Conditional probability tensor decompositions for multivariate categorical response regression

Aaron J. Molstad[†*] and Xin Zhang[‡*]
[†]University of Minnesota and [‡]Florida State University

Abstract

In many modern regression applications, the response consists of multiple categorical random variables whose probability mass is a function of a common set of predictors. In this article, we propose a new method for modeling such a probability mass function in settings where the number of response variables, the number of categories per response, and the dimension of the predictor are large. Our method relies on a functional probability tensor decomposition: a decomposition of a tensor-valued function such that its range is a restricted set of low-rank probability tensors. This decomposition is motivated by the connection between the conditional independence of responses, or lack thereof, and their probability tensor rank. We show that the model implied by such a low-rank functional probability tensor decomposition can be interpreted in terms of a mixture of regressions and can thus be fit using maximum likelihood. We derive an efficient and scalable penalized expectation maximization algorithm to fit this model and examine its statistical properties. We demonstrate the encouraging performance of our method through both simulation studies and an application to modeling the functional classes of genes.

Keywords: Dimension reduction; Generalized linear model; Latent variable model; Mixture regression; Tensor decomposition.

1 Introduction

We consider the problem of modeling the conditional distribution of multiple categorical response variables as a function of a p-dimensional vector of predictors, i.e., a multivariate categorical response regression analysis. Joint modeling of multiple, dependent categorical responses is crucial in applications such as characterizing the genetic basis of disease subtypes (Dahl and Zaitlen,

*The authors contributed equally to this work and are listed in alphabetical order. Corresponding author: Aaron J. Molstad (amolstad@umn.edu). Research for this paper was supported in part by grants DMS-2053697 (XZ), DMS-2113590 (XZ) and DMS-2113589 (AJM) from the U.S. National Science Foundation.

2020), which are often defined as a cross-classification of many distinct categorical variables. For example, breast cancer tumors can be characterized by the presence estrogen receptor (ER), human epidermal growth factor receptor 2 (HER2), and progestoerone receptor protens (i.e., 2^3 total subtypes). To make matters concrete, for an integer $M \geq 2$, let $\mathbf{Y} = (Y_1, \ldots, Y_M)$ be the multivariate categorical response, where each component Y_m has $c_m \geq 2$ many categories with numerically coded support $Y_m \in [c_m] = \{1, \ldots, c_m\}$ for all $m \in [M] = \{1, \ldots, M\}$. The regression models and methods developed in this article allow the predictor $\mathbf{X} \in \mathbb{R}^p$ to be either continuous or discrete (or mixed), and allow the predictors $\mathbf{x} \in \mathbb{R}^p$ to be either random or fixed. The regression problem is essentially the study of the conditional distribution $\mathbf{Y} \mid \mathbf{X}$ whose joint probability mass function consists of

$$P_{j_1 \ldots j_M}(\mathbf{x}) = \Pr(Y_1 = j_1, \ldots, Y_M = j_M \mid \mathbf{X} = \mathbf{x}) \geq 0, \quad j_m \in [c_m], \;\; m \in [M]. \tag{1}$$

From (1), we can define the Mth order tensor $\mathbf{P}(\mathbf{x}) \in \mathbb{R}^{c_1 \times \cdots \times c_M}$ whose $(j_1, \ldots, j_M)$th element is $P_{j_1 \ldots j_M}(\mathbf{x})$. This conditional probability tensor $\mathbf{P}(\mathbf{x})$ fully characterizes the conditional distribution of $\mathbf{Y} \mid \mathbf{X} = \mathbf{x}$ and is thus the quantity of interest in our study.

To fit (1), there are numerous approaches one could consider. At one extreme, each response could be modeled separately using, say, multinomial logistic regression. This approach is scalable to a large number of responses and high-dimensional predictors (Zhu and Hastie, 2004; Simon et al., 2013; Vincent and Hansen, 2014), but entirely ignores the dependence between response variables. On the other extreme, one could define a univariate categorical response variable based on the set of all possible category combinations and methods designed for a univariate categorical response could be applied. This is the regression analog of modeling counts in a M-way contigency table as a multinomial random variable (Agresti, 1992, Section 1.2). For example, in applications with only binary responses, this would require treating M binary response variables as a univariate categorical variable with 2^M categories. In Section 6, we consider a genomic application with $M = 14$ binary response variables, so this approach would lead to an unwieldy $2^M = 16384$ categories. This approach allows for arbitrary dependence among responses, but in so doing, treats

$\mathbf{P}(\mathbf{x})$ as a vector and thus fails to exploit its special tensor structure. Moreover, this approach would require an enormous amount of data for model fitting. If even a single category combination is not observed in the training data, this approach cannot be applied directly. Finally, when the number of categorical responses is even moderately large, model interpretation will be difficult because the number of parameters grows exponentially with the number of responses.

For the problem of modeling the conditional probability tensor function $\mathbf{P}$, we refer to these two approaches as *separate* modeling (of each response) and *vectorized* modeling (of the combined-category response), respectively. The objective of this article is to propose an alternative to these two approaches; a method which can model complex dependencies among response variables like vectorized modeling, yet provides fitted models which can be computed and interpreted with the ease and scalability of separate modeling. Our approach exploits the connection between the conditional independence of $Y_1, \ldots, Y_M$, or lack thereof, and the rank of the conditional probability tensor $\mathbf{P}(\mathbf{x})$. Later, we prove that separate modeling implicitly assumes $\mathbf{P}(\mathbf{x})$ is rank one, whereas vectorized modeling assumes no explicit upper bound on the rank of $\mathbf{P}(\mathbf{x})$. We may thus characterize models for which $\mathbf{P}(\mathbf{x})$ is low rank as intermediate to these two extremes. Neatly, we later show that both the separate model and vectorized model can be characterized as "edge-cases" of a rank-constrained $\mathbf{P}(\mathbf{x})$ when the rank is fixed at one or the rank is allowed to be as large as $\prod_{m=1}^{M} c_m / \max_{k \in [M]} c_k$, respectively. Section 2.3 contains comparisons of these approaches.

Motivated by this observation, in this article we propose a new method for multivariate categorical response which assumes the low-rankness of $\mathbf{P}(\mathbf{x})$ for all $\mathbf{x}$. We show that pursuing a low-rank decomposition of the conditional probability tensor $\mathbf{P}(\mathbf{x})$ provides a natural, intuitive, and scalable way to model the complex dependencies among responses. However, because $\mathbf{P}(\mathbf{x})$ consists of the probabilities from (1), there are intrinsic constraints (nonnegativity and "sum-to-one") not often encountered in standard tensor decomposition problems. To handle these difficulties, we assume that the conditional probability tensor function $\mathbf{P}$ can be decomposed into the weighted sum of rank one probability tensor functions. This assumption naturally allows $\mathbf{P}$ to be characterized as a

mixture of regressions model, and implies the low-rankness of $\mathbf{P}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$. Moreover, by exploiting a latent variable interpretation of the probability tensor function, we can fit our model using penalized maximum likelihood, which can accommodate large p, large M, and large c_m.

Before formally describing our proposed method and model, we first situate our work among existing methods for fitting (1).

In the statistical literature on categorical data analysis, existing methods for multivariate categorical response regression are primarily focused on developing parametric links between predictors and multiple categorical responses that allow for model interpretation in terms of marginal probabilities and higher-order associations (Molenberghs and Lesaffre, 1999; Glonek, 1996; Ekholm et al., 2000; McCullagh and Nelder, 1989). For example, one set of link functions correspond to log-linear models, and another to multivariate logistic models (Glonek and McCullagh, 1995). Other approaches propose nonparametric regression functions and allow for some response-specific predictors (Gao et al., 2001). In general, these works adopt the vectorized modeling approach and are often not feasible when $M \geq 3$ and p is large, where associations among response variables are difficult to parameterize.

More recently, in the high-dimensional regime, Molstad and Rothman (2023) proposed a novel penalty to enforce linear restrictions on the regression coefficient tensor under the vectorized model with multinomial link. Their penalty can lead to fitted models that can be interpreted in terms of which predictors affect only the marginal distributions of responses, the log odds ratios, or neither. However, their method does not easily generalize to more than two response variables. Theoretically, the estimation error bound for their method scales exponentially with M (more specifically, scales in $\prod_{m=1}^{M} c_m$) rather than linearly (such as $\sum_{m=1}^{M} c_m$) as does our method.

Of course, others have recognized the need for models and methods specifically designed for multivariate categorical response. One class of methods, based on the notion of binary relevance, comes from the literature on *multi-label classification* (Tsoumakas and Katakis, 2007) in machine learning. Many binary relevance methods fit separate univariate models, and thus fail to account

for dependence in the multivariate response (Dembczyński et al., 2012; Montañes et al., 2014; Zhang et al., 2018). One binary relevance approach that accounts for dependence uses classifier chains (Read et al., 2011; Senge et al., 2013). This approach fits univariate categorical response regressions for $Y_1, Y_2, \ldots, Y_M$ successively. For each univariate fit, the responses from previous fits are used as predictors in subsequent fits. For example, one would fit $Y_1 \mid \mathbf{X}$, then $Y_2 \mid \mathbf{X}, Y_1$, and so on. The fitted models are thus typically interpreted in terms of specific (univariate) conditional distributions for each response, rather than the joint distribution of interest. There is also an extensive literature on *unsupervised* modeling for multivariate categorical data (Fienberg, 2000; Dunson and Xing, 2009; Bhattacharya and Dunson, 2012), but this is not applicable to regression. Finally, we note that the problem of modeling $\mathbf{P}$ is not related to recent work on categorical data analysis focused on handling high-dimensional categorical predictors (e.g., Stokell et al., 2021).

Our model and method is related to—although fundamentally distinct from—existing research on tensor decompositions and regression. We give a very brief and selective survey in the following and refer the interested readers to Kolda and Bader (2009) and Bi et al. (2020). First, the idea of jointly modeling multiple responses using a (regularized) tensor decomposition is conceptually similar to that in tensor response regression (e.g., Li and Zhang, 2017), where the response is typically a continuous-valued tensor. However, our study is fundamentally different due to the discrete nature of the response. For example, continuous-valued tensor decompositions (e.g., Sun et al., 2017) are not applicable since they may not result in a valid probability tensor. Secondly, our regression problem is also distinct from recent studies focused on binary or categorical tensor decompositions (e.g., Wang and Li, 2020). Extensions of these unsupervised learning methods to our context is nontrivial, especially in settings where the predictor is high-dimensional. Notably, Yang and Dunson (2016) also used the term "conditional probability tensor", though their focus was on estimating conditional probabilities of a categorical Y on multivariate categorical predictor $\mathbf{X}$, which is fundamentally different from (1).

This paper has multiple contributions. First, we propose a general method for modeling

$\mathbf{P}$ which allows practitioners to consider alternatives to the separate and vectorized modeling approaches. Crucially, unlike existing approaches which do not assume conditional independence, our method is scalable to large p, large M, and large c_m's, without sacrificing flexibility or interpretability. The scalability allows for a broad range of potential applications, and the interpretability—in terms of both predictors selected and the estimated rank of the conditional probability tensor— allows practitioners to gain novel scientific insights.

Second, we introduce the notion of a *functional* tensor rank decomposition: a type of decomposition applicable to tensor-valued functions. Loosely, this type of decomposition assumes that the range of a tensor-valued function is a restricted set of low-rank tensors. Though we focus on its application to conditional probability tensor functions, this approach could also be applied in more general tensor response regression problems.

Third, in order to accommodate different scenarios, we propose new penalties to achieve highly interpretable global and local variable selection in mixture of regression models. We devise an efficient algorithm which we prove to produce a sequence of iterates that monotonically increase the penalized observed data log-likelihood. Statistical properties are also established to illustrate how, in an idealized setting, our method scales with respect to the number of responses, the number of categories per response, and the number of predictors.

