Journal of Machine Learning Research 25 (2024) 1-66 Submitted 7/23; Revised 8/24; Published 11/24

Multi-Response Linear Discriminant Analysis in High Dimensions

Kai Deng KDENG@FSU.EDU
Xin Zhang HENRY@STAT.FSU.EDU
Department of Statistics
Florida State University
Tallahassee, Florida, USA

Aaron J. Molstad AMOLSTAD@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, Minnesota, USA

Editor: Po-Ling Loh

Abstract

The problem of classifying multiple categorical responses is fundamental in modern machine learning and statistics, with diverse applications in fields such as bioinformatics and imaging. This manuscript investigates linear discriminant analysis (LDA) with high-dimensional predictors and multiple multi-class responses. Specifically, we first examine two different classification scenarios under the bivariate LDA model: joint classification of the two responses and conditional classification of one response while observing the other. To achieve optimal classification rules for both scenarios, we introduce two novel tensor formulations of the discriminant coefficients and corresponding regularization strategies. For joint classification, we propose an overlapping group lasso penalty and a blockwise coordinate descent algorithm to efficiently compute the joint discriminant coefficient tensors. For conditional classification, we utilize an alternating direction method of multipliers (ADMM) algorithm to compute the discriminant coefficient tensors under new constraints. We then extend our method and algorithms to general multivariate responses. Finally, we validate the effectiveness of our approach through simulation studies and applications to benchmark datasets.

Keywords: Convex Optimization, Discriminant Analysis, Group Lasso, Tensor

1. Introduction

The task of classifying multiple categorical responses is a frequent challenge in the fields of statistics and machine learning (Glonek and McCullagh, 1995; Lewis et al., 2004; Wang et al., 2016). In this article, we consider the classification of a multivariate categorical response $\mathbf{Y} = (Y_1, \ldots, Y_M)$, $M \geq 2$, based on a high-dimensional predictor $\mathbf{X} \in \mathbb{R}^p$. Each categorical response consists of multiple (numerically coded) categories, i.e., $Y_m \in \{1, 2, \ldots, c_m\}$ for each $m \in \{1, 2, \ldots, M\}$. Modeling the conditional associations among responses and dealing with the high dimensionality of the predictor pose immediate challenges. Furthermore, as the number of predictor variables p increases or the number of

response categories c_m exceeds two, the modeling and computational difficulties become more pronounced. Nevertheless, there exist many approaches to this problem.

A closely related problem is "multi-label" classification, where typically each $c_m = 2$. Multi-label classifiers are used for applications including text categorization and image annotation. There is a rich literature on multi-label classification in machine learning: see Tsoumakas and Katakis (2007) or Zhang and Zhou (2013) for an overview. Multi-label classification is a special instance of multi-task learning (Zhang and Yang, 2021) known as "homogeneous feature" multi-task learning. Many multi-label classifiers use heuristic modifications of existing classification methods to account for the associations among responses (Park and Lee, 2008; Wang et al., 2010). One prominent class of methods for multi-label classification is based on "classifier chains" (Read et al., 2009, 2011; Weng et al., 2020), which fit M classifiers sequentially. A chain of classifiers is obtained by fitting a model for $(Y_1|\mathbf{X})$, then $(Y_2|\mathbf{X}, Y_1)$, and so on until one fits a model for $(Y_M|\mathbf{X}, Y_1 \ldots, Y_{M-1})$. Using this sequential approach, classification chains can account for the dependence between responses by indirectly modeling $(Y_1, \ldots, Y_M \mid \mathbf{X})$. However, classification chains are difficult to interpret in terms of the distribution of interest, $(Y_1, \ldots, Y_M \mid \mathbf{X})$.

In the statistical literature, existing methods for multivariate categorical response regression aim to directly model the conditional distribution of the multivariate response as a function of the predictor. These methods develop new parametric links for generalized linear models that account for the conditional associations among responses (McCullagh and Nelder, 1989; Glonek, 1996). For example, Glonek and McCullagh (1995) use a multivariate logistic transform to parametrize the joint distribution of a bivariate response in terms of their marginal distributions and log-odds ratios. To handle high-dimensional predictors in a bivariate categorical response regression under the multinomial logistic link, Molstad and Rothman (2023) proposed a new regularization scheme that allows for the identification of predictors that affect the joint distribution of the responses, affect only the marginal distributions of the responses, or are irrelevant. Along different lines, Molstad and Zhang (2022) proposed a regularized variation of the latent class model (Ouyang and Xu, 2022), motivated by the assumption of low-rankness in the probability tensor function. Speaking broadly, these statistical methods are either too complex to handle a large number of responses M or are computationally burdensome. Other recently developed statistical methods utilize dimension reduction (e.g., by assuming that regression coefficients from $Y_1 \mid \mathbf{X}, \ldots, Y_M \mid \mathbf{X}$ span a low-dimensional subspace), but these methods tend to implicitly assume that responses are independent given $\mathbf{X}$ (Luo et al., 2018; Park et al., 2024).

Our method is based on the linear discriminant analysis (LDA) model that assumes $\mathbf{X}$ given $\mathbf{Y}$ is multivariate normal with the same covariance for all possible response category combinations. This model assumption is fundamentally different from existing multi-label classification methods and multivariate logistic regression models. Over the past two decades, extensive research has been conducted on high-dimensional LDA methods (Tibshirani et al., 2002; Fan and Fan, 2008; Witten and Tibshirani, 2011; Clemmensen et al., 2011; Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012, 2019). Although the LDA model assumption appears stringent, high-dimensional LDA methods consistently demonstrate robust classification performance (Hastie et al., 2009, Section 4.4.5). While high-dimensional extensions of linear discriminant analysis (LDA) have been widely studied in the literature, how to extend high-dimensional LDA from a univariate (often binary) response to multiple

multi-class responses remains unclear. In this article, we begin by focusing on the bivariate categorical response model. We introduce two novel regularization schemes for parameter estimation in high-dimensional settings. In particular, we introduce joint and conditional discriminant coefficient tensors, and corresponding new penalties for structured variable selection and regularization. We then extend our proposed method to settings with more than two responses. Efficient algorithms for computing our estimators are developed, and we study the statistical properties of our approach.

Our method is closely related to multi-class sparse discriminant analysis (MSDA; Mai et al., 2019), which assumes a group sparse structure across all discriminant vectors and employs a convex objective function. As we will later explain (Section 3.1), we adopt the same convex objective function as MSDA but employ novel penalties that account for the multivariate nature of the response. Specifically, we introduce two novel tensor formulations of the bivariate LDA estimation problem in high dimensions, dependent on the scenario of either joint or conditional classification. Existing multivariate response classification methods commonly focus on joint classification while disregarding conditional classification. Traditionally, when given the label information of one response, partial data matching the given label would be used for model training and require re-fitting with each new label given. In contrast, our method utilizes the entire dataset and requires only one fitting, which improves efficiency and results in more stable variable selection and classification performance.

The rest of this paper is organized as follows. Section 2 introduces the problem setup, including the bivariate linear discriminant analysis model and different classification rules. Section 3 proposes penalized objective functions for estimating the parameters of interest. In Section 4, we examine the statistical properties of our method, focusing on joint classification. Section 5 considers an alternative penalization strategy based on the latent overlapping group lasso. In Section 6, we develop efficient algorithms for computing our estimators. Section 7 extends our method to general multivariate response and semiparametric settings, and contrasts our method to some existing approaches. Simulation studies and real data examples are presented in Section 8 and Section 9, respectively. The Appendix contains additional extensions, algorithms, proofs, and numerical results.

2. Model

2.1 Bivariate LDA Model and Notation

We first introduce the bivariate linear discriminant analysis (LDA) model along with the relevant notation that will be used throughout this article.

We consider the classification problem of a bivariate, multi-class categorical response $\mathbf{Y} = (Y_1, Y_2)$, given a predictor $\mathbf{X} \in \mathbb{R}^p$. We use $[c_m]$ to denote the set $\{1, \ldots, c_m\}$ of possible response categories for Y_m, $m \in \{1, 2\}$, i.e., $Y_m \in [c_m]$ with $c_m \geq 2$. For every category combination (k_1, k_2), $k_1 \in [c_1]$, $k_2 \in [c_2]$, the bivariate LDA model assumes that

$$\mathbf{X}|(Y_1 = k_1, Y_2 = k_2) \sim \mathrm{N}_p(\boldsymbol{\mu}_{k_1 k_2}, \boldsymbol{\Sigma}), \quad \Pr(Y_1 = k_1, Y_2 = k_2) = \pi_{k_1 k_2} > 0, \tag{1}$$

where $\boldsymbol{\mu}_{k_1 k_2} \in \mathbb{R}^p$ is a mean vector, $\boldsymbol{\Sigma} \in \mathbb{S}_+^p$ is a $p \times p$ symmetric and positive definite covariance matrix, and $\sum_{k_1, k_2} \pi_{k_1 k_2} = 1$.

We adopt the following tensor notation. See Kolda and Bader (2009) for an overview of tensor operators. The *order* of a tensor is the number of dimensions, also known as ways or modes. *Fibers* are the higher-order analog of matrix rows and columns, defined by fixing every index but one. *Slices* are two-dimensional sections of a tensor, defined by fixing all but two indices. For a three-way tensor $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, we will use $\mathbf{A}_{[i_1,i_2,i_3]}$ to denote the $i_1 i_2 i_3$-th entry, and use $\mathbf{A}_{[:,i_2,i_3]}$, $\mathbf{A}_{[i_1,:,i_3]}$ and $\mathbf{A}_{[i_1,i_2,:]}$ to denote the mode-1, mode-2 and mode-3 fibers, respectively. For an M-way tensor $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, the mode-$m$ *matricization* arranges the mode-m fibers as the columns of the resulting matrix $\mathbf{A}_{(m)} \in \mathbb{R}^{p_m \times \prod_{k \neq m} p_k}$. The mode-$m$ *product* of tensor $\mathbf{A}$ with a matrix $\mathbf{G}_m \in \mathbb{R}^{d_m \times p_m}$ multiplies each mode-m fiber with the matrix $\mathbf{G}_m$, written as $\mathbf{A} \times_m \mathbf{G}_m$, resulting in an M-way tensor of dimension $p_1 \times \cdots \times p_{m-1} \times d_m \times p_{m+1} \times \cdots \times p_M$. For vectors $\boldsymbol{b}^{(m)} = (b_1^{(m)}, \ldots, b_{p_m}^{(m)})^\top \in \mathbb{R}^{p_m}$, $m \in [M]$, the *outer product* of the M vectors $\boldsymbol{b}^{(1)} \circ \cdots \circ \boldsymbol{b}^{(M)}$ gives an M-way tensor $\mathbf{B}$, such that element-wisely, $B_{i_1 \cdots i_M} = b_{i_1}^{(1)} \cdots b_{i_M}^{(M)}$ for all $i_m \in [p_m]$.

We will encounter three-way tensor parameters under the bivariate LDA model (1). For example, the group means can naturally be arranged into the mean tensor $\boldsymbol{\mu} \in \mathbb{R}^{p \times c_1 \times c_2}$ such that the set of all of mode-1 fibers is $\{\boldsymbol{\mu}_{11}, \boldsymbol{\mu}_{12}, \ldots, \boldsymbol{\mu}_{c_1 c_2}\}$. To simplify notation, we use $\mathbf{A}_{i_2 i_3}$ to denote the mode-1 fibers $\mathbf{A}_{[:,i_2,i_3]}$, and $\mathbf{A}_{\cdot i_3}$ to denote frontal slices $\mathbf{A}_{[:,:,i_3]}$, $\mathbf{A}_{i_2 \cdot}$ to denote lateral slices $\mathbf{A}_{[:,i_2,:]}$.

We let $\mathbf{1}_p = (1, \ldots, 1)^\top \in \mathbb{R}^p$, $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ be the identity matrix, and $I(\cdot)$ be the indicator function. For a matrix $\mathbf{C} = (c_{ij}) \in \mathbb{R}^{p \times q}$, define $\|\mathbf{C}\|_{2,1} = \sum_{i=1}^p \sqrt{\sum_{j=1}^q c_{ij}^2}$.

2.2 Bayes' Rules and Discriminant Coefficient Tensors

We consider two different classification problems. Joint classification is predicting the bivariate response $\mathbf{Y} = (Y_1, Y_2)$ jointly from the predictor $\mathbf{X}$. Conditional classification is predicting Y_1 from $(\mathbf{X}, Y_2)$ and Y_2 from $(\mathbf{X}, Y_1)$.

