3, 4\}$*,* $\mathcal{A}_{g,2} = \{5, 6\}$*,* $\mathcal{A}_{g,3} = \{7, 8\}$*; and for* $g = 3$*, define sets* $\mathcal{A}_{g,1} = \{5, 6\}$*,* $\mathcal{A}_{g,2} = \{11, 12\}$*,* $\mathcal{A}_{g,3} = \{17, 18\}$*. Then for each* $(g, m) \in \{1, \ldots, G\} \times \{1, 2, 3\}$*, the* $\widetilde{\mathbf{\Lambda}}_{g,m} = \bigodot_{j \in \mathcal{A}_{g,m}} \mathbf{\Lambda}_j$ *defined in Theorem 3 has size* $d^2 \times K$ *which is* 4×3*, similar to the structure in Example 2. Now from the derivation and discussion in Example 2, we know such a* $\widetilde{\mathbf{\Lambda}}_{g,m}$ *has full rank for almost all the valid parameters in the parameter space. So the conditions in Theorem 3 are easily satisfied, and for almost all the valid parameters of such a Gro-M^3, the identifiability conclusion follows.*

3.2 Generic Identifiability Conditions

Example 2 shows that the Khatri-Rao product of conditional probability tables easily has full column rank in a toy case, and Example 3 leverages this observation to establish identifiability for almost all parameters in the parameter space using Theorem 3. We next generalize this observation to derive more practical identifiability conditions, under the *generic identifiability* notion introduced by Allman et al. (2009). Generic identifiability generally means that the unidentifiable parameters belong to a set of Lebesgue measure

zero with respect to the parameter space. Its definition adapted to the current Gro-M^{3}s is given as follows.

Definition 4 (Generic Identifiability of Gro-M^{3}s) *Under a Gro-M^3, a parameter space $\mathcal{T}$ for $(\mathbf{\Lambda}, \mathbf{\Phi})$ is said to be generically identifiable, if there exists a subset $\mathcal{N} \subseteq \mathcal{T}$ that has Lebesgue measure zero with respect to $\mathcal{T}$ such that for any $(\mathbf{\Lambda}, \mathbf{\Phi}) \in \mathcal{T} \setminus \mathcal{N}$ and an associated $\mathbf{L}$ matrix, the following holds if and only if $(\mathbf{L}, \mathbf{\Lambda}, \mathbf{\Phi})$ and the alternative $(\overline{\mathbf{L}}, \overline{\mathbf{\Lambda}}, \overline{\mathbf{\Phi}})$ are identical up to permutations of the K extreme latent profiles and that of the G variable groups,*

$$\mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \mathbf{L}, \mathbf{\Lambda}, \mathbf{\Phi}) = \mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \overline{\mathbf{L}}, \overline{\mathbf{\Lambda}}, \overline{\mathbf{\Phi}}), \quad \forall \boldsymbol{c} \in \times_{j=1}^{p}[d_j].$$

Compared to the strict identifiability notion in Definition 1, the generic identifiability notion in Definition 4 is less stringent in allowing the existence of a measure zero set of parameters where identifiability does not hold; see the previous Example 2 for an instance of a measure-zero set. Such an identifiability notion usually suffices for real data analyses (Allman et al., 2009). In the following Theorem 5, we propose simple conditions to ensure generic identifiability of Gro-M^{3}s.

Theorem 5 *For the notation $\mathcal{A}_g = \{j \in [p] : \ell_{j,g} = 1\}$ defined in Theorem 3, suppose each $\mathcal{A}_g$ can be partitioned into three non-overlapping sets $\mathcal{A}_g = \cup_{m=1}^{3} \mathcal{A}_{g,m}$, such that for each g and m the following holds,*

$$\prod_{j \in \mathcal{A}_{g,m}} d_j \geq K. \tag{13}$$

Then the matrix $\bigodot_{j \in \mathcal{A}_{g,m}} \mathbf{\Lambda}_j$ has full column rank K for generic parameters. Further, the $\mathbf{\Lambda}$, $\mathbf{L}$, and $\mathbf{\Phi}$ are generically identifiable.

Compared to Theorem 3, Theorem 5 lifts the explicit full-rank requirements on *any* matrix. Rather, Theorem 5 only requires that certain products of d_j's should not be smaller than the number of extreme latent profiles, which in turn guarantees that the Khatri-Rao products of matrices have full column rank for generic parameters. Intuitively, the more variables belonging to a group and the more categories each variable has, the easier the identifiability conditions are to satisfy. This illustrates the benefit of dimension-grouping to model identifiability.

4. Dirichlet Gro-M^3 and Bayesian Inference

4.1 Dirichlet model and identifiability

The previous section studies identifiability of general Gro-M^{3}s, not restricting the distribution $D_{\boldsymbol{\alpha}}(\cdot)$ of the mixed membership scores to be a specific form. Next we focus on an interesting special case where $D_{\boldsymbol{\alpha}}(\cdot)$ is a Dirichlet distribution with unknown parameters $\boldsymbol{\alpha}$. Among all the possible distributions for the individual mixed-membership proportions, the Dirichlet distribution is the most popular. It is widely used in applications including social science survey data (Erosheva et al., 2007; Wang et al., 2015), topic modeling (Blei et al.,

2003; Griffiths and Steyvers, 2004), and data privacy (Manrique-Vallier and Reiter, 2012). We term the Gro-M^3 with $\boldsymbol{\pi}_i$ following a Dirichlet distribution the Dirichlet Gro-M^3, and propose a Bayesian inference procedure to estimate both the discrete variable groupings and the continuous parameters. Such a Dirichlet Gro-M^3 has the graphical model representation in Figure 1(d).

For an unknown vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k > 0$ for all $k \in [K]$, suppose

$$\text{Dirichlet Gro-M}^3\text{:} \qquad \boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,K}) \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\alpha_1, \ldots, \alpha_K). \tag{14}$$

The vector $\boldsymbol{\alpha}$ characterizes the distribution of membership scores. As $\alpha_k \to 0$, the model simplifies to a latent class model in which each individual belongs to a single latent class. For larger α_k's, there will tend to be multiple elements of $\boldsymbol{\pi}_i$ that are not close to 0 or 1.

Recall that the previous identifiability conclusions in Theorems 2–5 generally apply to $\mathbf{L}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Phi}$, where $\boldsymbol{\Phi}$ is the *core tensor* with K^G entries in our hybrid tensor decomposition. Now with the core tensor $\boldsymbol{\Phi}$ parameterized by the Dirichlet distribution in particular, we can further investigate the identifiability of the Dirichlet parameters $\boldsymbol{\alpha}$. The following proposition establishes the identifiability of $\boldsymbol{\alpha}$ for Dirichlet Gro-M^{3}s.

Proposition 6 *Consider a Dirichlet Gro-M^3. If $G \geq 2$, then following conclusions hold.*

(a) If the conditions in Theorem 2 or Theorem 3 are satisfied, then the Dirichlet parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are strictly identifiable.

(b) If the conditions in Theorem 5 are satisfied, then the Dirichlet parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are generically identifiable.

Remark 7 *Our identifiability results have implications for the collapsed Tucker (c-Tucker) decomposition for multivariate categorical data (Johndrow et al., 2017). Our assumption that the latent memberships underlying several variables are in one state is similar to that in c-Tucker. However, c-Tucker does not model mixed memberships, and the c-Tucker tensor core, $\boldsymbol{\Phi}$ in our notation, is assumed to arise from a CP decomposition (Goodman, 1974) with $\phi_{k_1,\ldots,k_G} = \sum_{v=1}^{r} w_v \prod_{g=1}^{G} \psi_{g,k_g,v}$. We can invoke the uniqueness of the CP decomposition (e.g., Kruskal, 1977; Allman et al., 2009) to obtain identifiability of parameters $\boldsymbol{w} = (w_v;\, v \in [r])$ and $\boldsymbol{\psi} = (\psi_{g,k,v};\, g \in [G],\, k \in [K], v \in [r])$. Hence, under our assumptions on the variable grouping structure in Section 3, imposing existing mild conditions on $\boldsymbol{w}$ and $\boldsymbol{\psi}$ will yield identifiability of all the c-Tucker parameters.*

4.2 Bayesian inference

Considering the complexity of our latent structure model, we adopt a Bayesian approach. We next describe the prior specification for $\mathbf{L}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\alpha}$ in Dirichlet Gro-M^{3}s. The number of variable groups G and number of extreme latent profiles K are assumed known; we relax this assumption in Section 5. Recall the indicators $s_1, \ldots, s_p \in [G]$ are defined as $s_j = g$ if and only if $\ell_{j,g} = 1$, so there is a one-to-one correspondence between the matrix $\mathbf{L}$ and the vector $\boldsymbol{s} = (s_1, \ldots, s_p)$. We adopt the following prior for the s_j's,

$$s_1, \ldots, s_p \overset{\text{i.i.d.}}{\sim} \text{Categorical}([G],\, \xi_1, \ldots, \xi_G),$$

where Categorical($[G]$, $\xi_1, \ldots, \xi_G$) is a categorical distribution over G categories with proportions $\xi_g \geq 0$ and $\sum_{g=1}^{G} \xi_g = 1$. We choose uniform priors over the probability simplex for $(\xi_1, \ldots, \xi_G)$ and each column of $\mathbf{\Lambda}_j$. We remark that if certain prior knowledge about the variable groups is available for the data, then it is also possible to employ informative priors such as those in Paganin et al. (2021) for the s_j's. For the Dirichlet parameters $\boldsymbol{\alpha}$, defining $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ and $\boldsymbol{\eta} = (\alpha_1/\alpha_0, \ldots, \alpha_K/\alpha_0)$, we choose the hyperpriors $\alpha_0 \sim \text{Gamma}(a_\alpha, b_\alpha)$ and $\boldsymbol{\eta}$ is uniform over the $(K-1)$-probability simplex.

Given a sample of size n, denote the observed data by $\mathbf{Y} = \{\boldsymbol{y}_i;\, i = 1, \ldots, n\}$. We propose a Metropolis-Hastings-within-Gibbs sampler and also a Gibbs sampler for posterior inference of $\mathbf{L}$, $\mathbf{\Lambda}$, and $\boldsymbol{\alpha}$ based on the data $\mathbf{Y}$.

Metropolis-Hastings-within-Gibbs Sampler. This sampler cycles through the following steps.

Step 1–3. Sample each column of the conditional probability tables $\mathbf{\Lambda}_j$'s, the individual mixed-membership proportions $\boldsymbol{\pi}_i$'s, and the individual latent assignments $z_{i,g}$'s from their full conditional posterior distributions. Define indicator variables $y_{i,j,c} = \mathbb{I}(y_{i,j} = c)$ and $z_{i,g,k} = \mathbb{I}(z_{i,g} = k)$. These posteriors are

$$\begin{aligned}
\{\boldsymbol{\lambda}_{j,:,k} \mid -\}_{s_j = g} &\sim \text{Dirichlet}\left(1 + \sum_{i=1}^{n} z_{i,g,k} y_{i,j,1},\ \ldots,\ 1 + \sum_{i=1}^{n} z_{i,g,k} y_{i,j,d_j}\right); \\
\boldsymbol{\pi}_i \mid - &\sim \text{Dirichlet}\left(\alpha_1 + \sum_{g=1}^{G} z_{i,g,1},\ \ldots,\ \alpha_K + \sum_{g=1}^{G} z_{i,g,K}\right); \\
\mathbb{P}(z_{i,g} = k \mid -) &= \frac{\pi_{i,k} \prod_{j:\, s_j = g} \prod_{c=1}^{d_j} \lambda_{j,c,k}^{y_{i,j,c}}}{\sum_{k'=1}^{K} \pi_{i,k'} \prod_{j:\, s_j = g} \prod_{c=1}^{d_j} \lambda_{j,c,k'}^{y_{i,j,c}}}, \quad k \in [K].
\end{aligned}$$

Step 4. Sample the variable grouping structure $(s_1, \ldots, s_p)$. The posterior of each s_j is

$$\mathbb{P}(s_j = g \mid -) = \frac{\xi_g \prod_{i=1}^{n} \lambda_{j,y_{i,j},z_{i,g}}}{\sum_{g'=1}^{G} \xi_{g'} \prod_{i=1}^{n} \lambda_{j,y_{i,j},z_{i,g'}}}, \quad g \in [G].$$

The posterior of $(\xi_1, \ldots, \xi_G)$ is

$$(\xi_1, \ldots, \xi_G) \mid - \sim \text{Dirichlet}\left(1 + \sum_{j=1}^{p} \mathbb{I}(s_j = 1), \ldots, 1 + \sum_{j=1}^{p} \mathbb{I}(s_j = G)\right).$$

Step 5. Sample the Dirichlet parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ via Metropolis-Hastings sampling. The conditional posterior distribution of $\boldsymbol{\alpha}$ (or equivalently, α_0 and $\boldsymbol{\eta}$) is

$$p(\boldsymbol{\alpha} \mid -) \propto \text{Gamma}(\alpha_0 \mid a, b) \times \text{Dirichlet}(\boldsymbol{\eta} \mid \mathbf{1}_K) \times \prod_{i=1}^{n} \text{Dirichlet}(\boldsymbol{\pi}_i \mid \boldsymbol{\alpha})$$

$$\propto \alpha_0^{a_\alpha - 1} \exp(-b_\alpha \alpha_0) \times \left[\frac{\Gamma(\alpha_0)}{\prod_{k=1}^K \Gamma(\alpha_k)}\right]^n \times \prod_{k=1}^K \left[\prod_{i=1}^n \pi_{i,k}\right]^{\alpha_k},$$

which is not an easy-to-sample-from distribution. We use a Metropolis-Hastings sampling strategy in Manrique-Vallier and Reiter (2012). The steps are detailed as follows.

- Sample each entry of $\boldsymbol{\alpha}^\star = (\alpha_1^\star, \ldots, \alpha_K^\star)$ from independent lognormal distributions (proposal distribution $g(\boldsymbol{\alpha}^\star \mid \boldsymbol{\alpha})$) as

$$\alpha_k^\star \overset{\text{ind.}}{\sim} \text{lognormal}(\log \alpha_k, \sigma_\alpha^2), \tag{15}$$

  where σ_α is a tuning parameter that affects the acceptance ratio of the draw. Based on our preliminary simulations, σ should be relatively small to avoid the acceptance ratio to be always too close to zero.

- Let $\alpha_0^\star = \sum_{k=1}^K \alpha_k^\star$. Define

$$\begin{aligned} r^\star &= \frac{p(\boldsymbol{\alpha}^\star \mid -) g(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star)}{p(\boldsymbol{\alpha} \mid -) g(\boldsymbol{\alpha}^\star \mid \boldsymbol{\alpha})} \\ &= \left(\frac{\alpha_0^\star}{\alpha_0}\right)^{a_\alpha - 1} \exp\left(-b_\alpha(\alpha_0^\star - \alpha_0)\right) \times \left(\frac{\Gamma(\alpha_0^*)}{\Gamma(\alpha_0)} \cdot \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k^\star)}\right)^n \\ &\quad \times \prod_{k=1}^K \left(\prod_{i=1}^n \pi_{i,k}\right)^{\alpha_k^\star - \alpha_k} \times \prod_{k=1}^K \frac{\alpha_k^\star}{\alpha_k} \end{aligned}$$

  The Metropolis-Hastings acceptance ratio of the proposed $\boldsymbol{\alpha}^\star$ is $r = \min\{1,\ r^\star\}$.

We track the acceptance ratio in the Metropolis-Hastings step along the MCMC iterations in a simulation study. Figure 3 shows the boxplots of the average acceptance ratios for various sample sizes in the same simulation as the later Table 3. This figure shows that the Metropolis-Hastings acceptance ratio is generally high and mostly exceeds 80%.

Gibbs Sampler. We also develop a fully Gibbs sampling algorithm for our Gro-M^3, leveraging the auxiliary variable method in Zhou (2018) to sample the Dirichlet parameters $\boldsymbol{\alpha}$. Especially, since we have proved in Proposition 6 that the entire Dirichlet parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is identifiable from the observed data distribution, we choose to freely and separately sample all the entries $\alpha_1, \ldots, \alpha_K$ instead of constraining these K entries to be equal as in Zhou (2018). Recall that for each subject i, $z_{i,g} \in [K]$ for $g \in [G]$ denotes the latent profile realization for the gth group of items. Introduce new notation $Z_{ik}^{\text{mult}} = \sum_{g=1}^G \mathbb{1}(z_{i,g} = k)$ for $i \in [N]$ and $k \in [K]$. Then $(Z_{i1}^{\text{mult}}, \ldots, Z_{i1}^{\text{mult}})$ follows the Dirichlet-Multinomial distribution with parameters G and $(\alpha_1, \ldots, \alpha_K)$. We introduce auxiliary Beta variables q_i for $i \in [N]$ and auxiliary Chinese Restaurant Table (CRT) variables t_{ik} for $i \in [N]$ and $k \in [K]$. Endowing the Dirichlet parameter α_k with the prior $\alpha_k \sim \text{Gamma}(a_0,\ b_0)$, we have the following Gibbs updates for sampling α_k.