2 Model

2.1 Decomposition of the conditional probability tensor function

For a positive integer M, an M-way tensor (also known as an Mth order tensor) is an array object $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ for positive integers $p_1, \ldots, p_M$. For example, a vector is a one-way tensor, and a matrix is a two-way tensor. The tensor rank decomposition, also known as the CANDECOMP/PARAFAC decomposition, is a generalization of the singular value decomposition for matrices (Hitchcock, 1927; Carroll and Chang, 1970). A rank-one tensor $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ can be written as the outer product of M vectors: $\mathbf{A} = \mathbf{a}^{(1)} \circ \cdots \circ \mathbf{a}^{(M)}$, which is defined element-wise

as $\mathbf{A}_{i_1,\ldots,i_M} = \mathbf{a}_{i_1}^{(1)} \cdots \mathbf{a}_{i_M}^{(M)}$ for all $i_m \in [p_m]$ and $m \in [M]$. In general, a rank-R tensor ($R \geq 2$) can be written as the sum of R rank-one tensors, each formed as the outer product of vectors $\mathbf{a}_r^{(m)},\ r \in [R],\ m \in [M]$: $\mathbf{A} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \circ \cdots \circ \mathbf{a}_r^{(M)}$. A tensor is said to be rank-R if it can be decomposed into R rank-one tensors but not into r rank-one tensors for any $r < R$.

The goal of this paper is to provide a statistical modeling framework for the M-th order tensor $\mathbf{P}(\mathbf{x}) \in \mathbb{R}^{c_1 \times \cdots \times c_M}$, whose $(j_1, \ldots, j_M)$th element is the conditional probability function $P_{j_1 \ldots j_M}(\mathbf{x})$ defined in (1). To that end, we first provide some characterizations of $\mathbf{P}(\mathbf{x})$ and its rank-R decomposition which motivate our modeling approach in Section 2.2.

First, given $\mathbf{x}$, there is no distinction between a conditional probability tensor $\mathbf{P}(\mathbf{x})$ and an unconditional probability tensor (i.e., a probability tensor which is not a function of predictors). Thus, many of the results in this section apply to unconditional probability tensors. We will explain shortly how these results motivate our model for the function $\mathbf{P}$.

Let $\mathcal{P}_{c_1,\ldots,c_M} \subset \mathbb{R}^{c_1 \times \cdots \times c_M}$ denote the set of M-way valid probability tensors, i.e., the set of M-way tensors satisfying non-negativity and sum-to-one constraints. We define the *probability tensor rank* based on the tensor rank decomposition restricted to the set $\mathcal{P}_{c_1,\ldots,c_M}$.

Definition 1. *The probability tensor rank of $\mathbf{A} \in \mathcal{P}_{c_1,\ldots,c_M}$ is the minimal number R such that $\mathbf{A}$ can be expressed as the weighted sum of R rank-one probability tensors, $\mathbf{A} = \sum_{r=1}^{R} \delta_r \mathbf{A}_r$ for some $\delta_r > 0$ and $\mathbf{A}_r \in \mathcal{P}^{(1)}_{c_1,\ldots,c_M}$, $r \in [R]$, where $\mathcal{P}^{(1)}_{c_1,\ldots,c_M}$ denotes the set of M-way rank-one probability tensors. Note that $\sum \delta_r = 1$ is guaranteed because $\mathbf{A}$ and $\mathbf{A}_r$, $r \in [R]$, are probability tensors.*

In Definition 1, a rank-one probability tensor is defined in the usual sense, i.e., it can be formed as the outer product of vectors. However, the probability tensor rank R is based on a more restrictive decomposition, in which the weights are positive and the tensors $\mathbf{A}_r$ are elements of $\mathcal{P}^{(1)}_{c_1,\ldots,c_M}$. These restrictions prompt meaningful statistical and probabilistic interpretation on the probability tensor decomposition, as we discuss later. Henceforth, we say a probability tensor is rank-R if its probability tensor rank is R.

Remark 1. *Because of the additional restrictions, a rank-R probability tensor decomposition in*

Definition 1 is also a valid rank-R CP decomposition. The probability tensor rank is always no less than the usual tensor CP rank, similar to the fact that the nonnegative rank of a nonnegative matrix (Cohen and Rothblum, 1993) is no less than its usual matrix rank. The uniqueness of a rank-R decomposition requires additional conditions, e.g., the sum of ranks of the M matricizations of the tensor is no less than $2R+M-1$ is a sufficient condition for the uniqueness up to permutation and. scaling (Sidiropoulos and Bro, 2000). Generalizing such results to probability tensors remains an open question. In this work, we focus on the existence of the decomposition and later tackle the non-uniqueness from the model identifiability perspective.

The following two propositions establish upper bounds on the probability tensor rank.

Proposition 1. *For any given* $\mathbf{x}$, $\mathbf{P}(\mathbf{x})$ *has probability tensor rank* $R \leq \prod_{m=1}^{M} c_m / \max_{k\in[M]} c_k$.

Proposition 1 implies that the CP rank is also less than or equals to $\prod_{m=1}^{M} c_m / \max_{k\in[M]} c_k$. This implication also extends Theorem 1 and Corollary 1 of Dunson and Xing (2009) by establishing the upper bound on the rank R while Dunson and Xing (2009) showed the existence of the rank. When $M = 2$, Proposition 1 implies that the singular value decomposition of $\mathbf{P}(\mathbf{x})$, as a $c_1 \times c_2$ matrix, holds for rank $R \leq \min(c_1, c_2)$ for any given $\mathbf{x}$. This well-known fact for the singular value decomposition of a matrix is thus extended to our decomposition with nonnegativity and sum-to-one constraints. The result of Proposition 1 formalizes the arguments outlined in equation (6) of Johndrow et al. (2017).

The upper bound in Proposition 1 assumes nothing about the dependence among the M responses. Indeed, if $\mathbf{P}(\mathbf{x})$ has rank $R = \prod_{m=1}^{M} c_m / \max_{k\in[M]} c_k$, the responses can be arbitrarily dependent. Parsimonious dependence structures, in contrast, can imply a tighter upper bound on the rank of $\mathbf{P}(\mathbf{x})$. We present one such example in the following proposition.

Proposition 2. *For a given* $\mathbf{x}$, *if the responses form* L *mutually independent groups—indexed by sets* $G_1, \ldots, G_L$ *where* $\cup_{l=1}^{L} G_l = [M]$ *and* $G_k \cap G'_k = \emptyset$ *for* $k \neq k'$*—then* $\mathbf{P}(\mathbf{x})$ *has probability rank* $R \leq \prod_{l=1}^{L} (\prod_{m\in G_l} c_m / \max_{k\in G_l} c_k)$.

Proposition 2 suggests that the rank of $\mathbf{P}(\mathbf{x})$ is related to the complexity of the dependence among $Y_1, \ldots, Y_M$ given $\mathbf{X} = \mathbf{x}$. To demonstrate this point, consider the application in Section 6 where $c_1 = \cdots = c_M = 2$, $M = 14$. The generic upper bound from Proposition 1 is $2^{13} = 8192$. Instead, suppose that at a given $\mathbf{x}$, eight of the responses were independent of all others, and the other six formed two groups of three responses which are mutually independent, i.e., $|G_1| = 3, |G_2| = 3, |G_4| = \cdots = |G_{10}| = 1$. In this case, Proposition 2 shows $R \leq 2^4 = 16$. In our application, we actually find that $R \approx 5$ yields the best results in terms of test set log-likelihood. Johndrow et al. (2017) also provide upper bounds on the probability tensor rank as a function of the sparsity in the log-linear model characterizing the joint distribution of the responses (in an unconditional setting), further reinforcing that parsimonious dependence structures imply probability tensor rank restrictions.

Considering the most extreme case of Proposition 2, where $L = M$ and $G_l = c_l$, we have the following well-known result about rank-one probability tensors as an direct consequence of Proposition 2 and Theorem 1.

Corollary 1. *For a given* $\mathbf{x}$*,* $\mathbf{P}(\mathbf{x})$ *is rank one if and only if the responses are independent.*

Of course, a probability tensor $\mathbf{P}(\mathbf{x})$ must have at least rank-one because rank-zero, which corresponds to the tensor of zeros, would not yield a valid probability tensor.

Up to this point, our results have applied to $\mathbf{P}(\mathbf{x})$ for a given $\mathbf{x}$. In full generality, $\mathbf{P}(\mathbf{x})$ may could have a distinct decomposition for each $\mathbf{x}$, where R also varies with $\mathbf{x}$, but allowing this degree of flexibility would make regression modeling impracticable. Instead, to achieve parsimony, it is reasonable to assume that $\mathbf{P}(\mathbf{x})$ will have the same (low) probability tensor rank for every $\mathbf{x}$, and moreover, the components of their decomposition will have the same functional form. To see how one could put this assumption to use, we consider the rank-one decomposition of $\mathbf{P}(\mathbf{x})$. Specifically, the following result establishes the connection between the rank-one conditional probability tensor and the conditional independence of the responses given the predictor.

Theorem 1. *For any given* $\mathbf{x}$*, if* $\mathbf{P}(\mathbf{x}) \in \mathcal{P}_{c_1,\ldots,c_M}$ *is rank-one, it can be decomposed uniquely as*

$$\mathbf{P}(\mathbf{x}) = \mathbf{p}_1(\mathbf{x}) \circ \cdots \circ \mathbf{p}_M(\mathbf{x}), \tag{2}$$

for $\mathbf{p}_m(\mathbf{x}) = \{\Pr(Y_m = 1 \mid \mathbf{X} = \mathbf{x}), \ldots, \Pr(Y_m = c_m \mid \mathbf{X} = \mathbf{x})\}^\top$*,* $m \in [M]$*. Furthermore, if* $\mathbf{P}(\mathbf{x})$ *is rank-one for all* $\mathbf{x}$*, then* $Y_1, \ldots, Y_M$ *are conditionally independent given* $\mathbf{X}$*, and vice versa.*

By definition, rank-one probability tensor means (2) holds for some $\mathbf{p}_m(\mathbf{x}) \in \mathbb{R}^{c_m}$, $m \in [M]$. While such a decomposition is not unique, as we show in Theorem 1, a rank-one probability tensor can always be decomposed uniquely into marginal probability vectors without loss of generality. This theorem thus gives a constructive and identifiable formulation of rank one probability tensor decomposition. As shown in Corollary 1, rank-one probability tensor *for a given* $\mathbf{x}$ is equivalent to independence of responses conditional on the event $\mathbf{X} = \mathbf{x}$. As shown in Theorem 1, rank-one probability tensor *for all* $\mathbf{x}$ is equivalent to conditional independence of responses given the random variable $\mathbf{X}$. The result of Theorem 1 suggests that if, for example, we assumed the responses were conditionally independent given $\mathbf{X}$, then we are equivalently assuming that $\mathbf{P}(\,\cdot\,) = \mathbf{p}_1(\,\cdot\,) \circ \cdots \circ \mathbf{p}_M(\,\cdot\,)$ for functions $\mathbf{p}_1, \ldots, \mathbf{p}_M$ such that $\mathbf{p}_m : \mathbb{R}^p \to \Delta^{c_m - 1}$ where $\Delta^{c_m-1} = \{\mathbf{u} \in \mathbb{R}^{c_m} : \mathbf{u}^\top \mathbf{1}_{c_m} = 1, \mathbf{u}_k \geq 0 \text{ for all } k \in [c_m]\}$. Under this assumption, we could thus model each $\mathbf{p}_m$ using standard regression models for Y_m on $\mathbf{X}$ separately. On the other hand, the low-rankness of $\mathbf{P}(\mathbf{x})$ for a specific value $\mathbf{x}$ would not reduce the population model complexity.

Thus, motivated by model parsimony, we introduce the *functional rank* of the probability tensor function $\mathbf{P} : \mathbb{R}^p \to \mathcal{P}_{c_1,\ldots,c_M}$ defining the conditional probability mass (1).

