Dimension-Grouped Mixed Membership Models for Multivariate Categorical Data

Yuqi Gu, Elena Erosheva, Gongjun Xu, David Dunson

Dimension-Grouped Mixed Membership Models for Multivariate Categorical Data

Yuqi Gu[1], Elena A. Erosheva[2], Gongjun Xu[3], and David B. Dunson[4]

[1]Department of Statistics, Columbia University
[2]Department of Statistics, School of Social Work, and the Center for Statistics and the Social Sciences, University of Washington
[3]Department of Statistics, University of Michigan
[4]Department of Statistical Science, Duke University

Dimension-Grouped Mixed Membership Models for Multivariate Categorical Data

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 an application to a functional disability 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 (Pritchard et al., 2000), biological and social networks (Airoldi et al., 2008), 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 assumption 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 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. The technical proofs of all the theoretical results are contained in the Supplementary Material.

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 p here 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) 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 mea-

surements 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

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

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. To see this, note topic models such as LDA are also called "bag of words" models, since given the individual topic proportion vector $\boldsymbol{\pi}_i$ of each document i, all the R words in the document have identical conditional distributions and hence are *exchangeable* replications. This means there is one single set of K topic-word distributions $\{\boldsymbol{\lambda}_k : k = 1, \ldots, K\}$, for which multiple words per document are observed. While in the considered GoM, each variable $j \in \{1, \ldots, p\}$ has its own set of K conditional distributions indexed by $\{\boldsymbol{\lambda}_{j,k} : k = 1, \ldots, K\}$, for each of which we observe *only one* response $y_{i,j}$ per subject i. This fact is made clear also in Figure 1(b), where for each $j \in [p]$ there is a population quantity, the parameter node $\boldsymbol{\Lambda}_{j,:,:}$ (also denoted by $\boldsymbol{\Lambda}_j$ for simplicity), that governs its distribution. Thus identifiability is a much greater challenge for GoM models.

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,

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

For each subject i, assuming $\boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,K}) \sim$ Distribution $D_{\boldsymbol{\alpha}}$, our key specification is the following generative process,

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

Hence, given the population parameters $(\mathbf{L}, \boldsymbol{\Lambda}, \boldsymbol{\alpha})$, the probability distribution of $\boldsymbol{y}_i$ can be written as

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

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.