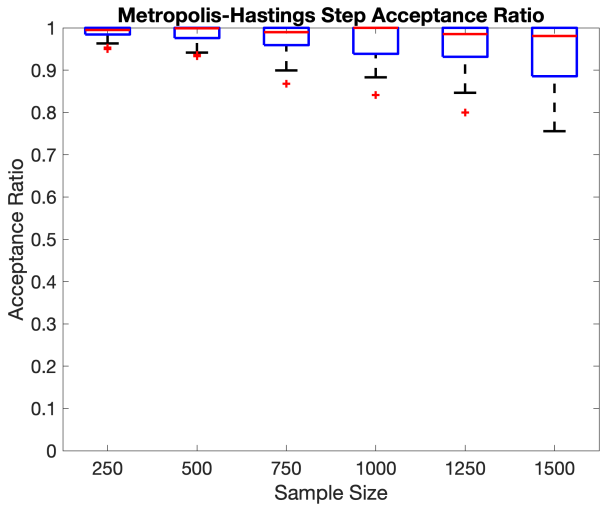


Figure 3: Metropolis-Hastings average acceptance ratio in the simulation setting $(p, G, K) = (30, 6, 4)$, corresponding to the first setting in Table 3 in the manuscript.

Step 5* Sample the auxiliary variables q_i, t_{ik} and the Dirichlet parameters α_k from the following full conditional posteriors:

$$q_i \sim \text{Beta}\left(\sum_{k=1}^{K} Z_{ik}^{\text{mult}},\ \sum_{k=1}^{K} \alpha_k\right), \qquad i \in [n];$$

$$t_{ik} \sim \text{CRT}(Z_{ik}^{\text{mult}},\ \alpha_k), \qquad i \in [n],\ k \in [K];$$

$$\alpha_k \sim \text{Gamma}\left(a_0 + \sum_{i=1}^{n} t_{ik},\ b_0 - \sum_{i=1}^{n} \log(1 - q_i)\right), \qquad k \in [K].$$

Replacing the previous Step 5 in the Metropolis-within-Gibbs sampler with the above Step 5* gives a fully Gibbs sampling algorithm for Gro-M^3.

Our simulations reveal the following empirical comparisons between the Gibbs sampler and the Metropolis-Hastings-within-Gibbs (MH-within-Gibbs) sampler. In terms of Markov chain mixing, the Gibbs sampler mixes faster than the MH-within-Gibbs sampler as expected, and requires fewer MCMC iterations to generate quality posterior samples *if initialized well*. However, in terms of estimation accuracy, we observe that the MH-within-Gibbs sampler tends to have better accuracy in estimating the identifiable model parameters. This is likely because that the MH-within-Gibbs sampler performs better on exploring the entire posterior space through the proposal distributions; whereas the Gibbs sampler tends to be more heavily influenced by the initial value of the parameters and can converge to suboptimal distributions if not initialized well. We next provide the experimental evidence behind the above observations.

Figure 4 provides typical traceplots for the MH-within-Gibbs sampler (left) and the Gibbs sampler (middle and right) in one simulation trial in the same setting as the later Table 3. The four horizontal lines in each panel denote the true parameter values $\boldsymbol{\alpha} =$

$(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.4, 0.5, 0.6, 0.7)$. The left and middle panels of Figure 4 are traceplots of α_k in MCMC chains initialized randomly with the same initial value, whereas the right panel corresponds to a chain initialized with the true parameter value $\boldsymbol{\alpha}$. Figure 4 shows that when initialized randomly with the same value, the MH-within-Gibbs chain converges to distributions much closer to the truth than the Gibbs sampler; in contrast, the Gibbs chain only manages to converge to the desirable posteriors when initialized with the true $\boldsymbol{\alpha}$. Furthermore, Figure 5 plots the root mean squared error quantitles (25%, 50%, 75%) of $\boldsymbol{\alpha}$ estimated using the two samplers from the 50 simulation replicates in each setting. The parameter initialization in each replicate for the two samplers is random and identical. Figure 5 clearly shows that the MH-within-Gibbs sampler has lower estimation error for $\boldsymbol{\alpha}$. In summary, when *initialized randomly using the same mechanism*, the MH-within-Gibbs sampler has higher parameter estimation accuracy despite that the Gibbs sampler mixes faster. Therefore, we choose to present the estimation results of the MH-within-Gibbs sampler in the later Section 5.

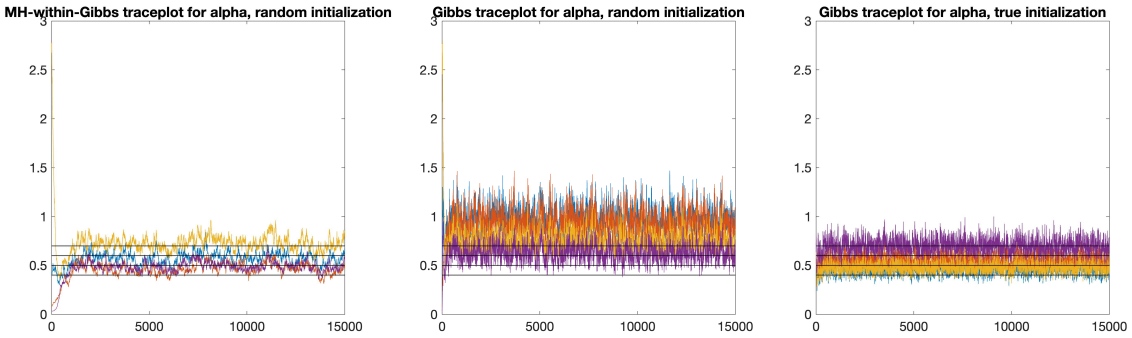


Figure 4: Traceplots of the MH-within-Gibbs sampler (left) and the Gibbs sampler (middle and right) applied to one simulated dataset with $(n, p, G, K) = (500, 30, 6, 4)$. The horizontal lines in each panel denote the true $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.4, 0.5, 0.6, 0.7)$. The left and middle panels correspond to chains initialized randomly with the same initial value, whereas the right panel corresponds to a chain initialized with the true parameter value $\boldsymbol{\alpha}$.

After collecting posterior samples from the output of the MCMC algorithm, for those continuous parameters in the model we can calculate their posterior means as point estimates. As for the discrete variable grouping structure, we can obtain the posterior modes of each s_j. That is, given the T posterior samples of $\boldsymbol{s}^{(t)} = (s_1^{(t)}, \ldots, s_p^{(t)})$ for $t = 1, \ldots, T$, we define point estimates $\overline{\boldsymbol{s}}$ and $\overline{\mathbf{L}}$ with entries

$$\overline{s}_j = \arg\max_{g \in [G]} \sum_{t=1}^{T} \mathbb{I}(s_j^{(t)} = g); \qquad \overline{\ell}_{j,g} = \begin{cases} 1, & \text{if } \overline{s}_j = g; \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

5. Simulation Studies

In this section, we carry out simulation studies to assess the performance of the proposed Bayesian estimation approach, while verifying that identifiable parameters are indeed estimated more accurately as sample size grows. In Section 5.1, we perform a simulation study

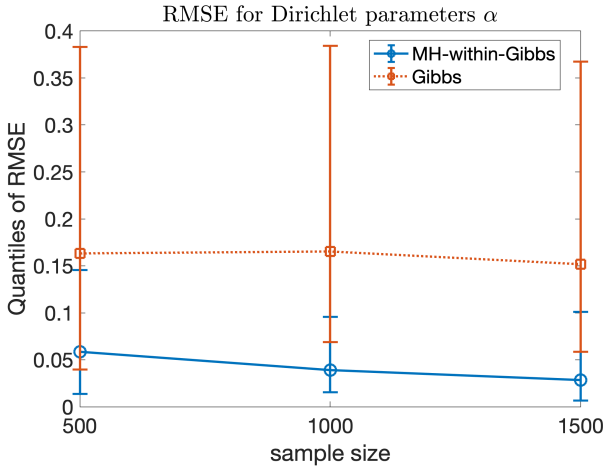


Figure 5: Root mean squared errors (RMSE) quantiles (25%, 50%, 75%) for the MH-within-Gibbs sampler and the Gibbs sampler obtained from 50 simulation replicates for each sample size. In each simulation replicate, the initializations of the Gibbs chain and the MH-within-Gibbs chain are identical.

to assess the estimation accuracy of the model parameters, assuming the number of extreme latent profiles K and the number of variable groups G are known. This is the same assumption as in many existing estimation methods of traditional MMMs (e.g., Manrique-Vallier and Reiter, 2012). In Section 5.2, to facilitate the use of our estimation method in applications, we propose data-driven criteria to select K and G and perform a corresponding simulation study.

5.1 Estimation of Grouping Structure and Model Parameters

In this simulation study, we assess the proposed algorithm's performance in estimating the $(\mathbf{L}, \mathbf{\Lambda}, \boldsymbol{\alpha})$ in Dirichlet Gro-M^{3}s. We consider various simulation settings, with $K = 2, 3$, or 4, and $(p, G) = (30, 6)$, $(60, 12)$, or $(90, 15)$. The number of categories of each y_j is specified to be three, i.e., $d_1 = \cdots = d_p = 3$. The true $\mathbf{\Lambda}$-parameters are specified as follows: in the most challenging case with $K = 4$ and $(p, G) = (90, 15)$, for $u = 0, 1, \ldots, p/6 - 1$ we specify

$$\mathbf{\Lambda}_{6u+1} = \begin{pmatrix} 0.1 & 0.7 & 0.3 & 0.1 \\ 0.8 & 0.2 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.3 & 0.8 \end{pmatrix}; \ \mathbf{\Lambda}_{6u+2} = \begin{pmatrix} 0.1 & 0.8 & 0.1 & 0.2 \\ 0.2 & 0.1 & 0.6 & 0.5 \\ 0.7 & 0.1 & 0.3 & 0.3 \end{pmatrix}; \ \mathbf{\Lambda}_{6u+3} = \begin{pmatrix} 0.1 & 0.8 & 0.2 & 0.9 \\ 0.2 & 0.1 & 0.5 & 0.05 \\ 0.7 & 0.1 & 0.3 & 0.05 \end{pmatrix};$$

$$\mathbf{\Lambda}_{6u+4} = \begin{pmatrix} 0.1 & 0.1 & 0.8 & 0.3 \\ 0.8 & 0.2 & 0.1 & 0.6 \\ 0.1 & 0.7 & 0.1 & 0.1 \end{pmatrix}; \ \mathbf{\Lambda}_{6u+5} = \begin{pmatrix} 0.2 & 0.7 & 0.3 & 0.1 \\ 0.6 & 0.2 & 0.4 & 0.1 \\ 0.2 & 0.1 & 0.3 & 0.8 \end{pmatrix}; \ \mathbf{\Lambda}_{6u+6} = \begin{pmatrix} 0.1 & 0.8 & 0.1 & 0.2 \\ 0.2 & 0.1 & 0.1 & 0.6 \\ 0.7 & 0.1 & 0.8 & 0.2 \end{pmatrix}.$$

As for other simulation settings with smaller K and (p, G), we specify the $\mathbf{\Lambda}_j$'s by taking a subset of the above matrices and retaining a subset of columns from each of these matrices. The true Dirichlet parameters $\boldsymbol{\alpha}$ are set to $(0.4, 0.5)$ for $K = 2$, $(0.4, 0.5, 0.6)$ for $K = 3$, and $(0.4, 0.5, 0.6, 0.7)$ for $K = 4$. The true grouping matrix $\mathbf{L}$ of size $p \times G$ is specified to

containing p/G copies of identity submatrices $\mathbf{I}_G$ up to a row permutation. Under these specifications, our identifiability conditions in Theorem 3 are satisfied. We consider sample sizes $n = 250, 500, 1000, 1500$. In each scenario, 50 independent datasets are generated and fitted with the proposed MCMC algorithm described in Section 4. In our MCMC algorithm under all simulation settings, we take hyperparameters to be $(a_\alpha, b_\alpha) = (2, 1)$ and $\sigma_\alpha = 0.02$. The MCMC sampler is run for 15000 iterations, with the first 10000 iterations as burn-in and every fifth sample is collected after burn-in to thin the chain.

We observed good mixing and convergence behaviors of the model parameters from examining the trace plots. In particular, simulations show that the estimation of the discrete variable grouping structure in matrix $\mathbf{L}$ (equivalently, vector $\boldsymbol{s}$) is quite accurate in general, and the posterior means of the continuous $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}$ are also close to their truth. Next we first present details of two typical simulation trials as an illustration, before presenting summaries across the independent simulation replicates.

Two random simulation trials were taken from the settings $(n, p, G, K) = (500, 30, 6, 2)$ and $(n, p, G, K) = (500, 90, 15, 2)$. All the parameters were randomly initialized from their prior distributions. In Figure 6, the left three plots in each of the first two rows show the sampled $\mathbf{L}_{\text{iter.}}$ in the MCMC algorithm, after the 1st, 201st, and 401st iterations, respectively; the fourth plot show the posterior mode $\overline{\mathbf{L}}$ defined in (16), and the last plot shows the simulation truth $\mathbf{L}$. If an $\widetilde{\mathbf{L}}$ equals the true $\mathbf{L}$ after a column permutation then it indicates $\widetilde{\mathbf{L}}$ and $\mathbf{L}$ induce identical variable groupings. The bottom two plots in Figure 6 show the Adjusted Rand Index (ARI, Rand, 1971) of the variable groupings of $\mathbf{L}_{\text{iter.}}$ ($\boldsymbol{s}_{\text{iter.}}$) with respect to the true $\mathbf{L}$ (true $\boldsymbol{s}$) along the first 1000 MCMC iterations. The ARI measures the similarity between two clusterings, and it is appropriate to compare a true $\boldsymbol{s}$ and an estimated $\overline{\boldsymbol{s}}$ because they each summarizes a clustering of the p variables into G groups. The ARI is at most 1, with ARI $= 1$ indicating perfect agreement between two clusterings. The bottom row of Figure 6 shows that in each simulation trial, the ARI measure starts with values around 0 due to the random MCMC initialization, and within a few hundred iterations the ARI increases to a distribution over much larger values. For the simulation with $(n, p, G, K) = (500, 90, 15, 2)$, the posterior mode of $\mathbf{L}$ exactly equals the truth, and the corresponding plot on the bottom right of Figure 6 shows the ARI is distributed very close to 1 after just about 500 MCMC iterations. In general, our MCMC algorithm has excellent performance in inferring the $\mathbf{L}$ from randomly initialized simulations; also see the later Tables 1–3 for more details.

We next present estimation accuracy results of both $\mathbf{L}$ and $(\boldsymbol{\Lambda}, \boldsymbol{\alpha})$ summarized across 50 simulation replicates in each setting. For continuous parameters $(\boldsymbol{\Lambda}, \boldsymbol{\alpha})$, we calculate their Root Mean Squared Errors (RMSEs) to evaluate the estimation accuracy. To obtain the estimation error of $(\boldsymbol{\Lambda}, \boldsymbol{\alpha})$ after collecting posterior samples, we need to find an appropriate permutation of the K extreme latent profiles in order to compare the $(\overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\alpha}})$ and the true $(\boldsymbol{\Lambda}, \boldsymbol{\alpha})$. To this end, we first reshape each of $\overline{\boldsymbol{\Lambda}}$ and $\boldsymbol{\Lambda}$ to a $(\sum_{j=1}^{p} d_j) \times K$ matrix $\overline{\boldsymbol{\Lambda}}_{\text{mat}}$ and $\boldsymbol{\Lambda}_{\text{mat}}$, calculate the inner product matrix $(\boldsymbol{\Lambda}_{\text{mat}})^\top \overline{\boldsymbol{\Lambda}}_{\text{mat}}$, and then find the index i_k of the largest entry in each kth row of the inner product matrix. Such a vector of indices $(i_1, \ldots, i_K)$ gives a permutation of the K profiles, and we will compare $\overline{\boldsymbol{\Lambda}}_{j,:,(i_1,\ldots,i_K)}$ to $\boldsymbol{\Lambda}_j$ and compare $\overline{\boldsymbol{\alpha}}_{(i_1,\ldots,i_K)}$ to $\boldsymbol{\alpha}$. In Tables 1–3, we present the RMSEs of $(\boldsymbol{\Lambda}, \boldsymbol{\alpha})$ and the ARIs of $\mathbf{L}$ under the aforementioned 36 different simulation settings. The median and interquartile range of the ARIs or RMSEs across the simulation replicates are shown in these tables.