Definition 2. *The functional rank of the probability tensor function* $\mathbf{P}(\cdot)$ *is the minimal number* R *such that* $\mathbf{P}(\cdot) = \sum_{r=1}^{R} \delta_r \mathbf{P}_r(\cdot)$ *for* $\delta_r > 0$ *and* $\mathbf{P}_r : \mathbb{R}^p \to \mathcal{P}^{(1)}_{c_1,\ldots,c_M}$*,* $r \in [R]$*.*

From the above definition, $\mathbf{P}_r(\cdot)$ has functional rank one and $\mathbf{P}_r(\mathbf{x}) \in \mathcal{P}^{(1)}_{c_1,\ldots,c_M}$ is a rank-one probability tensor for all $\mathbf{x}$. We now consider generalizing the rank-one structure to an arbitrary rank-R. Motivated by Theorem 1 and Definition 2, the next theorem considers the decomposition of $\mathbf{P}(\cdot)$ into a sum of R rank-one probability tensor functions (see Figure 1).

Theorem 2. *If the functional rank of* $\mathbf{P} : \mathbb{R}^p \to \mathcal{P}_{c_1,\ldots,c_M}$ *is* $R \geq 1$*, then it can be written as*

$$\mathbf{P}(\cdot) = \sum_{r=1}^{R} \delta_r \mathbf{P}_r(\cdot) = \sum_{r=1}^{R} \delta_r \mathbf{p}_{1r}(\cdot) \circ \cdots \circ \mathbf{p}_{Mr}(\cdot), \tag{3}$$

where $\delta_r > 0$, $\sum_{r=1}^{R} \delta_r = 1$ *and* $\mathbf{p}_{mr} : \mathbb{R}^p \mapsto \Delta^{c_m - 1}$ *for* $(m, r) \in [M] \times [R]$*. Moreover, if the decomposition of the function* $\mathbf{P}$ *in* (3) *holds, then there exists a categorical random variable* Z *independent of* $\mathbf{X}$ *such that* $\Pr(Z = r) = \delta_r$ *and* $\mathbf{p}_{mr}(\mathbf{x}) = \{\Pr(Y_m = 1 \mid \mathbf{X} = \mathbf{x}, Z = r), \ldots, \Pr(Y_m = c_m \mid \mathbf{X} = \mathbf{x}, Z = r)\}^\top$ *for* $r \in [R]$*. Consequently,* (3) *is equivalent to the conditional independence of* $Y_1, \ldots, Y_M$ *given* $(\mathbf{X}, Z)$*.*

Theorem 2 shows that a conditional probability tensor with functional rank-R can always be decomposed with additional constraints that $\sum_r \delta_r = 1$ and $\mathbf{p}_{mr} : \mathbb{R}^p \to \Delta^{c_m - 1}$, without loss of generality. Recall that $\Delta^{c_m - 1}$ represents the sets of valid probability mass functions of c_m-categorical random variable. Then the rank-R decomposition can be interpreted as the product of marginal probabilities $\Pr(Z = r) = \delta_r$ and conditional probability tensors $\mathbf{P}_r(\mathbf{x})$, which is defined as the probability function of $\mathbf{Y} \mid (\mathbf{X} = \mathbf{x}, Z = r)$. When (3) holds, such a categorical variable Z always exists by this construction. The rank-one probability tensor $\mathbf{P}_r(\mathbf{x})$ can then be decomposed into the product of $\mathbf{p}_{mr}(\mathbf{x}) = \{\Pr(Y_m = 1 \mid \mathbf{X} = \mathbf{x}, Z = r), \ldots, \Pr(Y_m = c_m \mid \mathbf{X} = \mathbf{x}, Z = r)\}^\top$ similar to the decomposition in Theorem 1. The rank-R decomposition in (3) is the key modeling assumption of our approach. The result of Theorem 2 suggests a natural population-level decomposition of $\mathbf{P}(\cdot)$ as illustrated in Figure 1 for $R = 3$. The functional rank-R decomposition is not always unique. Nevertheless, the rank and the existence of the decomposition is well-defined and guaranteed based on Definition 2 and Theorem 2, which motivates us to introduce a finite mixture of regressions model in the sequel.

2.2 Finite mixture of regressions model

Recall that for a $\mathbf{P}$ with rank-R functional probability tensor decomposition (3), there exists latent categorical variable $Z \in [R] = \{1, \ldots, R\}$ independent of $\mathbf{X}$ such that $Y_1, \ldots, Y_M$ are conditionally independent given $\mathbf{X}$ and Z. Specifically, we have $\delta_r = \Pr(Z = r \mid \mathbf{X} = \mathbf{x}) = \Pr(Z = r)$

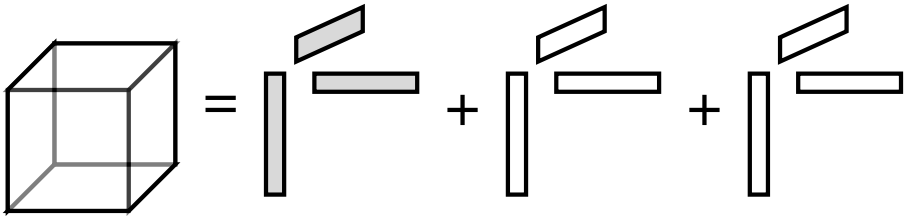

Figure 1: Diagram of the decomposition of the probability tensor $\mathbf{P}(\mathbf{x})$ (white cube) into the sum of $R = 3$ rank-one tensors. The gray rank-one tensor represents, for example, $\delta_1 \mathbf{p}_{11}(\mathbf{x}) \circ \cdots \circ \mathbf{p}_{M1}(\mathbf{x})$ from the decomposition in (3).

and $\mathbf{p}_{mr}(\mathbf{x}) = \Pr(Y_m = j_m \mid \mathbf{X} = \mathbf{x}, Z = r)$ such that (3) is satisfied in Theorem 2. It follows that

$$P_{j_1 \ldots j_M}(\mathbf{x}) = \sum_{r=1}^{R} \big\{ \Pr(Z = r) \prod_{m=1}^{M} \Pr(Y_m = j_m \mid \mathbf{X} = \mathbf{x}, Z = r) \big\}. \tag{4}$$

This connection between (3) and (4) naturally leads to a finite mixture of regression model, which implies the population level tensor decomposition (3) for all values of $\mathbf{x}$ by assuming a parametric link for the conditional probability function $\Pr(Y_m = j_m \mid \mathbf{X} = \mathbf{x}, Z = r)$ for $(m, r) \in [M] \times [R]$. With this latent categorical variable $Z \in [R]$, our proposed model is thus

$$\Pr(Z = r) = \delta_r, \quad f(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, Z = r, \boldsymbol{\theta}) = \prod_{m=1}^{M} f_r(Y_m \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_{mr}), \tag{5}$$

where f and f_r denote generic probability mass functions and $\boldsymbol{\beta}_{mr}$ (to be specified later) is the model parameter for the regression of Y_m on $\mathbf{X}$ given $Z = r$. The parameters in this model are denoted $\boldsymbol{\theta} = \{\delta_r, \boldsymbol{\theta}_r\}_{r \in [R]} = \{\delta_r, \boldsymbol{\beta}_{1r}, \ldots, \boldsymbol{\beta}_{Mr}\}_{r \in [R]}$. Of course, we can write the joint probability mass function of interest as a mixture of regressions without Z as

$$f(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) = \sum_{r=1}^{R} \delta_r f_r(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\theta}_r) = \sum_{r=1}^{R} \big\{ \delta_r \prod_{m=1}^{M} f_r(Y_m \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_{mr}) \big\}. \tag{6}$$

The model (5) has an intuitive interpretation: there are R latent states indexed by Z, and conditional on the latent state and $\mathbf{X}$, the response variables are independent. For each value of the latent variable Z, we have a distinct sub-model with conditionally independent responses (i.e., separate models). Each category of Z corresponds to a rank-one probability tensor (e.g., the gray rank-one tensor from Figure 1). The vectors $\mathbf{p}_{mr}(\mathbf{x})$ in (3) naturally consist of probabilities $\Pr(Y_m = 1 \mid Z = r, \mathbf{X} = \mathbf{x}), \ldots, \Pr(Y_m = c_m \mid Z = r, \mathbf{X} = \mathbf{x})$. Notably, maximum likelihood

estimates of $\boldsymbol{\theta}$, and consequently, of $\mathbf{P}$, can be obtained even when all category combinations are not observed in the training data. This is because each set of parameters $\boldsymbol{\theta}_r$ corresponds to marginal conditional probabilities for each response.

For the remainder, we assume a multinomial logistic regression model for each $Y_m \mid (\mathbf{X} = \mathbf{x}, Z = r)$. The regression coefficients $\boldsymbol{\beta}_{mr} \in \mathbb{R}^{p \times c_m}$ thus characterize the c_m possible outcomes of a multinomial random variable Y_m (based on a single trial), whose mass function $f_r(Y_m \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_{mr})$ consists of the probabilities

$$\Pr(Y_m = j \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_{mr}) = \frac{\exp(\boldsymbol{\beta}_{mrj}^\top \mathbf{x})}{\sum_{k=1}^{c_m} \exp(\boldsymbol{\beta}_{mrk}^\top \mathbf{x})}, \quad j \in [c_m], \tag{7}$$

where $\boldsymbol{\beta}_{mrj} \in \mathbb{R}^p$ is the jth column of $\boldsymbol{\beta}_{mr} \in \mathbb{R}^{p \times c_m}$.

Remark 2. *Under our model assumptions* (6) *and* (7)*, the identifiability of the model parameters δ_r's and $\boldsymbol{\beta}_{mr}$'s implies that the rank decomposition of function* $\mathbf{P}$ *in* (3) *is unique. However, the identifiability problem is non-trivial and analogous to the identifiability in the finite mixture of regression models and the latent class models, which are fundamental problems still not yet completely understood (Ouyang and Xu, 2022; Do et al., 2025). Even within each mixture component ($Z = r$), the $\boldsymbol{\beta}_{mr}$ are not identifiable without imposing an identifiability constraint, e.g., a "sum-to-zero" constraint $\sum_{j=1}^{c_m} \boldsymbol{\beta}_{mrj} = 0$. Later, we explain that when using our proposed penalties, the sum-to-zero constraint is enforced automatically. For formal definitions of the parameter space and identifiability under the finite mixture of regressions model, see Khalili and Chen (2007, Definitions 1 and 2). Similar to Khalili and Chen (2007), we remark that the identifiability depends on many factors and simply assume that the model under consideration in this paper is identifiable.*

Our functional probability tensor decomposition is related to the latent class model, which has been extensively studied and applied in psychological and epidemiological research. In latent class models, multiple binary or categorical variables ($\mathbf{Y}$) are measured as surrogates for estimation and inference on the unobservable definitive categorical outcome (Z), which characterizes underlying population heterogeneity and the subjects' class memberships. In particular, Bandeen-Roche et al.

(1997), Huang and Bandeen-Roche (2004), and Ouyang and Xu (2022)—among others—consider the regression extension of latent class models by including covariates (e.g., $\mathbf{X}$) and modeling Z given $\mathbf{X}$, and $\mathbf{Y}$ given Z (or more generally, $\mathbf{Y}$ given $Z, \mathbf{X}$) with generalized linear models. Based on this connection, our work provides a new perspective on latent class models with covariates, and a means for application thereof with high-dimensional predictors (Section 3).

Mathematically, our model can be transformed into the regression-based latent class model of Huang and Bandeen-Roche (2004) by replacing $\delta_r = \Pr(Z = r)$ with $\delta_r(\mathbf{x}) = \Pr(Z = r \mid \mathbf{X} = \mathbf{x})$. We do not purse this direction for multiple reasons. First, there is a philosophical difference between the latent class model and the multivariate categorical response regression model. Allowing Z (instead of, or in addition to, $\mathbf{Y}$) to be dependent on $\mathbf{X}$ is crucial in latent class models because Z is the definitive outcome and $\mathbf{Y}$ is a surrogate thereof. For example, Bandeen-Roche et al. (1997) assume $\mathbf{Y}$ is independent of $\mathbf{X}$ given the latent variable Z. However, this is unnecessary in our context where the outcome of interest is $\mathbf{Y}$ and Z is only used to introduce dependence among the Y_m. The fundamentally different applications lead to different model assumptions.