For joint classification, the classification error of a classifier $\widehat{\mathbf{Y}}(\mathbf{X}) = \phi_{\mathbf{Y}}(\mathbf{X})$ is defined as $\mathrm{E}\{I(\widehat{\mathbf{Y}}(\mathbf{X}) \neq \mathbf{Y})\}$. The Bayes' rule for joint classification $\phi_{\mathbf{Y}} : \mathbb{R}^p \to [c_1] \times [c_2]$ is the classifier that achieves the lowest possible classification error and can be written as

$$\phi_{\mathbf{Y}}(\mathbf{X}) = \underset{k_1 \in [c_1], k_2 \in [c_2]}{\operatorname{argmax}} \Pr(Y_1 = k_1, Y_2 = k_2 \mid \mathbf{X}). \tag{2}$$

For conditional classification of Y_1 given $(\mathbf{X}, Y_2)$, the classification error of a classifier $\widehat{Y}_1(\mathbf{X}, Y_2) = \phi_{Y_1}(\mathbf{X}, Y_2)$ is $\mathrm{E}\{I(\widehat{Y}_1(\mathbf{X}, Y_2) \neq Y_1)\}$. The Bayes' rule for conditional classification $\phi_{Y_1} : \mathbb{R}^p \times [c_2] \to [c_1]$ achieves the lowest possible classification error and can be written as

$$\phi_{Y_1}(\mathbf{X}, Y_2) = \underset{k_1 \in [c_1]}{\operatorname{argmax}} \Pr(Y_1 = k_1 \mid \mathbf{X}, Y_2). \tag{3}$$

Under the bivariate LDA model (1), both joint and conditional Bayes' rules can be derived straightforwardly. For joint classification, the Bayes' rule is

$$\phi_{\mathbf{Y}}(\mathbf{X}) = \underset{k_1 \in [c_1], k_2 \in [c_2]}{\operatorname{argmax}} \left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_{k_1 k_2} + \boldsymbol{\mu}_{11}}{2} \right)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{k_1 k_2} - \boldsymbol{\mu}_{11}) + \log \pi_{k_1 k_2} \right\}. \tag{4}$$

When the response Y_2 is observed as $Y_2 = k_2$, the conditional Bayes' rule $\phi_{Y_1}(\mathbf{X}, Y_2)$ becomes

$$\phi_{Y_1}(\mathbf{X}, Y_2 = k_2) = \underset{k_1 \in [c_1]}{\operatorname{argmax}} \left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_{k_1 k_2} + \boldsymbol{\mu}_{1 k_2}}{2} \right)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{k_1 k_2} - \boldsymbol{\mu}_{1 k_2}) + \log \pi_{k_1 k_2} \right\}. \quad (5)$$

From (4), we note that the optimal joint classifier depends on $\mathbf{X}$ only through $(c_1 c_2 - 1)$ linear combinations: $\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{k_1 k_2} - \boldsymbol{\mu}_{11})$, $k_1 \in [c_1]$ and $k_2 \in [c_2]$. Recall that $\boldsymbol{\mu} \in \mathbb{R}^{p \times c_1 \times c_2}$ is the tensor of all group means. By defining $\boldsymbol{\delta} = \boldsymbol{\mu} - \boldsymbol{\mu}_{11} \circ \mathbf{1}_{c_1} \circ \mathbf{1}_{c_2} \in \mathbb{R}^{p \times c_1 \times c_2}$ and $\boldsymbol{\beta} = \boldsymbol{\delta} \times_1 \boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{p \times c_1 \times c_2}$, we have that the mode-1 fibers of $\boldsymbol{\beta}$ satisfy $\boldsymbol{\beta}_{k_1 k_2} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{k_1 k_2} - \boldsymbol{\mu}_{11})$, $\boldsymbol{\beta}_{11} = 0$. Clearly, $\boldsymbol{\beta}$ is the discriminant coefficient tensor needed for the joint Bayes' rule (4).

For conditional classification, the parameter of interest is no longer $\boldsymbol{\beta}$. The Bayes' rule (5) for conditional classification of Y_1 given $Y_2 = k_2$ suggests that the $(c_1 - 1)$ discriminant directions are $(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{2 k_2} - \boldsymbol{\mu}_{1 k_2}), \ldots, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{c_1 k_2} - \boldsymbol{\mu}_{1 k_2})) = (\boldsymbol{\beta}_{2 k_2} - \boldsymbol{\beta}_{1 k_2}, \ldots, \boldsymbol{\beta}_{c_1 k_2} - \boldsymbol{\beta}_{1 k_2}) \in \mathbb{R}^{p \times (c_1 - 1)}$. Defining $\mathbf{A}^{c_1} = (-\mathbf{1}_{c_1 - 1}, \mathbf{I}_{c_1 - 1}) \in \mathbb{R}^{(c_1 - 1) \times c_1}$, and using that the k_2-th frontal slice of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_{\cdot k_2} = (\boldsymbol{\beta}_{1 k_2}, \ldots, \boldsymbol{\beta}_{c_1 k_2}) \in \mathbb{R}^{p \times c_1}$, one can see that $\boldsymbol{\beta}_{\cdot k_2}(\mathbf{A}^{c_1})^\top \in \mathbb{R}^{p \times (c_1 - 1)}$ are the discriminant directions for conditional classification of Y_1 when $Y_2 = k_2$. For all c_2 possible values of Y_2, we let $\boldsymbol{\theta}^r := \boldsymbol{\theta}^r(\boldsymbol{\beta}) = \boldsymbol{\beta} \times_2 \mathbf{A}^{c_1} \in \mathbb{R}^{p \times (c_1 - 1) \times c_2}$ denote tensor parameter of interest for conditional classification of Y_1 given Y_2 and $\mathbf{X}$.

Similarly, when given $Y_1 = k_1$ for any $k_1 \in [c_1]$, the conditional discriminant directions for Y_2 become $(\boldsymbol{\beta}_{k_1 2} - \boldsymbol{\beta}_{k_1 1}, \ldots, \boldsymbol{\beta}_{k_1 c_2} - \boldsymbol{\beta}_{k_1 1}) = \boldsymbol{\beta}_{k_1 \cdot}(\mathbf{A}^{c_2})^\top \in \mathbb{R}^{p \times (c_2 - 1)}$, where $\mathbf{A}^{c_2} = (-\mathbf{1}_{c_2 - 1}, \mathbf{I}_{c_2 - 1}) \in \mathbb{R}^{(c_2 - 1) \times c_2}$ and $\boldsymbol{\beta}_{k_1 \cdot} = (\boldsymbol{\beta}_{k_1 1}, \ldots, \boldsymbol{\beta}_{k_1 c_2})$ is the k_1-th lateral slice of $\boldsymbol{\beta}$. Thus, the parameter of interest for conditional classification under model (1) includes the two conditional discriminant coefficient tensors $\boldsymbol{\theta} := \{\boldsymbol{\theta}^r(\boldsymbol{\beta}), \boldsymbol{\theta}^c(\boldsymbol{\beta})\}$, where $\boldsymbol{\theta}^r(\boldsymbol{\beta}) = \boldsymbol{\beta} \times_2 \mathbf{A}^{c_1} \in \mathbb{R}^{p \times (c_1 - 1) \times c_2}$ and $\boldsymbol{\theta}^c(\boldsymbol{\beta}) = \boldsymbol{\beta} \times_3 \mathbf{A}^{c_2} \in \mathbb{R}^{p \times c_1 \times (c_2 - 1)}$.

The total number of free parameters in the discriminant directions, to be estimated, is $p(c_1 c_2 - 1)$ for joint classification and $pc_1(c_2 - 1) + p(c_1 - 1)c_2$ for conditional classification.

2.3 Connections to Classical LDA and Marginal Classification

When the focus is joint classification, we can transform model (1) to a univariate LDA model that combines all the categories (k_1, k_2) for $k_1 \in [c_1]$, $k_2 \in [c_2]$ to a univariate categorical response $\widetilde{Y} \in [K]$, $K = c_1 c_2$. Specifically, for some bijective function $h : [c_1] \times [c_2] \to [K]$, $\widetilde{Y}$ satisfies $\Pr(\widetilde{Y} = h(k_1, k_2) \mid \mathbf{X}) = \Pr(Y_1 = k_1, Y_2 = k_2 \mid \mathbf{X})$ for all $k_1 \in [c_1], k_2 \in [c_2]$. Then the joint classification of $\mathbf{Y}$ is equivalent to the classification of $\widetilde{Y}$. Let $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_{(1)} \in \mathbb{R}^{p \times K}$ be the mode-1 matricization of the mean tensor $\boldsymbol{\mu}$ with class means arranged along K columns, where $\widetilde{\boldsymbol{\mu}}_k$ is the k-th column of $\widetilde{\boldsymbol{\mu}}$ that represents the mean of the k-th category, and π_k be the membership weight for the k-th category. The Bayes' rule $\phi_{\widetilde{Y}} : \mathbb{R}^p \to [K]$ for the category-combined LDA is

$$\phi_{\widetilde{Y}}(\mathbf{X}) = \underset{k \in [K]}{\operatorname{argmax}} \left\{ \left(\mathbf{X} - \frac{\widetilde{\boldsymbol{\mu}}_k + \widetilde{\boldsymbol{\mu}}_1}{2} \right)^\top \boldsymbol{\Sigma}^{-1} (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1) + \log \pi_k \right\}. \quad (6)$$

The dependence between the Bayes' rule and predictor $\mathbf{X}$ is captured linearly by $\boldsymbol{\Sigma}^{-1}(\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)$ for $k = 2, \ldots, K$, which we refer to as the discriminant directions. Thus, the parameter of interest for joint classification is $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}(\widetilde{\boldsymbol{\mu}}_2 - \widetilde{\boldsymbol{\mu}}_1, \ldots, \widetilde{\boldsymbol{\mu}}_K - \widetilde{\boldsymbol{\mu}}_1) \in \mathbb{R}^{p \times (K-1)}$.

At the population level, the joint classification Bayes' rule (4) is equivalent to the category combined Bayes rule (6). However, combining the categories in $\widetilde{Y}$ ignores the fact that the two responses Y_1 and Y_2 are distinct. Recognizing the two distinct responses, we propose novel group-wise regularization on the discriminant coefficient tensor for more effective variable selection and prediction. In Section 3.2, particularly Figure 1, we show that the proposed new group penalization accounts for the difference in selecting relevant variables in two responses' classification while varying different groups of the other response; in Section 3.3, particularly Figure 2, we develop a new parameterization and a group penalty for conditional classification, which is based on a different Bayes' rule and employs different variables than joint classification. Simply combining the categories in $\widetilde{Y}$ would fail to exploit the special group structures arising from multiple responses.

While joint and conditional classification are the focus of this paper, we briefly discuss the separate classification of each response. Under model (1), the marginal Bayes' rule $\phi_{Y_1} : \mathbb{R}^p \to [c_1]$ for classification of Y_1 can be written as

$$\phi_{Y_1}(\mathbf{X}) = \underset{k_1 \in [c_1]}{\operatorname{argmax}} \sum_{k_2=1}^{c_2} \pi_{k_1 k_2} \exp\left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_{k_1 k_2} + \boldsymbol{\mu}_{11}}{2} \right)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{k_1 k_2} - \boldsymbol{\mu}_{11}) \right\}, \tag{7}$$

whose decision boundary is nonlinear unlike the linear decisions from the joint and conditional Bayes' rules. The marginal Bayes' rule (7) may give drastically different classification results than the joint Bayes' rule (6) or the conditional Bayes' rule (5). For example, for a given predictor $\mathbf{X}$, suppose that the joint distribution of $\mathbf{Y} = (Y_1, Y_2)$ is given by $\Pr(Y_1 = 1, Y_2 = 1 \mid \mathbf{X}) = 0.4$, $\Pr(Y_1 = 1, Y_2 = 2 \mid \mathbf{X}) = 0$, $\Pr(Y_1 = 2, Y_2 = 1 \mid \mathbf{X}) = 0.3$ and $\Pr(Y_1 = 2, Y_2 = 2 \mid \mathbf{X}) = 0.3$. Then the joint Bayes' rule (6) gives $\widehat{Y}_1 = 1$, and the conditional Bayes' rule (5) gives $\widehat{Y}_1 = 1$ or $\widehat{Y}_1 = 2$ depending on Y_2. However, the marginal Bayes' rule (7) results in $\widehat{Y}_1 = 2$. This toy example illustrates that the marginal Bayes' rule (7) is potentially misleading for both joint and conditional classification. In addition, variable selection based on the marginal Bayes' rule in general also differs from those of joint and conditional classifications. In other words, whenever we observe multiple categorical responses, it is inappropriate to separately predict each response ignoring the information from all other responses.

3. Methodology

3.1 Convex Objective Function

Existing approaches for high-dimensional LDA focus on penalizing the discriminant directions or within class means. Our formulation is an extension of the multi-class sparse discriminant analysis method (MSDA, Mai et al., 2019) that modifies its group lasso penalty on the discriminant directions $\lambda \|\widetilde{\boldsymbol{\beta}}\|_{2,1}$ to exploit the multivariate construction of the response. Recall that $\widetilde{\boldsymbol{\beta}}$ is from the Bayes' rule (6), where $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p \times (c_1 c_2 - 1)}$ is essentially the mode-1 matricization of the discriminant coefficient tensor $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{(1)} = (\mathbf{0}_p, \widetilde{\boldsymbol{\beta}})$ with $\mathbf{0}_p$ being the p-dimensional vector of zeros.

Following the same strategy as MSDA, we recognize that the discriminant coefficient tensor $\boldsymbol{\beta}$ is the minimizer of the following convex function $g : \mathbb{R}^{p\times c_1\times c_2} \to \mathbb{R}$,

$$g(\boldsymbol{\beta}) = \sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\left\{\frac{1}{2}\boldsymbol{\beta}_{k_1k_2}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_{k_1k_2} - \boldsymbol{\delta}_{k_1k_2}^\top\boldsymbol{\beta}_{k_1k_2}\right\}, \tag{8}$$

where $\boldsymbol{\delta}_{k_1k_2} = \boldsymbol{\mu}_{k_1k_2} - \boldsymbol{\mu}_{11}$. The formulation (8) does not require an inverse covariance matrix and, with appropriate penalization, serves as an objective function for obtaining sparse estimates of the Bayes' rule discriminant directions. Next, we propose different penalties to be added to g with the focuses on joint and conditional classification in high-dimensions.