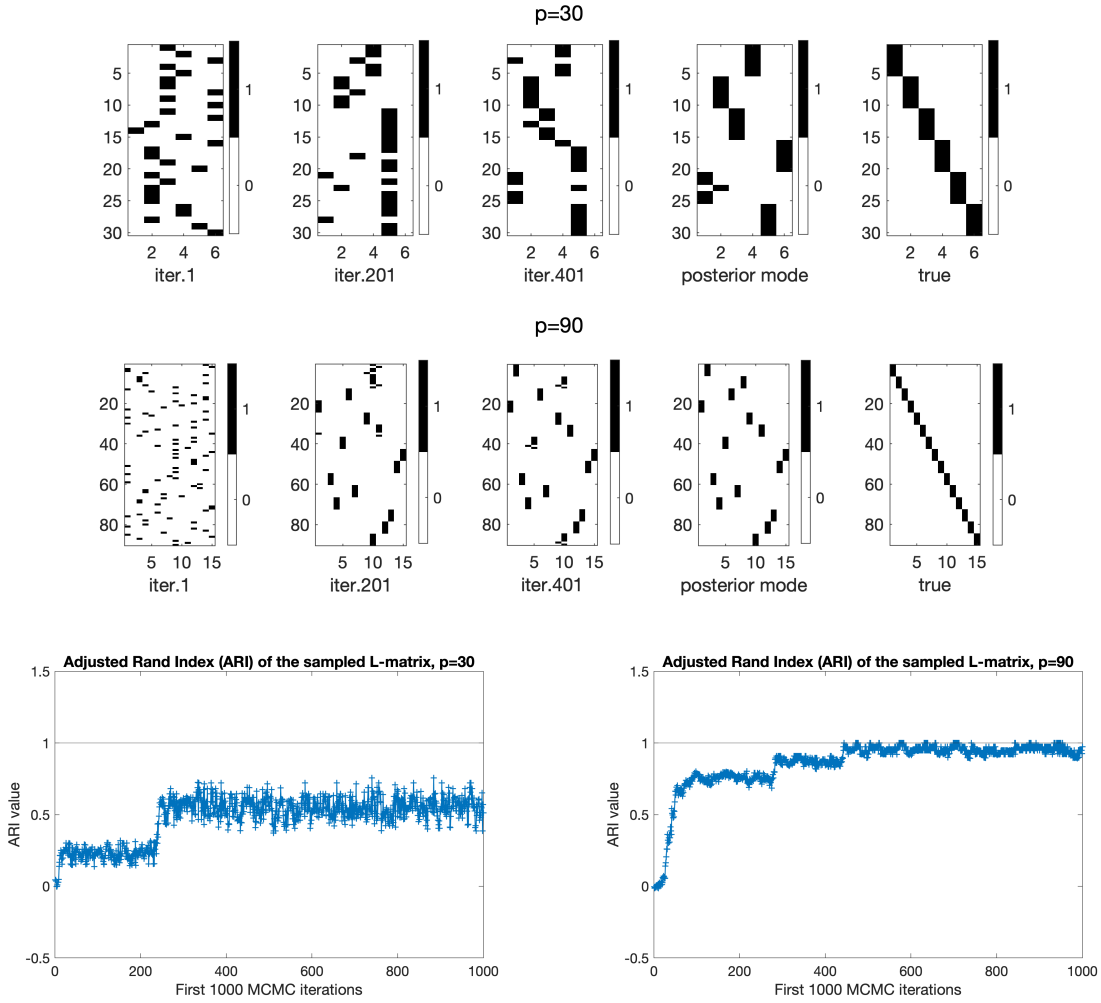


Figure 6: Estimation of $\mathbf{L}$ (from $\boldsymbol{s}$) in two random simulation trials, one under $(n, p, G, K) = (500, 30, 6, 2)$ and the other under $(n, p, G, K) = (500, 90, 15, 2)$. In each of the first two rows, the left three plots record the sampled $\mathbf{L}_{\text{iter.}}$ after the 1st, 201st, and 401st MCMC iteration, respectively. The fourth plot shows the posterior mode $\overline{\mathbf{L}}$ and the last shows the true $\mathbf{L}$. The two plots in the bottom row record the ARI of the clustering of p variables given by $\mathbf{L}_{\text{iter.}}$ along the first 1000 MCMC iterations, for each of the two simulation scenarios.

Tables 1–3 show that under each setting of true parameters with a fixed (p, G, K), the ARIs of the variable grouping $\mathbf{L}$ generally increase as sample size n increases, and the RMSEs of $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}$ decreases as n increases. This shows the increased estimation accuracy with an increased sample size. In particular, the estimation accuracy of the variable grouping structure is quite high across the considered settings. The estimation errors are slightly larger for larger values of K in Table 3 compared to smaller values of K in Tables 1 and 2. Overall, the simulation results empirically confirm the identifiability and estimability of the model parameters in our Dirichlet Gro-M^3.

	$\{p, G\}$	n	ARI of $\mathbf{L}$		RMSE of $\mathbf{\Lambda}$		RMSE of $\boldsymbol{\alpha}$	
			Median	(IQR)	Median	(IQR)	Median	(IQR)
$K = 2$	(30, 6)	250	0.74	(0.18)	0.042	(0.005)	0.064	(0.056)
		500	0.88	(0.17)	0.030	(0.004)	0.031	(0.043)
		1000	0.91	(0.29)	0.023	(0.014)	0.027	(0.028)
		1500	0.91	(0.31)	0.018	(0.022)	0.026	(0.045)
	(60, 12)	250	0.73	(0.13)	0.042	(0.004)	0.039	(0.041)
		500	0.79	(0.14)	0.032	(0.003)	0.031	(0.021)
		1000	0.85	(0.20)	0.027	(0.010)	0.018	(0.029)
		1500	0.81	(0.21)	0.028	(0.016)	0.024	(0.025)
	(90, 15)	250	0.95	(0.05)	0.042	(0.003)	0.045	(0.045)
		500	1.00	(0.00)	0.026	(0.002)	0.032	(0.023)
		1000	1.00	(0.00)	0.018	(0.001)	0.019	(0.021)
		1500	1.00	(0.08)	0.015	(0.010)	0.017	(0.017)

Table 1: Simulation results of the Dirichlet Gro-M^3 for $K = 2$. "ARI" of $\mathbf{L}$ is the Adjusted Rand Index of the estimated variable groupings with respect to the truth. "RMSE" of $\mathbf{\Lambda}$ and $\boldsymbol{\alpha}$ are Root Mean Squared Errors. "Median" and "IQR" are based on 50 replicates in each simulation setting.

	(p, G)	n	ARI of $\mathbf{L}$		RMSE of $\mathbf{\Lambda}$		RMSE of $\boldsymbol{\alpha}$	
			Median	(IQR)	Median	(IQR)	Median	(IQR)
$K = 3$	(30, 6)	250	1.00	(0.00)	0.045	(0.004)	0.046	(0.048)
		500	1.00	(0.00)	0.033	(0.003)	0.046	(0.059)
		1000	1.00	(0.00)	0.023	(0.022)	0.039	(0.037)
		1500	1.00	(0.00)	0.019	(0.023)	0.029	(0.032)
	(60, 12)	250	1.00	(0.00)	0.045	(0.004)	0.044	(0.030)
		500	1.00	(0.00)	0.032	(0.002)	0.030	(0.018)
		1000	1.00	(0.00)	0.023	(0.002)	0.021	(0.017)
		1500	1.00	(0.00)	0.018	(0.002)	0.020	(0.017)
	(90, 15)	250	1.00	(0.00)	0.045	(0.002)	0.047	(0.036)
		500	1.00	(0.00)	0.031	(0.002)	0.026	(0.022)
		1000	1.00	(0.00)	0.022	(0.001)	0.021	(0.013)
		1500	1.00	(0.21)	0.019	(0.024)	0.024	(0.023)

Table 2: Simulation results of the Dirichlet Gro-M^3 for $K = 3$. See the caption of Table 1 for the meanings of columns.

Our MCMC algorithm can be viewed as a novel algorithm for Bayesian factorization of probability tensors. To see this, note that the observed response vector ranges in the p-way

	$(p,\, G)$	n	ARI of $\mathbf{L}$		RMSE of $\mathbf{\Lambda}$		RMSE of $\boldsymbol{\alpha}$	
			Median	(IQR)	Median	(IQR)	Median	(IQR)
$K = 4$	(30, 6)	250	1.00	(0.00)	0.064	(0.007)	0.078	(0.056)
		500	1.00	(0.00)	0.046	(0.006)	0.062	(0.072)
		1000	1.00	(0.00)	0.032	(0.004)	0.043	(0.046)
		1500	1.00	(0.00)	0.026	(0.004)	0.032	(0.036)
	(60, 12)	250	1.00	(0.00)	0.064	(0.005)	0.060	(0.031)
		500	1.00	(0.00)	0.043	(0.003)	0.047	(0.027)
		1000	1.00	(0.00)	0.031	(0.002)	0.032	(0.014)
		1500	1.00	(0.00)	0.025	(0.001)	0.023	(0.017)
	(90, 15)	250	1.00	(0.00)	0.046	(0.004)	0.053	(0.036)
		500	1.00	(0.00)	0.041	(0.003)	0.037	(0.022)
		1000	1.00	(0.00)	0.029	(0.001)	0.026	(0.027)
		1500	1.00	(0.00)	0.024	(0.001)	0.026	(0.020)

Table 3: Simulation results of the Dirichlet Gro-M^3 for $K = 4$. See the caption of Table 1 for the meanings of columns.

contingency table $\boldsymbol{y}_i \in [d_1] \times [d_2] \cdots \times [d_p]$, and the marginal probabilities of a random vector $\boldsymbol{y}_i$ falling each of the $\prod_{j=1}^{p} d_j$ cells therefore form a probability tensor with p modes. Our Gro-M^3 model provides a general and interpretable hybrid tensor factorization; it reduces to the nonnegative CP decomposition when the grouping matrix equals the $p \times 1$ one-vector and reduces to the nonnegative Tucker decomposition when the grouping matrix equals the $p \times p$ identity matrix. Specifically, our estimated Dirichlet parameters $\boldsymbol{\alpha}$ help define the tensor core and our estimated conditional probability parameters $\boldsymbol{\lambda}_{j,k}$ constitute the tensor arms. In this regard, we view our proposed MCMC algorithm as contributing a new tensor factorization method with nice uniqueness guarantee (i.e., identifiability guarantee) and good empirical performance.

We conduct a simulation study to empirically verify the theoretical identifiability results. Specifically, in the simulation setting $(p, G, K) = (30, 6, 4)$, corresponding to the first setting in Table 3, we now consider more sample sizes $n \in \{250,\ 500,\ 750,\ 1000,\ 1250,\ 1500\}$. For each sample size, we conducted 50 independent simulation replications and calculated the average root mean squared errors (RMSEs) of the model parameters $\mathbf{\Lambda}$ and $\boldsymbol{\alpha}$. Figure 7 plots the RMSEs versus the sample size n and shows that as n increases, the RMSEs decrease gradually. This trend provides an empirical verification of identifiability, and corroborates the conclusion that under an identifiable model, the model parameters can be estimated increasingly accurately as one collects more and more samples.

5.2 Selecting G and K from Data

In Section 3, model identifiability is established under the assumption that G and K are known, like many other latent structure models; for example, generic identifiability of latent

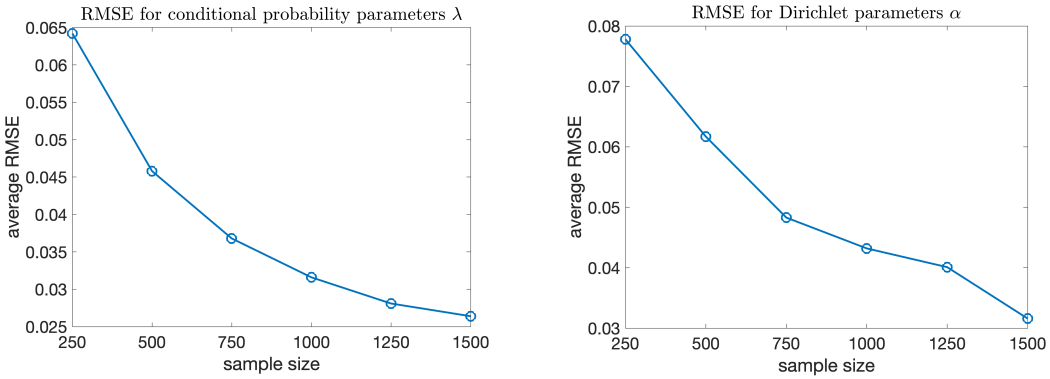


Figure 7: Empirical verification of identifiability. Root mean square errors (RMSEs) of model parameters averaged across simulation replicates decrease as sample size increases. The simulation setting is $(p, G, K) = (30, 6, 4)$, which is the first setting in Table 3.

class models in Allman et al. (2009) is established assuming the number of latent classes is known. But in order to provide a practical estimation pipeline applicable to real-world applications, we next briefly discuss how to select G and K in a data-driven way.

Our basic rationale is to use a practically useful criterion that favors a model with good out-of-sample predictive performance while remaining parsimonious. Gelman et al. (2014) contains a comprehensive review of various predictive information criteria for evaluating Bayesian models. We first considered using the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002), a traditional model selection criteria for Bayesian models. However, our preliminary simulations imply that DIC does not work well for selecting the latent dimensions in Gro-M^{3}s. In particular, we observed that DIC sometimes severely overselects the latent dimensions in our model, while that the WAIC (Widely Applicable Information Criterion, Watanabe, 2010) has better performance in our simulation studies (see the next paragraph for details). Our observation about DIC agrees with previous studies on the inconsistency of DIC in several different settings (Gelman et al., 2014; Hooten and Hobbs, 2015; Piironen and Vehtari, 2017).

Watanabe (2010) proved that WAIC is asymptotically equal to Bayesian leave-one-out cross validation and provided a solid theoretical justification for using WAIC to choose models with relatively good predictive ability. WAIC is particularly useful for models with hierarchical and mixture structures, making it well suited to selecting the latent profile dimension K and variable group dimension G in our proposed model. Denote the posterior samples by $\boldsymbol{\theta}^{(t)}$, $t = 1, \ldots, T$. For each $i \in [n]$ and $t \in [T]$, denote

$$p(\boldsymbol{y}_i \mid \boldsymbol{\theta}^{(t)}) = \prod_{m=1}^{G} \left[\sum_{k=1}^{K} \pi_{ik}^{(t)} \prod_{\ell_{j,m}^{(t)}=1} \prod_{c=1}^{d_j} \left(\lambda_{j,c,k}^{(t)} \right)^{y_{i,j,c}} \right].$$

In particular, Gelman et al. (2014) recommended using the following version of the WAIC, where "lppd" refers to *log pointwise predictive density* and p_{WAIC_2} measures the model

complexity through the variance,

$$\begin{aligned}\text{WAIC} &= -2\left(\text{lppd} - p_{\text{WAIC}_2}\right) \\ &= -2\sum_{i=1}^{n}\log\left(\frac{1}{T}\sum_{t=1}^{T}p(\boldsymbol{y}_i \mid \boldsymbol{\theta}^{(t)})\right) + 2\sum_{i=1}^{n}\text{var}_{t=1}^{T}\left(\log p\left(\boldsymbol{y}_i \mid \boldsymbol{\theta}^{(t)}\right)\right),\end{aligned} \tag{17}$$

where $\text{var}_{t=1}^{T}$ refers to the variance based on T posterior samples, with definition $\text{var}_{t=1}^{T}(a_t) = 1/(T-1)\sum_{t=1}^{T}\left(a_t - \sum_{t'=1}^{T}a_{t'}/T\right)^2$. Based on the above definition, the WAIC can be easily calculated based on posterior samples. The model with a smaller WAIC is favored.