Secondly, assuming Z to be independent of $\mathbf{X}$ simplifies the model interpretation. As seen in Figure 2, our latent variable model can be viewed as a mixture of generalized linear regressions, where $\delta_r = \Pr(Z = r) = \Pr(Z = r \mid \mathbf{X})$, $r \in [R]$, are non-stochastic weights of the R mixtures. The assumption of non-stochastic weights is widely adopted in the study of the finite mixture of regression (FMR) model. Additionally, our assumptions of independence between Z and $\mathbf{X}$ helps parameter identifiability, a fundamental issue in latent class models (Ouyang and Xu, 2022).

Finally, we note that it is relatively straightforward to generalize our method by modeling $\delta_r(\mathbf{x})$ as a multinomial logistic regression of $Z \mid \mathbf{X}$. This extension is almost identical to the extension from the FMR model to the mixture of experts model (Jacobs et al., 1991). However, because (6) is already very flexible due to the large number of regression parameters, the extension from δ_r to $\delta_r(\mathbf{x})$ sometimes offers little improvement in predicting $\mathbf{Y}$. We conducted a simulation example to illustrate this point: see our discussion thereof in Section 5.

2.3 Comparison with alternative approaches

In this section, we compare our model assumptions (6) and (7) to the two other approaches for fitting M categorical responses $\mathbf{Y} = (Y_1, \ldots, Y_M)$ on the predictor $\mathbf{X} \in \mathbb{R}^p$. For the sake of comparison, we use multinomial logistic links for all approaches.

The first, and most naive, direct approach is separate modeling. This model assumes that

$$\Pr(Y_m = j \mid \mathbf{X} = \mathbf{x}) = \frac{\exp(\boldsymbol{\eta}_{mj}^\top \mathbf{x})}{\sum_{k=1}^{c_m} \exp(\boldsymbol{\eta}_{mk}^\top \mathbf{x})}, \quad j \in [c_m],\ m \in [M], \tag{8}$$

where $\boldsymbol{\eta}_{mj} \in \mathbb{R}^p$ for $j \in [c_m]$ and $m \in [M]$. This model is equivalent to our model with $R = 1$. If (8) is true, then our model with $R > 1$ becomes over-parameterized but can still provide consistent estimates of $\boldsymbol{\eta}_{mj}$'s (with some asymptotic efficiency loss). If (6) and (7) are true for $R > 1$, then separate fitting based on (8) will not only lose the interrelationship between responses, but will be incorrect for the marginal probabilities $\Pr(Y_m \mid \mathbf{X} = \mathbf{x})$, $m \in [M]$. This may be somewhat surprising, but can be seen from the latent variable representation

$$\Pr(Y_m = j \mid \mathbf{X} = \mathbf{x}) = \sum_{r=1}^{R} \delta_r \Pr(Y_m = j \mid \mathbf{X} = \mathbf{x}, Z = r) = \sum_{r=1}^{R} \delta_r \frac{\exp(\boldsymbol{\beta}_{mrj}^\top \mathbf{x})}{\sum_{k=1}^{c_m} \exp(\boldsymbol{\beta}_{mrk}^\top \mathbf{x})},$$

which can not be rewritten as proportional to $\exp(\boldsymbol{\eta}_{mj}^\top \mathbf{x})$. Intuitively, the latent variable Z introduces heterogeneity, and hence nonlinearity, in the conditional probability function $\Pr(Y_m = j \mid \mathbf{X} = \mathbf{x})$. As a result, separate model fitting is insufficient even when there is only one response variable. This is analogous to fitting a linear model to a mixture of regressions with heterogeneous sub-populations.

The second direct approach is vectorized modeling. As mentioned, this approach transforms $\mathbf{Y}$ into a univariate categorical response $\boldsymbol{Y}_*$ with $c_* = \prod_{m=1}^{M} c_m$ categories. The corresponding multinomial logistic regression model assumes

$$\Pr(\boldsymbol{Y}_* = j \mid \mathbf{X} = \mathbf{x}) = \frac{\exp(\boldsymbol{\gamma}_j^\top \mathbf{x})}{\sum_{k=1}^{c_*} \exp(\boldsymbol{\gamma}_k^\top \mathbf{x})}, \quad j \in [c_*], \tag{9}$$

where $\boldsymbol{\gamma}_j \in \mathbb{R}^p$ for $j \in [c_*]$. For example, Molstad and Rothman (2023) assume (9) and impose linear restrictions on the matrix of $\boldsymbol{\gamma}_j$'s. Similar to the separate fitting approach, if (6) and (7)

are true for $R > 1$, then this joint fitting based on (9) will also be incorrect even for marginal probabilities $\Pr(Y_m \mid \mathbf{X} = \mathbf{x})$, $m \in [M]$. On the other hand, if (9) is correct, then the rank R in our model may be as large as c_*. That is, we have c_* rank-one tensors that each consists of one element of the probability tensor $\mathbf{P}(\mathbf{x})$. The number of free parameters is $p(c_* - 1) = p(\prod_m^M c_m - 1)$ for (9) and $(R-1)+pR\sum_m(c_m-1)$ for our latent variable model (6) and (7). For example, consider the scenario where $c = c_1 = \cdots = c_M$. Then the number of free parameters becomes $p(c^M - 1)$ for the vectorized model (9) and $(R-1) + pR(c-1)M$ for the rank R version of our model. As the number of responses M increases, the complexity of (9) increases exponentially the order of $O(pc^M)$ while our model's complexity increases linearly in the order of $O(pcMR)$. As the number of categories for each response c increases, our model's complexity still increases linearly in c but the complexity of (9) increases more rapidly as c^M as $M \geq 2$. To gain further intuition, when $c = M = 4$ and $p = 100$ (as in our simulation studies), the joint model has 25500 free parameters and our model has 2401 when $R = 2$ or 3602 when $R = 3$. Finally, note that the primary benefit of the mixture of regressions model is a reduction in the number of parameters related to the response dimension. In a subsequent section, we introduce new regularization schemes to address high-dimesionality of the predictor.

3 Penalized maximum likelihood estimation

Let the observed data be $\{(\mathbf{Y}_i, \mathbf{x}_i)\}_{i=1}^n$ where $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{Mi})$ for $i \in [n] = \{1, \ldots, n\}$. Recall that $\boldsymbol{\theta}$ denotes all of the unknown parameters $\{(\delta_r, \boldsymbol{\beta}_{1r}, \ldots, \boldsymbol{\beta}_{Mr})\}_{r\in[R]} \in \mathcal{D}^R$ where $\mathcal{D} = (0,1) \times \mathbb{R}^{p\times c_1} \times \cdots \times \mathbb{R}^{p\times c_M}$ with $\sum_{r=1}^R \delta_r = 1$. The conditional log-likelihood of $\mathbf{Y} \mid \mathbf{X}$ evaluated at $\boldsymbol{\theta}$ is

$$\sum_{i=1}^n \log\left[\sum_{r=1}^R \delta_r \Big\{ \prod_{m=1}^M f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_{mr})\Big\}\right]. \tag{10}$$

In this section, we first describe the standard EM algorithm for maximizing (10) over $\mathcal{D}^R$. The standard EM algorithm, which iterates between the expectation (E) step and the maximization

(M) step, is only applicable in the classical low-dimensional setting. To address settings with high-dimensional predictors, we later discuss how to maximize a penalized version of (10). In particular, we devise a computational algorithm that replaces the penalized M-step with an approximation guaranteed to monotonically increase the objective function.

3.1 The EM algorithm and its parallel M-step

The standard EM algorithm will deal with the complete-data log-likelihood; that is, the log-likelihood of $(\mathbf{Y}, Z) \mid \mathbf{X}$, treating Z as if it were observable. Let $Z_{ir} = \mathbf{1}(Z_i = r)$. Recalling that Z is independent of $\mathbf{X}$ with $\Pr(Z = r \mid \mathbf{X}) = \Pr(Z = r) = \delta_r$, the log-likelihood of $(\mathbf{Y}, Z) \mid \mathbf{X}$ evaluated at $\boldsymbol{\theta}$ is thus

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{r=1}^{R} Z_{ir} \log \{f_r(\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\theta}_r)\} + \sum_{i=1}^{n} \sum_{r=1}^{R} Z_{ir} \log(\delta_r), \tag{11}$$

where $f_r(\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\theta}_r) = \prod_{m=1}^{M} f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_{mr})$ by definition. Each iteration of the EM algorithm, indexed by $t = 0, 1, 2, \ldots$, consists of two steps. In the E-step, we compute the $\mathcal{Q}$-function at tth iterate $\boldsymbol{\theta}^{(t)}$, $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathrm{E}\left[\mathcal{L}(\boldsymbol{\theta}) \mid \{(\mathbf{Y}_i, \mathbf{x}_i)\}_{i=1}^{n}, \boldsymbol{\theta}^{(t)}\right]$. To do so, we first compute the conditional estimate of $\pi_{ir} = \mathrm{E}(Z_{ir})$, the probability the $Z_i = r$, given the observed data and $\boldsymbol{\theta}^{(t)}$,

$$\pi_{ir}^{(t)} = \frac{\delta_r^{(t)} f_r(\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\theta}_r^{(t)})}{\sum_{s=1}^{R} \delta_s^{(t)} f_s(\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\theta}_r^{(t)})}. \tag{12}$$

Then we can express the $\mathcal{Q}$-function as

$$\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \sum_{r=1}^{R} \left[\pi_{ir}^{(t)} \sum_{m=1}^{M} \log\{f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_{mr})\} + \pi_{ir}^{(t)} \log(\delta_r) \right]. \tag{13}$$

In the M-step, we compute $\boldsymbol{\theta}^{(t+1)}$, which we define as the maximizer of (13) with respect to $\boldsymbol{\theta}$. One can verify that $\delta_r^{(t+1)} = n^{-1} \sum_{i=1}^{n} \pi_{ir}^{(t)}$, so the main challenge is maximizing the $\mathcal{Q}$-function with respect to the regression coefficients $\boldsymbol{\beta}_{mr} \in \mathbb{R}^{p \times c_m}$ for $(m, r) \in [M] \times [R]$.

From the first term in the $\mathcal{Q}$-function, one can see that the maximization with respect to the $\boldsymbol{\beta}_{mr}$ is separable across each (m, r) combination. Therefore, for $(m, r) \in [M] \times [R]$ in an embarassingly

parallel fashion, we need only compute

$$\begin{aligned}\boldsymbol{\beta}_{mr}^{(t+1)} &= \underset{\boldsymbol{\beta}_{mr}\in\mathbb{R}^{p\times c_m}}{\operatorname{argmax}}\ \ell_{mr}(\boldsymbol{\beta}_{mr} \mid \boldsymbol{\theta}^{(t)}), \\ \ell_{mr}(\boldsymbol{\beta}_{mr} \mid \boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n} \pi_{ir}^{(t)} \log\{f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_{mr})\}.\end{aligned} \tag{14}$$

The solution to the above optimization problem is obtained by fitting a weighted multinomial logistic regression model of Y_m on $\mathbf{X}$. This could be done using a modified version of the standard computational approaches, e.g., a quasi-Newton algorithm. This special structure naturally lends itself to settings with large M and large R.

The update (14) reinforces the generality of our latent variable model. We could replace the assumption of multinomial logistic link in (7) with a different assumption on $f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_{mr})$. The only necessary modification of the estimation procedure is in (14).