3.2 Penalty for Joint Classification

For the joint classification of (Y_1, Y_2), the joint Bayes' rule (4) indicates that the j-th predictor X_j, the j-th component of $\mathbf{X}$, does not affect the joint classification if and only if $\boldsymbol{\beta}_{[j,k_1,k_2]} = 0$ for all $k_1 \in [c_1]$ and $k_2 \in [c_2]$. We define the discriminant set $\mathcal{S} = \{j : \boldsymbol{\beta}_{[j,k_1,k_2]} \neq 0 \text{ for some } k_1 \in [c_1], k_2 \in [c_2]\}$ and assume the number of important variables $|\mathcal{S}| \ll p$. In order to take advantage of the tensor structure of $\boldsymbol{\beta}$ and distinguish Y_1 and Y_2 in their joint classification, we view $\boldsymbol{\beta}$ in two directions (mode-2 and mode-3) for the classification of Y_1 and Y_2, respectively. For Y_1 classification, we fix $k_2 \in [c_2]$ and let

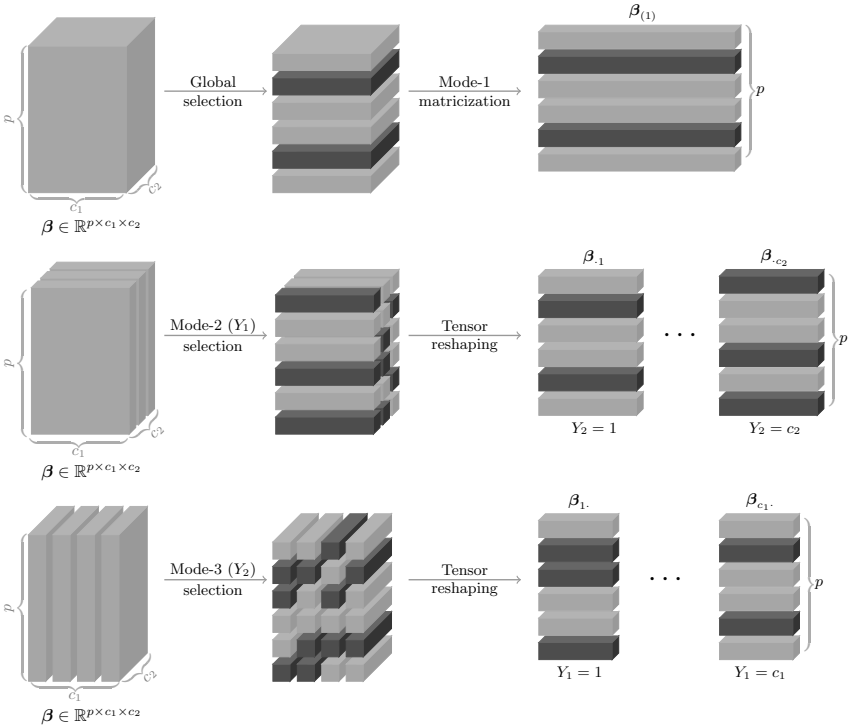

Figure 1: Visualization of different variable selection mechanisms. Top row: variable selection using category combined response $\widetilde{Y}$ and the MSDA method. Middle and bottom rows: variable selection focusing on mode-2 (Y_1) and mode-3 (Y_2) of the discriminant coefficient tensors by the proposed method.

$\mathcal{S}_{\cdot k_2} = \{j : \boldsymbol{\beta}_{[j,k_1,k_2]} \neq 0 \text{ for some } k_1 \in [c_1]\}$. It is straightforward to see $\mathcal{S} = \cup_{k_2=1}^{c_2} \mathcal{S}_{\cdot k_2}$. Therefore, by assuming the sparsity of each frontal slice $\boldsymbol{\beta}_{\cdot k_2}$ for $k_2 \in [c_2]$, we can achieve variable selection for joint classification. Similarly, for Y_2 classification, we assume lateral slices $\boldsymbol{\beta}_{k_1 \cdot}$ to be sparse.

We propose to penalize the frontal and lateral slices $\boldsymbol{\beta}_{\cdot k_2}$ and $\boldsymbol{\beta}_{k_1 \cdot}$ of the coefficient tensor $\boldsymbol{\beta}$ to select the mode-2 and mode-3 fibers in $\boldsymbol{\beta}$. These two directions of the discriminant coefficient tensor are associated with important variables for classifying Y_1 and for classifying Y_2, respectively. In this way, we account for the distinct response variables in a more nuanced way than simply combining them into a synthetic univariate response $\widetilde{Y}$. As a result, we achieve more efficient estimation of $\boldsymbol{\beta}$, more interpretable variable selection, and higher classification accuracy (see Section 4 for some theoretical insights). Specifically, we propose the following penalty for joint classification,

$$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{k_2=1}^{c_2} \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1} + \lambda_2 \sum_{k_1=1}^{c_1} \|\boldsymbol{\beta}_{k_1 \cdot}\|_{2,1}, \tag{9}$$

where $\lambda = (\lambda_1, \lambda_2)^\top \in \mathbb{R}_+^2$ is the tuning parameter and $\sum_{k_2=1}^{c_2} \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}$ is the mode-2 penalty that encourages some predictor variables to be irrelevant for classification of Y_1. Similarly, $\sum_{k_1=1}^{c_1} \|\boldsymbol{\beta}_{k_1 \cdot}\|_{2,1}$ is the mode-3 penalty that focuses on Y_2 classification. The penalty $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is a special case of the group lasso penalty (Yuan and Lin, 2006), where we encourage only a subset of p predictors to be important in the bivariate LDA problem for (Y_1, Y_2) classification.

The sparse structure encouraged by $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is illustrated in Figure 1. It can be seen that this new penalty is flexible by selecting different variables for one response's classification while varying different groups of the other response.

3.3 Penalty for Conditional Classification

As discussed in Section 2.2, the tensor parameters of interest for conditional classification are $\boldsymbol{\theta}^r \in \mathbb{R}^{p \times (c_1-1) \times c_2}$ and $\boldsymbol{\theta}^c \in \mathbb{R}^{p \times c_1 \times (c_2-1)}$, which are transformed from $\boldsymbol{\beta}$ by multiplying with constant contrast matrices to reduce redundancies in the parameters. As a direct consequence of this change, the important variables for conditional classification are also different from the joint classification. For example, if the j-th predictor is identified as important for joint classification, this does not necessarily imply it is important for conditional classification given $Y_2 = k_2$. The conditional Bayes' rule (5) suggests that if $\boldsymbol{\beta}_{[j,1,k_2]} = \boldsymbol{\beta}_{[j,2,k_2]} = \cdots = \boldsymbol{\beta}_{[j,c_1,k_2]} = c$ for some constant $c \neq 0$, the j-th predictor X_j is not involved in conditional classification of Y_1 since $\boldsymbol{\theta}^r_{[j,k_1,k_2]} = 0$ for $k_1 \in [c_1 - 1]$. Thus, when conditional on $Y_2 = k_2$, we define the j-th predictor as unimportant if and only if $\boldsymbol{\theta}^r_{[j,k_1,k_2]} = \boldsymbol{\beta}_{[j,k_1+1,k_2]} - \boldsymbol{\beta}_{[j,1,k_2]} = 0$ for all $k_1 \in [c_1 - 1]$ and define the discriminant set for conditional classification of Y_1 as $\mathcal{A}_{\cdot k_2} = \{j : \boldsymbol{\theta}^r_{[j,k_1,k_2]} \neq 0 \text{ for some } k_1 \in [c_1 - 1]\}$. Therefore, we directly penalize the conditional discriminant coefficient tensors $\boldsymbol{\theta}^r$ and $\boldsymbol{\theta}^c$ for conditional variable selection and classification.

We propose the penalty for conditional classification of both Y_1 and Y_2 as

$$\mathcal{H}_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{k_2=1}^{c_2} \|\boldsymbol{\theta}^r_{\cdot k_2}\|_{2,1} + \lambda_2 \sum_{k_1=1}^{c_1} \|\boldsymbol{\theta}^c_{k_1 \cdot}\|_{2,1}, \tag{10}$$

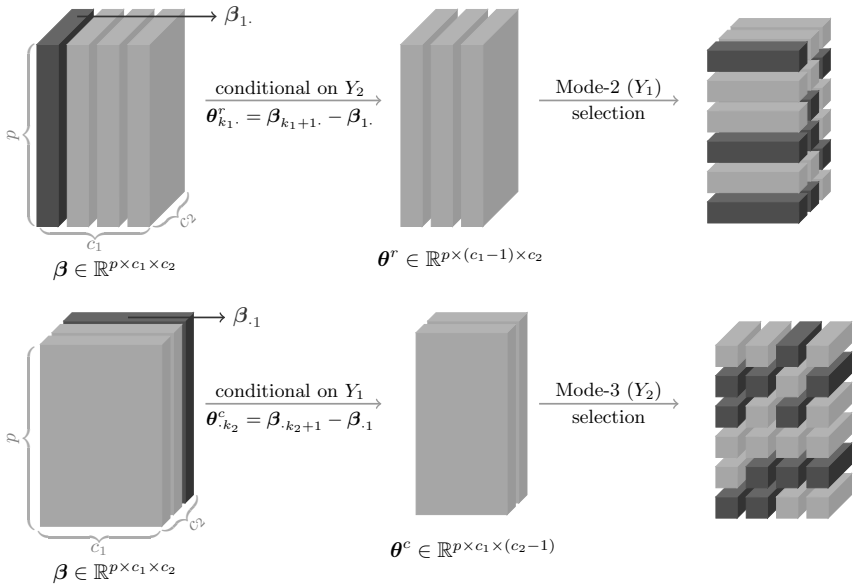

Figure 2: Visualization of variable selection for conditional classification. We focus on the conditional classification of Y_1 and Y_2 by directly imposing sparse structures on conditional discriminant parameters $\boldsymbol{\theta}^r$ and $\boldsymbol{\theta}^c$ transformed from the joint discriminant parameter $\boldsymbol{\beta}$.

where $\boldsymbol{\theta}^r = \boldsymbol{\beta} \times_2 \mathbf{A}^{c_1} \in \mathbb{R}^{p\times(c_1-1)\times c_2}$ and $\boldsymbol{\theta}^c = \boldsymbol{\beta} \times_3 \mathbf{A}^{c_2} \in \mathbb{R}^{p\times c_1\times(c_2-1)}$. The parameterization from $\boldsymbol{\beta}$ to $\boldsymbol{\theta}^r$ and $\boldsymbol{\theta}^c$ and the sparse structure encouraged by $\mathcal{H}_\lambda(\boldsymbol{\beta})$ are intuitively visualized in Figure 2. The mode-2 selection on $\boldsymbol{\theta}^r$ and the mode-3 selection on $\boldsymbol{\theta}^c$ are analogous to the mode-2 and mode-3 selection on $\boldsymbol{\beta}$for join classification (Figure 1).

Finally, it is important to note that $\mathcal{A}_{\cdot k_2} \subseteq \mathcal{S}_{\cdot k_2}$. This implies that our variable selection for conditional classification is a refinement step to the variable selection for joint classification. In practice, when the task is joint classification, we use $\mathcal{P}_\lambda(\boldsymbol{\beta})$ to regularize the joint discriminant tensor $\boldsymbol{\beta}$; when the task also includes conditional classification, we further regularize the conditional discriminant tensor $(\boldsymbol{\theta}^r(\boldsymbol{\beta}), \boldsymbol{\theta}^c(\boldsymbol{\beta}))$ based on the subset of variables that are already selected for joint classification.

3.4 Proposed Estimation Criteria

Suppose we have independent observations $\{\mathbf{X}_i, Y_{1i}, Y_{2i}\}_{i=1}^n$ from the bivariate LDA (1). Then the unpenalized sample version of g, denoted g_n, is given by

$$g_n(\boldsymbol{\beta}) = \sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\left\{\frac{1}{2}\boldsymbol{\beta}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \boldsymbol{\beta}_{k_1k_2}\right\}, \tag{11}$$

where $\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^n (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_{1i}Y_{2i}})(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_{1i}Y_{2i}})^\top/n$, $\widehat{\boldsymbol{\delta}}_{k_1k_2} = \widehat{\boldsymbol{\mu}}_{k_1k_2} - \widehat{\boldsymbol{\mu}}_{11}$, and $\widehat{\boldsymbol{\mu}}_{k_1k_2} = \sum_{i=1}^n \mathbf{X}_i I(Y_{1i}=k_1, Y_{2i}=k_2)/\sum_{i=1}^n I(Y_{1i}=k_1, Y_{2i}=k_2)$ for each $(k_1,k_2) \in [c_1]\times[c_2]$.

For joint classification, we solve the following regularized optimization problem to obtain sparse estimates of the joint discriminant coefficient tensor $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} \in \underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^{p\times c_1\times c_2}\\ \boldsymbol{\beta}_{11}=0}}{\operatorname{argmin}} \Big\{ g_n(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}) \Big\}. \tag{12}$$

For conditional classification, we impose the penalty $\mathcal{H}_\lambda(\boldsymbol{\beta})$ that targets conditional discriminant coefficient tensors $\boldsymbol{\theta}^r$ and $\boldsymbol{\theta}^c$ and solve

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} \in \underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^{p\times c_1\times c_2}\\ \boldsymbol{\beta}_{11}=0}}{\operatorname{argmin}} \Big\{ g_n(\boldsymbol{\beta}) + \mathcal{H}_\lambda(\boldsymbol{\beta}) \Big\} \quad \text{subject to } \boldsymbol{\theta}^r = \boldsymbol{\beta} \times_2 \mathbf{A}^{c_1}, \boldsymbol{\theta}^c = \boldsymbol{\beta} \times_3 \mathbf{A}^{c_2}. \tag{13}$$

Although the loss function g_n and optimization parameter $\boldsymbol{\beta}$ in regularization problems (12) and (13) are the same, the focus of (13) is no longer obtaining $\boldsymbol{\beta}$ but the reduced tensor parameters $\boldsymbol{\theta}^r$ and $\boldsymbol{\theta}^c$. Different algorithms are used to solve the two optimization problems.