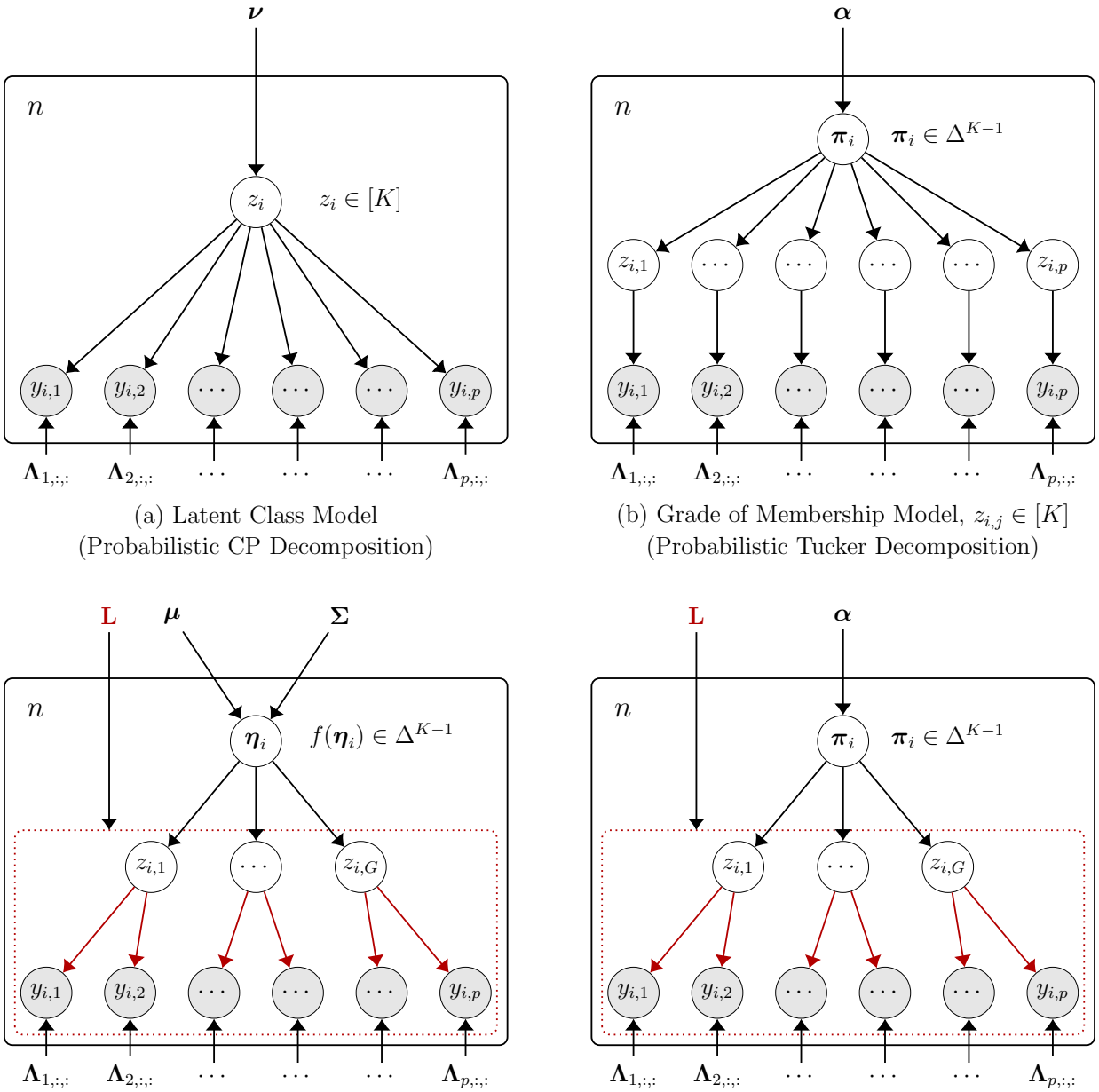
We visualize the proposed model as a probabilistic graphical model to highlight connections to and differences from existing latent structure models for multivariate categorical data. In Figure 1, we show the graphical model representations of two popular latent structure models for multivariate categorical data in (a) and (b), and for the newly proposed Gro-M^3 in (c) and (d). The $\boldsymbol{\Lambda}_j$ for $j \in [p]$ denotes a $d_j \times K$ matrix with entries $\lambda_{j,c_j,k}$. Each column of $\boldsymbol{\Lambda}_j$ 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 $\boldsymbol{\Lambda}_j$ 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 $\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_p$. 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 helping to establish model identifiability (see Section 3), our dimension-grouping modeling assumption is also motivated by real-world applications. In Section 6, we apply our methodology to a widely analyzed survey dataset, the NLTCS functional disability data initially analyzed with the grade of membership (GoM) model Erosheva et al. (2007). To gain some insight into the possible dependence structures among the $p = 16$ item responses, we calculate the sample mutual information measures between each pair of items. Figure 2 plots the pairwise mutual information matrix for the NLTCS data. By visual inspection of the left plot in Figure 2, variables 1-6, all belonging to the "Activities of Daily Living" type of survey questions, have stronger dependence among themselves than with other variables, and variables 7-11 form a similar block. Details on calculating the pairwise mutual information between categorical variables are provided in the Supplementary Material, and there we also plot pairwise mutual information matrices for datasets simulated under our model. The simulated data pairwise mutual information plots show a somewhat similar grouping pattern to that for the NLTCS data. We conjecture that many real world datasets in other applied domains exhibit similar grouped dependence structures.

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 $\{p\left(y_{i,1} = c_1, \ldots, y_{i,p} = c_p \mid -\right);\, c_j \in [d_j] \text{ for each } j \in [p]\}$ 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



(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 $\mathbf{L}$.

and they sum up to one, so we call it a *probability tensor*. We next describe in detail the tensor decomposition perspective of our model; such a perspective will turn out to be useful