3.2 Penalties on the regression coefficient tensor

To address the $p > n$ case, we propose to maximize a penalized version of (10). Imposing penalties on each $\boldsymbol{\beta}_{mr}$ separately is possible, but may lead to fitted models which are difficult to interpret. Moreover, by imposing penalties across both mixture and response components, efficiency can be greatly improved. To achieve this, first organize all $\boldsymbol{\beta}_{mr} \in \mathbb{R}^{p\times c_m}$ into a tensor parameter $\mathbf{B} \in \mathbb{R}^{p\times R\times C}$, where $C = \sum_{m=1}^{M} c_m$ and define the rth *mode-2 slice* of $\mathbf{B}$ as $\mathbf{B}_{[:,r,:]} = (\boldsymbol{\beta}_{1r}, \ldots, \boldsymbol{\beta}_{Mr}) \in \mathbb{R}^{p\times C}$ for $r \in [R]$. We propose to estimate the parameters $\boldsymbol{\theta}$ using

$$\underset{\boldsymbol{\theta}\in\mathcal{D}^R}{\operatorname{argmax}} \left\{\mathcal{F}(\boldsymbol{\theta}) - \mathcal{P}_\lambda(\mathbf{B})\right\}, \tag{15}$$

where $\mathcal{F}$ is the observed data conditional log-likelihood in (10) and $\mathcal{P}_\lambda$ is a sparsity-inducing penalty by the tuning parameter $\lambda > 0$. To compute (15), we need only modify the M-step of the EM algorithm from Section 3.1 to be replaced with the following joint optimization problem

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times R\times C}}{\operatorname{argmax}}\ \mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)}), \quad \mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)}) = \sum_{m=1}^{M}\sum_{r=1}^{R} \ell_{mr}(\boldsymbol{\beta}_{mr} \mid \boldsymbol{\theta}^{(t)}) - \mathcal{P}_\lambda(\mathbf{B}), \tag{16}$$

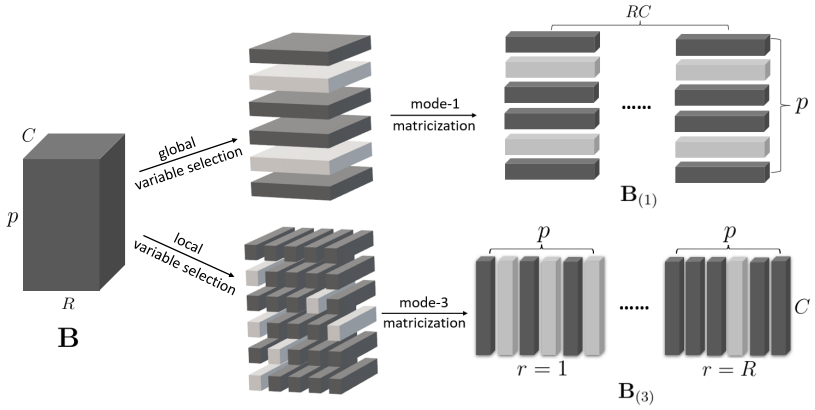

Figure 2: Visualizing the global and local variable selections and the corresponding sparsity patterns in the tensor regression parameter $\mathbf{B}$. The two plots in middle respectively show the mode-1 slices and the mode-3 fibers to be selected. Our blockwise proximal gradient descent algorithm updates one row of $\mathbf{B}_{(1)}$ at a time for the penalty $\mathcal{G}_\lambda$, and updates one column of $\mathbf{B}_{(3)}$ at a time for the penalty $\mathcal{H}_\lambda$.

where $\ell_{mr}(\cdot \mid \boldsymbol{\theta}^{(t)})$ is defined in (14) and $\mathcal{P}_\lambda$ is a sparsity-inducing penalty with tuning parameter $\lambda > 0$. If $\lambda = 0$, (16) would reduce to (14). However, when $\lambda > 0$, the penalized objective function $\mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$ may not be separable across responses and mixture components depending on the choice of penalty $\mathcal{P}_\lambda$: we propose two such penalties which correspond to distinct types of variable selection.

First, we consider global variable selection. We say that the jth predictor is irrelevant if a change in the jth component of $\mathbf{x}$ does not change $\mathbf{P}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$. Under our model assumption on $\mathbf{P}$, for the jth variable to be irrelevant it must be that $\mathbf{B}_{[j,:,:]} = (d_{j11} \cdot \mathbf{1}_{c_1}^\top, d_{j12} \cdot \mathbf{1}_{c_1}^\top, \ldots, d_{jMR} \cdot \mathbf{1}_{c_M}^\top)^\top \in \mathbb{R}^{MC}$ for constants $d_{jmr} \in \mathbb{R}$, $(m, r) \in [M] \times [R]$. Recall that we have over-parameterized $\mathbf{B}$ in the sense that each $\boldsymbol{\beta}_{mr} \in \mathbb{R}^{p \times c_m}$ has only $(c_m - 1)$ identifiable columns (see the "sum-to-zero" constraint under equation (7)). This means that we may replace the d_{jmr}'s with 0 without loss of generality and ensure parameter identifiability. This equivalence between predictor irrelevance and sparsity in $\mathbf{B}$ is discussed more in the theoretical analysis (Section 4).

For such global variable selection, we propose the following penalty term as $\mathcal{P}_\lambda$ in (16),

$$\mathcal{G}_\lambda(\mathbf{B}) = \lambda \sum_{j=1}^{p} \|\mathbf{B}_{[j,:,:]}\|_F = \lambda \sum_{j=1}^{p} \sqrt{\sum_{r=1}^{R} \sum_{m=1}^{M} \|[\boldsymbol{\beta}_{mr}]_{j,:}\|_2^2}, \tag{17}$$

where $\lambda > 0$ is a user-specified tuning parameter and $[\boldsymbol{\beta}_{mr}]_{j,:}$ is the jth row of $\boldsymbol{\beta}_{mr}$. The penalty $\mathcal{G}_\lambda$ is nondifferentiable when for some $j \in [p] = \{1, \ldots, p\}$, $[\boldsymbol{\beta}_{mr}]_{j,k} = 0$ for all $k \in [c_m]$ and $(m, r) \in [M] \times [R]$. This penalty thus links the $\boldsymbol{\beta}_{mr}$ across both latent states and response variables. For large values of λ, this penalty will encourage estimates of $\mathbf{P}$ such that many predictors are estimated to be irrelevant by encouraging zeros across the same rows of all RM coefficient matrices.

Although $\mathcal{G}_\lambda$ can achieve a highly interpretable global form of variable selection, an alternative penalty, $\mathcal{H}_\lambda$, allows for variable selection specific to each latent state (i.e., local variable selection). Specifically, we also propose the penalty

$$\mathcal{H}_\lambda(\mathbf{B}) = \lambda \sum_{j=1}^{p} \sum_{r=1}^{R} \|\mathbf{B}_{[j,r,:]}\|_2 = \lambda \sum_{j=1}^{p} \sum_{r=1}^{R} \sqrt{\sum_{m=1}^{M} \|[\boldsymbol{\beta}_{mr}]_{j,:}\|_2^2}. \tag{18}$$

In contrast to $\mathcal{G}_\lambda$, the penalty $\mathcal{H}_\lambda$ assumes that for a particular value of the latent variable Z, a possibly unique set of predictors are important. This penalty allows practitioners to characterize the categories of latent variable Z in terms of the predictors selected as relevant or not.

Figure 2 provides a visualization of the global and local variable selection in terms of the sparsity of $\mathbf{B}$. The global penalty acts on entire mode-1 slices of the tensor parameter $\mathbf{B}$. As shown in the plot, the mode-1 matricization $\mathbf{B}_{(1)} \in \mathbb{R}^{p \times RC}$ transforms each mode-1 slice into a row vector. The penalty $\mathcal{G}_\lambda(\mathbf{B})$ is thus the group lasso penalty on rows of $\mathbf{B}_{(1)}$. On the other hand, the local penalty is targeting on a more refined sparsity pattern that is shown as mode-3 fibers of $\mathbf{B}$. Analogously, the mode-3 matricization $\mathbf{B}_{(3)} \in \mathbb{R}^{C \times pR}$ aligns the fibers across $r \in [R]$. The penalty $\mathcal{H}_\lambda(\mathbf{B})$ is thus the group lasso penalty on columns of $\mathbf{B}_{(3)}$.

For concreteness, we focus on computing (15) using the penalty $\mathcal{G}_\lambda$; only trivial modifications of our algorithm are needed to accommodate $\mathcal{H}_\lambda$.

3.3 Penalized EM algorithm

In order to compute our estimator efficiently, we do not solve (16) exactly at each iteration. Rather, we approximate the solution to (16) by maximizing a minorizing quadratic approximation to $\mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$ (Lange, 2016). Our minorizing quadratic approximation is

$$\begin{aligned}\widetilde{\mathcal{M}}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)}) = \quad & \mathcal{M}_0(\mathbf{B}^{(t)} \mid \boldsymbol{\theta}^{(t)}) - \mathcal{P}_\lambda(\mathbf{B}) - \tfrac{1}{2\tau}\textstyle\sum_{j=1}^p \|\mathbf{B}_{[j,:,:]} - \mathbf{B}^{(t)}_{[j,:,:]}\|_F^2 \\ & + \textstyle\sum_{j=1}^p \operatorname{tr}\big\{\nabla_{\mathbf{B}_{[j,:,:]}} \mathcal{M}(\mathbf{B}^{(t)} \mid \boldsymbol{\theta}^{(t)})^\top (\mathbf{B}_{[j,:,:]} - \mathbf{B}^{(t)}_{[j,:,:]})\big\}, \end{aligned} \tag{19}$$

for which, with a sufficiently small step size $\tau > 0$, we have the following result.

Proposition 3. *Let $\boldsymbol{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top \in \mathbb{R}^{n \times p}$. If $\tau > 0$ is chosen so that $RM\|\boldsymbol{X}\|_F^2 \max_{k \in [M]} \sqrt{c_k} \leq \tau^{-1} < \infty$, then $\widetilde{\mathcal{M}}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)}) \geq \mathcal{M}_\lambda(\mathbf{B}^{(t)} \mid \boldsymbol{\theta}^{(t)})$ for all $\mathbf{B} \in \mathbb{R}^{p \times R \times C}$. Thus, if we define $\mathbf{B}^{(t+1)} = \operatorname{argmax}_{\mathbf{B} \in \mathbb{R}^{p \times R \times C}} \widetilde{\mathcal{M}}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$, then we are guaranteed that $\mathcal{M}_\lambda(\mathbf{B}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq \mathcal{M}_\lambda(\mathbf{B}^{(t)} \mid \boldsymbol{\theta}^{(t)})$ by the minorize-maximize principle (Lange, 2016). In addition, when $\mathcal{P}_\lambda = \mathcal{G}_\lambda$ the maximizer of $\widetilde{\mathcal{M}}_\lambda(\cdot \mid \boldsymbol{\theta}^{(t)})$ has the closed form*

$$\mathbf{B}^{(t+1)}_{[j,:,:]} = \max\left(1 - \frac{\lambda\tau}{\|\mathbf{U}_j^{(t)}\|_F}, 0\right) \mathbf{U}_j^{(t)}, \quad j \in [p] \tag{20}$$

where $\mathbf{U}_j^{(t)} = \mathbf{B}^{(t)}_{[j,:,:]} + \tau \nabla_{\mathbf{B}_{[j,:,:]}} \mathcal{M}_0(\mathbf{B}^{(t)}_{[j,:,:]} \mid \boldsymbol{\theta}^{(t)})$. When $\mathcal{P}_\lambda = \mathcal{H}_\lambda$, the maximizer of $\widetilde{\mathcal{M}}_\lambda(\cdot \mid \boldsymbol{\theta}^{(t)})$ has the closed form

$$\mathbf{B}^{(t+1)}_{[j,r,:]} = \max\left(1 - \frac{\lambda\tau}{\|\mathbf{v}_{jr}^{(t)}\|_2}, 0\right) \mathbf{v}_{jr}^{(t)}, \quad j \in [p], r \in [R], \tag{21}$$

where $\mathbf{v}_{jr}^{(t)} = \mathbf{B}^{(t)}_{[j,r,:]} + \tau \nabla_{\mathbf{B}_{[j,r,:]}} \mathcal{M}_0(\mathbf{B}^{(t)}_{[j,r,:]} \mid \boldsymbol{\theta}^{(t)})$.