We carried out a simulation study to evaluate how WAIC performs on selecting G and K, focusing on the previous setting where 50 independent datasets are generated from $(n, p, G, K) = (1000, 30, 6, 3)$. When fixing the candidate K to the truth $K = 3$ and varying the candidate $G_{\text{candi}} \in \{4, 5, 6, 7, 8\}$, the percentages of the datasets that each of $G = 4, 5, 6, 7, 8$ is selected are 0%, 0%, 74% (true G), 20%, 6%, respectively. When fixing the candidate G to the truth $G = 6$ and varying $K_{\text{candi}} \in \{2, 3, 4, 5, 6\}$, the percentages of the datasets that each of $K = 2, 3, 4, 5, 6$ is selected are 0%, 80% (true K), 6%, 4%, 10%, respectively. Further, when varying (K, G) in the grid of 25 possible pairs $\{2, 3, 4, 5, 6\} \times \{4, 5, 6, 7, 8\}$, the percentage of the datasets for which the true pair $(K, G) = (3, 6)$ is selected by WAIC is 58% and neither K nor G ever gets underselected. In general, our simulations show that the WAIC does not tend to underselect the latent dimensions K and G, and that it generally has a reasonably good accuracy of selecting the truth. We remark that here our goal was to pick a practical selection criterion that can be readily applied in real-world applications. To develop a selection strategy for deciding on the number of latent dimensions with rigorous theoretical guarantees under the proposed models would need future investigations.

6. Real Data Applications

6.1 NLTCS Disability Survey Data

In this section we apply Gro-M^3 methodology to a functional disability dataset extracted from the National Long Term Care Survey (NLTCS), created by the former Center for Demographic Studies at Duke University. This dataset has been widely analyzed, both with mixed membership models (Erosheva et al., 2007; Manrique-Vallier, 2014), and with other models for multivariate categorical data (Dobra and Lenkoski, 2011; Johndrow et al., 2017). Here we reanalyze this dataset as an illustration of our dimension-grouped mixed membership approach.

The `NLTCS` dataset was downloaded from at `http://lib.stat.cmu.edu/datasets/`. It is an extract containing responses from $n = 21574$ community-dwelling elderly Americans aged 65 and above, pooled over 1982, 1984, 1989, and 1994 survey waves. The disability survey contains $p = 16$ items, with respondents being either coded as healthy (level 0) or as disabled (level 1) for each item. Each respondent provides a 16-dimensional response vector $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,16}) \in \{0, 1\} \times \cdots \times \{0, 1\}$, where each variable $y_{i,j}$ follows a special categorical distribution with two categories, i.e., a Bernoulli distribution, with parameters specific to item j. Among the $p = 16$ NLTCS disability items, functional disability researchers

distinguish six activities of daily living (ADLs) and ten instrumental activities of daily living (IADLs). Specifically, the first six ADL items are more basic and relate to hygiene and personal care: eating, getting in/out of bed, getting around inside, dressing, bathing, and getting to the bathroom or using a toilet. The remaining ten IADL items are related to activities needed to live without dedicated professional care: doing heavy house work, doing light house work, doing laundry, cooking, grocery shopping, getting about outside, travelling, managing money, taking medicine, and telephoning.

Here, we apply the MCMC algorithm developed for the Dirichlet Gro-M^3 to the data; the Dirichlet distribution was also used to model the mixed membership scores in Erosheva et al. (2007). Our preliminary analysis of the NLTCS data indicates the Dirichlet parameters $\boldsymbol{\alpha}$ are relatively small, so we adopt a small $\sigma_\alpha = 0.002$ in the lognormal proposal distribution in Eq. (15) in the Metropolis-Hastings sampling step. For each setting of (G, K), we run the MCMC for 40000 iterations and consider the first 20000 as burn-in to be conservative. We retain every 10th sample after the burn-in. The candidate values for the (G, K) are all the combinations of $G \in \{2, 3, \ldots, 15, 16\}$ and $K \in \{6, 7, \ldots, 11, 12\}$.

For selecting the values of latent dimensions (G, K) in practice, we recommend picking the $(G^\star, K^\star)$ that provide the lowest WAIC value and also do not contain any empty groups of variables. In particular, for certain pairs of (G, K) (in our case, for all $G > 10$) under the NLTCS data, we observe that the posterior mode of the grouping matrix, $\overline{\mathbf{L}}$, has some all-zero columns. If $\widetilde{G}$ denotes the number of not-all-zero columns in $\overline{\mathbf{L}}$, this means after model fitting, the number of groups occupied by the p variables is $\widetilde{G} < G$. Models with $\widetilde{G} < G$ are difficult to interpret because empty groups that do not contain any variables cannot be assigned meaning. Therefore, we focus only on models where $\overline{\mathbf{L}}$ does not contain any all-zero columns and pick the one with the smallest WAIC among these models. Using this criterion, for the NLTCS data, the model with $G^\star = 10$ and $K^\star = 9$ is selected. We have observed reasonably good convergence and mixing of our MCMC algorithm for the NLTCS data. The proposed new dimension-grouping model provides a better fit in terms of WAIC and a parsimonious alternative to traditional MMMs.

We provide the estimated $\overline{\mathbf{L}}$ under the selected model with $G^\star = 10$ and $K^\star = 9$ in Figure 8. The estimated variable groupings are given in Figure 8. Out of the $G^\star = 10$ groups, there are three groups that contain multiple items. In Figure 8, the item labels of these three groups are colored in blue ($j = 1, 2, 4, 5$), red ($j = 9, 10, 16$), and yellow ($j = 12, 13$) for better visualization. These groupings obtained by our model lead to several observations. First, four out of six ADL variables ($j = 1, 2, 4, 5$) are categorized into one group. This group of items are basic self-care activities that require limited mobility. Second, the three IADL variables ($j = 9, 10, 16$) in one group may be related to traditional gender roles – these items correspond to activities performed more frequently by women than by men. Finally, the two items $j = 12$ "getting about outside" and $j = 13$ "traveling" that require high level of mobility form another group. Note that such a model-based grouping of the items is different than the established groups (ADL and IADL), and could not have been obtained by applying previous mixed membership models (Erosheva et al., 2007).

In addition to the variable grouping structures, we plot posterior means of the positive response probabilities $\overline{\boldsymbol{\Lambda}}_{:,1,:}$ in Figure 9 for the selected model. For each survey item $j \in [p]$ and each extreme latent profile $k \in [K]$, the $\boldsymbol{\Lambda}_{j,1,k}$ records the conditional probability of giving a positive response of being disabled on this item conditional on possessing the kth

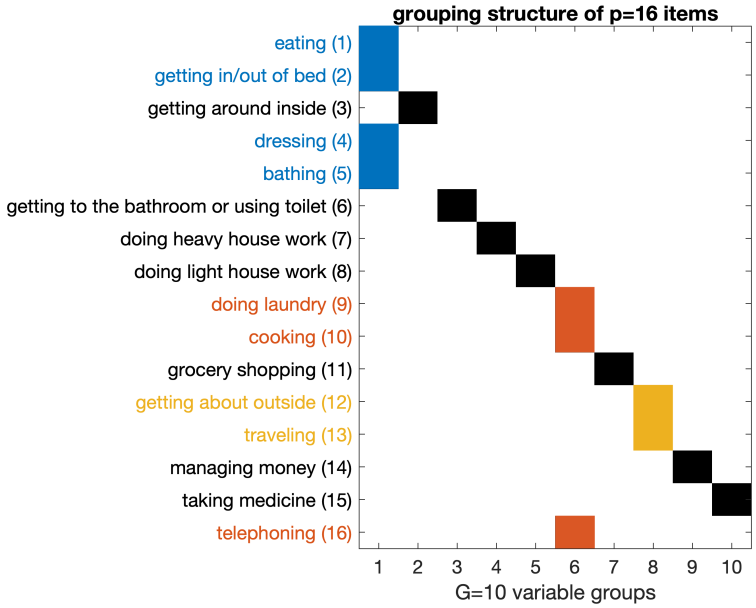


Figure 8: Estimated variable grouping structure $\boldsymbol{s}$ (i.e., $\mathbf{L}$) for the NLTCS data with $(G^\star, K^\star) = (10, 9)$. The first six items are ADL "activities of daily living" and the remaining ten items are IADL "instrumental activities of daily living". Out of the $G^\star = 10$ variable groups, the three groups containing multiple items are colored coded in blue ($j = 1, 2, 4, 5$), red ($j = 9, 10, 16$), and yellow ($j = 12, 13$) for better visualization.

latent profile. The $K^\star = 9$ profiles are quite well separated and can be interpreted as usual in mixed membership analysis. For example, in Figure 9, the leftmost column for $k = 1$ represents a relatively healthy latent profile, the rightmost column for $k = 9$ represents a relatively severely disabled latent profile. As for the Dirichlet parameters $\boldsymbol{\alpha}$, their posterior means are $\overline{\boldsymbol{\alpha}} = (0.0245,\ 0.0289,\ 0.0074,\ 0.0176,\ 0.0231,\ 0.0193,\ 0.0001,\ 0.0001,\ 0.0242)$. Such small values of the Dirichlet parameters imply that membership score vectors tend to be dominated by one component for a majority of individuals. This observation is consistent with Erosheva et al. (2007). Meanwhile, here we obtain a simpler model than that in Erosheva et al. (2007) as each subject can partially belong to up to G latent profiles according to the grouping of variables, rather than $p = 16$ ones as in the traditional MMMs.

We emphasize again that the bag-of-words topic models make the exchangeability assumption, which is fundamentally different from, and actually more restrictive than, our Gro-M^3 when modeling non-exchangeable item response data. Specifically, the exchangeability assumption would force all the item parameters $\{\boldsymbol{\lambda}_{j,k} \in \mathbb{R}^d : \ k \in [K]\}$ across all the items $j \in [p]$ to be identical, which is unrealistic for the survey response data (or the personality test data to be analyed in Section 6.2) in which different items clearly have different characteristics. For example, if one were to use a topic model such as LDA to analyze the NLTCS disability survey data, then a plot of the $16 \times K$ conditional probability table like Figure 9 would not have been possible, because all the $p = 16$ items would share the same K-dimensional vector of conditional Bernoulli probabilities.

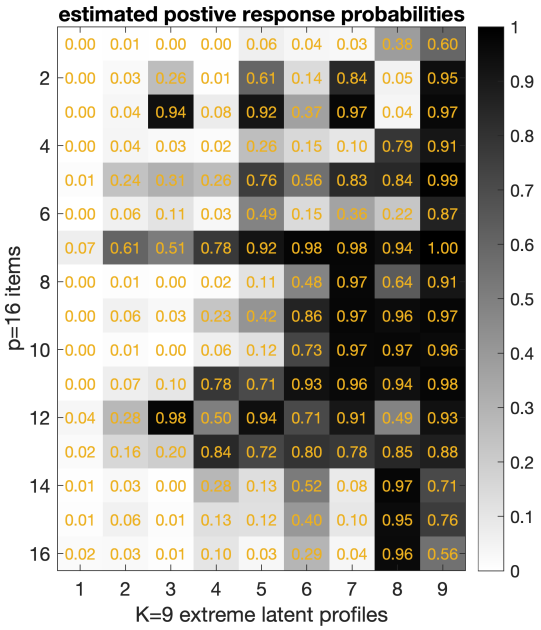


Figure 9: Estimated positive response probabilities $\mathbf{\Lambda}_{:,1,:}$ for the NLTCS data with $(G^\star, K^\star) = (10, 9)$. Each column represents one extreme latent profile. Entries are conditional probabilities of giving a positive response (1 = disabled) to each item given that latent profile.

6.2 International Personality Item Pool (IPIP) Personality Test Data

We also apply the proposed method to analyze a personality test dataset containing multivariate *polytomous* responses: the International Personality Item Pool (IPIP) personality test data. This dataset is publicly available from the Open-Source Psychometrics Project website `https://openpsychometrics.org/_rawdata/`. The dataset contains $n_{\text{all}} = 1005$ subjects' responses to $p = 40$ Likert rated personality test items in the International Personality Item Pool. After dropping those subjects who have missing entries in their responses, there are $n = 901$ complete response vectors left. Each subject's observed response vector is 40-dimensional, where each dimension ranges in $\{1, 2, 3, 4, 5\}$ with $d_1 = d_2 = \cdots = d_p = 5$ categories. Each of these 40 items was designed to measure one of the four personality factors: Assertiveness (short as "AS"), Social confidence (short as "SC"), Adventurousness (short as "AD"), and Dominance (short as "DO"). Specifically, items 1-10 measure AS, items 11-20 measure SC, items 21-30 measure AD, and items 31-40 measure DO. The responses of certain reversely-termed items (i.e., items 7–10, 16–20, 25–30) are preprocessed to be $\widetilde{y}_{ij} = 6 - y_{ij}$. We apply our new model to analyze this dataset for various numbers of variable groups $G \in \{3, 4, 5, 6, 7\}$ and $K = 4$ extreme latent profiles, and the WAIC selects the model with $G = 5$ groups. We plot the posterior mode of the estimated grouping matrix in Figure 10, together with the content of each item.

Figure 10 shows that our new modeling component of variable grouping is able to reveal the item blocks that measure different personality factors in a totally unsupervised man-

ner. Moreover, the estimated variable grouping cuts across the four established personality factors to uncover a more nuanced structure. For example, the group of items {AS1, SC4, SC10} concerns the verbal expression aspect of a person; the group of items {AS3–AS10, SC5, SC7} concerns a person's intention to lead and influence other people. In summary, for this new personality test dataset, the proposed Gro-M^3 not only provides better model fit than the usual GoM model (since $G = 5 \ll p = 40$ is selected by WAIC), but also enjoys interpretability and uncovers meaningful subgroups of the observed variables.

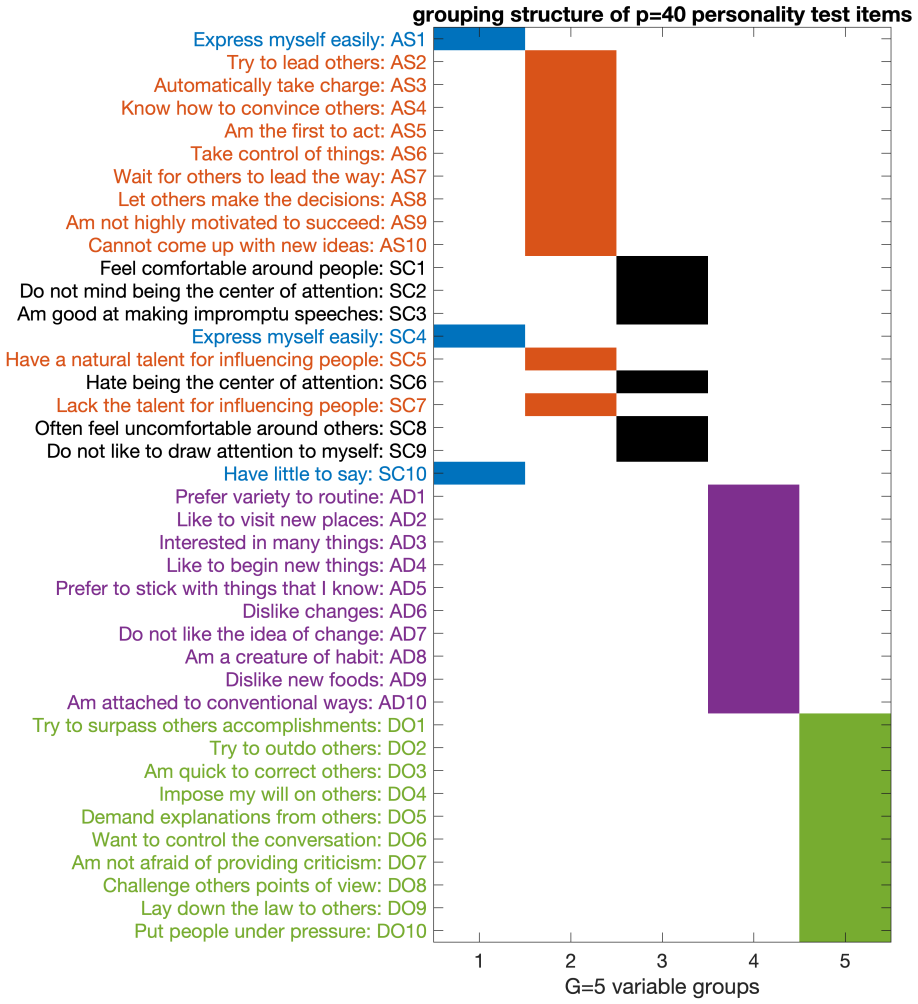


Figure 10: IPIP personality test items grouping structure estimated from our Gro-M^3. Item type abbreviations are: "AS" represents "Assertiveness", "SC" represents "Social confidence", "AD" represents "Adventurousness", and "DO" represents "Dominance".