The intrinsic constraint $\boldsymbol{\beta}_{11} = 0$ does not complicate the optimization problems. Recall that $\boldsymbol{\beta}_{11} = 0$ by definition, so this is simply a mode-1 fiber of $\boldsymbol{\beta}$ that does not need to be updated in the algorithms.

4. Statistical Properties

We next consider the statistical properties of the proposed estimator for joint classification. Specifically, we establish a high-probability error bound for $\widehat{\boldsymbol{\beta}}$, defined in (12), that allows n and p to both diverge, while allowing p to diverge much faster than n.

We begin by stating the assumptions under which our error bound holds. For a symmetric positive definite matrix $\mathbf{M}$, its largest and smallest eigenvalues are denoted by $\varphi_{\max}(\mathbf{M})$ and $\varphi_{\min}(\mathbf{M})$.

A1. There exists a constant $v_\varphi > 0$ such that $v_\varphi^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}) \leq \varphi_{\max}(\boldsymbol{\Sigma}) \leq v_\varphi$.

A2. There exists a constant $v_\pi > 0$ such that $\pi_{k_1k_2} \geq v_\pi$ for all $k_1 \in [c_1]$ and $k_2 \in [c_2]$.

A3. There exists a constant $v_\Delta > 0$ such that $v_\Delta \leq (\boldsymbol{\mu}_{k_1k_2} - \boldsymbol{\mu}_{11})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{k_1k_2} - \boldsymbol{\mu}_{11}) \leq 3v_\Delta$ for all $k_1 \in [c_1]$ and $k_2 \in [c_2]$ with $(k_1, k_2) \neq (1, 1)$.

Assumptions **A1** and **A2** are common in high-dimensional linear discriminant analysis (e.g., Cai and Liu, 2011; Mai et al., 2019). Assumption **A2** implicitly requires that the number of response categories c_1 and c_2 are bounded. As such, our error bound treats c_1 and c_2 as fixed constants. Assumption **A3** concerns the separability between the pair of classes $(Y_1, Y_2) = (k_1, k_2)$ and $(Y_1, Y_2) = (1, 1)$. This assumption is also common in high-dimensional linear discriminant analysis (Min et al., 2023).

Recall from the beginning of Section 3.2 that $\mathcal{S}_{\cdot k_2} = \{j : \boldsymbol{\beta}_{[j,k_1,k_2]} \neq 0 \text{ for some } k_1 \in [c_1]\}$ and $\mathcal{S}_{k_1\cdot} = \{j : \boldsymbol{\beta}_{[j,k_1,k_2]} \neq 0 \text{ for some } k_2 \in [c_2]\}$. Define $s_{2,k_2} = |\mathcal{S}_{\cdot k_2}|$ as the cardinality of $\mathcal{S}_{\cdot k_2}$, define $s_{1,k_1} = |\mathcal{S}_{k_1\cdot}|$ as the cardinality of $\mathcal{S}_{k_1\cdot}$, and define $s_\ell^* = \max\{s_{\ell,1}, \ldots, s_{\ell,c_\ell}\}$ for $\ell \in \{1, 2\}$. Implicitly, we assume $s_\ell^* \geq 1$ for $\ell \in \{1, 2\}$. For a constant $\alpha \in [0, 1]$, define $(a, b)_\alpha = \alpha^2 a + (1 - \alpha)^2 b$.

With this notation in hand, we are ready to state our main result.

Theorem 1 *Suppose the data are generated from model* (1)*, Assumptions* ***A1****–****A3*** *hold, and* $\max(s_1^*, s_2^*)\log(p)/n \to 0$ *as* $n \to \infty$*. Let* $\phi \in [0,1]$ *be a fixed constant. If* $\lambda_1 = \phi M\{c_2\log(p)/n\}^{1/2}$ *and* $\lambda_2 = (1-\phi)M\{c_1\log(p)/n\}^{1/2}$ *for constant* $M > 0$ *sufficiently large, then there exists a constant* $v_1 \in (0,\infty)$ *such that*

$$\Pr\left\{\sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \le v_1(s_1^*, s_2^*)_\phi \frac{\log(p)}{n}\right\} \ge 1 - O(p^{-1})$$

for n *sufficiently large.*

The result of Theorem 1 reveals that the error bound is determined by $n^{-1}\log(p)$ and $(s_1^*, s_2^*)_\phi$, a linear combination of the number of important predictors for classifying each of the two responses over all categories of the other. The effect of each mode's sparsity can be modified by the choice of ϕ. The optimal ϕ is $\phi^* = s_2^*/(s_1^* + s_2^*)$, for which it can be verified that $(s_1^*, s_2^*)_{\phi^*} < \min(s_1^*, s_2^*) \le \max(s_1^*, s_2^*)$.

The error bound in Theorem 1 should be compared to the error bound which could be obtained by ignoring the bivariate construction of the response (Y_1, Y_2). Specifically, let $\bar{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p\times c_1\times c_2}, \boldsymbol{\beta}_{11}=0}\{g_n(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}_{(1)}\|_{2,1}\}$ be the MSDA estimator from Mai et al. (2019) applied to the response $\widetilde{Y}$ with c_1c_2 many response categories. Under the assumptions **A1**–**A3**, if the tuning parameter was selected appropriately, the asymptotic error bound would be $\sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\|\bar{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \le v_1'|\mathcal{S}|\log(p)/n$ where $v_1' \in (0,\infty)$ is a constant and $\mathcal{S} = \{j : \boldsymbol{\beta}_{[j,k_1,k_2]} \ne 0 \text{ for some } (k_1,k_2) \in [c_1]\times[c_2]\}$. This result follows directly from Theorem 3.3 of Min et al. (2023) with their $M = 1$.

We can also establish a result on how well our estimator recovers an optimal classification rule. Define

$$D_{k_1,k_2}(\mathbf{X}) = \left\{\left(\mathbf{X} - \frac{\boldsymbol{\mu}_{k_1k_2} + \boldsymbol{\mu}_{11}}{2}\right)^\top \boldsymbol{\beta}_{k_1k_2} + \log\pi_{k_1k_2}\right\}$$

and let $\widehat{D}_{k_1,k_2}(\mathbf{X})$ be the version of $D_{k_1,k_2}(\mathbf{X})$ with $\boldsymbol{\beta}_{k_1k_2}$ replaced with $\widehat{\boldsymbol{\beta}}_{k_1k_2}$, and $(\boldsymbol{\mu}_{k_1k_2}, \pi_{k_1k_2})$ replaced with their maximum likelihood estimators.

Following Min et al. (2023), we define the optimal "strong misclassification rate" as

$$R_{\boldsymbol{\Theta}}^{\text{opt}} = \sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\sum_{(u_1,u_2)\ne(k_1,k_2)}\pi_{k_1k_2}\Pr\nolimits_{\boldsymbol{\Theta}}\{D_{k_1,k_2}(\mathbf{X}) < D_{u_1,u_2}(\mathbf{X}) \mid \text{labels}(\mathbf{X}) = (k_1,k_2)\}$$

where $\boldsymbol{\Theta} = (\pi_{11},\ldots,\pi_{c_1c_2},\boldsymbol{\mu}_{11},\ldots,\boldsymbol{\mu}_{c_1c_2},\boldsymbol{\Sigma})$. The strong misclassification rate upper bounds the misclassification rate (which replaces $\sum_{(u_1,u_2)\ne(k_1,k_2)}\pi_{k_1k_2}\Pr_{\boldsymbol{\Theta}}\{D_{k_1,k_2}(\mathbf{X}) < D_{u_1,u_2}(\mathbf{X}) \mid \text{labels}(\mathbf{X}) = (k_1,k_2)\}$ with $\pi_{k_1k_2}\Pr_{\boldsymbol{\Theta}}\{D_{k_1,k_2}(\mathbf{X}) < D_{u_1,u_2}(\mathbf{X})$ for some $(u_1u_2) \ne (k_1,k_2) \mid \text{labels}(\mathbf{X}) = (k_1,k_2)\}$), and serves as a measure of classification accuracy. Here, labels$(\mathbf{X})$ denotes the true response category combination of $\mathbf{X}$. The strong misclassification rate for our estimator is

$$R_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\Theta}}_{\text{MLDA}}) = \sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\sum_{(u_1,u_2)\ne(k_1,k_2)}\pi_{k_1k_2}\Pr\nolimits_{\boldsymbol{\Theta}}\{\widehat{D}_{k_1,k_2}(\mathbf{X}) < \widehat{D}_{u_1,u_2}(\mathbf{X}) \mid \text{labels}(\mathbf{X}) = (k_1,k_2)\}$$

where $\widehat{\boldsymbol{\Theta}}_{\text{MLDA}}$ denotes the set of estimators defining $\widehat{D}$.

We have the following proposition, which follows in part from the result of Theorem 1, and from the proof of Theorem 3.3 of Min et al. (2023).

Proposition 1 *Let* $\mathcal{G}(p, v_\pi, v_\varphi, v_\Delta, s_1^*, s_2^*) = \{\boldsymbol{\Theta} = (\pi_{11}, \ldots, \pi_{c_1c_2}, \boldsymbol{\mu}_{11}, \ldots, \boldsymbol{\mu}_{c_1c_2}, \boldsymbol{\Sigma}) : v_\varphi^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}) \leq \varphi_{\max}(\boldsymbol{\Sigma}) \leq v_\varphi, \pi_{k_1k_2} \geq v_\pi$ *for each* (k_1, k_2) *and* $v_\Delta \leq (\boldsymbol{\mu}_{k_1k_2} - \boldsymbol{\mu}_{11})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{k_1k_2} - \boldsymbol{\mu}_{11}) \leq 3v_\Delta$ *for each* $(k_1, k_2) \neq (1, 1), \max_{k \in [c_2]} |\mathcal{S}_{\cdot k}| \leq s_2^*, \max_{k \in [c_1]} |\mathcal{S}_{k\cdot}| \leq s_1^*\}$. *Under the conditions of Theorem 1, there exists a constant* $v_2 \in (0, \infty)$ *such that*

$$\inf_{\boldsymbol{\Theta} \in \mathcal{G}} \Pr\left\{ R_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\Theta}}_{\text{MLDA}}) - R_{\boldsymbol{\Theta}}^{\text{opt}} \leq v_2 \max(s_1^*, s_2^*) \frac{\log(p)}{n} \right\} \geq 1 - O(p^{-1}),$$

for n *sufficiently large, where* $\mathcal{G} = \mathcal{G}(p, v_\pi, v_\varphi, v_\Delta, s_1^*, s_2^*)$.

The proof of both results in this section can be found in the Appendix.

5. Alternative Regularization Strategy

To compute (12), we recommend using an accelerated proximal gradient descent algorithm (Parikh and Boyd, 2014, Chapter 4.3). Given the k-th and $(k-1)$-th iterates of $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\beta}^{(k-1)}$, respectively, the $(k+1)$-th iterate of this algorithm is given by

$$\boldsymbol{\beta}^{(k+1)} = \underset{\substack{\boldsymbol{\beta} \in \mathbb{R}^{p \times c_1 \times c_2} \\ \boldsymbol{\beta}_{11} = 0}}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{k_1=1}^{c_1} \sum_{k_2=1}^{c_2} \|\boldsymbol{\beta}_{k_1k_2} - \{\boldsymbol{\beta}_{k_1k_2}^{(k,k-1)} - \alpha \nabla_{\boldsymbol{\beta}_{k_1k_2}} g_n(\boldsymbol{\beta}^{(k,k-1)})\}\|_2^2 + \mathcal{P}_{\alpha\lambda}(\boldsymbol{\beta}) \right],$$

where $\alpha > 0$ is a sufficiently small step size, $\boldsymbol{\beta}^{(k,k-1)} = \boldsymbol{\beta}^{(k)} + \frac{k}{k+3}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)})$, and $\nabla_{\boldsymbol{\beta}_{k_1k_2}} g_n$ is the gradient of g_n with respect to $\boldsymbol{\beta}_{k_1k_2}$. The optimization problem to compute $\boldsymbol{\beta}^{(k+1)}$ is separable across the first mode of $\boldsymbol{\beta}$, so this reduces to p separate optimizations problems, each of the form $\operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{c_1 \times c_2}} \left\{ \frac{1}{2}\|\mathbf{A} - \mathbf{B}\|_F^2 + \mathcal{P}_\lambda(\mathbf{A}) \right\}$, which may be solved using existing iterative algorithms (e.g., Yuan et al., 2011). Because each iterate of this accelerated proximal gradient descent scheme will itself require p parallel iterative algorithms, we next propose an alternative formulation based on latent overlapping groups.

The estimator proposed in Section 3 utilizes a variation of the group lasso penalty with overlapping groups (Yan and Bien, 2017). As such, this penalty can impose a specific type of sparsity in estimates of the discriminant vectors $\boldsymbol{\beta}_{k_1k_2}$. Namely, this penalty allows nonzero coefficient estimates to occur in the complement of unions of groups of coefficients. For example, in the case that $c_1 = c_2 = 3$ and $p = 1$, the overlapping group lasso penalty allows $\widehat{\boldsymbol{\beta}}_{[1,2,2]} \neq 0$ with all other components zero. However, for the sake of interpretability and computational efficiency, it may be preferable to only allow for entire groups to enter the model as zero or nonzero. For this, we can utilize an alternative variation of the group lasso penalty with overlapping groups: the latent overlapping group lasso penalty (Jacob et al., 2009; Obozinski et al., 2011; Yan and Bien, 2017).