The theoretical range for τ is not used in our implementation. Instead, we select τ using an Armijo-type backtracking line search which allows us to consider larger step sizes τ while maintaining the ascent property described in Proposition 3. It is important to emphasize that by defining $\mathbf{B}^{(t+1)}$ as in Proposition 3, $\mathbf{B}^{(t+1)}$ is not, in general, the argument maximizing $\mathcal{M}_\lambda(\cdot \mid \boldsymbol{\theta}^{(t)})$. However, we found this approximation scheme to be more computationally efficient than solving the M-step exactly at each iteration, and we can easily verify that it ensures ascent.

Remark 3. *As long as each τ is chosen according to Proposition 3 (or by backtracking line search), the objective function from* (15) *evaluated at $\boldsymbol{\theta}^{(t+1)}$ is guaranteed to be no less than the objective function from* (15) *evaluated at $\boldsymbol{\theta}^{(t)}$. That is, the sequence of iterates $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ generated by Algorithm S1 (Supplementary Material) monotonically increase the value of the objective function from* (15).

Remark 3 relies on the fact that our algorithm is an instance of the expectation conditional-maximization algorithm (Meng and Rubin, 1993). We summarize the entire algorithm in Algorithm S1 of the Supplementary Material.

We provide details about our implementation (e.g., initialization scheme, algorithm, and convergence criteria) in Supplementary Material Section S7. Regarding the choice of R, we found that cross-validation may not be necessary. In both our simulation studies and real data example, we found that overspecifying R often led to no worse performance than did selecting R by cross-validation. This can be partly explained by the fact that when using our penalties, the penalized EM algorithm automatically forces some δ_r estimates to be close to zero (e.g., less than 10^{-8}) when λ is sufficiently large. Consequently, computing the solution path for our estimator explores both varying levels of sparsity in the regression coefficients and implicitly, various values of R. See Section 5.3 for further details.

4 Statistical analysis of exact penalized M-step

To better understand the performance of our method, we study the statistical error involved in the penalized M-step of Algorithm S1. In the finite-sample analysis of the maximizer of $\mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$, it is very challenging to establish uniform concentration inequalities about the stochastic objective function $\mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$ and its gradient $\nabla\mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$, which depend on the estimates $\boldsymbol{\theta}^{(t)}$. For example, a theoretical study of the EM algorithm may require sample splitting; given a total of n samples and T iterations, the sample-splitting EM algorithm would use T subsets of size n/T to break the dependence of $\mathbf{B}^{(t+1)} = \operatorname{argmax}_{\mathbf{B}} \mathcal{M}_\lambda(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$ on $\boldsymbol{\theta}^{(t)}$. See Balakrishnan

et al. (2017) and Zhang et al. (2020) for examples. Because of this challenge, we leave the finite-sample statistical analysis of our penalized EM algorithm as future research. Instead, we study an idealized estimator $\widehat{\mathbf{B}}^{\dagger}$ from a modified version of $\mathcal{M}_{\lambda}(\mathbf{B} \mid \boldsymbol{\theta}^{(t)})$ which replaces the $\pi_{ir}^{(t)}$ with the "oracle" data Z_{ir}. That is, by analyzing the maximizer of the penalized conditional log-likelihood of $(\mathbf{Y}, Z) \mid \mathbf{X}$ in (11), we derive bounds that are meant to illustrate how $p, c_1, \ldots, c_M$, and the sparsity of the $\boldsymbol{\beta}_{mr}$ affect estimation of the regression parameter $\mathbf{B}$ in an idealized scenario.

Throughout this section, let $|\mathcal{A}|$ denote the cardinality of a set $\mathcal{A}$. Similarly, let $\|\mathbf{A}\|_{1,2} = \sum_k \|\mathbf{A}_{k,:}\|_2$ be the norm which sums the Euclidean norms of the rows of its matrix-valued argument. To simplify matters, we treat $\boldsymbol{X} = (\mathbf{x}_1, \cdots \mathbf{x}_n)^{\top} \in \mathbb{R}^{n \times p}$ as nonrandom and standardized such that $\sum_{i=1}^{n} \boldsymbol{X}_{i,j}^2 = n$ for $j \in [p]$. We focus on the penalized estimator using the global variable selection penalty $\mathcal{G}_{\lambda}(\mathbf{B})$. In the Supplementary Material Section S6, we discuss how similar results could be obtained under the penalty $\mathcal{H}_{\lambda}$. To avoid cumbersome tensor notation and operators, we redefine $\mathbf{B} = \mathbf{B}_{(1)} \in \mathbb{R}^{p \times RC}$ as the matrix parameter in this section.

The estimator we study, $\widehat{\mathbf{B}}^{\dagger}$, is defined formally as

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times RC}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{R} \left[Z_{ir} \sum_{m=1}^{M} \log \left\{ f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_{mr}) \right\} \right] + \mathcal{G}_{\lambda}(\mathbf{B}) \right\}. \tag{22}$$

where $Z_{ir} = \mathbf{1}(Z_i = r)$ for $(i, r) \in [n] \times [R]$. The above estimator hence does not depend on $\boldsymbol{\theta}^{(t)}$. We will treat R as a fixed and allow M, $c_1, \ldots, c_m$, p, and n to tend to infinity. Our objective is to establish an error bound on $\widehat{\mathbf{B}}^{\dagger} - \mathbf{B}^{\dagger}$ where we define $\mathbf{B}^{\dagger} = \operatorname{argmin}_{\mathbf{B} \in \boldsymbol{\pi}^*} \mathcal{G}_{\lambda}(\mathbf{B})$ with $\boldsymbol{\pi}^*$ denoting the set of all $\mathbf{B}$ which lead to the true probabilities $\mathbf{P}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$. In the Supplementary Material, we show that $\mathbf{B}^{\dagger}$ is uniquely defined, does not depend on λ, and that for each irrelevant predictor, the corresponding row $\mathbf{B}_{j,:}^{\dagger} \in \mathbb{R}^{RC}$ is zero. Thus, define $\mathcal{S} = \{j : \mathbf{B}_{j,:}^{\dagger} \neq 0, j \in [p]\}$ and $\mathcal{S}^c := \{j : \mathbf{B}_{j,:}^{\dagger} = 0, j \in [p]\}$ as the set of relevant and irrelevant predictors, respectively.

Before we describe the needed conditions and assumptions, a few comments are in order. First, the estimator in (22) is distinct from the estimator which estimates each $\boldsymbol{\beta}_{mr}$ or even $(\boldsymbol{\beta}_{1r}, \ldots, \boldsymbol{\beta}_{Mr})$ separately across $r \in [R]$. The penalty we use, $\mathcal{G}_{\lambda}$, necessarily ties the estimators together so that

even in the case that the Z_i are known, all components are estimated jointly. Second, the Z_i's are random so our statistical analysis must consider the joint distribution of the Z_i and the responses. Finally, we mention that the estimator in (22) is also relevant to the emerging literature on regression with heterogeneous sub-populations. In that context, Z_i is the indicator that a subject belongs to a particular sub-population (hence observable), and then our estimator from (22) is adjusting for Z in order to harmonize the data (Fortin et al., 2017) while our penalty $\mathcal{G}_\lambda$ can identify homogeneous structures across sub-population (Tang and Song, 2016).

Our first assumption is a new version of the restricted strong convexity condition that applies to data generated from a mixture of regressions model. This assumption will depend on $n_{\min} = \min_{r\in[R]} \sum_{i=1}^n Z_{ir}$, the minimum number of samples observed across the R latent states. Let $\Delta = (\tilde{\Delta}_{11}, \ldots, \tilde{\Delta}_{M1}, \tilde{\Delta}_{12}, \ldots, \tilde{\Delta}_{MR}) \in \mathbb{R}^{p\times RC}$ with $\tilde{\Delta}_{mr} \in \mathbb{R}^{p\times c_m}$ for $(r, m) \in [R] \times [M]$, and let $\{\sum_{i=1}^n \Psi^\dagger_{mr}(\mathbf{x}_i)\otimes\mathbf{x}_i\mathbf{x}_i^\top\}$ be the Hessian of $f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}^\dagger_{mr})$ with respect to the vectorization of its argument at $\boldsymbol{\beta}^\dagger_{mr}$. The exact form of the positive semidefinite matrix $\Psi^\dagger_{mr}(\mathbf{x}_i) \in \mathbb{R}^{c_m\times c_m}$ is given in the Supplementary Material. In addition, let $\mathcal{N}(n_{\min}) = \{\mathcal{N}_1, \ldots, \mathcal{N}_R\}$ denote the set of R-element partitions of $[n]$ where $\min_{r\in[R]} |\mathcal{N}_r| \geq n_{\min}$.

A1. (Restricted strong convexity) There exists a constant $k > 0$ such that

$$0 < k \leq \kappa(\mathcal{S}, n_{\min}) = \inf_{(\Delta,\mathcal{A})\in\mathbb{C}(\mathcal{S})\times\mathcal{N}(n_{\min})} \sum_{m=1}^M \sum_{r=1}^R \frac{\tilde{\Delta}_{mr}^\top \left[\sum_{i\in\mathcal{A}_r}\{\Psi^\dagger_{mr}(\mathbf{x}_i)\otimes\mathbf{x}_i\mathbf{x}_i^\top\}\right]\tilde{\Delta}_{mr}}{n\|\Delta\|_F^2},$$

where $\mathbb{C}(\mathcal{S}) = \{\Delta \in \mathbb{R}^{p\times R\sum_{m=1}^M c_m} : \Delta \neq 0, \|\Delta_{\mathcal{S}^c,:}\|_{1,2} \leq 3\|\Delta_{\mathcal{S},:}\|_{1,2}\}$.

Loosely speaking, one could expect the restricted strong convexity condition to hold if (i) $n_{\min}$ is not too small relative to n, (ii) that the probabilities $\Pr(Y_m = j_m \mid Z = r, \mathbf{X} = \mathbf{x})$ for $j_m \in [c_m]$ are bounded away from zero and one for all $(m, r) \in [M] \times [R]$ and all $\mathbf{x} \in \mathbb{R}^p$, and (iii) the relevant predictors (indexed by $\mathcal{S}$) are neither too correlated with one another, nor with unimportant predictors.

Explicitly, the quantity $\kappa(\mathcal{S}, n_{\min})$ is a function of $n_{\min}$ and $\mathcal{S}$. The minimum sample size $n_{\min}$ determines the space over which the infimum is taken with respect to $\mathcal{A}$ as it defines $\mathcal{N}(n_{\min})$. If

Δ were fixed, taking the infimum with respect to $\mathcal{A}$ would correspond to finding the worst possible – in the sense of minimizing $\kappa(\mathcal{S}, n_{\min})$ – partition of the subjects into the R latent states with each state having at least $n_{\min}$ subjects. The set $\mathcal{S}$, in contrast, defines the set $\mathbb{C}(\mathcal{S})$. Implicitly, $\kappa(\mathcal{S}, n_{\min})$ is a function of R, M, and the δ_r's. In fact, to use this quantity in our proofs, we need the following additional assumption.

A2. (δ_r bounds) There exists a constant $v \in (0, 1/2)$ such that $v \leq \delta_r \leq 1 - v$ for $r \in [R]$.

Lastly, we also make an assumption about the data generating process.

A3. (Data generating process) The data $\{(Z_i, \mathbf{Y}_i)\}_{i=1}^n$ are independent with $\Pr(Z_i = r) = \delta_r$, and $f(\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}_i, Z_i = r, \boldsymbol{\theta}) = \prod_{m=1}^M f_r(Y_{mi} \mid \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_{mr}^\dagger)$ where f_r is as defined in (7) for $r \in [R]$.

We are ready to state our main result, which will depend on a sample size condition, **C1**.