We also conduct experiments to compare our probabilistic hybrid decomposition Gro-M^3 with the probabilistic CP decomposition (the latent class model in Dunson and Xing (2009)) and the probabilistic Tucker decomposition (the GoM model in Erosheva et al. (2007)) on the IPIP personality test data. After fitting each tensor decomposition method to the data, we calculate the model-based Cramer's V measure between each pair of items

and see how different methods perform on recovering meaningful item dependence structure. Figure 11 presents the model-based pairwise Cramer's V calculated using the three tensor decompositions, along with the model-free Cramer's V calculated directly from data. Figure 11 shows that our Gro-M^3 decomposition clearly outperforms probabilistic CP and Tucker decomposition in recovering the meaningful block structure of the personality test items.

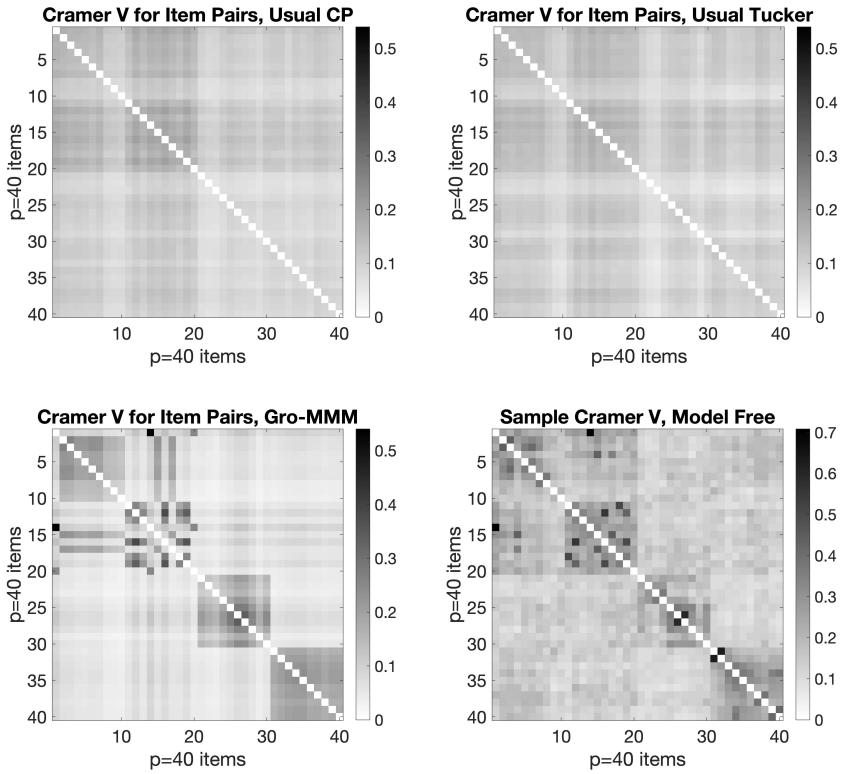


Figure 11: Upper two panels: Cramer's V posterior means for item pairs obtained using the usual CP decomposition (latent class model) and the usual Tucker decomposition (grade of membership model). Bottom left: Cramer's V posterior means for item pairs obtained using the Gro-M^3. Bottom right: Sample Cramer's V for item pairs calculated directly from data.

7. Discussion

We have proposed a new class of mixed membership models for multivariate categorical data, dimension-grouped mixed membership models (Gro-M^{3}s), studied its model identifiability, and developed a Bayesian inference procedure for Dirichlet Gro-M^{3}s. On the methodology side, the new model strikes a nice balance between model flexibility and model parsimony. Considering popular existing latent structure models for multivariate categorical data, the Gro-M^3 bridges the parsimonious yet insufficiently flexible Latent Class Model (corresponding to CP decomposition of probability tensors) and the very flexible yet not parsimonious Grade of Membership Model (corresponding to Tucker decomposition of probability tensors). On the theory side, we establish the identifiability of population parameters that

govern the distribution of Gro-M^{3}s. The quantities shown to be identifiable include not only the continuous model parameters, but also the key discrete structure – how the variables' latent assignments are partitioned into groups. The obtained identifiability conclusions lay a solid foundation for reliable statistical analysis and real-world applications. We have performed Bayesian estimation for the new model using a Metropolis-Hastings-within-Gibbs sampler. Numerical studies show that the method can accurately estimate the quantities of interest, empirically validating the identifiability results.

For the special case of binary responses with $d_1 = \cdots = d_p = 2$, as pointed out by a reviewer, models with Bernoulli-to-latent-Poisson link in Zhou et al. (2016) and the Bernoulli-to-latent-Gaussian link in multivariate item response theory models in Embretson and Reise (2013) are useful tools that can capture certain lower-dimensional latent constructs. Our model differs from these models in terms of statistical and practical interpretation. In our Gro-M^3, each subject's latent variables are a mixed membership vector $\boldsymbol{\pi}_i$ ranging in the probability simplex Δ^{K-1}, and can be interpreted as that each subject is a partial member of each of the K extreme latent profiles. For $k \in [K]$, the kth extreme latent profile also can be directly interpreted by inspecting the estimated item parameters $\{\boldsymbol{\lambda}_{j,k} : \ j \in [p]\}$. Geometrically, the entry π_{ik} captures the relative proximity of each subject to the kth extreme latent behavioral profile. Such an interpretation of individual-level mixtures are highly desirable in applications such as social science surveys (Erosheva, 2003) and medical diagnosis (Woodbury et al., 1978), where each extreme latent profile represents a prototypical response pattern. Therefore, in these applications, the mixed membership modeling is more interpretable and preferable to using a nonlinear transformation of certain underlying Gaussian or Poisson latent variables to model binary matrix data (such as the Bernoulli-to-latent-Poisson or Bernoulli-to-latent-Gaussian link).

We remark that our proposed Gro-M^3 covers the usual GoM model as a special case. In fact, the GoM model can be readily recovered by setting our grouping matrix $\mathbf{L}$ to be the $p \times p$ identity matrix (i.e., $\mathbf{L} = \mathbf{I}_p$). In terms of practical estimation, we can simply fix $\mathbf{L} = \mathbf{I}_p$ throughout our MCMC iterations and estimate other quantities in the same way as in our current algorithm. Using this approach, we have compared the performance of our flexible Gro-M^3 and the classical GoM model in the real data analyses. Specifically, for both the NLTCS disability survey data and the IPIP personality test data, fixing $\mathbf{L} = \mathbf{I}_p$ with $G = p$ variable groups gives larger WAIC values than the selected more parsimonious model with $G \ll p$. This indicates that the traditional GoM model is not favored by the information criterion and gives a poorer model fit to the data. We also point out that our MCMC algorithm can be viewed as a novel Bayesian factorization algorithm for probability tensors, in a similar spirit to the existing Bayesian tensor factorization methods such as Dunson and Xing (2009) and Zhou et al. (2015). Our Bayesian Gro-M^3 factorization outperforms usual probabilistic tensor factorizations in recovering the item dependence structure in the IPIP personality data analysis. Therefore, we view our proposed MCMC algorithm as contributing a new type of tensor factorization approach with nice uniqueness guarantee (i.e., identifiability guarantee) and a Bayesian factorization procedure with good empirical performance.

Our modeling assumption of the variable grouping structure can be useful to other related models. For example, Manrique-Vallier (2014) proposed a longitudinal MMM to capture heterogeneous pathways of disability and cognitive trajectories of elderly population

for disability survey data. The proposed dimension-grouping assumption can provide an interesting new interpretation to such longitudinal settings. Specifically, when survey items are answered in multiple time points, it may be plausible to assume that a subject's latent profile locally persists for a block of items, before potentially switching to a different profile for the next block of items. This can be readily accommodated by the dimension-grouping modeling assumption, with the slight modification that items belonging to the same group should be forced to be close in time. Our identifiability results can be applied to this setup. Similar computational procedures can also be developed. Furthermore, although this work focuses on modeling multivariate categorical data, the applicability of the new dimension-grouping assumption is not limited to such data. A similar assumption may be made in other mixed membership models; examples include the generalized latent Dirichlet models for mixed data types studied in Zhao et al. (2018).

In terms of identifiability, the current work has focused on the population quantities, including the variable grouping matrix $\mathbf{L}$, the conditional probability tables $\mathbf{\Lambda}$, and the Dirichlet parameters $\boldsymbol{\alpha}$. In addition to these *population parameters*, an interesting future question is the identification of individual mixed membership proportions $\{\boldsymbol{\pi}_i;\, i = 1, \ldots, n\}$ for subjects *in the sample*. Studying the identification and accurate estimation of $\boldsymbol{\pi}_i$'s presumably requires quite different conditions from ours. A recent work (Mao et al., 2020) considered a similar problem for mixed membership stochastic block models for network data. Finally, in terms of estimation procedures, in this work we have employed a Bayesian approach to Dirichlet Gro-M^{3}s, and the developed MCMC sampler shows excellent computational performance. In the future, it would also be interesting to consider method-of-moments estimation for the proposed models related to Zhao et al. (2018) and Tan and Mukherjee (2017).

This work has focused on proposing a new interpretable and identifiable mixed membership model for multivariate categorical data, and our MCMC algorithm has satisfactory performance in real data applications. In the future, it would be interesting to develop scalable and online variational inference methods, which would make the model more applicable to massive-scale real-world datasets. We expect that it is possible to develop variational inference algorithms similar in spirit to Blei et al. (2003) for topic models and Airoldi et al. (2008a) for mixed membership stochastic block models to scale up computation. In addition, just as the hierarchical Dirichlet process (Teh et al., 2006) is a natural nonparametric generalization of the parametric latent Dirichlet allocation (Blei et al., 2003) model, it would also be interesting to generalize our Gro-M^3 to the nonparametric Bayesian setting to automatically infer K and G. Developing a method to automatically infer K and G will be of great practical value, because in many situations there might not be enough prior knowledge for these quantatities. We leave these directions for future work.

Acknowledgements

This work was partially supported by NIH grants R01ES027498 and R01ES028804, NSF grants DMS-2210796, SES-2150601 and SES-1846747, and IES grant R305D200015. This project also received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No. 856506). The

authors thank the Action Editor Dr. Mingyuan Zhou and three anonymous reviewers for constructive and helpful comments that helped improve this manuscript.

References

Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. *Advances in Neural Information Processing Systems*, 21, 2008a.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008b.

Edoardo M Airoldi, David Blei, Elena A Erosheva, and Stephen E Fienberg. *Handbook of mixed membership models and their applications*. CRC press, 2014.

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A): 3099–3132, 2009.

Animashree Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

Animashree Anandkumar, Daniel Hsu, Majid Janzamin, and Sham Kakade. When are overcomplete topic models identifiable? Uniqueness of tensor tucker decompositions with structured sparsity. *The Journal of Machine Learning Research*, 16(1):2643–2694, 2015.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE, 2012.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98, 2003.

Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics*, 5(2A):969–993, 2011.

David B Dunson and Chuanhua Xing. Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

Susan E Embretson and Steven P Reise. *Item response theory*. Psychology Press, 2013.

Elena A Erosheva. Bayesian estimation of the grade of membership model. *Bayesian Statistics*, 7:501–510, 2003.

Elena A Erosheva, Stephen E Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1): 5220–5227, 2004.

Elena A Erosheva, Stephen E Fienberg, and Cyrille Joutard. Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*, 1(2):346, 2007.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

Leo A Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

Gérard Govaert and Mohamed Nadif. *Co-clustering: models, algorithms and applications*. John Wiley & Sons, 2013.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

Mevin B Hooten and N Thompson Hobbs. A guide to Bayesian model selection for ecologists. *Ecological monographs*, 85(1):3–28, 2015.

James E Johndrow, Anirban Bhattacharya, and David B Dunson. Tensor decompositions and sparse log-linear models. *Annals of Statistics*, 45(1):1–38, 2017.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, 1977.

Lester W Mackey, David J Weiss, and Michael I Jordan. Mixed membership matrix factorization. In *International Conference on Machine Learning*, 2010.

Daniel Manrique-Vallier. Longitudinal mixed membership trajectory models for disability survey data. *Annals of Applied Statistics*, 8(4):2268, 2014.

Daniel Manrique-Vallier and Jerome P Reiter. Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107(500): 1385–1394, 2012.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, pages 1–13, 2020. doi: 10.1080/01621459.2020.1751645.

Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

XuanLong Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21(1):618–646, 2015.

Sally Paganin, Amy H Herring, Andrew F Olshan, and David B Dunson. Centered partition processes: Informative priors for clustering (with discussion). *Bayesian Analysis*, 16(1): 301–370, 2021.

Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Hachem Saddiki, Jon McAuliffe, and Patrick Flaherty. GLAD: a mixed-membership model for heterogeneous tumor subtype classification. *Bioinformatics*, 31(2):225–232, 2015.

Zhuoran Shang, Elena A Erosheva, and Gongjun Xu. Partial-mastery cognitive diagnosis models. *Annals of Applied Statistics*, 15(3):1529–1555, 2021.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Zilong Tan and Sayan Mukherjee. Partitioned tensor factorizations for learning mixed membership models. In *International Conference on Machine Learning*, pages 3358–3367, 2017.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. doi: 10.1198/016214506000000302. URL https://doi.org/10.1198/016214506000000302.

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

Y. Samuel Wang, Ross L. Matsueda, and Elena A. Erosheva. A variational em method for mixed membership models with multivariate rank data: an analysis of public policy preferences. *arXiv: Methodology*, pages 1452–1480, 2015.

Yining Wang. Convergence rates of latent topic models under relaxed identifiability conditions. *Electronic Journal of Statistics*, 13(1):37–66, 2019.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12), 2010.

Max A Woodbury, Jonathan Clive, and Arthur Garson Jr. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11(3):277–298, 1978.

Shiwen Zhao, Barbara E Engelhardt, Sayan Mukherjee, and David B Dunson. Fast moment estimation for generalized latent Dirichlet models. *Journal of the American Statistical Association*, 113(524):1528–1540, 2018.

Jing Zhou, Anirban Bhattacharya, Amy H Herring, and David B Dunson. Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512): 1562–1576, 2015.

Mingyuan Zhou. Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis*, 13(4):1065–1093, 2018.

Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable Gamma belief networks. *Journal of Machine Learning Research*, 17(1):5656–5699, 2016.

Supplementary Material

This Supplementary Material contains two sections. The first section Supplement A contains the proofs of all the theoretical results in the paper. The second section Supplement B presents a note on the pairwise mutual information measures between categorical variables.

Supplement A: Proofs of Theoretical Results

7.1 Proof of Theorem 2

For notational simplicity, from now on we will omit the subscript i of subject-specific random variables without loss of generality; all such variables including $\boldsymbol{y}$ and $\boldsymbol{z}$ should be understood as associated with a random subject. Denote by $\boldsymbol{z} = (z_1, \ldots, z_G) \in [K]^G$ a configuration of the latent profiles realized for the G groups of variables. Recall that given a fixed grouping matrix $\mathbf{L}$, the associated group assignment vector $\boldsymbol{s} = (s_1, \ldots, s_p)$ is defined as $s_j = g$ if and only if $\ell_{j,g} = 1$. We next introduce a new notation. For each variable $j \in [p]$, each category $c_j \in [d_j]$, and each possible latent profile configuration $\boldsymbol{z} \in \{1, \ldots, K\}^G$, define a new parameter $\gamma_{j,c_j,\boldsymbol{z}}$ to be

$$\gamma_{j,c_j,\boldsymbol{z}} = \lambda_{j,c_j,z_{s_j}}. \tag{18}$$

Collect all the γ-parameters in $\mathbf{\Gamma} = (\gamma_{j,c_j,\boldsymbol{z}})$, then $\mathbf{\Gamma}$ is a three-way array (which is a tensor of size $p \times d \times K^G$ if $d_1 = \cdots = d_p = d$) since $j \in [p]$, $c_j \in [d_j]$, and $\boldsymbol{z} \in [K]^G$. For each $j \in [p]$, we will denote the $d_j \times K^G$ matrix $\mathbf{\Gamma}_{j,:,:}$ by $\mathbf{\Gamma}_j$ for simplicity. The representation in (18) implies that many entries in $\mathbf{\Gamma}$ are equal. Specifically, for two arbitrary latent assignment vectors $\boldsymbol{z} = (z_1, \ldots, z_G)$ and $\boldsymbol{z}' = (z'_1, \ldots, z'_G)$ with $\boldsymbol{z} \neq \boldsymbol{z}'$, as long as $z_{s_j} = z'_{s_j}$ there would be $\gamma_{j,c_j,\boldsymbol{z}} = \gamma_{j,c_j,\boldsymbol{z}'}$. We choose to use the over-parameterization in (18) since this notation facilitates the study of identifiability through the underlying tensor decomposition structure, as will be revealed soon. In particular, the $\mathbf{\Gamma}_j$'s have the following property.