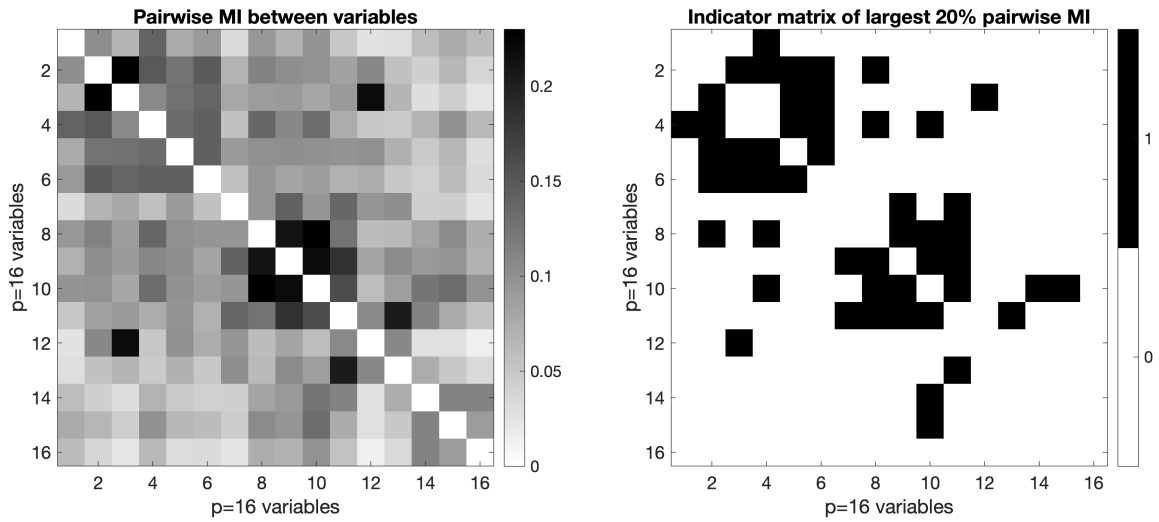


Figure 2: Pairwise Mutual Information (MI) measures between the $p = 16$ item responses for the NLTCS disability survey data. See Section 6 for details of the dataset, including the specific content of each item in Figure 4. The left plot shows the calculated MI for each pair of items, while the right plot shows the binary indicator matrix, each entry indicating whether the corresponding pair of items has the largest 20% MI among all the pairs.

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,

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

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

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 the sum of its entries is one. Viewed from a tensor decomposition perspective, this is the popular Tucker decomposition (Tucker, 1966); more specifically this is the non-negative and probabilistic version of the Tucker decomposition. The $\boldsymbol{\Phi}^{\mathrm{GoM}}$ represents the Tucker tensor core, and the product of the $\{\lambda_{j,c_j,k}\}$ 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

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

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 $\boldsymbol{\nu}^{\text{LC}} = (\nu_k;\, k \in [K])$, 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,

$$\begin{aligned} p^{\text{Gro-M}^3}\left(y_{i,1},\ldots,y_{i,p} \mid \mathbf{L},\boldsymbol{\Lambda},\boldsymbol{\alpha}\right) &= \int_{\Delta^{K-1}}\prod_{g=1}^{G}\left[\sum_{k=1}^{K}\pi_{i,k}\prod_{j:\,\ell_{j,g}=1}f(y_{i,j}\mid\lambda_{j,c_j,k})\right]dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i) \\ &= \sum_{k_1=1}^{K}\cdots\sum_{k_p=1}^{K}\prod_{g=1}^{G}\prod_{j:\,\ell_{j,g}=1}f(y_{i,j}\mid\lambda_{j,c_j,k_j})\underbrace{\int_{\Delta^{K-1}}\pi_{i,k_1}\cdots\pi_{i,k_G}dD_{\boldsymbol{\alpha}}(\boldsymbol{\pi}_i)}_{=:\,\phi^{\text{Gro-M}^3}_{k_1,\ldots,k_G}}. \end{aligned} \quad (7)$$

In this case, $\boldsymbol{\Phi}^{\text{Gro-M}^3} := \left(\phi^{\text{Gro-M}^3}_{k_1,\ldots,k_G};\; k_g \in [K]\right)$ 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 $\boldsymbol{\Phi}$, and did not consider identifiability. Nonetheless and interestingly, our identifiability results can be applied to establish identifiability of c-Tucker decomposition (see Remark 1 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 $R > 1$).

Recently, Russo et al. (2021) proposed a class of multivariate mixed membership models. Their model is motivated by applications in which it is natural to partition variables into groups prior to analysis. For example, in a malaria risk study based on household survey data on the Brazilian Amazon frontier, there are two groups of variables, corresponding to environmental and behavioral factors. Within each group of variables $b \in \{1, \ldots, G\}$, Russo et al. (2021) endow each subject i with a group-specific mixed-membership proportion vector $\boldsymbol{\pi}_i^{(b)}$. Our dimension-grouping assumption is different in the following two aspects: each subject is still endowed with a single vector of membership scores instead of G different vectors, and the grouping information is not fixed in advance but is treated as unknown here.