Theorem 3. *Let $k_1 > 2$ and $k_2 > 2$ be fixed constants. Suppose **A1–A3** hold. If $\lambda = \{\sum_{m=1}^M c_m/n\}^{1/2} + \{4Mk_2 \log p/n\}^{1/2}$, then for n sufficiently large such that condition **C1** holds,*

$$\|\hat{\mathbf{B}}^\dagger - \mathbf{B}^\dagger\|_F \leq \frac{3k_1}{\kappa(\mathcal{S}, n_{\min})} \left(\sqrt{\frac{|\mathcal{S}| \sum_{m=1}^M c_m}{4n}} + \sqrt{\frac{k_2 |\mathcal{S}| M \log p}{n}} \right)$$

with probability at least $1 - p^{1-k_2/2} - \sum_{r=1}^R \exp[-2n\{\delta_r - (n_{\min} - 1)/n\}^2]$.

The condition on the sample size (**C1**) needed for the result of Theorem 3 is as follows.

C1. (Sample size) Let $\phi_n = \max_{i \in [n]} \|\mathbf{x}_i\|_2 k_1 \lambda \sqrt{54|\mathcal{S}|}/\{2\kappa(\mathcal{S}, n_{\min})\}$ where λ is as prescribed in Theorem 3. The sample size n is sufficiently large with respect to $p, c_1, \ldots, c_M, M, k_1, k_2$, and $|\mathcal{S}|$ such that $e^{-\phi_n} + \phi_n - 1 - \phi_n^2/k_1 > 0$.

Our error bound illustrates how our idealized M-step scales with respect to the number of response variables and number of categories per response. It is especially instructive to compare the error bound in Theorem 3 to that from Molstad and Rothman (2023). Recall that Molstad and Rothman (2023) propose an estimator of the regression coefficient tensor under the vectorized

model. Unsurprisingly, their error bound scales in $\prod_{m=1}^{M} c_m$ rather than $\sum_{m=1}^{M} c_m$. This can be explained by the fact that the regression coefficient tensor of interest under the vectorized model has $p \prod_{m=1}^{M} c_m$ elements whereas under our model assumptions, there are only $Rp \sum_{m=1}^{M} c_m$ unknown regression coefficients.

5 Simulation studies

5.1 Data generating models and competing methods

In this section, we compare our method for estimating conditional probability tensors to numerous alternative approaches. To distinguish between model parameters and tuning parameters, in this section, we use R^* to denote the functional rank of the true probability tensor function, and use R to denote the rank used for model fitting. We consider data generating models with R^*, the training sample size, the mixture probabilities, and the magnitude of entries in the $\boldsymbol{\beta}_{mr}$'s varying. For a given R^*, we generate n independent realizations of the predictor $\mathbf{X} \sim \mathrm{N}_{100}(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{j,k} = 0.5^{|j-k|}$ for $(j,k) \in [p] \times [p]$ and generate $Y_1, \ldots, Y_4$ from

$$\Pr(Y_1 = j_1, \ldots, Y_4 = j_4 \mid \mathbf{X} = \mathbf{x}) = \sum_{r=1}^{R^*} \delta_r \Big\{ \prod_{m=1}^{4} \Pr(Y_m = j_m \mid \mathbf{X} = \mathbf{x}, Z = r) \Big\}$$

where $\Pr(Y_m = j_m \mid \mathbf{X} = \mathbf{x}, Z = r) = \exp(\mathbf{x}^\top \boldsymbol{\beta}_{mrj_m}) / \{\sum_{k=1}^{4} \exp(\mathbf{x}^\top \boldsymbol{\beta}_{mrk})\}$ for $m \in [4], j_m \in [4]$, and $r \in [R]$. We set $M = 4$ and $c_1 = \cdots = c_4 = 4$ throughout. When $R^* = 2$, we set $\delta_2 = 1 - \delta_1$, whereas when $R^* = 5$, we set $(\delta_1, \ldots, \delta_5) = (\delta_1, (1-\delta_1)/4, \ldots (1-\delta_1)/4)$. In both cases, we consider various values of δ_1. In each setting we consider, we randomly select five predictors to affect the conditional probability tensor. Each of the five corresponding rows of the $\boldsymbol{\beta}_{mr}$'s has entries which are drawn independently from $\mathrm{N}(0, \sigma_\beta^2)$. To select tuning parameters, we generate a validation set of size 200 from the same data generating model. To quantify performance of the various estimators, we also generate a testing set of size 1000.

We consider multiple versions of our method with candidate $R \in [4]$ (denoted `Mix-1` through `Mix-4`) and $\mathcal{P}_\lambda$ taken to be $\mathcal{G}_\lambda$. We compare these variations of our method to two methods

which implicitly assume independence: fitting separate multinomial logistic regression models to each response with (a) group-lasso penalty on the rows of each of the regression coefficient matrices (`Sep-Group`) and (b) the L_1-norm penalty applied the regression coefficient matrices (`Sep-L1`). `Sep-Group` is equivalent to our method with $R = 1$ and $\mathcal{G}_\lambda(\mathbf{B})$ replaced with $\sum_{m=1}^{M} \lambda_m \sum_{j=1}^{p} \|[\boldsymbol{\beta}_{m1}]_{j,:}\|_2$. We omit the method of Molstad and Rothman (2023) from our comparisons in this section because their software cannot be applied with $M > 3$. Moreover, their method is meant to identify which predictors are irrelevant, only affect the response marginal distributions, or affect high-order associations.

For all considered methods, tuning parameters are chosen to minimize the negative log-likelihood evaluated on the validation set. For the methods which fit separate models, tuning parameters are chosen for each response separately. We evaluate the performance by calculating the square-root average Kullback-Leibler divergence on the testing set.

Many additional simulation studies can be found in Section S9 of the Supplementary Materials. There, we consider larger p and larger M, and we compare our method to various "vectorized" modeling approaches. We also perform studies comparing the two proposed penalties $\mathcal{H}_\lambda$ and $\mathcal{G}_\lambda$, and compared our method to that of Molstad and Rothman (2023) in a setting where $M = 2$.

5.2 Results

In Figure 3, we display the square-root average Kullback-Leibler divergences on the testing set for each of the six methods with $R^* = 2$. Focusing first on the top row, where coefficients tend to have smaller magnitude, we see that when $n = 75$, there is only minor differences between methods. When $n = 150$ or $n = 300$, however, we notice that `Mix-2`, `Mix-3`, and `Mix-4` all tend to outperform the other competitors. This is not surprising given that each of these methods contains the model with the true number of components as a special case. Another interesting aspect of these results is that as δ_1 approaches 0.5, the differences in performance become more apparent. This too is not surprising because when $\delta_1 = 0.1$, `Mix-1` could serve as a reasonable

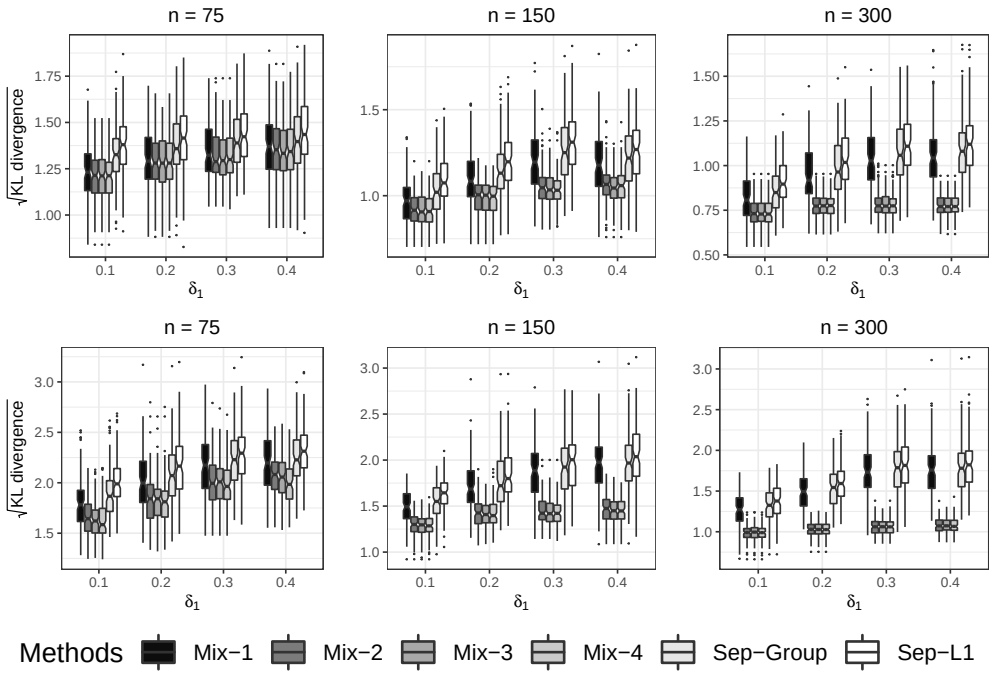

Figure 3: The square-root average Kullback-Leibler divergence evaluated on the testing set for the six considered methods. Results displayed are from 100 independent replications with (top row) $(\sigma_\beta, R^*) = (1, 2)$ and (bottom row) $(\sigma_\beta, R^*) = (2, 2)$.

approximation to $\mathbf{P}$. Results are effectively the same when $\sigma_\beta = 2$, except that differences between the methods are more pronounced even in the case that $n = 75$ or $n = 150$.

We display analogous results for the data generating model with $R^* = 5$ in Figure 4. Overall, we see results similar to those when $R^* = 2$. Specifically, when $n = 300$ and coefficient magnitudes are (relatively) small, there is little distinction between the methods. As $n = 450$, we see a more clear seperation between the versions of our method with different numbers of mixture components. Notably, we see that as δ_1 increases, `Mix-2` performs more similarly to `Mix-5` and `Mix-7`. This makes sense as when $\delta_1 = 0.67$, four of the five components make relatively small contributions to $\mathbf{P}$. Just as in the case with $R^* = 2$, the magnitude of the regression coefficients $\boldsymbol{\beta}_{mr}$ appears to have an effect as well. Specifically, when $\sigma_\beta = 2$, the difference between estimators is much greater than under similar settings with $\sigma_\beta = 1$.

In Section S3 of the Supplementary Material, we present results for the same set of competitors

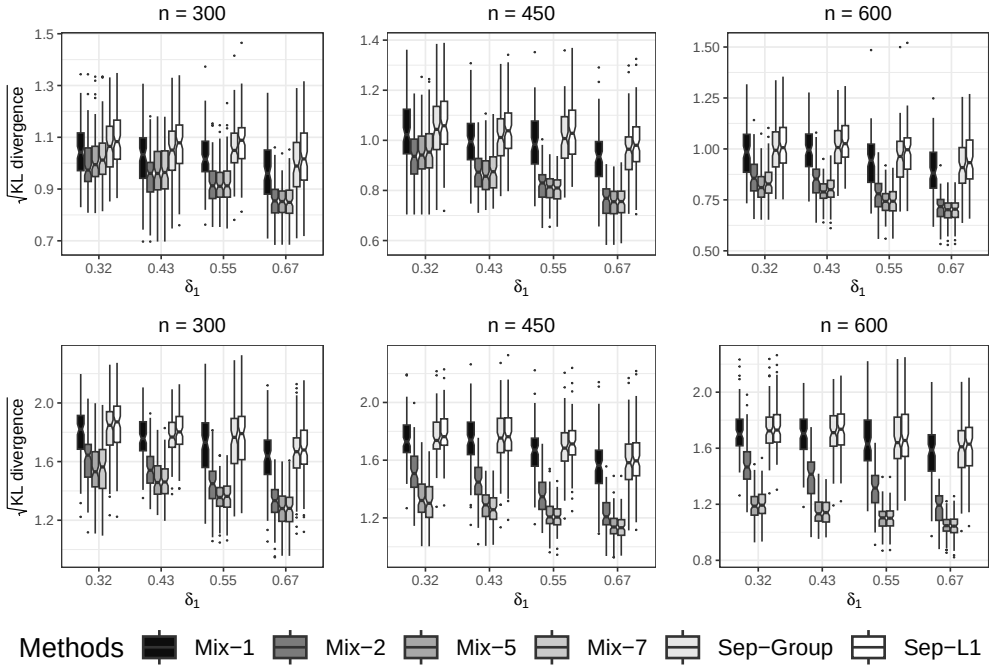

Figure 4: The square-root average Kullback-Leibler divergence evaluated on the testing set for the six considered methods. Results displayed are from 100 independent replications with (top row) $(\sigma_\beta, R^*) = (1, 5)$ and (bottom row) $(\sigma_\beta, R^*) = (2, 5)$.

under identical data generating models, but using Hellinger distance as a performance metric. In brief, relative performances are very similar to those based on KL divergence.