Lemma 8 *Under the definition in (18), for any set of indices $S \subseteq [p]$ such that $\{s_j : j \in S\} \supseteq [G]$, there is*

$$\bigotimes_{g=1}^{G} \left(\bigodot_{j \in S:\, s_j = g} \mathbf{\Lambda}_j \right) = \bigodot_{j \in S} \mathbf{\Gamma}_j. \tag{19}$$

Now we can equivalently rewrite the previous model specification (7) as follows,

$$\begin{aligned}
&\mathbb{P}^{\text{C-M}^3}(y_1 = c_1, \ldots, y_p = c_p \mid \mathbf{L}, \mathbf{\Lambda}, \mathbf{\Phi}) = \pi_{c_1,\ldots,c_p} \\
&= \sum_{z_1=1}^{K} \cdots \sum_{z_G=1}^{K} \phi_{z_1,\ldots,z_G} \prod_{j=1}^{p} \lambda_{j,c_j,z_{s_j}} = \sum_{z_1=1}^{K} \cdots \sum_{z_G=1}^{K} \phi_{z_1,\ldots,z_G} \prod_{j=1}^{p} \gamma_{j,c_j,\boldsymbol{z}} \\
&= \sum_{\boldsymbol{z} \in [K]^G} \phi_{\boldsymbol{z}} \prod_{j=1}^{p} \gamma_{j,c_j,\boldsymbol{z}} = \mathbb{P}(y_1 = c_1, \ldots, y_p = c_p \mid \mathbf{\Gamma}, \mathbf{\Phi}),
\end{aligned} \tag{20}$$

where $\boldsymbol{c} = (c_1, \ldots, c_p) \in \times_{j=1}^{p}[d_j]$. Denote by $\boldsymbol{\Phi} = (\phi_{z_1,\ldots,z_G})$ a G-th order tensor of size $K \times \cdots \times K$. Denote by $\mathrm{vec}(\boldsymbol{\Pi})$ the vectorized version of $\boldsymbol{\Pi}$, so $\mathrm{vec}(\boldsymbol{\Pi})$ is a vector of length $\prod_{j=1}^{p} d_j$; in particular, this vector has entries defined as follows,

$$\mathrm{vec}(\boldsymbol{\Pi})_{c_1+(c_2-1)d_1+\cdots+(c_p-1)d_1\cdots d_{p-1}} = \pi_{c_1,c_2,\cdots,c_p} \tag{21}$$

for any $\boldsymbol{c} = (c_1, \ldots, c_p) \in \times_{j=1}^{p}[d_j]$. Suppose alternative parameters $(\overline{\mathbf{L}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\Phi}})$ lead to the same distribution of the observed variables; that is $\mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \mathbf{L}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}) = \mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \overline{\mathbf{L}}, \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\Phi}})$ holds for each possible response pattern $\boldsymbol{c} \in \times_{j=1}^{p}[d_j]$. Then by the equivalence in (20), we also have $\mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \boldsymbol{\Gamma}, \boldsymbol{\Phi}) = \mathbb{P}(\boldsymbol{y} = \boldsymbol{c} \mid \overline{\boldsymbol{\Gamma}}, \overline{\boldsymbol{\Phi}})$ for all $\boldsymbol{c} \in \times_{j=1}^{p}[d_j]$.

The following two lemmas will be useful.

Lemma 9 *Without loss of generality, suppose the first $\sum_{j=1}^{p} \ell_{j,1}$ variables belong to the first group, the second $\sum_{j=1}^{p} \ell_{j,2}$ variables belong to the second group, etc. That is, the matrix $\mathbf{L}$ takes a block-diagonal form. The Gro-M^3 in (20) implies the following identity*

$$\mathrm{vec}(\boldsymbol{\Pi}) = \left\{ \bigotimes_{g=1}^{G} \bigodot_{j:\, \ell_{j,g}=1} \boldsymbol{\Lambda}_j \right\} \cdot \mathrm{vec}(\boldsymbol{\Phi}). \tag{22}$$

Lemma 10 *Suppose there are two disjoint sets of G observed variables $S^{(1)} = \{j_1^{(1)}, \ldots, j_G^{(1)}\}$ and $S^{(2)} = \{j_1^{(2)}, \ldots, j_G^{(2)}\}$ satisfying $s_{j_g^{(1)}} = s_{j_g^{(2)}} = g$ for each $g = 1, \ldots, G$. Then*

$$\bigotimes_{g=1}^{G} \left\{ \boldsymbol{\Lambda}_{j_g^{(1)}} \bigodot \boldsymbol{\Lambda}_{j_g^{(2)}} \right\} = \left\{ \bigotimes_{g=1}^{G} \boldsymbol{\Lambda}_{j_g^{(1)}} \right\} \bigodot \left\{ \bigotimes_{g=1}^{G} \boldsymbol{\Lambda}_{j_g^{(2)}} \right\} \quad \text{up to a permutation of rows.} \tag{23}$$

If there further is $s_{j_G^{(3)}} = G$ for some $j_G^{(3)} \in [p]$, then up to a permutation of rows there is

$$\begin{aligned} &\bigotimes_{g=1}^{G-1} \left\{ \boldsymbol{\Lambda}_{j_g^{(1)}} \bigodot \boldsymbol{\Lambda}_{j_g^{(2)}} \right\} \bigotimes \left\{ \boldsymbol{\Lambda}_{j_G^{(1)}} \bigodot \boldsymbol{\Lambda}_{j_G^{(2)}} \bigodot \boldsymbol{\Lambda}_{j_G^{(3)}} \right\} \\ &= \left\{ \bigotimes_{g=1}^{G} \boldsymbol{\Lambda}_{j_g^{(1)}} \right\} \bigodot \left\{ \bigotimes_{g=1}^{G-1} \boldsymbol{\Lambda}_{j_g^{(2)}} \bigotimes \left(\boldsymbol{\Lambda}_{j_G^{(2)}} \bigodot \boldsymbol{\Lambda}_{j_G^{(3)}} \right) \right\}. \end{aligned} \tag{24}$$

We continue with the proof of Theorem 2. Under the conditions of the theorem, without loss of generality we can assume that for each $g \in [G]$, the first three variables (among the p ones) belonging to the gth group have their corresponding $\boldsymbol{\Lambda}_j$ full-column-rank; denote the indices of these three variables by $j_g^{(1)}, j_g^{(2)}, j_g^{(3)}$. For example, if $\mathbf{L}$ takes the block diagonal form, then for $g = 1$ such three variables are indexed by $\left\{ j_1^{(1)}, j_1^{(2)}, j_1^{(3)} \right\} = \{1, 2, 3\}$; for $g = 2$ they are indexed by $\left\{ j_2^{(1)}, j_2^{(2)}, j_2^{(3)} \right\} = \left\{ \sum_{j=1}^{p} \ell_{j,1} + 1, \sum_{j=1}^{p} \ell_{j,1} + 2, \sum_{j=1}^{p} \ell_{j,1} + 3 \right\}$,

etc. Define the following sets of variable indices,

$$S^{(m)} = \left\{ j_1^{(m)}, \ldots, j_G^{(m)} \right\} \text{ for } m = 1, 2, 3; \quad S^{(0)} = \{1, \ldots, p\} \setminus \bigcup_{m=1}^{3} S^{(m)}.$$

Lemmas 9 and 10 imply that we can write $\mathrm{vec}(\mathbf{\Pi})$ under the true parameters as follows

$$\begin{aligned}
\mathrm{vec}(\mathbf{\Pi}) &= \left\{ \bigotimes_{g=1}^{G} \bigodot_{j:\, \ell_{j,g}=1} \mathbf{\Lambda}_j \right\} \cdot \mathrm{vec}(\mathbf{\Phi}) \\
&= \left[\left\{ \bigotimes_{g=1}^{G} \mathbf{\Lambda}_{j_g^{(1)}} \right\} \bigodot \left\{ \bigotimes_{g=1}^{G} \mathbf{\Lambda}_{j_g^{(2)}} \right\} \bigodot \left\{ \bigotimes_{g=1}^{G} \left(\mathbf{\Lambda}_{j_g^{(3)}} \bigodot \left(\bigodot_{j \in S^{(0)}:\, \ell_{j,g}=1} \mathbf{\Lambda}_j \right) \right) \right\} \right] \cdot \mathrm{vec}(\mathbf{\Phi}) \\
&\overset{(\star)}{=} \left[\left\{ \bigodot_{g=1}^{G} \mathbf{\Gamma}_{j_g^{(1)}} \right\} \bigodot \left\{ \bigodot_{g=1}^{G} \mathbf{\Gamma}_{j_g^{(2)}} \right\} \bigodot \left\{ \bigodot_{j \in S^{(3)} \cup S^{(0)}} \mathbf{\Gamma}_j \right\} \right] \cdot \mathrm{vec}(\mathbf{\Phi}). \qquad (25)
\end{aligned}$$

The last equality $(\star)$ above follows from Lemma 8, by noting that for each $m = 1, 2, 3$, the index set $\{j_1^{(m)}, \ldots, j_G^{(m)}\} = [K]$. The last equality in the above display results from the property of the Khatri-Rao product. Define

$$f^{(1)}(\mathbf{\Gamma}) := \bigodot_{g=1}^{G} \mathbf{\Gamma}_{j_g^{(1)}} = \bigodot_{j \in S^{(1)}} \mathbf{\Gamma}_j, \quad f^{(2)}(\mathbf{\Gamma}) := \bigodot_{g=1}^{G} \mathbf{\Gamma}_{j_g^{(2)}} = \bigodot_{j \in S^{(2)}} \mathbf{\Gamma}_j, \quad f^{(3)}(\mathbf{\Gamma}) := \bigodot_{j \in S^{(3)} \cup S^{(0)}} \mathbf{\Gamma}_j.$$

It can be seen that the definitions of the above three functions $f^{(1)}(\cdot)$, $f^{(2)}(\cdot)$, $f^{(3)}(\cdot)$ of $\mathbf{\Gamma}$ only depend on the two sets of variable indices $S^{(1)}$ and $S^{(2)}$, which in turn are determined by the true grouping matrix $\mathbf{L}$. Now (25) can be further written as

$$\mathrm{vec}(\mathbf{\Pi}) = \left(f^{(1)}(\mathbf{\Gamma}) \bigodot f^{(2)}(\mathbf{\Gamma}) \bigodot f^{(3)}(\mathbf{\Gamma}) \right) \cdot \mathrm{vec}(\mathbf{\Phi}) = \bigodot_{j=1}^{p} \mathbf{\Gamma}_j \cdot \mathrm{vec}(\mathbf{\Phi}).$$

So for true parameters $(\mathbf{\Gamma}, \mathbf{\Phi})$ and alternative parameters $(\overline{\mathbf{\Gamma}}, \overline{\mathbf{\Phi}})$ that lead to the same distribution of the observed $\boldsymbol{y}$, we have

$$\begin{aligned}
\mathrm{vec}(\mathbf{\Pi}) &= \left(f^{(1)}(\mathbf{\Gamma}) \bigodot f^{(2)}(\mathbf{\Gamma}) \bigodot f^{(3)}(\mathbf{\Gamma}) \right) \cdot \mathrm{vec}(\mathbf{\Phi}) \qquad (26) \\
&= \left(f^{(1)}(\overline{\mathbf{\Gamma}}) \bigodot f^{(2)}(\overline{\mathbf{\Gamma}}) \bigodot f^{(3)}(\overline{\mathbf{\Gamma}}) \right) \cdot \mathrm{vec}(\overline{\mathbf{\Phi}}).
\end{aligned}$$

Recall that under the assumptions in the theorem and the current notation, for each $j \in S^{(1)} \cup S^{(2)} \cup S^{(3)}$ the matrix $\mathbf{\Lambda}_j$ has full column rank K. According to the property of the Kronecker product, the matrices $\bigotimes_{g=1}^{G} \mathbf{\Lambda}_{j_g^{(1)}}$ and $\bigotimes_{g=1}^{G} \mathbf{\Lambda}_{j_g^{(2)}}$ each has full column rank K^G. Further, since $\mathbf{\Lambda}_{j_g^{(3)}}$ has full column rank K, the Khatri-Rao product $\mathbf{\Lambda}_{j_g^{(3)}} \bigodot \left(\bigodot_{j \in S^{(0)}:\, \ell_{j,g}=1} \mathbf{\Lambda}_j \right)$ must have full column rank K. Therefore the matrix

$\left\{\bigotimes_{g=1}^{G}\left(\mathbf{\Lambda}_{j_g^{(3)}} \odot \left(\bigodot_{j\in S^{(0)}:\ell_{j,g}=1}\mathbf{\Lambda}_j\right)\right)\right\}$ also has full column rank K^G. By definition of $f^{(m)}(\mathbf{\Gamma})$'s, the above full-rank assertions indeed mean that $f^{(1)}(\mathbf{\Gamma})$, $f^{(2)}(\mathbf{\Gamma})$, and $f^{(3)}(\mathbf{\Gamma})$ all have full column rank K^G.

We next invoke a useful lemma on the uniqueness of three-way tensor decompositions, the Kruskal's theorem established in Kruskal (1977), and then proceed similarly as the proof procedures in Allman et al. (2009). For a matrix $\mathbf{M}$, its Kruskal rank is defined to be the largest number r such that any r columns of $\mathbf{M}$ are linearly independent. Denote the Kruskal rank of matrix $\mathbf{M}$ by $\text{rank}_K(\mathbf{M})$.

Lemma 11 (Kruskal's Theorem) *Suppose $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ are three matrices of dimension $a_m \times K$ for $m = 1, 2, 3$, $\mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3$ are three matrices each with K columns, and $\bigodot_{m=1}^{3}\mathbf{M}_m = \bigodot_{m=1}^{3}\mathbf{N}_m$. If $\text{rank}_K(\mathbf{M}_1)+\text{rank}_K(\mathbf{M}_2)+\text{rank}_K(\mathbf{M}_3) \geq 2K+2$, then there exists a permutation matrix $\mathbf{P}$ and three invertible diagonal matrices $\mathbf{D}_m$ with $\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 = \mathbf{I}_K$ and $\mathbf{N}_m = \mathbf{M}_m\mathbf{D}_m\mathbf{P}$ for each $m = 1, 2, 3$.*

If a matrix has full column rank K, then it must also have Kruskal rank K by definition. As a corrolary of Lemma 11, if the three matrices $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$ all have full column rank K, then the condition $\text{rank}_K(\mathbf{M}_1) + \text{rank}_K(\mathbf{M}_2) + \text{rank}_K(\mathbf{M}_3) = 3K \geq 2K + 2$ is satisfied and the uniqueness conclusion follows. We now take $\mathbf{M}_m = f^{(m)}(\mathbf{\Gamma})$, $\mathbf{N}_m = f^{(m)}(\overline{\mathbf{\Gamma}})$ for $m = 1, 2$, and define

$$\mathbf{M}_3 = f^{(3)}(\mathbf{\Gamma}) \cdot \text{diag}(\text{vec}(\mathbf{\Phi})), \quad \mathbf{N}_3 = f^{(3)}(\overline{\mathbf{\Gamma}}) \cdot \text{diag}(\text{vec}(\overline{\mathbf{\Phi}})),$$

then there is $\text{vec}(\mathbf{\Pi}) = \bigodot_{m=1}^{3}\mathbf{M}_m = \bigodot_{m=1}^{3}\mathbf{N}_m$. According to our argument right after (26), $\text{rank}_K(\mathbf{M}_m) = \text{rank}_K(f^{(m)}(\mathbf{\Gamma})) = K$ for $m = 1, 2$. As for $\mathbf{M}_3$, since $f^{(3)}(\mathbf{\Gamma})$ has full column rank K and the entries of $\mathbf{\Phi}$ are positive, the $\mathbf{M}_3$ also has full column rank K. Therefore, we can invoke Lemma 11 to establish that there exists a permutation matrix $\mathbf{P}$ and three invertible diagonal matrices $\mathbf{D}_m$ with $\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 = \mathbf{I}_K$ such that

$$f^{(m)}(\overline{\mathbf{\Gamma}}) = \mathbf{N}_m = \mathbf{M}_m\mathbf{D}_m\mathbf{P} = f^{(m)}(\mathbf{\Gamma})\mathbf{D}_m\mathbf{P}$$

for $m = 1, 2, 3$.