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 groupings 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. 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 1. *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 1 hold, so* $\mathbf{\Lambda}$*,* $\mathbf{L}$ *and* $\mathbf{\Phi}$ *are identifiable. Theorem 1 implies that if* $\mathbf{L}$ *contains any additional row vectors other than those in (9) the model is still identifiable.*

Theorem 1 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 1 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 1 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 2 exploits the grouping structure in $\mathbf{L}$ to relax the identifiability conditions in the previous Theorem 1.

Theorem 2. *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 1, Theorem 2 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

$$\boldsymbol{\Lambda}_1 \bigodot \boldsymbol{\Lambda}_2 = \begin{pmatrix} a_1b_1 & a_2b_2 & a_3b_3 \\ a_1(1-b_1) & a_2(1-b_2) & a_3(1-b_3) \\ (1-a_1)b_1 & (1-a_2)b_2 & (1-a_3)b_3 \\ (1-a_1)(1-b_1) & (1-a_2)(1-b_2) & (1-a_3)(1-b_3) \end{pmatrix}.$$

Indeed, $\boldsymbol{\Lambda}_1 \bigodot \boldsymbol{\Lambda}_2$ *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* $\boldsymbol{\Lambda}_1 \bigodot \boldsymbol{\Lambda}_2$ *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* $(\boldsymbol{\Lambda}_1 \bigodot \boldsymbol{\Lambda}_2)\boldsymbol{x} = 0$*, then*

$$\begin{cases} a_1b_1x_1 + a_2b_2x_2 + a_3b_3x_3 = 0; \\ a_1(1-b_1)x_1 + a_2(1-b_2)x_2 + a_3(1-b_3)x_3 = 0; \\ (1-a_1)b_1x_1 + (1-a_2)b_2x_2 + (1-a_3)b_3x_3 = 0; \\ (1-a_1)(1-b_1)x_1 + (1-a_2)(1-b_2)x_2 + (1-a_3)(1-b_3)x_3 = 0; \end{cases}$$

$$\overset{\text{invertible transform}}{\Longleftrightarrow} \quad \begin{cases} a_1b_1x_1 + a_2b_2x_2 + a_3b_3x_3 = 0; \\ a_1x_1 + a_2x_2 + a_3x_3 = 0; \\ b_1x_1 + b_2x_2 + b_3x_3 = 0; \\ x_1 + x_2 + x_3 = 0. \end{cases}$$

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 $\{(a_1, a_2, a_3, b_1, b_2, b_3) \in [0,1]^6 :$ *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* $(\boldsymbol{\Lambda}_1 \bigodot \boldsymbol{\Lambda}_2)\boldsymbol{x} = \mathbf{0}$ *implies* $\boldsymbol{x} = \mathbf{0}$ *which means* $\boldsymbol{\Lambda}_1 \bigodot \boldsymbol{\Lambda}_2$ *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 2 are much weaker than those in Theorem 1 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 2, 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 2 (instead of Theorem 1) 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 2 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 2 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 2. 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 2 (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 $(\mathbf{\Lambda}, \mathbf{\Phi}) \in \mathcal{T} \backslash \mathcal{N}$ 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 2 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 3, we propose simple conditions to ensure generic identifiability of Gro-M^{3}s.

Theorem 3. *For the notation $\mathcal{A}_g = \{j \in [p] : \ell_{j,g} = 1\}$ defined in Theorem 2, 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 2, Theorem 3 lifts the explicit full-rank requirements on *any* matrix. Rather, Theorem 3 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), 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 1–3 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 1. *Consider a Dirichlet Gro-M^3. If $G \geq 2$, then following conclusions hold.*

(a) If the conditions in Theorem 1 or Theorem 2 are satisfied, then the Dirichlet parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are strictly identifiable.

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

Remark 1. *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 $\text{Categorical}([G],\, \xi_1, \ldots, \xi_G)$ 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 $\boldsymbol{\Lambda}_j$. 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 $\mathbf{Y} = \{\boldsymbol{y}_i;\, i = 1, \ldots, n\}$. Given $\mathbf{Y}$, we propose a Metropolis-Hastings-within-Gibbs sampler for posterior computation of $\mathbf{L}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\alpha}$. The MCMC sampler cycles through the following steps.