When using the penalty $\mathcal{P}_\lambda$, there are pc_1c_2 parameters in total to be estimated in $\boldsymbol{\beta}$ and p groups of parameters to be selected in each $\boldsymbol{\beta}_{\cdot k_2}$ and $\boldsymbol{\beta}_{k_1\cdot}$, indicating $p(c_1 + c_2)$ groups of parameters in total are included in the group lasso penalty $\mathcal{P}_\lambda(\boldsymbol{\beta})$. However, each group $\boldsymbol{\beta}_{[j,:,k_2]}$ in $\boldsymbol{\beta}_{\cdot k_2}$ is overlapped with other c_1 groups $\boldsymbol{\beta}_{[j,1,:]}, \ldots, \boldsymbol{\beta}_{[j,c_1,:]}$. Figure 1 demonstrates this overlapping structure of $\mathcal{P}_\lambda(\boldsymbol{\beta})$: each mode-2 fiber in the middle tensor

of middle row is overlapped with c_1 mode-3 fibers in the middle tensor of bottom row. Similarly, each group $\boldsymbol{\beta}_{[j,k_1,:]}$ in $\boldsymbol{\beta}_{k_1,\cdot}$ is overlapped with other c_2 groups $\boldsymbol{\beta}_{[j,:,1]}, \ldots, \boldsymbol{\beta}_{[j,:,c_2]}$. Thus, the penalty is not separable and efficient algorithms—such as a blockwise coordinate descent algorithm—cannot be applied straightforwardly.

The latent overlapping group lasso penalty circumvents this issue of nonseperability. Let $G = G_1 \cup G_2$ be the index set of $p(c_1 + c_2)$ groups of parameters in $\boldsymbol{\beta}$, where $G_1 = \{g = (j, k_2, 1) \mid j \in [p], k_2 \in [c_2]\}$ and $G_2 = \{g = (j, k_1, 2) \mid j \in [p], k_1 \in [c_1]\}$ are the index sets of mode-2 selection groups and mode-3 selection groups, respectively. That is, for all tuples $g = (g_1, g_2, g_3) \in G$, if the third element $g_3 = 1$, then $g \in G_1$ and we use $\boldsymbol{\beta}_g \in \mathbb{R}^{c_1}$ to denote the mode-2 fiber $\boldsymbol{\beta}_{[j,:,k_2]}$ as illustrated in the middle tensor of the middle row in Figure 1; if the third element $g_3 = 2$, then $g \in G_2$ and we use $\boldsymbol{\beta}_g \in \mathbb{R}^{c_2}$ to denote the mode-3 fiber $\boldsymbol{\beta}_{[j,k_1,:]}$ as illustrated in the middle tensor of the bottom row in Figure 1. For all $g \in G$, let $\boldsymbol{\nu}^{(g)} \in \mathbb{R}^{p \times c_1 \times c_2}$ be a tensor such that $\boldsymbol{\nu}^{(g)}_\ell = 0$ if $\ell \neq g$ and $\mathcal{V}^{(g)} \subseteq \mathbb{R}^{p \times c_1 \times c_2}$ be the subspace of such tensors. The overlapping group lasso seeks to find $\boldsymbol{\nu}^{(g)}$'s such that $\boldsymbol{\beta} = \sum_{g \in G} \boldsymbol{\nu}^{(g)}$. Then for a group g, instead of directly penalizing $\boldsymbol{\beta}_g$, we penalize the corresponding $\boldsymbol{\nu}^{(g)}$, or equivalently the subset of the tensor $\boldsymbol{\nu}^{(g)}_g$. As such, the overlapped entries in $\mathcal{P}_\lambda$ is separated among the $\boldsymbol{\nu}^{(g)}$ so that we can penalize the $\boldsymbol{\nu}^{(\ell)}$ separately and obtain sparse structures similar to those that $\mathcal{P}_\lambda$ encourages. Specifically, letting $\boldsymbol{\nu} = \{\boldsymbol{\nu}^{(g)}\}_{g \in G}$, we seek to obtain $\boldsymbol{\nu}$ by solving

$$\widehat{\boldsymbol{\nu}} \in \underset{\boldsymbol{\nu}^{(g)} \in \mathcal{V}^{(g)}, g \in G}{\operatorname{argmin}} \left[\sum_{k_1, k_2} \left\{ \frac{1}{2} \Big(\sum_{g \in G} \boldsymbol{\nu}^{(g)}_{k_1 k_2} \Big)^\top \widehat{\boldsymbol{\Sigma}} \Big(\sum_{g \in G} \boldsymbol{\nu}^{(g)}_{k_1 k_2} \Big) - \widehat{\boldsymbol{\delta}}^\top_{k_1 k_2} \Big(\sum_{g \in G} \boldsymbol{\nu}^{(g)}_{k_1 k_2} \Big) \right\} + \sum_{g \in G} \lambda_g \|\boldsymbol{\nu}^{(g)}_g\|_2 \right], \tag{14}$$

where $\boldsymbol{\nu}^{(g)}_{k_1 k_2} \in \mathbb{R}^p$ is the mode-1 fiber $\boldsymbol{\nu}^{(g)}_{[:,k_1,k_2]}$, $\lambda_g = \lambda_1$ if $g \in G_1$ and $\lambda_g = \lambda_2$ if $g \in G_2$.

The latent overlapping group penalty can be defined in terms of $\boldsymbol{\beta}$ as

$$\mathcal{P}_{\mathcal{V},\lambda}(\boldsymbol{\beta}) := \inf_{\substack{\boldsymbol{\nu}^{(g)} \in \mathcal{V}^{(g)}, \\ \boldsymbol{\beta} = \sum_{g \in G} \boldsymbol{\nu}^{(g)}}} \sum_{g \in G} \lambda_g \|\boldsymbol{\nu}^{(g)}\|_2. \tag{15}$$

We present a visualization of the decomposition of $\boldsymbol{\beta}$ into the $\boldsymbol{\nu}^{(g)}$ in Figure 3. It can be shown (Jacob et al., 2009) that solving the optimization (14) provides a solution to

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{p \times c_1 \times c_2}}{\operatorname{argmin}} \Big\{ g_n(\boldsymbol{\beta}) + \mathcal{P}_{\mathcal{V},\lambda}(\boldsymbol{\beta}) \Big\}. \tag{16}$$

As we alluded to earlier in this section, the latent overlapping group formulation in (16) may be preferable to (12) for two main reasons. The sparsity pattern induced by (16) encourages the support of important variables to be the union of important groups and the optimization can be solved with an efficient blockwise coordinate descent algorithm that we describe in the next section. Nevertheless, we note that if the group lasso estimator in (12) is needed in practice, the proximal gradient descent algorithm outlined at the beginning of this section can be used to compute (12).

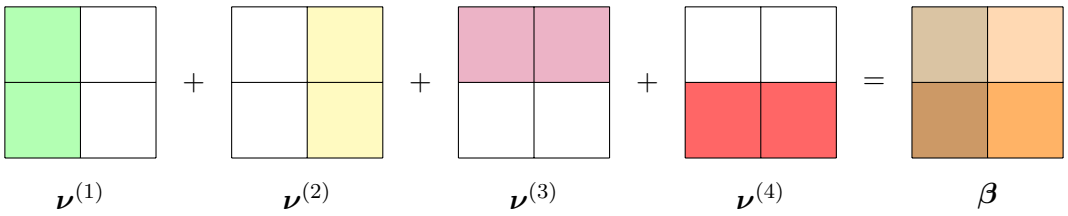

Figure 3: Visualization of how the overlapping group lasso penalty (15) decomposes components of $\boldsymbol{\beta}$ into components $\boldsymbol{\nu}^{(1)}, \ldots, \boldsymbol{\nu}^{(4)}$ in the special case that $c_1 = c_2 = 2$ and $p = 1$. In each matrix, colored coefficients can be nonzero, whereas white coefficients are zero by definition.

6. Estimation

6.1 Algorithm for Computing (16)

To apply the blockwise coordinate descent algorithm, we need only update one $\nu^{(g)}$ at a time with all others held fixed. For that, we have the following Proposition 2.

Proposition 2 *For all $g \in G$, if given $\boldsymbol{\nu}^{(\ell)}$ for $\ell \in G$ with $\ell \neq g$, then the solution of (14) is given by the proximal mapping*

$$\widehat{\boldsymbol{\nu}}_g^{(g)} = \underset{\boldsymbol{\nu}_g^{(g)}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{Z}^{(g)} - \sum_{\ell \in G} \boldsymbol{\nu}_g^{(\ell)}\|_2^2 + \frac{\lambda_g}{\widehat{\sigma}_{jj}} \|\boldsymbol{\nu}_g^{(g)}\|_2 \right\}, \tag{17}$$

where $\mathbf{Z}^{(g)}$ is an intermediate estimator for $\boldsymbol{\beta}_g$, $\widehat{\sigma}_{ij}$ is the (i,j)-th entry of $\widehat{\boldsymbol{\Sigma}}$. When $g \in G_1$ (mode-2 selection), $\mathbf{Z}^{(g)} = (\widehat{\boldsymbol{\delta}}_g - \boldsymbol{\beta}_{[:,:,k_2]}^\top \widehat{\boldsymbol{\Sigma}}_{\cdot j})/\widehat{\sigma}_{jj} + \boldsymbol{\beta}_g$ with $j = j(g) \in [p]$ and $k_2 = k_2(g) \in [c_2]$; when $g \in G_2$ (mode-3 selection), $\mathbf{Z}^{(g)} = (\widehat{\boldsymbol{\delta}}_g - \boldsymbol{\beta}_{[:,k_1,:]}^\top \widehat{\boldsymbol{\Sigma}}_{\cdot j})/\widehat{\sigma}_{jj} + \boldsymbol{\beta}_g$ with $j = j(g) \in [p]$ and $k_1 = k_1(g) \in [c_1]$.

Based on Proposition 2, we now summarize the estimation procedure for high-dimensional bivariate LDA in Algorithm 1. Note that here and elsewhere, the function $(a)_+ = \max(a, 0)$ is applied elementwise to its argument.

Proposition 3 *If $\widehat{\boldsymbol{\Sigma}} \succeq 0$, then Algorithm 1 converges to a global minimizer of (16). Furthermore, if $\widehat{\boldsymbol{\Sigma}} \succ 0$, then the global minimizer is unique.*

Proposition 3 ensures that Algorithm 1 converges to a global minimizer. When $\widehat{\boldsymbol{\Sigma}}$ is positive definite, (16) is strictly convex and the global minimizer is unique. In practice, if $\widehat{\boldsymbol{\Sigma}}$ is positive semidefinite, one can replace the estimate with $\widehat{\boldsymbol{\Sigma}}_\gamma = \widehat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p$ (Ledoit and Wolf, 2004) to ensure the uniqueness of the solution, where $\gamma > 0$ is a small constant.

After obtaining the estimated discriminant coefficient $\widehat{\boldsymbol{\beta}}$ from Algorithm 1, we get the joint prediction $\widehat{\mathbf{Y}}(\mathbf{X})$ by plugging the estimates $(\widehat{\boldsymbol{\beta}}_{k_1 k_2}, \widehat{\boldsymbol{\mu}}_{k_1 k_2}, \widehat{\pi}_{k_1 k_2})$, $k_1 \in [c_1]$, $k_2 \in [c_2]$ into the joint Bayes' rule (6). If the true label $Y_2 = k_2$ is observed, we obtain the conditional prediction $\widehat{Y}_1(\mathbf{X}, Y_2 = k_2)$ by plugging $(\widehat{\boldsymbol{\beta}}_{k_1 k_2}, \widehat{\boldsymbol{\mu}}_{k_1 k_2}, \widehat{\pi}_{k_1 k_2})$, $k_1 \in [c_1]$ into the conditional Bayes' rule (5).

6.2 Algorithm for Computing (13)

Let $\boldsymbol{w}^r \in \mathbb{R}^{p\times(c_1-1)\times c_2}$ and $\boldsymbol{w}^c \in \mathbb{R}^{p\times c_1\times(c_2-1)}$ be the Lagrangian multipliers associated with the two constraints in (13). And let $\boldsymbol{\theta} = (\mathrm{vec}(\boldsymbol{\theta}^r)^\top, \mathrm{vec}(\boldsymbol{\theta}^c)^\top)^\top \in \mathbb{R}^{p((c_1-1)c_2+c_1(c_2-1))}$, $\boldsymbol{w} = (\mathrm{vec}(\boldsymbol{w}^r)^\top, \mathrm{vec}(\boldsymbol{w}^c)^\top)^\top \in \mathbb{R}^{p((c_1-1)c_2+c_1(c_2-1))}$. Then for $\rho > 0$, the augmented Lagrangian of the objective function in (13) is

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{w}) = &\sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\left\{\frac{1}{2}\boldsymbol{\beta}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \boldsymbol{\beta}_{k_1k_2}\right\} + \lambda_1 \sum_{k_2=1}^{c_2}\|\boldsymbol{\theta}^r_{\cdot k_2}\|_{2,1} + \lambda_2 \sum_{k_1=1}^{c_1}\|\boldsymbol{\theta}^c_{k_1\cdot}\|_{2,1} \\ &+ \frac{\rho}{2}\|\mathbf{A}\mathrm{vec}(\boldsymbol{\beta}) - \boldsymbol{\theta} + \boldsymbol{w}\|_2^2 - \frac{\rho}{2}\|\boldsymbol{w}\|_2^2, \end{aligned} \tag{18}$$

where $\mathbf{A} = (\mathbf{I}_{c_2} \otimes (\mathbf{A}^{c_1})^\top \otimes \mathbf{I}_p, (\mathbf{A}^{c_2})^\top \otimes \mathbf{I}_{c_1} \otimes \mathbf{I}_p)^\top \in \mathbb{R}^{p((c_1-1)c_2+c_1(c_2-1))\times pc_1c_2}$.