The next step is to show that the diagonal matrices $\mathbf{D}_i$ are all identity matrices. Note that each column of the $\prod_{j\in S^{(1)}} d_j \times K$ matrix $f^{(1)}(\mathbf{\Gamma}) = \bigodot_{g=1}^{G}\mathbf{\Gamma}_{j_g^{(1)}} = \bigotimes_{g=1}^{G}\mathbf{\Lambda}_{j_g^{(1)}}$ characterizes the conditional joint distribution of $\{y_j : j \in S^{(1)}\}$ given the latent assignment vector $\boldsymbol{z} \in [K]^G$ under the true $\mathbf{\Lambda}$-parameters. And similarly, each column of $f^{(1)}(\overline{\mathbf{\Gamma}}) = \bigodot_{g=1}^{G}\overline{\mathbf{\Gamma}}_{j_g^{(1)}} = \bigotimes_{g=1}^{G}\overline{\mathbf{\Lambda}}_{j_g^{(1)}}$ characterizes the conditional joint distribution of $\{y_j : j \in S^{(1)}\}$ given $\boldsymbol{z} \in [K]^G$ under the alternative $\overline{\mathbf{\Lambda}}$. Therefore the sum of each column of $f^{(1)}(\mathbf{\Gamma})$ or that of $f^{(1)}(\overline{\mathbf{\Gamma}})$ equals one, which implies the diagonal matrix $\mathbf{D}_m$ is an identity matrix for $m = 1$ or 2. Since Lemma 11 ensures $\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 = \mathbf{I}_K$, we also obtain $\mathbf{D}_3 = \mathbf{I}_K$. By far we have obtained $f^{(m)}(\overline{\mathbf{\Gamma}}) = f^{(m)}(\mathbf{\Gamma})\mathbf{P}$ for $m = 1, 2$ and

$$f^{(3)}(\overline{\mathbf{\Gamma}}) \cdot \text{diag}(\text{vec}(\overline{\mathbf{\Phi}})) = f^{(3)}(\mathbf{\Gamma}) \cdot \text{diag}(\text{vec}(\mathbf{\Phi}))\mathbf{P}. \tag{27}$$

Note that the permutation matrix $\mathbf{P}$ has rows and columns both indexed by latent assignment vectors $\boldsymbol{z} \in [K]^G$. For $m = 3$, consider an arbitrary $\boldsymbol{z}_1 \in [K]^G$ and assume without

loss of generality that $\boldsymbol{z}_2 \in [K]^G$ satisfies that the $(\boldsymbol{z}_2, \boldsymbol{z}_1)$th entry of matrix $\mathbf{P}$ is $\mathbf{P}_{\boldsymbol{z}_2,\boldsymbol{z}_1} = 1$. Then the $\boldsymbol{z}_1$th column of the matrix equality $f^{(3)}(\overline{\mathbf{\Gamma}})\cdot\mathrm{diag}(\mathrm{vec}(\overline{\mathbf{\Phi}})) = f^{(3)}(\mathbf{\Gamma})\cdot\mathrm{diag}(\mathrm{vec}(\mathbf{\Phi}))\mathbf{P}$ takes the form

$$f^{(3)}(\overline{\mathbf{\Gamma}})_{\boldsymbol{c},\boldsymbol{z}_1} \cdot \mathrm{vec}(\overline{\mathbf{\Phi}})_{\boldsymbol{z}_1} = f^{(3)}(\mathbf{\Gamma})_{\boldsymbol{c},\boldsymbol{z}_2} \cdot \mathrm{vec}(\mathbf{\Phi})_{\boldsymbol{z}_2}$$

for each $\boldsymbol{c} \in \times_{j=1}^{p}[d_j]$; summing the above equality over the index $\boldsymbol{c} \in \times_{j\in\mathcal{A}_1}[d_j]$ gives $\mathrm{vec}(\overline{\mathbf{\Phi}})_{\boldsymbol{z}_1} = \mathrm{vec}(\mathbf{\Phi})_{\boldsymbol{z}_2}$. Note we have generally established $\mathrm{vec}(\overline{\mathbf{\Phi}})_{\boldsymbol{z}_1} = \mathrm{vec}(\mathbf{\Phi})_{\boldsymbol{z}_2}$ whenever $\mathbf{P}_{\boldsymbol{z}_2,\boldsymbol{z}_1} = 1$, which essentially implies $\mathrm{vec}(\overline{\mathbf{\Phi}})^\top = \mathrm{vec}(\mathbf{\Phi})^\top \cdot \mathbf{P}$. Note that this shows the identifiability of the tensor core in our hybrid tensor decomposition formulation of the GroM^3. Further, $\mathrm{vec}(\overline{\mathbf{\Phi}})^\top = \mathrm{vec}(\mathbf{\Phi})^\top \cdot \mathbf{P}$ implies that

$$\mathrm{diag}(\mathrm{vec}(\overline{\mathbf{\Phi}})) = \mathrm{diag}(\mathrm{vec}(\mathbf{\Phi}) \cdot \mathbf{P}) = \mathrm{diag}(\mathrm{vec}(\mathbf{\Phi})) \cdot \mathbf{P}.$$

Combining the above display to the previous (27), since $\mathrm{diag}(\mathrm{vec}(\overline{\mathbf{\Phi}}))$ is a diagonal matrix with positive diagonal entries, we can right multiply the inverse of this matrix with the LHS of (27) and meanwhile right multiply the inverse of $\mathrm{diag}(\mathrm{vec}(\mathbf{\Phi})) \cdot \mathbf{P}$ with the RHS of (27); this gives $f^{(3)}(\overline{\mathbf{\Gamma}}) = f^{(3)}(\mathbf{\Gamma})\mathbf{P}$.

Our final step of proving the theorem is to show that the established $f^{(m)}(\overline{\mathbf{\Gamma}}) = f^{(m)}(\mathbf{\Gamma})\mathbf{P}$ for $m = 1, 2, 3$ implies the identifiability of $\mathbf{\Lambda}$ and $\mathbf{L}$. First, since $f^{(m)}(\mathbf{\Gamma})$ is defined as certain Khatri-Rao products of the individual $\mathbf{\Gamma}_j$'s, we claim that the $f^{(m)}(\overline{\mathbf{\Gamma}}) = f^{(m)}(\mathbf{\Gamma})\mathbf{P}$ indeed implies that $\overline{\mathbf{\Gamma}}_j = \mathbf{\Gamma}_j\mathbf{P}$ for each $j \in [p]$. To see this, note that each column of $f^{(m)}(\overline{\mathbf{\Gamma}})$ and $f^{(m)}(\mathbf{\Gamma})\mathbf{P}$ characterizes the conditional joint distribution of variables $\{y_j : j \in S^{(m)}\}$ given the $\boldsymbol{z}$. So the conditional marginal distribution $\mathbf{\Gamma}_j$ can be obtained by summing up appropriate row vectors of the matrices $f^{(m)}(\overline{\mathbf{\Gamma}})$ and $f^{(m)}(\mathbf{\Gamma})\mathbf{P}$, corresponding to marginalizing out other variables except the jth one. Now without loss of generality we can assume that $\mathbf{P} = \mathbf{I}_{K^G}$, then $\overline{\mathbf{\Gamma}}_j = \mathbf{\Gamma}_j\mathbf{P}$ gives $\bar{\gamma}_{j,c_j,\boldsymbol{z}} = \gamma_{j,c_j,\boldsymbol{z}}$ for all $\boldsymbol{z} \in [K]^G$. Now it only remains to show that $\mathbf{\Gamma}$ uniquely determines $\mathbf{\Lambda}$ and $\mathbf{L}$. By definition (18) there is $\gamma_{j,c_j,\boldsymbol{z}} = \lambda_{j,c_j,z_{s_j}}$. For arbitrary s_j and $\overline{s}_j$, we first consider an arbitrary latent assignment $\boldsymbol{z} \in [K]^G$ such that $z_{s_j} = z_{\overline{s}_j}$, then

$$\lambda_{j,c_j,k} = \lambda_{j,c_j,z_{s_j}} = \gamma_{j,c_j,\boldsymbol{z}} = \overline{\gamma}_{j,c_j,\boldsymbol{z}} = \overline{\lambda}_{j,c_j,z_{\overline{s}_j}} = \overline{\lambda}_{j,c_j,k}.$$

The above reasoning proves the identifiability of $\mathbf{\Lambda}$. Thus far we have proved part (a) of Theorem 2.

We next prove part (b) of the theorem. We use proof by contradiction to show the identifiability of the grouping matrix $\mathbf{L}$ (or equivalently, the identifiability of the vector $\boldsymbol{s}$). If there exists some $j \in [p]$ such that the jth rows of $\mathbf{L}$ and $\overline{\mathbf{L}}$ are different, then $s_j \neq \overline{s}_j$; denote $s_j =: g$ and $\overline{s}_j =: g'$. Next for arbitrary two different indices $k, k' \in [K]$ and $k \neq k'$, we consider a latent assignment $\boldsymbol{z} \in [K]^G$ such that $z_g = k$ and $z_g = k'$. Then there are

$$\lambda_{j,c_j,k} = \lambda_{j,c_j,z_{s_j}} = \gamma_{j,c_j,\boldsymbol{z}} = \overline{\gamma}_{j,c_j,\boldsymbol{z}} = \overline{\lambda}_{j,c_j,z_{\overline{s}_j}} = \overline{\lambda}_{j,c_j,k'} \quad \text{for all} \quad c_j \in [d_j].$$

Since $k \neq k'$, the above equality means the kth and k'th columns of $\mathbf{\Lambda}_j$ are identical. Since k and k' are two arbitrary indices, this means all the column vectors in the matrix $\mathbf{\Lambda}_j$ are identical. This contradicts the assumption in part (b) of the theorem. Therefore we have

shown that $s_j = \overline{s}_j$ must hold for an arbitrary $j \in [p]$. This proves the identifiability of $\mathbf{L}$ from $\mathbf{\Gamma}$. This completes the proof of Theorem 2.

7.2 Proof of Theorem 3

The proof of this theorem is similar in spirit to the previous Theorem 2 by exploiting the inherent tensor decomposition structure, but differing in taking advantage of the more dimension-grouping structure under the assumptions here. Recall $\mathcal{A}_g = \{g \in [p] : \ell_{j,g} = 1\} = \cup_{m=1}^{3} \mathcal{A}_{g,m}$. We write $\mathrm{vec}(\mathbf{\Pi})$ under the true parameters as

$$\begin{aligned}
\mathrm{vec}(\mathbf{\Pi}) &= \left\{ \bigotimes_{g=1}^{G} \bigodot_{j \in \mathcal{A}_g} \mathbf{\Lambda}_j \right\} \cdot \mathrm{vec}(\mathbf{\Phi}) \\
&= \left\{ \bigotimes_{g=1}^{G} \left(\bigodot_{m=1}^{3} \bigodot_{j \in \mathcal{A}_{g,m}} \mathbf{\Lambda}_j \right) \right\} \cdot \mathrm{vec}(\mathbf{\Phi}) \\
&= \left[\left\{ \bigotimes_{g=1}^{G} \left(\bigodot_{j \in \mathcal{A}_{g,1}} \mathbf{\Lambda}_j \right) \right\} \bigodot \left\{ \bigotimes_{g=1}^{G} \left(\bigodot_{j \in \mathcal{A}_{g,2}} \mathbf{\Lambda}_j \right) \right\} \bigodot \left\{ \bigotimes_{g=1}^{G} \left(\bigodot_{j \in \mathcal{A}_{g,3}} \mathbf{\Lambda}_j \right) \right\} \right] \cdot \mathrm{vec}(\mathbf{\Phi}).
\end{aligned}$$

Since each $\mathcal{A}_{g,m}$ is nonempty under the assumption in the theorem and $\cup_{g=1}^{G} \mathcal{A}_{g,m} \supseteq [G]$, we can use Lemma 8 to obtain that

$$\bigodot_{j \in \cup_{g=1}^{G} \mathcal{A}_{g,m}} \mathbf{\Gamma}_j = \bigotimes_{g=1}^{G} \left(\bigodot_{j \in \mathcal{A}_{g,m}} \mathbf{\Lambda}_j \right) = \bigotimes_{g=1}^{G} \widetilde{\mathbf{\Lambda}}_{g,m},$$

where the second equality above follows from the definition of $\widetilde{\mathbf{\Lambda}}_{g,m}$ in the theorem. We now define

$$f^{(m)}(\mathbf{\Gamma}) := \bigodot_{j \in \cup_{g=1}^{G} \mathcal{A}_{g,m}} \mathbf{\Gamma}_j, \quad m = 1, 2, 3.$$

Since the theorem has the assumption that each $\widetilde{\mathbf{\Lambda}}_{g,m}$ has full column rank K, we have that $f^{(m)}(\mathbf{\Gamma})$ has full rank K. Note that each $f^{(m)}(\cdot)$ is the Khatri-Rao product of certain $\mathbf{\Gamma}_j$'s and $f^{(m)}$ depends on the true grouping matrix $\mathbf{L}$. Also note that there is

$$\begin{aligned}
\mathrm{vec}(\mathbf{\Pi}) &= \left(f^{(1)}(\mathbf{\Gamma}) \bigodot f^{(2)}(\mathbf{\Gamma}) \bigodot f^{(3)}(\mathbf{\Gamma}) \right) \cdot \mathrm{vec}(\mathbf{\Phi}) \\
&= \left(f^{(1)}(\overline{\mathbf{\Gamma}}) \bigodot f^{(2)}(\overline{\mathbf{\Gamma}}) \bigodot f^{(3)}(\overline{\mathbf{\Gamma}}) \right) \cdot \mathrm{vec}(\overline{\mathbf{\Phi}}).
\end{aligned}$$

Now the problem is in exactly the same formulation as that in the proof of Theorem 2, so we can proceed in the same way to establish the identifiability of $\mathbf{\Phi}$ and individual $\mathbf{\Gamma}_j$'s. The identifiability of $\mathbf{\Gamma}$ further gives the identifiability of $\mathbf{\Lambda}$ and $\mathbf{L}$. This finishes the proof of Theorem 3. □

7.3 Proof of Theorem 5

We first prove the following claim:

Claim 1. Under condition (13) that $\prod_{j\in\mathcal{A}_{g,m}} d_j \geq K$ in the theorem, the following matrix $\widetilde{\mathbf{\Lambda}}_{g,m}$ has full column rank K for generic parameters,

$$\widetilde{\mathbf{\Lambda}}_{g,m} = \bigodot_{j\in\mathcal{A}_{g,m}} \mathbf{\Lambda}_j.$$

The $\widetilde{\mathbf{\Lambda}}_{g,m}$ above has the same definition as that in Theorem 3. The proof of this claim is similar in spirit to that of Lemma 13 in Allman et al. (2009). Note that the statement that the $\prod_{j\in\mathcal{A}_{g,m}} d_j \times K$ matrix $\widetilde{\mathbf{\Lambda}}_{g,m}$ does not have full column rank is equivalent to the statement that the maps sending $\widetilde{\mathbf{\Lambda}}_{g,m}$ to its $K \times K$ minors are all zero maps. There are

$$\binom{\prod_{j\in\mathcal{A}_{g,m}} d_j}{K}$$

such maps, and each of this map is a polynomial with indeterminants $\lambda_{j,c_j,k}$'s. To show that $\widetilde{\mathbf{\Lambda}}_{g,m}$ has full column rank K for generic parameters, we just need to show that these maps are not all zero polynomials. According to the property of the polynomial maps, it indeed suffices to find one particular set of $\{\mathbf{\Lambda}_j;\, j \in \mathcal{A}_{g,m}\}$ such that the resulting Khatri-Rao product $\widetilde{\mathbf{\Lambda}}_{g,m}$ has full column rank.