Step 1–3. Sample each column of the conditional probability tables $\boldsymbol{\Lambda}_j$'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 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. The conditional posterior distribution of $\boldsymbol{\alpha}$ (or equivalently, α_0 and $\boldsymbol{\eta}$) is

$$\begin{aligned}
p(\boldsymbol{\alpha} \mid -) &\propto \text{Gamma}(\alpha_0 \mid a, b) \times \text{Dirichlet}(\boldsymbol{\eta} \mid \mathbf{1}_K) \times \prod_{i=1}^{n} \text{Dirichlet}(\boldsymbol{\pi}_i \mid \boldsymbol{\alpha}) \\
&\propto \alpha_0^{a_\alpha - 1} \exp(-b_\alpha \alpha_0) \times \left[\frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}\right]^n \times \prod_{k=1}^{K} \left[\prod_{i=1}^{n} \pi_{i,k}\right]^{\alpha_k},
\end{aligned}$$

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 $r = \min\{1,\ r^\star\}$.

After collecting posterior samples from the output of the above 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 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 setting

as considered in our identifiability theory as well 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 2 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 $\mathbf{\Lambda}$ 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 3, 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 3 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 3 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 3 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 $(\mathbf{\Lambda}, \boldsymbol{\alpha})$ summarized across 50 simulation replicates in each setting. For continuous parameters $(\mathbf{\Lambda}, \boldsymbol{\alpha})$, we calculate their Root Mean Squared Errors (RMSEs) to evaluate the estimation accuracy. To obtain the estimation error of $(\mathbf{\Lambda}, \boldsymbol{\alpha})$ after collecting posterior samples, we need to find an appropriate permutation of the K extreme latent profiles in order to compare the $(\overline{\mathbf{\Lambda}}, \overline{\boldsymbol{\alpha}})$ and the true $(\mathbf{\Lambda}, \boldsymbol{\alpha})$. To this end, we first reshape each of $\overline{\mathbf{\Lambda}}$ and $\mathbf{\Lambda}$ to a $(\sum_{j=1}^{p} d_j) \times K$ matrix $\overline{\mathbf{\Lambda}}_{\text{mat}}$ and $\mathbf{\Lambda}_{\text{mat}}$, calculate the inner product matrix $(\mathbf{\Lambda}_{\text{mat}})^{\top} \overline{\mathbf{\Lambda}}_{\text{mat}}$, 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