Next, we apply the alternating direction method of multipliers (ADMM) to solve the optimization (13). Following Boyd et al. (2011), we iteratively update $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{w}$ in the Lagrangian (18) while fixing others to obtain $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{w}^{(t)}$ at step $t = 0, 1, 2, \ldots$ as

$$\boldsymbol{\beta}^{(t)} = \underset{\boldsymbol{\beta}, \boldsymbol{\beta}_{11}=0}{\mathrm{argmin}}\, L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{w}^{(t-1)}), \tag{19}$$

$$\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta}^r, \boldsymbol{\theta}^c}{\mathrm{argmin}}\, L_\rho(\boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}, \boldsymbol{w}^{(t-1)}), \tag{20}$$

$$\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(t-1)} + \mathbf{A}\mathrm{vec}(\boldsymbol{\beta}^{(t)}) - \boldsymbol{\theta}^{(t)}. \tag{21}$$

In the following Proposition, we provide simplified expressions for optimization problems (19) and (20), which lead to closed-form updates in the ADMM algorithm.

Algorithm 1 Blockwise coordinate descent algorithm for joint classification (16).

1. **Input**: Sample estimation $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\delta}}$, parameter groups G.
2. **Initialize**: $(\widehat{\boldsymbol{\nu}}^{(g)})^{(0)} = 0$ for all $g \in G$.
3. **Iterate**: For steps $t = 1, 2, \ldots,$ do the following until convergence.
   For all $g \in G$:
   (a) Compute $(\mathbf{Z}^{(g)})^{(t-1)}$ as defined in Proposition 2.
   (b) Update $(\widehat{\boldsymbol{\nu}}^{(g)})^{(t)}$ by solving (17):

$$(\widehat{\boldsymbol{\nu}}_g^{(g)})^{(t)} \longleftarrow \left(1 - \frac{\lambda_g/\widehat{\sigma}_{jj}}{\|(\mathbf{Z}^{(g)})^{(t-1)} - \sum_{\substack{\ell\in G,\\ \ell\neq g}}(\widehat{\boldsymbol{\nu}}_g^{(\ell)})^{(t-1)}\|_2}\right)_+ \left\{(\mathbf{Z}^{(g)})^{(t-1)} - \sum_{\substack{\ell\in G,\\ \ell\neq g}}(\widehat{\boldsymbol{\nu}}_g^{(\ell)})^{(t-1)}\right\}.$$

   (c) Update $\boldsymbol{\beta}^{(t)} = \sum_{g\in G}(\widehat{\boldsymbol{\nu}}^{(g)})^{(t)}$.
4. **Output**: $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$.

Proposition 4 *The vectorized minimizer of* (19) *is* $\text{vec}(\boldsymbol{\beta}^{(t)}) = (\mathbf{0}_p^\top, (\boldsymbol{\beta}_{-1}^{(t)})^\top)^\top$, *where* $\mathbf{0}_p$ *is a* p*-dimensional vector of zeros and* $\boldsymbol{\beta}_{-1}^{(t)} \in \mathbb{R}^{p(c_1c_2-1)}$ *is given by*

$$\boldsymbol{\beta}_{-1}^{(t)} = \underset{\boldsymbol{\beta}_{-1}}{\text{argmin}} \left\{ \frac{1}{2}\boldsymbol{\beta}_{-1}^\top \big(\mathbf{I}_{c_1c_2-1} \otimes \widehat{\boldsymbol{\Sigma}} + \rho \mathbf{A}_{-p}^\top \mathbf{A}_{-p}\big)\boldsymbol{\beta}_{-1} - \big(\rho \mathbf{A}_{-p}^\top (\boldsymbol{\theta}^{(t-1)} - \boldsymbol{w}^{(t-1)}) + \widehat{\boldsymbol{\delta}}_{-1}\big)^\top \boldsymbol{\beta}_{-1} \right\}, \tag{22}$$

with $\mathbf{A}_{-p}$ *being the submatrix of* $\mathbf{A}$ *with the first* p *columns removed and* $\widehat{\boldsymbol{\delta}}_{-1} \in \mathbb{R}^{(c_1c_2-1)p}$ *being the vectorized* $\widehat{\boldsymbol{\delta}}$ *excluding the zero* $\widehat{\boldsymbol{\delta}}_{11}$.

The minimizer of (20) *is the solution to*

$$(\boldsymbol{\theta}_{\cdot k_2}^r)^{(t)} = \underset{\boldsymbol{\theta}_{\cdot k_2}^r \in \mathbb{R}^{p\times(c_1-1)}}{\text{argmin}} \left\{ \frac{\rho}{2}\|\boldsymbol{\theta}_{\cdot k_2}^r - (\boldsymbol{\beta}_{\cdot k_2}^{(t)}(\mathbf{A}^{c_1})^\top + (\boldsymbol{w}_{\cdot k_2}^r)^{(t-1)})\|_F^2 + \lambda_1 \|\boldsymbol{\theta}_{\cdot k_2}^r\|_{2,1} \right\},\ k_2 \in [c_2]; \tag{23}$$

and

$$(\boldsymbol{\theta}_{k_1\cdot}^r)^{(t)} = \underset{\boldsymbol{\theta}_{k_1\cdot}^c \in \mathbb{R}^{p\times(c_2-1)}}{\text{argmin}} \left\{ \frac{\rho}{2}\|\boldsymbol{\theta}_{k_1\cdot}^c - (\boldsymbol{\beta}_{k_1\cdot}^{(t)}(\mathbf{A}^{c_2})^\top + (\boldsymbol{w}_{k_1\cdot}^c)^{(t-1)})\|_F^2 + \lambda_2 \|\boldsymbol{\theta}_{k_1\cdot}^c\|_{2,1} \right\},\ k_1 \in [c_1]. \tag{24}$$

Both (23) *and* (24) *can be solved in closed form using group-wise soft-thresholding.*

Based on Proposition 4, we can easily solve all the necessary steps of the ADMM algorithm for the conditional classification version of the high-dimensional bivariate LDA. We summarize the detailed algorithm in Appendix D (Algorithm A.1).

After the estimated conditional discriminant coefficient $\widehat{\boldsymbol{\theta}}^r$ and $\widehat{\boldsymbol{\theta}}^c$ are obtained from Algorithm A.1, the conditional prediction $\widehat{Y}_1(\mathbf{X}, Y_2 = k_2)$ is obtained by plugging the estimates $\widehat{\boldsymbol{\theta}}^r$, $\widehat{\boldsymbol{\mu}}$ and $\widehat{\pi}_{k_1k_2}$, $k_1 \in [c_1]$ into the conditional Bayes' rule (5). Similarly, the conditional $\widehat{Y}_2(\mathbf{X}, Y_1 = k_1)$ depends on $\widehat{\boldsymbol{\theta}}^c$, $\widehat{\boldsymbol{\mu}}$ and $\widehat{\pi}_{k_1k_2}$, $k_2 \in [c_2]$.

As discussed in Section 3.3, we consider the problem (13) for conditional classification as an additional step after solving problem (12) for joint classification. Specifically, Let $\widehat{\mathcal{S}} = \{j : \widehat{\boldsymbol{\beta}}_{[j,k_1,k_2]} \neq 0 \text{ for some } k_1 \in [c_1], k_2 \in [c_2]\}$ be the estimated set of important variables from Algorithm 1. We then apply Algorithm A.1 on the reduced data $\mathbf{X}_i^{\widehat{\mathcal{S}}} \in \mathbb{R}^{|\widehat{\mathcal{S}}|}$, $i \in [n]$, for conditional classification variable selection and estimation. This implementation also ensures that the solution to the optimization problem (22) for updating $\boldsymbol{\beta}^{(t)}$ in the ADMM algorithm to be always well-defined.

Finally, we have the following result.

Proposition 5 *Let* $(\boldsymbol{\beta}^{(t)}, (\boldsymbol{\theta}^r)^{(t)}, (\boldsymbol{\theta}^c)^{(t)})$ *denote the output of Algorithm A.1 after* t *iterations, and let* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}^r, \widehat{\boldsymbol{\theta}}^c)$ *be a global minimizer of* (13). *Then, as* $t \to \infty$, *we have* $\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}\|_F \to 0$, $\|(\boldsymbol{\theta}^r)^{(t)} - \widehat{\boldsymbol{\theta}}^r\|_F \to 0$ *and* $\|(\boldsymbol{\theta}^c)^{(t)} - \widehat{\boldsymbol{\theta}}^c\|_F \to 0$.

6.3 Tuning Parameter Selection

For both joint classification (12) and conditional classification coefficient tensor (13), and for their latent overlapping group lasso variations, the tuning parameters $\lambda = (\lambda_1, \lambda_2)^\top$ are selected by minimizing the 5-fold cross-validated joint or conditional classification error.

For joint classification, let $(\lambda^{11}, \ldots, \lambda^{K_1K_2}) = (\lambda_1^1, \ldots, \lambda_1^{K_1}) \times (\lambda_2^1, \ldots, \lambda_2^{K_2})$ denote $K_1 \times K_2$ candidate tuning parameters and $(\mathbf{X}^1, \mathbf{Y}^1), \ldots, (\mathbf{X}^5, \mathbf{Y}^5)$ denote the 5 data folds. For each candidate tuning parameter $\lambda^{k\ell}$ and testing set $(\mathbf{X}^m, \mathbf{Y}^m)$ of size n_m, we obtain the estimate $\widehat{\boldsymbol{\beta}}_{\lambda^{k\ell}}^{-m}$ (using data from all folds except the mth) and corresponding joint classification error $\mathrm{Err}_m^{k\ell} = \sum_{i=1}^{n_m} I(\widehat{\mathbf{Y}}_i^m(\mathbf{X}_i^m) \neq \mathbf{Y}_i^m)/n_m$ based on $\widehat{\boldsymbol{\beta}}_{\lambda^{k\ell}}^{-m}$. Then, the optimal tuning parameter $\widehat{\lambda}$ minimizes the following cross-validated joint classification error. That is, $\widehat{\lambda}$ is defined as $\widehat{\lambda} = \mathrm{argmin}_{\lambda^{k\ell}} \sum_{m=1}^5 \mathrm{Err}_m^{k\ell}$.

For conditional classification, similar to joint classification, we have $K_1 \times K_2$ candidate tuning parameters and 5 data folds. For each candidate tuning parameter $\lambda^{k\ell}$ and testing set $(\mathbf{X}^m, \mathbf{Y}^m)$ of size n_m, we obtain the estimates of conditional discriminant parameters $(\widehat{\boldsymbol{\theta}}^r, \widehat{\boldsymbol{\theta}}^c)_{\lambda^{k\ell}}^{-m}$ along with the corresponding conditional classification errors $\mathrm{Err}_{Y_1^m|Y_2^m}^{k\ell} = \sum_{k_2=1}^{c_2} \sum_{Y_{2i}^m=k_2} I(\widehat{Y}_{1i}^m(\mathbf{X}_i^m, Y_{2i}^m) \neq Y_{1i}^m)/n_m$ and $\mathrm{Err}_{Y_2^m|Y_1^m}^{k\ell} = \sum_{k_1=1}^{c_1} \sum_{Y_{1i}^m=k_1} I(\widehat{Y}_{2i}^m(\mathbf{X}_i^m, Y_{1i}^m) \neq Y_{2i}^m)/n_m$. Then, the optimal tuning parameter $\widehat{\lambda}$ minimizes the following cross-validated conditional classification error. That is, $\widehat{\lambda} = \mathrm{argmin}_{\lambda^{k\ell}} \sum_{m=1}^5 \left(\mathrm{Err}_{Y_1^m|Y_2^m}^{k\ell} + \mathrm{Err}_{Y_2^m|Y_1^m}^{k\ell}\right)$.

As we describe in the next section, with response $\mathbf{Y} = (Y_1, \ldots, Y_M)$ there are M tuning parameters $\lambda = (\lambda_1, \ldots, \lambda_M)^\top$ to be selected for both joint and conditional classification. To simplify the parameter selection process in practice, we consider K parameter candidates $\lambda^1, \ldots, \lambda^K$, and for each candidate λ^k we assume $\lambda_1^k = \lambda_2^k = \cdots = \lambda_M^k$. Then the tuning parameters are selected following the same procedure described above.

7. Extensions and Comparisons

7.1 Extension to Three or More Responses

In this section, we briefly discuss the extensions of Algorithm 1 (joint classification) and Algorithm A.1 (conditional classification) from the bivariate response setting (1) to multivariate response setting (25). More details are included in the Appendix.

Suppose we have multivariate response $\mathbf{Y} = (Y_1, \ldots, Y_M)$, where $Y_m \in [c_m]$ for each $m \in [M]$, and predictor $\mathbf{X} \in \mathbb{R}^p$. For every category combination $(k_1, \ldots, k_M)$, $k_m \in [c_m]$, $m \in [M]$, the multivariate categorical response LDA model assumes that

$$\mathbf{X}|(Y_1 = k_1, \ldots, Y_M = k_M) \sim \mathrm{N}_p(\boldsymbol{\mu}_{k_1 \ldots k_M}, \boldsymbol{\Sigma}), \tag{25}$$

where $\Pr(Y_1 = k_1, \ldots, Y_M = k_M) = \pi_{k_1 \ldots k_M}$, $\boldsymbol{\mu}_{k_1 \ldots k_M} \in \mathbb{R}^p$ is the mean vector and $\pi_{k_1 \ldots k_M} > 0$ is a probability such that $\sum_{k_1, \ldots, k_M} \pi_{k_1 \ldots k_M} = 1$.