Consider a set of distinct prime numbers denoted by $\{a_{j,c};\, j = 1,\ldots,p,\, c = 1,\ldots,d_j\}$. Define

$$\mathbf{\Lambda}_j^\star = \begin{pmatrix} 1 & a_{j,1} & a_{j,1}^2 & \cdots & a_{j,1}^{K-1} \\ 1 & a_{j,2} & a_{j,2}^2 & \cdots & a_{j,2}^{K-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & a_{j,d_j} & a_{j,d_j}^2 & \cdots & a_{j,d_j}^{K-1} \end{pmatrix}, \tag{28}$$

then $\mathbf{\Lambda}_j^\star$ is a $d_j \times K$ Vandermonde matrix. Generally, for a d-dimensional vector $\boldsymbol{b}$, let $\mathrm{VDM}(\boldsymbol{b}) = \mathrm{VDM}(b_1,\ldots,b_d)$ denote the the $d \times d$ Vandermonde matrix with the (i,c)th entry being b_i^{c-1}, so the $\mathbf{\Lambda}_j^\star$ defined in (28) can be written as $\mathbf{\Lambda}_j^\star = \mathrm{VDM}(a_{j,1},\ldots,a_{j,d_j})$. Now consider a $\widetilde{\mathbf{\Lambda}}_{g,m}^\star$ defined as

$$\widetilde{\mathbf{\Lambda}}_{g,m}^\star = \bigodot_{j\in\mathcal{A}_{g,m}} \mathbf{\Lambda}_j^\star.$$

Under the assumption (13) in the theorem that $\prod_{j\in\mathcal{A}_{g,m}} d_j \geq K$, the K columns of $\widetilde{\mathbf{\Lambda}}_{g,m}$ are indeed the first K columns in the following Vandermonde matrix

$$\mathbf{V}_{g,m} = \mathrm{VDM}\left(\prod_{j\in\mathcal{A}_{g,m}} a_{j,1},\ldots,\prod_{j\in\mathcal{A}_{g,m}} a_{j,d_j}\right).$$

Since by construction the $a_{j,c}$'s are distinct prime numbers, for each $j \in S_m$ the d_j products $\prod_{j\in\mathcal{A}_{g,m}} a_{j,1},\ \ldots,\ \prod_{j\in\mathcal{A}_{g,m}} a_{j,d_j}$ are also distinct numbers. Therefore the $\mathbf{V}_{g,m}$ defined

above has full rank $\prod_{j\in S_m} d_j$. Since $\prod_{j\in\mathcal{A}_{g,m}} d_j \geq K$ and $\widetilde{\mathbf{\Lambda}}^{\star}_{g,m}$ has columns from the first K columns of $\mathbf{V}_{g,m}$, we have that $\widetilde{\mathbf{\Lambda}}^{\star}_{g,m}$ has full column rank K for this particular choice of parameters. Note that the $\mathbf{\Lambda}^{\star}_j$ defined in (28) does not have each column summing up to one, as is required in the parameterization of the probability tensor. But performing a positive rescaling of the each column of $\mathbf{\Lambda}^{\star}_j$ to a conditional probability table $\mathbf{\Lambda}_j$ would not change the above reasoning and conclusion about matrix rank; so we have proved the earlier Claim 1 that each $\widetilde{\mathbf{\Lambda}}_{g,m}$ has full column rank K for generic parameters. Given this conclusion, for generic parameters in the parameter space the situation is reduced back to that under Theorem 3. So the identifiability condition in Theorem 2 carries over, and we can obtain the conclusion that identifiability holds here for generic parameters. This completes the proof of Theorem 5. □

7.4 Proof of Proposition 6

Recall that under the conditions of the previous theorem, we already have the conclusion that $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ are identifiable. Next the question boils down to whether $(\alpha_1, \ldots, \alpha_K)$ are identifiable from $\mathbf{\Phi} = (\phi_{k_1,\ldots,k_G})$. By the definition, we have

$$\phi_{k_1,\ldots,k_G} = \mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_{k_1}\cdots\pi_{k_G}\right] = \int_{\Delta^{K-1}} \pi_{k_1}\cdots\pi_{k_G} d\mathrm{Dir}_{\boldsymbol{\alpha}}(\boldsymbol{\pi}).$$

First we consider the case of $G = 2$. Denote $\alpha_0 = \sum_{k=1}^{K}\alpha_k$. Then according to the moment property of the Dirichlet distribution, there is

$$\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_k\pi_\ell\right] = \begin{cases} \dfrac{\alpha_k\alpha_\ell}{\alpha_0(\alpha_0+1)}, & \text{if } k\neq\ell; \\ \dfrac{\alpha_k(\alpha_k+1)}{\alpha_0(\alpha_0+1)}, & \text{if } k=\ell. \end{cases}$$

Therefore for $k \neq \ell$, consider x and y defined as follows,

$$x := \frac{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_k^2\right]}{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_k\pi_\ell\right]} = \frac{\alpha_k+1}{\alpha_\ell},$$

$$y := \frac{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_\ell^2\right]}{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_k\pi_\ell\right]} = \frac{\alpha_\ell+1}{\alpha_k}.$$

Since x and y are already identified, then we can solve for α_k and α_ℓ as follows

$$\alpha_k = \frac{x+1}{xy-1}, \quad \alpha_\ell = \frac{y+1}{xy-1}.$$

Since the above reasoning holds for arbitrary pairs of (k,ℓ) with $k \neq \ell$, we have obtained the identifiability of the entire vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$.

Next we consider the general case of $G > 2$. For arbitrary $1 \leq k \neq \ell \leq K$, consider two sequences $(k, k, k_3, \ldots, k_G)$, $(k, \ell, k_3, \ldots, k_G) \in [K]^G$. According to the property of the

Dirichlet distribution, we have

$$\frac{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_k\pi_k\pi_{k_3}\cdots\pi_{k_G}\right]}{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_k\pi_\ell\pi_{k_3}\cdots\pi_{k_G}\right]}=\frac{\alpha_k+1}{\alpha_\ell}, \tag{29}$$

$$\frac{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_\ell\pi_\ell\pi_{k_3}\cdots\pi_{k_G}\right]}{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\pi_k\pi_\ell\pi_{k_3}\cdots\pi_{k_G}\right]}=\frac{\alpha_\ell+1}{\alpha_k}. \tag{30}$$

Now that the left hand sides of the above two equations are identified by the previous theorem, we denote them by $u :=$ LHS of (29) and $v :=$ LHS of (30). The u and v are identified constants. Solving for α_k and α_ℓ gives

$$\alpha_k=\frac{u+1}{uv-1},\quad \alpha_\ell=\frac{v+1}{uv-1}.$$

Since k,ℓ are arbitrary, we have shown that the entire vector $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_K)$ is identifiable. This completes the proof of Proposition 6. □

7.5 Proof of Supporting Lemmas

Proof [Proof of Lemma 8] First note that the $\bigodot_{j\in S}\boldsymbol{\Gamma}_j$ on the right hand side of (19) has size $\prod_{j\in S}d_j\times K^G$. Further, since $\{s_j : j\in S\}\supseteq[G]$, the set $\{j\in S : s_j=g\}$ is nonempty. So the $\bigodot_{j\in S:\,s_j=g}\boldsymbol{\Lambda}_j$ has K columns and hence the left hand side of (19) also has size $\prod_{j\in S}d_j\times K^G$. Without loss of generality, suppose $S=\{1,2,\ldots,|S|\}$, where $|S|$ denotes the cardinality of the set S. The $(c_1+(c_2-1)d_1+\cdots+(c_{|S|}-1)d_1\cdots d_{|S|-1},\ z_1+(z_2-1)K+\cdots+(z_G-1)K^{G-1})$th entry of the RHS of (19) is $\prod_{j\in S}\gamma_{j,c_j,\boldsymbol{z}}$, which by definition equals $\prod_{j\in S}\lambda_{j,c_j,z_{s_j}}=\prod_{g=1}^G\prod_{j\in S:\,s_j=g}\lambda_{j,c_j,z_g}$; this is exactly the $(c_1+(c_2-1)d_1+\cdots+(c_{|S|}-1)d_1\cdots d_{|S|-1},\ z_1+(z_2-1)K+\cdots+(z_G-1)K^{G-1})$th entry of the LHS of (19). This completes the proof of Lemma 8. ■

Proof [Proof of Lemma 9] First note that both hand sides of (22) are vectors of size $\prod_{j=1}^p d_j\times 1$. To see this for the right hand side of (22), note the matrix $\bigodot_{j:\,\ell_{j,g}=1}\boldsymbol{\Lambda}_j$ has size $\prod_{j:\,\ell_{j,g}=1}d_j\times K^{\sum_{j=1}^p\ell_{j,g}}$, and hence the matrix $\bigotimes_{g=1}^G\bigodot_{j:\,\ell_{j,g}=1}\boldsymbol{\Lambda}_j$ has size

$$\prod_{g=1}^G\prod_{j:\,\ell_{j,g}=1}d_j\times K^{\sum_{g=1}^G\sum_{j=1}^p\ell_{j,g}},$$

which is just $\prod_{j=1}^p d_j\times K^G$. Further note that the vector $\mathrm{vec}(\boldsymbol{\Phi})$ has size $K^G\times 1$, so the $\left\{\bigotimes_{g=1}^G\bigodot_{j:\,\ell_{j,g}=1}\boldsymbol{\Lambda}_j\right\}\mathrm{vec}(\boldsymbol{\Phi})$ on the right hand side of (22) has size $\prod_{j=1}^p d_j\times 1$, matching the size of the left hand side. Next consider the individual entries of both hand sides of (22). First, by definition of the vec() operator, the $[c_1+(c_2-1)d_1+\cdots+(c_p-1)d_1\cdots d_{p-1}]$-th entry of the left hand side of (22) is $\pi_{c_1,\ldots,c_p}$. Next, according to (20), the $\pi_{c_1,\ldots,c_p}$ can be

written in the following way,

$$\begin{aligned}
\pi_{c_1,\ldots,c_p} &= \sum_{z_1=1}^{K} \cdots \sum_{z_G=1}^{K} \phi_{z_1,\ldots,z_G} \prod_{j=1}^{p} \gamma_{j,c_j,\boldsymbol{z}} \\
&= \sum_{z_1=1}^{K} \cdots \sum_{z_G=1}^{K} \mathrm{vec}(\boldsymbol{\Phi})_{z_1+(z_2-1)K+\cdots+(z_G-1)K^{G-1}} \times \prod_{g=1}^{G} \prod_{j:\,\ell_{j,g}=1} \lambda_{j,c_j,z_g} \\
&= \sum_{z_1=1}^{K} \cdots \sum_{z_G=1}^{K} \mathrm{vec}(\boldsymbol{\Phi})_{z_1+(z_2-1)K+\cdots+(z_G-1)K^{G-1}} \\
&\qquad \times \left\{ \bigotimes_{g=1}^{G} \bigodot_{j:\,\ell_{j,g}=1} \boldsymbol{\Lambda}_j \right\}_{c_1+(c_2-1)d_1+\cdots+(c_p-1)d_1\cdots d_{p-1},\, z_1+(z_2-1)K+\cdots+(z_G-1)K^{G-1}} \\
&= \mathrm{vec}(\boldsymbol{\Phi})^\top \cdot \left\{ \bigotimes_{g=1}^{G} \bigodot_{j:\,\ell_{j,g}=1} \boldsymbol{\Lambda}_j \right\}^\top_{c_1+(c_2-1)d_1+\cdots+(c_p-1)d_1\cdots d_{p-1},\,:} .
\end{aligned}$$

The last row in the above display exactly equals the $[c_1+(c_2-1)d_1+\cdots+(c_p-1)d_1\cdots d_{p-1}]$-th entry of the RHS of (22). This proves the equality in (22) and completes the proof of Lemma 9. ■

Proof [Proof of Lemma 10] In the LHS of (23), the term $\boldsymbol{\Lambda}_{j_g^{(1)}} \bigodot \boldsymbol{\Lambda}_{j_g^{(2)}}$ has size $d_{j_g^{(1)}} d_{j_g^{(2)}} \times K$ and hence the Kronecker product $\bigotimes_{g=1}^{G} \left\{ \boldsymbol{\Lambda}_{j_g^{(1)}} \bigodot \boldsymbol{\Lambda}_{j_g^{(2)}} \right\}$ has size $\prod_{j\in S^{(1)}\cup S^{(1)}} d_j \times K^G$. In the RHS of (23), the term $\left\{ \bigotimes_{g=1}^{G} \boldsymbol{\Lambda}_{j_g^{(1)}} \right\}$ has size $\prod_{g=1}^{G} d_{j_g^{(1)}} \times K^G$, and hence the Khatri-Rao product of two such terms

$$\left\{ \bigotimes_{g=1}^{G} \boldsymbol{\Lambda}_{j_g^{(1)}} \right\} \bigodot \left\{ \bigotimes_{g=1}^{G} \boldsymbol{\Lambda}_{j_g^{(2)}} \right\}$$

has size $\left(\prod_{g=1}^{G} d_{j_g^{(1)}} d_{j_g^{(2)}} \right) \times K^G$. So both hand sides of (23) has size $\prod_{j\in S^{(1)}\cup S^{(2)}} d_j \times K^G$. The equality (23) can be similarly shown as in the proof of Lemma 9 by writing out and checking the individual elements of the two matrices on the LHS and RHS of (23).

Similarly, the LHS and RHS of (24) both have size $\prod_{j\in S^{(1)}\cup S^{(1)}\cup \left\{ j_G^{(3)} \right\}} d_j \times K^G$ and the equality can be similarly shown as in the proof of Lemma 9. ■

8. Supplement B: Pairwise Cramer's V between Categorical Variables

According to the definition of mutual information in information theory, for two discrete variables $y_j \in [d_j]$ and $y_m \in [d_m]$, their Cramer's V is

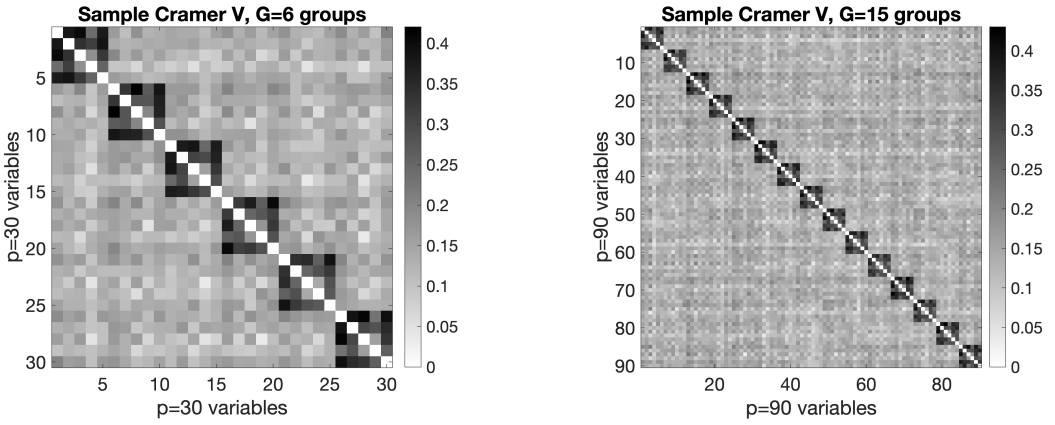
$$\mathrm{CRV}(y_j, y_m) = \left\{ \frac{1}{\min(d_j,\ d_m)} \sum_{c_1 \in [d_j]} \sum_{c_2 \in [d_m]} \frac{\Big(p_{(y_j,y_m)}(c_1,c_2) - p_{y_j}(c_1) p_{y_m}(c_2)\Big)^2}{p_{y_j}(c_1) p_{y_m}(c_2)} \right\}^{1/2} \tag{31}$$

where $p_{y_j}(c_1) = \mathbb{P}(y_j = c_1)$ denotes the marginal distribution of y_j and $p_{(y_j,y_m)}(c_1, c_2) = \mathbb{P}(y_j = c_1, y_m = c_2)$ denotes the joint distribution of y_j and y_m. The Cramer's V measures the the inherent dependence expressed in the joint distribution of two variables relative to their marginal distributions under the independence assumption. Therefore, Cramer's V measures the dependence between variables and it equals zero if and only of the two variables are independent; otherwise Cramer's V is positive.

The expression of Cramer's V in (31) is the population version. Given a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ with $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,p})$, the population quantities of the marginal and joint distributions in (31) can be replaced by their sample estimates. That is, the previous $p_{y_j}(c_1)$ and $p_{(y_j,y_m)}(c_1, c_2)$ are replaced by the following,

$$p^{\mathrm{samp}}_{y_j}(c_1) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y_{i,j} = c_1), \quad p^{\mathrm{samp}}_{(y_j,y_m)}(c_1, c_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y_{i,j} = c_1, y_{i,m} = c_2).$$

Using the sample-based Cramer's V measure, we calculate the Cramer's V for all the pairs of variables when j and m each range from 1 to p. For two randomly chosen simulated datasets from the simulations settings $p = 30,\ G = 6,\ K = 3,\ n = 1000$ and $p = 90,\ G = 15,\ K = 3,\ n = 1000$ described in Section 5 in the main text, their pairwise Cramer's V plots are displayed in Figure 12.



(a) $p = 30,\ G = 6,\ K = 3,\ n = 1000.$ (b) $p = 90,\ G = 15,\ K = 3,\ n = 1000.$

Figure 12: Cramer's V of item pairs for two simulated datasets.

By visual inspection, Figure 12 shows a block-diagonal structure of the $p \times p$ pairwise Cramer's V matrix for both simulation settings. In each of these settings, the true grouping matrix $\mathbf{L}$ used to generate data takes the form that the first p/G variables belong to a same group, the second p/G variables belong to another same group, etc. Therefore, Figure 12 implies in the simulations, the variables belonging to the same group tend to show higher dependence than those variables belonging to different groups.