In the Supplementary Material Section S2 and S3, we include results for numerous additional simulation studies. These include simulation studies with $R^* \in \{6, 7, 8, 9\}$, and simulation studies under two types of model misspecification (including when Z depends on $\mathbf{X}$). In brief, our method appears reasonably robust against misspecification.

5.3 Effect of overspecifying R

In the simulation results displayed in Figures 3 and 4, we see that overspecifying the number of mixture components seems to have little effect on estimation. In Figure 4, we see that when $n = 450$, `Mix-7` even tended to perform slightly better than `Mix-5`, which has the correctly specified number of components. To explore whether more extreme overspecification has a similar

effect, we repeated the simulations displayed in the bottom panel of Figure 4 and added three additional versions of our method with $R \in \{5, 10, 15\}$, which we call `Mix-5`, `Mix-10`, and `Mix-15`, respectively. We display results in Figure S6 of the Supplementary Material. Based on the results, it seems that extreme overspecification of R may even improve estimation accuracy when n is small: `Mix-15` seemed to perform as well or better than all other methods in each scenario. However, it is important to note that for the particular tuning parameters chosen for, say, `Mix-15`, there were often only two nonzero estimates of the δ_r's. This is one particularly appealing feature of the solution path for fitting (15): for large values of λ, even when R is large, the solution will have one $\delta_r = 1$ and $\delta_{r'} = 0$ for $r' \neq r$. As λ decreases, eventually a second δ_r becomes nonzero so that the fitted model effectively has $R = 2$. This continues with each additional δ_r becoming nonzero as $\lambda \to 0$. We illustrate this phenomenon in Figure 5 of Section 6, and discuss this feature of our method in more depth in Supplementary Materials Section S4.

6 Modeling functional classes of genes

In this section, we apply our method to the problem of modeling a gene's functional classes based on both the gene's expression and phylogenetic profile. The dataset we analyzed, which was collected on yeast, was originally studied in Elisseeff and Weston (2001) and can be downloaded from `https://www.uco.es/kdis/mllresources/`. The predictors consist of $p = 103$ components, which are the collection of both the gene's expression and phylogenetic profile. In these data, there are $M = 14$ functional classes (including metabolism, energy, protein synthesis, etc.). Each of the $n = 2417$ genes can be characterized as belonging to multiple functional classes. For example, a gene may affect both metabolism and protein synthesis. Thus, it is natural to treat each functional class assignment as a binary response so that we have $c_1 = \cdots = c_{14} = 2$.

Because of the relatively large number of response variables, we used the version of our method with the penalty $\mathcal{H}_\lambda$ described in (18). This penalty allows different components of the predictor to be relevant for different values of the latent variable Z. In the case that $R = 1$, this is equivalent

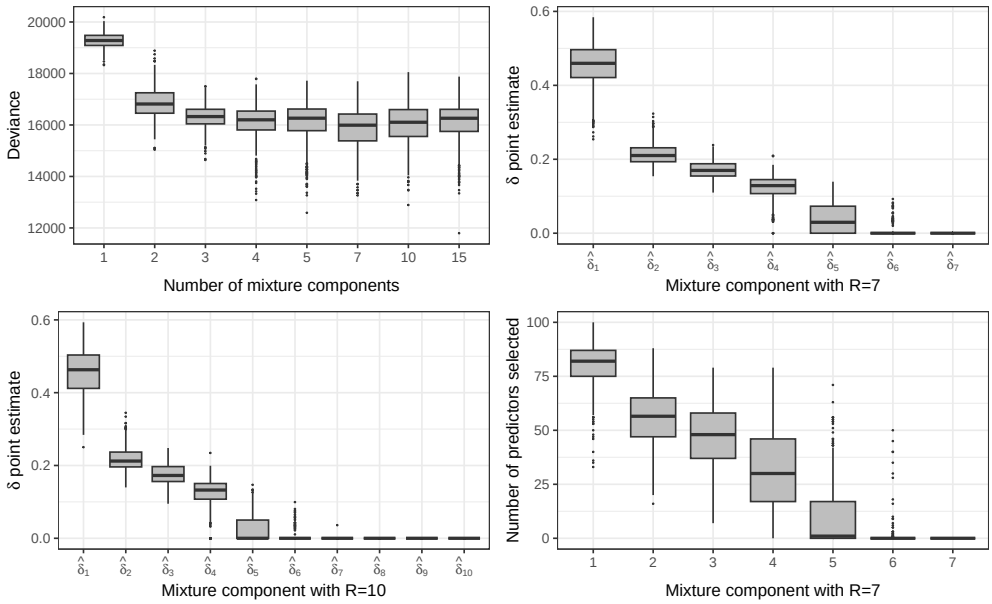

Figure 5: Results based on 500 replications: (top left) test set deviance for eight versions of the mixture model with $R \in \{1, 2, 3, 4, 5, 7, 10, 15\}$, (top right, bottom left) ordered estimates of the δ_r for the version of our method with $R = 7$ and $R = 10$, respectively, and (bottom right) the number of predictors selected for each component mixture (in descending order according to the $\hat{\delta}_r$) with $R = 7$.

to the penalty $\mathcal{G}_\lambda$, but differs for $R > 1$.

To compare our method's performance across different choices of R, for 500 independent replications we split the data into training, validation, and testing sets of size 1500, 500, and 417, respectively. For $R \in \{1, 2, 3, 4, 5, 7, 10, 15\}$ separately, we fit the model to the training set and selected the tuning parameter λ by minimizing the negative log-likelihood evaluated on the validation set. Then, we compute the deviance and misclassification accuracy on the testing set for the selected model. In Table 1 of Section S8, we display the average joint error rate and the average testing set deviance for our method with the various number of components. Evidently, $R = 7$ performed best in terms of deviance, and was only slight worse (and not significantly different from) $R = 15$ in terms of classification error. With $R \geq 4$, there is relatively little difference in terms of performance.

We display the test set deviances in the top left panel of Figure 5. Note that the version of our

method assuming independent responses ($R = 1$) performs much worse than the versions with $R > 1$; adding even the second mixture component ($R = 2$) decreased testing set deviance by nearly 15%. In the two rightmost panels, we display the estimated δ_r and the number of predictors selected in the corresponding $\boldsymbol{\beta}_{mr}$'s with $R = 7$. We see that in general, when $R = 7$, the selected model effectively has 5 or fewer $\hat{\delta}_r$ nonzero. We also notice that the mixture component with the highest probability often had 75 or more predictors included in the model, whereas those mixture components with smaller probabilities tended to have fewer. Examining the results with $R = 10$ in the bottom leftmost panel, we again see often only five of fewer $\hat{\delta}_r$ are estimated to be nonzero. Additional results can be found in Section S8 of the Supplementary Material.

7 Discussion

In this article, we propose a general modeling strategy for the regression of a multivariate categorical response on a high-dimensional predictor based on the population-level tensor rank decomposition of the conditional probability tensor function. Numerically, our method is shown to perform well with a large number of response variables and large number of categories per response—a setting where many existing competitors fail or cannot be applied. Our results also suggest that our method is insensitive to over-specification of the number of mixtures R.

Based on our theoretical analysis of an idealized and simplified estimator, we conjecture that the convergence rate of the penalized EM algorithm, under suitable assumptions (see, for example, Balakrishnan et al., 2017; Cai et al., 2019), is $\sqrt{|\mathcal{S}|M \log p/n}$ when treating R and c_m's as constants. We leave this as a future theoretical study. Promising future research directions also include (i) initialization of the algorithm and (ii) alternative convex penalties on the tensor $\mathbf{B}$. Regarding (i), methods such as Sedghi et al. (2016) are applicable to our setting and could potentially improve over the random initialization that we are currently using. For (ii), recent advances on tensor nuclear norm penalties (e.g., Raskutti et al., 2019) neatly fit in our optimization framework and can take advantage of the tensor construction of $\mathbf{B}$.

References

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical science*, 7(1):131–153.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EMalgorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Bhattacharya, A. and Dunson, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107(497):362–377.

Bi, X., Tang, X., Yuan, Y., Zhang, Y., and Qu, A. (2020). Tensors in statistics. *Annual Review of Statistics and Its Application*, 8.

Cai, T. T., Ma, J., and Zhang, L. (2019). Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319.

Cohen, J. E. and Rothblum, U. G. (1993). Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168.

Dahl, A. and Zaitlen, N. (2020). Genetic influences on disease subtypes. *Annual review of genomics and human genetics*, 21:413–435.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. (2012). On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45.

Do, D., Do, L., and Nguyen, X. (2025). Strong identifiability and parameter learning in regression with heterogeneous response. *Electronic Journal of Statistics*, 19(1):131–203.

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

Ekholm, A., McDonald, J. W., and Smith, P. W. (2000). Association models for a multivariate binary response. *Biometrics*, 56(3):712–718.

Elisseeff, A. and Weston, J. (2001). A kernel method for multi-labelled classification. *Advances in neural information processing systems*, 14:681–687.

Fienberg, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments. *Journal of the American Statistical Association*, 95(450):643–647.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., and Gur, R. E. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170.

Gao, F., Wahba, G., Klein, R., and Klein, B. (2001). Smoothing spline anova for multivariate bernoulli observations with application to ophthalmology data. *Journal of the American Statistical Association*, 96(453):127–160.

Glonek, G. F. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, 83(1):15–28.

Glonek, G. F. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):533–546.

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.

Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

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

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038.

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

Lange, K. (2016). *MM optimization algorithms*. SIAM.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC Press.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278.

Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, 18(17-18):2237–2255.

Molstad, A. J. and Rothman, A. J. (2023). A likelihood-based approach for multivariate categorical response regression in high dimensions. *Journal of the American Statistical Association*, 118(542):1402–1414.

Montañes, E., Senge, R., Barranquero, J., Quevedo, J. R., del Coz, J. J., and Hüllermeier, E. (2014). Dependent binary relevance models for multi-label classification. *Pattern Recognition*, 47(3):1494–1508.

Ouyang, J. and Xu, G. (2022). Identifiability of latent class models with covariates. *Psychometrika*, 87(4):1343–1360.

Raskutti, G., Yuan, M., and Chen, H. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333.

Sedghi, H., Janzamin, M., and Anandkumar, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231.

Senge, R., Coz Velasco, J. J. d., and Hüllermeier, E. (2013). Rectifying classifier chains for multi-label classification. *Space, 2 (8)*.

Sidiropoulos, N. D. and Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3):229–239.

Simon, N., Friedman, J., and Hastie, T. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*.

Stokell, B. G., Shah, R. D., and Tibshirani, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(3):579–611.

Sun, W. W., Lu, J., Liu, H., and Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916.

Tang, L. and Song, P. X. (2016). Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research*, 17(1):3915–3937.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Vincent, M. and Hansen, N. R. (2014). Sparse group lasso and high dimensional multinomial classification. *Computational Statistics and Data Analysis*, 71:771–786.

Wang, M. and Li, L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *The Journal of Machine Learning Research*, 21:154–1.

Yang, Y. and Dunson, D. B. (2016). Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association*, 111(514):656–669.

Zhang, L., Ma, R., Cai, T. T., and Li, H. (2020). Estimation, confidence intervals, and large-scale hypotheses testing for high-dimensional mixed linear regression. *arXiv preprint arXiv:2011.03598*.

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., and Geng, X. (2018). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443.