We arrange the group means into the mean tensor $\boldsymbol{\mu} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$ such that $\boldsymbol{\mu}_{[:, k_1, \ldots, k_M]}$ is $\boldsymbol{\mu}_{k_1 \ldots k_M}$. Then, we define $\boldsymbol{\delta} = \boldsymbol{\mu} - \boldsymbol{\mu}_{1 \ldots 1} \circ \mathbf{1}_{c_1} \cdots \circ \mathbf{1}_{c_M} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$ and $\boldsymbol{\beta} = \boldsymbol{\delta} \times_1 \boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$ so that $\boldsymbol{\beta}_{k_1 \ldots k_M} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{k_1 \ldots k_M} - \boldsymbol{\mu}_{1 \ldots 1})$, $\boldsymbol{\beta}_{1 \ldots 1} = 0$. Thus, $\boldsymbol{\beta}$ is the tensor discriminant coefficient identified by the joint Bayes' rule derived in Appendix E, where we provide details of the extension from bivariate response to $M \geq 3$. Moreover, Section (E.4) propose a simple pairwise likelihood approach that is built upon our bivariate LDA method.

Similar to the bivariate LDA (1), we are interested in two classification problems: joint classification of multivariate response $\mathbf{Y} = (Y_1, \ldots, Y_M)$ from predictor $\mathbf{X}$ and the conditional classification of a univariate response Y_s from $(\mathbf{X}, \mathbf{Y}_{-s})$, where $\mathbf{Y}_{-s}$ is the $(M-1)$-dimensional response vector excluding Y_s. For conditional classification of more than one

response simultaneously, we would combine these response into a univariate Y_s and still let $\mathbf{Y}_{-s}$ denote the rest of the responses. We discuss the specifics of our method for conditional classification in the Appendix, and illustrate this method in our simulations studies.

After collecting samples $\{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^n$ from the multivariate LDA (25), the unpenalized sample objective function that extends (12) to multi-response is given by

$$\widetilde{g}_n(\boldsymbol{\beta}) = \sum_{k_1,\ldots,k_M} \left\{ \frac{1}{2} \boldsymbol{\beta}_{k_1\ldots k_M}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{k_1\ldots k_M} - \widehat{\boldsymbol{\delta}}_{k_1\ldots k_M}^\top \boldsymbol{\beta}_{k_1\ldots k_M} \right\}, \tag{26}$$

where $\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^n (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_{1i}\ldots Y_{Mi}})(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_{1i}\ldots Y_{Mi}})^\top / n$, $\widehat{\boldsymbol{\delta}}_{k_1\ldots k_M} = \widehat{\boldsymbol{\mu}}_{k_1\ldots k_M} - \widehat{\boldsymbol{\mu}}_{1\ldots 1}$, $\widehat{\boldsymbol{\mu}}_{k_1\ldots k_M} = \sum_{i=1}^n \mathbf{X}_i \cdot I(Y_{1i} = k_1, \ldots, Y_{Mi} = k_M) / \sum_{i=1}^n I(Y_{1i} = k_1, \ldots, Y_{Mi} = k_M)$. To exploit the high-order generalization of the type of sparsity we assume when $M = 2$, we propose a generalization of the penalty $\mathcal{P}_\lambda$, which we denote $\widetilde{\mathcal{P}}_\lambda$. The idea can be explain from Figure 1 middle row: for mode-2 (Y_1) selection, we aggregate the categories of all the other responses and the aggregated response from $\mathbf{Y}_{-1}$, which has $\prod_{m\neq 1} c_m$ categories and replaces the single response Y_2 in the bivariate case. Let $\boldsymbol{\beta}_{j,(m)} \in \mathbb{R}^{c_m \times \prod_{\ell\neq m} c_\ell}$ be the mode-m matricization of $\boldsymbol{\beta}_{[j,:,:,\ldots,:]} \in \mathbb{R}^{c_1\times\cdots\times c_M}$. We can thus define $\widetilde{\mathcal{P}}_\lambda(\boldsymbol{\beta}) = \sum_{m=1}^M \lambda_m \sum_{j=1}^p \|\boldsymbol{\beta}_{j,(m)}^\top\|_{2,1}$ as the natural generalization of the $\mathcal{P}_\lambda$. The generalization of $\mathcal{P}_{\mathcal{V},\lambda}$ follows a similar logic.

7.2 Extension to Semiparametric LDA Model

As a semiparametric generalization of the normal-theory discriminant analysis, Lin and Jeon (2003) proposed a more flexible model based on marginal transformation of each predictor variable. This semiparametric approach is further extended to high-dimensional settings by Mai and Zou (2015); Jiang and Leng (2016). Extending our proposed model and method to semiparametric discriminant analysis can be developed analogous to these existing approaches. Specifically, the semiparametric generalization of our model (25) can be written as,

$$(h_1(X_1), \ldots, h_p(X_p))^\top | (Y_1 = k_1, \ldots, Y_M = k_M) \sim \mathrm{N}_p(\boldsymbol{\mu}_{k_1\ldots k_M}, \boldsymbol{\Sigma}), \tag{27}$$

where $h_j(\cdot)$ is a univariate monotone transformation function for the j-th predictor and needs to be estimated. Following Mai and Zou (2015); Jiang and Leng (2016), the sparse estimation under (27) has two steps. First, we estimate the p univariate transformation functions by a pooled normal score transformation, weighted-averaging over each class combination of the responses. See Mai et al. (2023) for background and more recent developments of such normal score transformation. Then in the second step, the proposed estimation procedure in this paper would be directly applied to the transformed data. We expect this semiparametric extension to be beneficial in many applications and will investigate its performance empirically and theoretically in the future.

7.3 Comparison to Past Work

In this section, we compare and contrast our proposed method to those of Mai et al. (2019) and Molstad and Rothman (2023), mainly from the aspect of computational feasibility and methodological differences.

In the context of univariate-response linear discriminant analysis in high dimensions, Mai et al. (2019) proposed the so-called "multiclass sparse discriminant analysis" estimator. Under the univariate response linear discriminant analysis model,

$$\mathbf{X}|Y_1 = k_1 \sim \mathrm{N}_p(\boldsymbol{\mu}_{k_1}, \boldsymbol{\Sigma}), \quad \Pr(Y_1 = k_1) = \pi_{k_1} > 0, \quad k_1 \in [c_1], \tag{28}$$

they proposed to estimate the discriminant vectors $\boldsymbol{\beta}_{k_1} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{k_1} - \boldsymbol{\mu}_1) \in \mathbb{R}^p$, $k_1 \in [c_1]$, by minimizing

$$\underset{\boldsymbol{\beta}, \boldsymbol{\beta}_1 = 0}{\operatorname{argmin}} \left\{ \sum_{k_1=1}^{c_1} \left(\frac{1}{2} \boldsymbol{\beta}_{k_1}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{k_1} - \widehat{\boldsymbol{\delta}}_{k_1}^\top \boldsymbol{\beta}_{k_1} \right) + \lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\beta}_{[j,:]}\|_2 \right\}. \tag{29}$$

Evidently, (29) is a special case of our method when $c_2 = 1$ and $\lambda_2 = 0$. However, how to generalize (29) to accommodate multi-response settings (i.e., $c_1 \geq 2$ and $c_2 \geq 2$) is not obvious. For example, a natural generalization of Mai et al. (2019) to the multi-response setting simply replaces Y_1 in (29) with $\widetilde{Y}$, which has $c_1 c_2$ many response categories. Like our method, this approach allows Y_1 and Y_2 to be arbitrarily dependent, and performs variable selection. As discussed in Section 4, when only a subset of the important predictors affect each response across each category of the other response, our proposed method may perform better theoretically.

Besides Mai et al. (2019), a referee asked that we compare to the method of Molstad and Rothman (2023). Our method is fundamentally different from that proposed in Molstad and Rothman (2023). First, we assume a linear discriminant analysis model, whereas Molstad and Rothman (2023) assume a multinomial logistic regression model for the combined-category response $\widetilde{Y}$. The differences between LDA and logistic regression model are well understood (Hastie et al., 2009, Chapter 4.4.5): LDA can be more efficient due to the additional assumptions on the distribution of predictors. Beyond the different model assumptions, conceptually, the goals of our method's regularization scheme is fundamentally different from that in Molstad and Rothman (2023). In Molstad and Rothman (2023), the authors aim to identify which predictors (i) affect both marginal distributions and the joint distribution of the responses, (ii) affect only the marginal distributions, or (iii) are irrelevant. In contrast, our method aims to identify which predictors are irrelevant, and which predictors are relevant for, say, Y_1 when $Y_2 = k_2$ (or vice versa). In general, our method cannot identify variables that only affect the marginal distributions of the responses; whereas the method of Molstad and Rothman (2023) cannot identify which variables are relevant for Y_1 when $Y_2 = k_2$.

8. Simulations

8.1 Simulation Setup

In the following simulation studies, we first create $\boldsymbol{\beta} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$ and covariance matrix $\boldsymbol{\Sigma}$, then the tensor of class means $\boldsymbol{\mu}$ is obtained through the equality $\boldsymbol{\mu} = \boldsymbol{\beta} \times_1 \boldsymbol{\Sigma}$ (or equivalently, $\boldsymbol{\mu}_{k_1 \ldots k_M} = \boldsymbol{\Sigma} \boldsymbol{\beta}_{k_1 \ldots k_M}$). The discriminant coefficient tensor $\boldsymbol{\beta}$ are then redefined as $\boldsymbol{\beta} = (\boldsymbol{\mu} - \boldsymbol{\mu}_{11\ldots1} \circ \mathbf{1}_{c_1} \circ \cdots \circ \mathbf{1}_{c_M}) \times_1 \boldsymbol{\Sigma}^{-1}$ to satisfy our specific parameterization in Section 2.2. For 100 independent replications, classification performance is evaluated on a testing data of size 1000. All tuning parameters are selected by 5-fold cross-validation.

We demonstrate the proposed method under six different models. Models **M1**–**M4** are bivariate response models where $c_1 = c_2 = 3$ and the discriminant coefficient tensor $\boldsymbol{\beta}$ is designed with different structures. Namely, under **M1**, $\boldsymbol{\beta}$ is very sparse; under **M2**, the signals in $\boldsymbol{\beta}_{\cdot k_2}$ increase with k_2 and the sets of important variables $\mathcal{S}_{\cdot k_2}$'s are disjoint for different k_2; under **M3**, $\boldsymbol{\beta}$ has a mix of both large and small signals; and under **M4**, $\boldsymbol{\beta}$ has a larger number of small signals. Figure 4 displays the structure of three-way discriminant coefficient tensor $\boldsymbol{\beta}$ for models **M1**–**M4** using 3-dimensional cubes, where vertical line with circles represents discriminant direction $\boldsymbol{\beta}_{k_1 k_2}$, $k_1 \in [c_1]$, $k_2 \in [c_2]$, gray circles represent nonzero entries with signal $\boldsymbol{\beta}_{[j,k_1,k_2]}$ marked by the circles, and white circles represent zero entries. Note that all coefficients not displayed in Figure 4 are zero. Models **M5** and **M6** are general multivariate categorical responses models with $M = 4$ and $c_1 = \cdots = c_4 = 2$. Detailed settings of other parameters are summarized below, where we use $\mathrm{CS}(\rho)$ and $\mathrm{AR}(\rho)$, $-1 < \rho < 1$ to denote correlation matrices with compound symmetric (i.e. ρ on all off-diagonals) and auto-regressive (i.e. $\rho^{|i-j|}$ for the (i,j)-th element) structures, respectively. For all the simulated models, we set number of predictors $p = 1000$, sample size $n = 30K$ where $K = \prod_{m=1}^{M} c_m$, and set the covariance $\boldsymbol{\Sigma}$ to be $\mathrm{AR}(0.3)$, unless specified otherwise. When $M = 2$, let $\boldsymbol{\pi}$ be the matrix with (k_1, k_2)-th entry $\pi_{k_1 k_2}$. The other model parameters are set as follows.

- **M1**: $\pi_{k_1 k_2} = 1/9$ for $k_1 \in [3]$, $k_2 \in [3]$.
- **M2**: $\boldsymbol{\pi} = (0.3, 0.4, 0.3)^\top \circ (0.6, 0.2, 0.2)$, and $n = 60K$.
- **M3** and **M4**: $\boldsymbol{\pi} = (0.3, 0.4, 0.3)^\top \circ (1/3, 1/3, 1/3)$, $n = 60K$, and $\boldsymbol{\Sigma} = \mathrm{CS}(0.3)$
- **M5**: $\pi_{k_1 \ldots k_4} = 1/16$ for $k_m \in [2]$, $m \in [4]$. The matricized coefficient tensor $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_{(1)}$ is generated as $\widetilde{\boldsymbol{\beta}}_{jk} = 2$ for $j = 2k-1, 2k$, $k \in \{2, \ldots, K\}$ and $\widetilde{\boldsymbol{\beta}}_{jk} = 0$ otherwise.
- **M6**: Same as **M5**, except the nonzero elements in the coefficient tensor are set as $\boldsymbol{\beta}_{[j,k_1,k_2,k_3,k_4]} = 1.5k_1 + U$ for $k_m \in [c_m]$, $m = 1, \ldots, 4$, $j = 3k_{-1} - 2, 3k_{-1} - 1, 3k_{-1}$, where $k_{-1} \in [2^3]$ is the univariate category transformed from (k_2, k_3, k_4), and U follows an independent uniform distribution over $[-0.5, 0.5]$, and $\boldsymbol{\Sigma} = \mathrm{CS}(0.5)$.