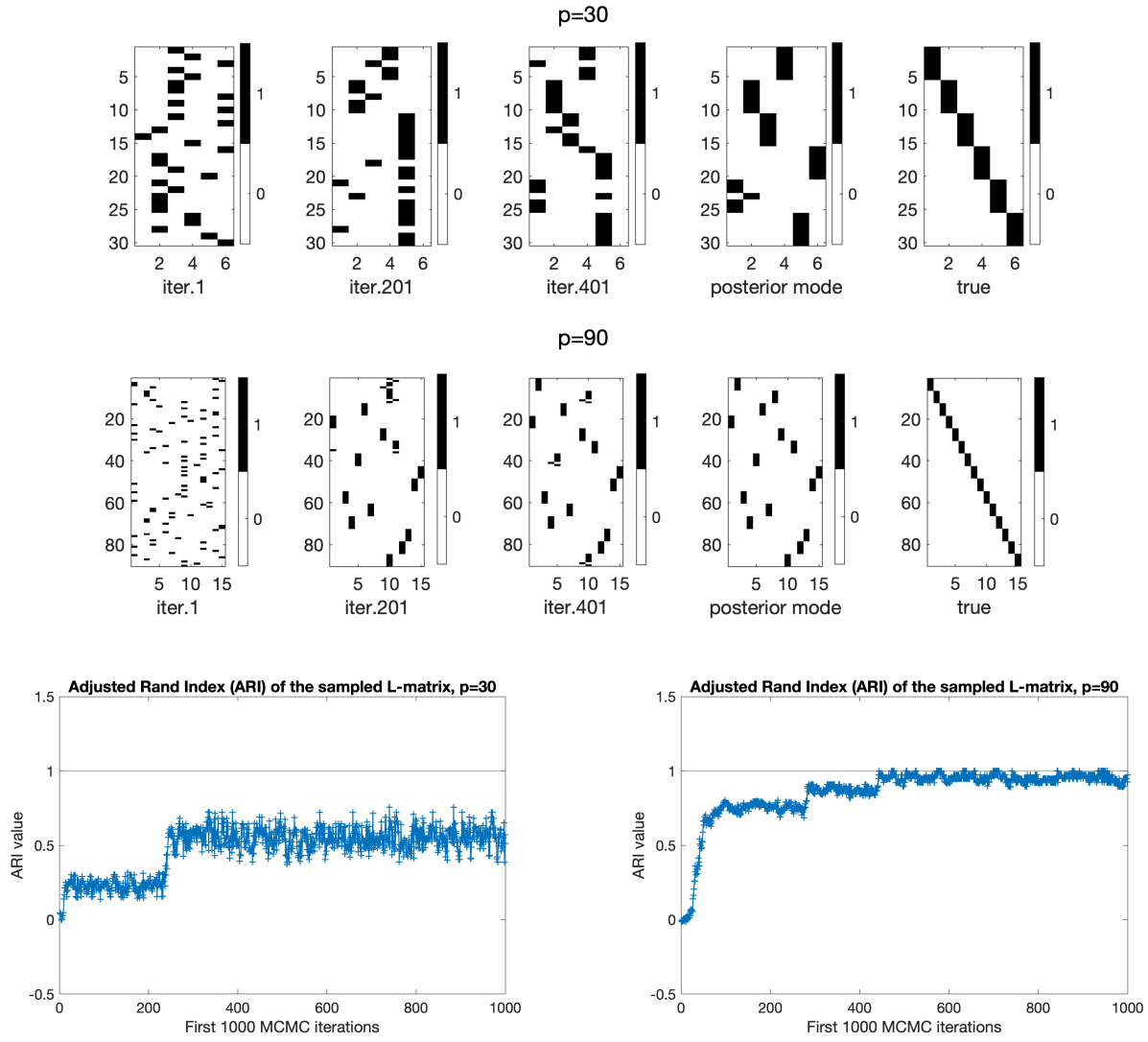


Figure 3: Estimation of $\mathbf{L}$ (from $\boldsymbol{s}$) in two random simulation trials, one under $(n, p, G, K) = (500, 30, 6, 2)$ and the other under $(n, p, G, K) = (500, 90, 15, 2)$. In each of the first two rows, the left three plots record the sampled $\mathbf{L}_{\text{iter.}}$ after the 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.

$(i_1, \ldots, i_K)$ 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.

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 $\mathbf{\Lambda}$ 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 accurate 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.

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 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

	$(p,\, G)$	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.

	$(p,\, G)$	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.

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. This simulation study shows 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 Application to 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 dataset was downloaded from `NLTCS` at `http://lib.stat.cmu.edu/datasets/`. It is an extract from the NLTCS 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. 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 4. The estimated variable groupings are given in Figure 4. Out of the $G^\star = 10$ groups, there are three groups that contain multiple items. In Figure 4, 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 ralated to traditional gender roles – these items correspond to activities performed more frequently by women than by men. Finally, the

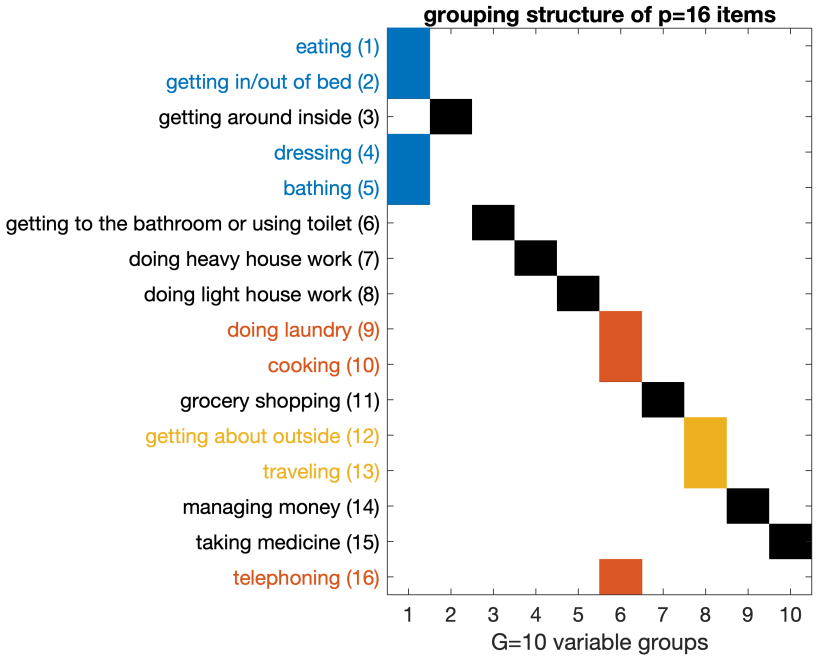


Figure 4: Estimated variable grouping structure $\boldsymbol{s}$ (i.e., $\mathbf{L}$) for the NLTCS data with $(G^\star, K^\star) = (10, 9)$. The first six items are ADL "activities of daily living" and the remaining ten items are IADL "instrumental activities of daily living". Out of the $G^\star = 10$ 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.

two items $j = 12$ "getting about outside" and $j = 13$ "traveling" that require high level of mobility form another group. The variable grouping structure we obtained using the Gro-M^3 model partly resembles the dependence structure shown in the previous Figure 2 on pairwise mutual information. We emphasize that such a model-based grouping of the items could not have been obtained applying previous mixed membership modeling approaches to the NLTCS disability data (Erosheva et al., 2007).

In addition to the variable grouping structures, we plot posterior means of the positive response probabilities $\overline{\mathbf{\Lambda}}_{:,1,:}$ in Figure 5 for the selected model. For each survey item $j \in [p]$ and each extreme latent profile $k \in [K]$, the $\mathbf{\Lambda}_{j,1,k}$ records the conditional probability of giving a positive response of being disabled on this item conditional on possessing the kth 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 5, 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.

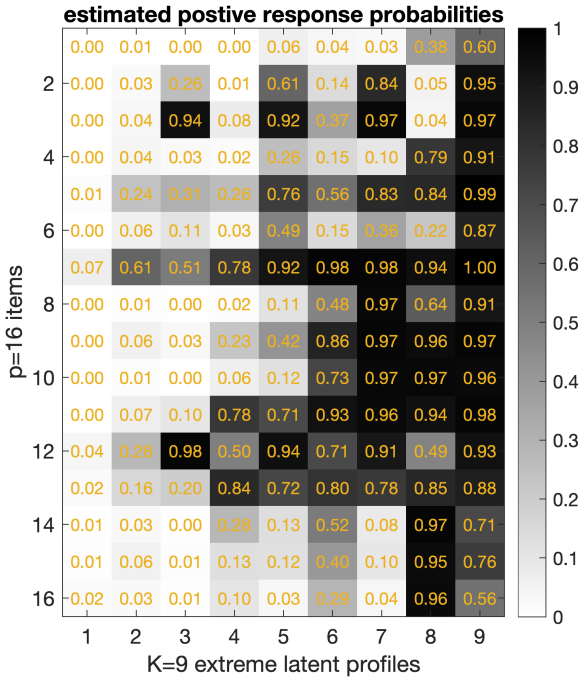


Figure 5: Estimated positive response probabilities $\boldsymbol{\Lambda}_{:,1,:}$ for the NLTCS data with $(G^\star, K^\star) = (10, 9)$. Each column represents one extreme latent profile. Entries are conditional probabilities of giving a positive response (1 = disabled) to each item given that latent profile.

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 (correspond-

ing 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.

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).

References

Airoldi, E. M., Blei, D., Erosheva, E. A., and Fienberg, S. E. (2014). *Handbook of mixed membership models and their applications*. CRC press.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

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

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y.-K. (2012). A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925.

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

Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models–going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE.

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

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

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

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

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

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

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

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

Govaert, G. and Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

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

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

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

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

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

Kruskal, J. B. (1977). 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.

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

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

Mao, X., Sarkar, P., and Chakrabarti, D. (2020). Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, pages 1–13.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

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

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

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

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

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

Russo, M., Singer, B. H., and Dunson, D. B. (2021). Multivariate mixed membership modeling: Inferring domain-specific risk profiles. *Annals of Applied Statistics*, Online.

Shang, Z., Erosheva, E. A., and Xu, G. (2021). Partial-mastery cognitive diagnosis models. *The Annals of Applied Statistics, to appear*.

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

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

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

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

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

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

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