For joint and marginal classification, we compare our high-dimensional multivariate LDA (MLDA) with existing methods including the multi-class discriminant analysis (MSDA, Mai et al., 2019) fitted on all K classes, penalized LDA (PLDA, Witten and Tibshirani, 2011)

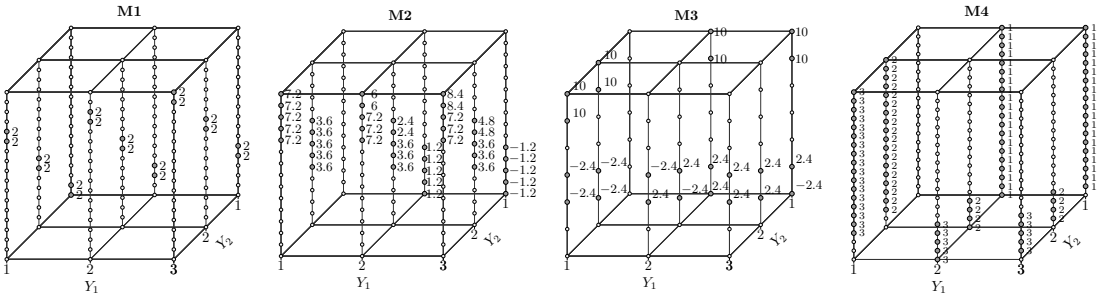

Figure 4: Illustration of the sparsity pattern of $\boldsymbol{\beta}$ for bivariate models **M1**–**M4**.

fitted on all K classes, and sparse logistic regression (Logistic, Friedman et al., 2010) fitted on each response marginally. We also include the Bayes' rule with true parameters plugged in as a benchmark (Oracle). The above competitors are implemented by `R` packages `msda`, `penalizedLDA` and `glmnet`. For conditional classification, in addition to MLDA and MSDA, we include the conditional MLDA from Algorithm A.1 or Algorithm A.3, which is applied on the reduced data $\mathbf{X}^{\widehat{\mathcal{S}}}$ with only variables selected from the first step MLDA (MLDA-C), along with both MSDA and sparse logistic regression fit on partial data assuming the label information of one response or multiple responses is given (MSDA-S and Logistic, respectively).

8.2 Joint Classification Results

For each model, we report the joint classification error $\sum_{i=1}^{n} I(\widehat{\mathbf{Y}}_i \neq \mathbf{Y}_i)/n$ and summarize the results in Figure 5. It is clear that our method, MLDA, achieved the lowest joint classification error across all six models. Under model **M1**, we have a very sparse setting where each $\boldsymbol{\beta}_{k_1k_2}$ contains two nonzero entries with same signal strength, but with different sets of important variables so that the total number of important variables is $|\mathcal{S}| = 18$ in $\boldsymbol{\beta}$, while $\boldsymbol{\beta}_{\cdot k_2}$ and $\boldsymbol{\beta}_{k_1\cdot}$ both contain 6 important variables. With both mode-2 and mode-3 selection, MLDA obtained more efficient estimation than MSDA and thus performed better in classification. MLDA also had an advantage over other methods under Model **M2**, where the signal strength of $\boldsymbol{\beta}_{\cdot k_2}$ increases as k_2 increases, and the sets of important variables are disjoint for distinct k_2. When fitting MSDA on all $K = 9$ classes, it tends to overpenalize $\boldsymbol{\beta}_{\cdot 1}$, and fails to identify some important variables and incorrectly selects some unimportant

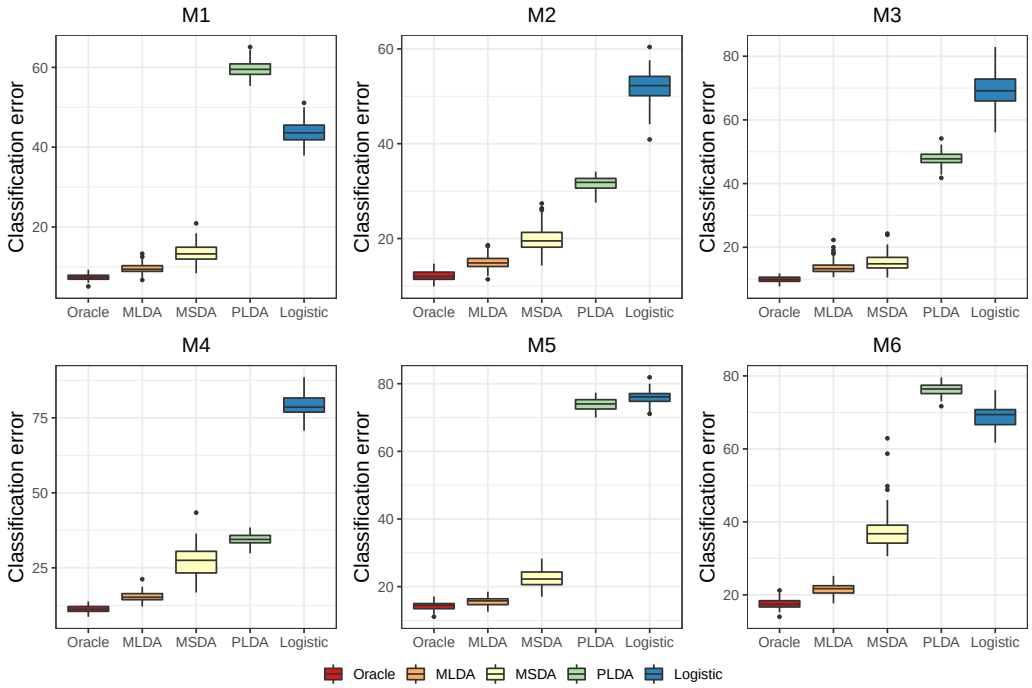

Figure 5: Joint classification error (%) under **M1**–**M6**.

variables. Under model **M3**, the discriminant tensor $\boldsymbol{\beta}$ has mixed signal strength, with most of the signals coming from variables $j \in \{1, 2, 3, 4\}$, and others having relatively small signal strength. As shown in Figure 5, MLDA performs only slightly better than MSDA in joint classification, which is expected since the global selection offered by MSDA should be able to recover most of the important predictors in this setting. Model **M4** illustrates that MLDA is more robust to more complicated settings, e.g. with dense signals, by distinguishing Y_1 and Y_2 with mode-2 and mode-3 variable selection. Under model **M5**, where we have $M = 4$ responses, MLDA achieves a joint classification error that is close to the Oracle error, while other competing methods fail to deal with a large number of categories K and high sparsity level. Model **M6** also has $M = 4$, but with more random coefficient entries and more highly correlated predictors so that it becomes even harder for MSDA and other competing methods to estimate the sparsity structure in the discriminant coefficient tensor $\boldsymbol{\beta}$. This led to larger classification errors, while MLDA still performed similarly to Oracle.

Overall, MLDA significantly outperformed the competing methods in joint classification, especially when $\boldsymbol{\beta}$ has more complicated structures such as dense signals and mixed signal strengths, and when $\boldsymbol{\beta}$ has a larger number of important variables jointly than conditionally.

8.3 Conditional Classification Results

In this section, we demonstrate the conditional classification performance of our method implemented with Algorithm A.1 and Algorithm A.3, for bivariate response and multivariate response ($M > 2$) settings, respectively. We use the bivariate response model **M2** and multivariate response model **M5** for illustration. We report the classification error of conditional predictions. Specifically, the conditional classification error of $\widehat{Y}_1(\mathbf{X}, Y_2)$ given $Y_2 = k_2$ is defined as $\sum_{Y_{2i}=k_2} I(\widehat{Y}_{1i}(\mathbf{X}_i, Y_{2i}) \neq Y_{1i}) / \sum_{Y_{2i}=k_2} 1$.

Figure 6 summarizes the conditional classification errors of the proposed and competing methods under **M2**. From Figure 4, we see that **M2** has an increasing signal strength across the $\boldsymbol{\beta}_{\cdot k_2}$, and that the sets of important variables are disjoint for different k_2. As previously

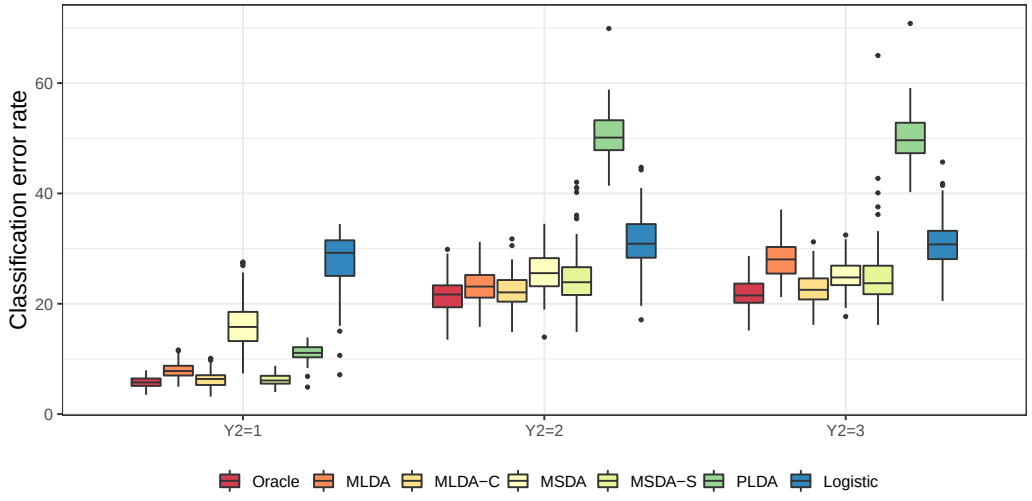

Figure 6: Conditional classification of Y_1 from $(\mathbf{X}, Y_2)$ under **M2**.

mentioned, while fitting MSDA on all $K = 9$ classes, it tends to overpenalize $\boldsymbol{\beta}_{\cdot 1}$, and has poor variable selection performance. As shown in Figure 7, MLDA yielded better variable selection than MSDA as MLDA penalizes each mode-3 slices $\boldsymbol{\beta}_{\cdot k_2}$ whereas MSDA globally penalizes slices $\boldsymbol{\beta}_{[j,:,:]}$ without account for the distinct directions corresponding to Y_1 and Y_2. Not surprisingly, MLDA-C achieved best classification performance, which improves upon MLDA, especially when given $Y_2 = 3$. There are two potential reasons for the observed improvement. First, some important variable j selected by MLDA in $\boldsymbol{\beta}_{\cdot k_2}$ is not useful for conditional prediction if $\boldsymbol{\beta}_{[j,1,k_2]} = \boldsymbol{\beta}_{[j,2,k_2]} = \cdots = \boldsymbol{\beta}_{[j,c_1,k_2]}$, and MLDA-C can successfully identify these variables. Second, the joint signal $\boldsymbol{\beta}_{\cdot k_2}$ increases with k_2, so MLDA may overpenalize $\boldsymbol{\beta}_{\cdot 1}$ or underpenalize $\boldsymbol{\beta}_{\cdot 3}$ so that variable selection performance is poor for all three frontal slices simultaneously. This is why the false positive rate of MLDA increased from around 0% to 18% when the label of Y_2 changed from $k_2 = 2$ to $k_2 = 3$. Also, MLDA-C achieved best false positive rate close to 0% for all $k_2 \in [3]$.

Figures 8 summarizes the conditional classification of $Y_1|(Y_2, Y_3, Y_4)$ and $(Y_1, Y_2)|(Y_3, Y_4)$ under the multivariate response model **M5**. Again, we see that MLDA and MLDA-C achieved the lowest conditional classification error with MLDA-C performing slightly better than MLDA. In terms of variable selection performance, MLDA-C attained false positive rate close to 0%, which is significantly better than other competing methods except Logistic. Logistic obtained 0% false positive rate in both cases. However, as shown in Tables A.14 & A.15 in Appendix, Logistic failed to recognize important variables with a small true positive rate, while MLDA and MLDA-C achieved 100% true positive rate.

Overall, we conclude that when the focus is conditional classification, one can always use Algorithm A.1 as a refinement to the output from Algorithm 1 for better classification and variable selection performance conditionally. However, it is notable that MLDA performed reasonably well in terms of conditional classification, even without this refinement.

In Appendix B, we summarize the classification error of each response by the estimated joint Bayes' rule (4), the classification error of Y_s conditional on different labels of the other response, and variable selection performance evaluated by true and false positive rate. From Tables A.1–A.12 in the Appendix, we can see MLDA also achieves best conditional variable

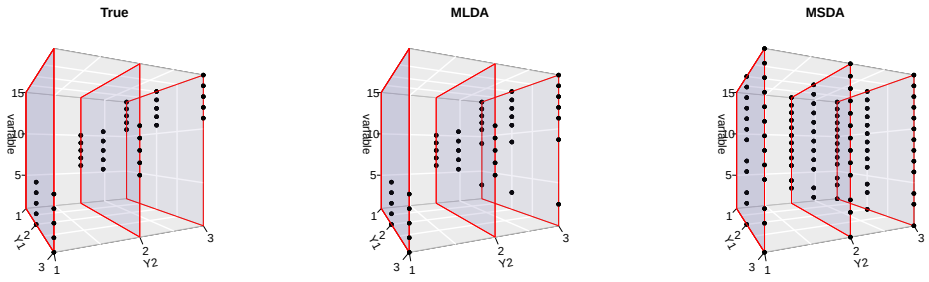

Figure 7: Visualization of variable selection by $\widehat{\boldsymbol{\beta}}$ under **M2**. Black points represent nonzero signals. From left to right: true signals in $\boldsymbol{\beta}$, signals recognized by MLDA and signals recognized by MSDA.

selection performance in all scenarios. Also, MSDA-S fit on partial data for conditional classification can fail to recover the important predictors with smaller signals, thus resulting in high conditional classification error (see Tables A.7–A.9 for **M3**). Although MSDA-S outperformed MSDA in the conditional classification under models **M2** and **M5**, it can perform much worse than MSDA in other settings (see Tables A.7–A.12 for models **M3** and **M4** in the Appendix), especially in real applications where samples are limited and categories are unbalanced.

9. Application to Benchmark Datasets

In this section, we demonstrate the proposed methods in several benchmark multivariate response classification datasets. For joint classification, MLDA is implemented using Algorithm A.1, which computes the $M \geq 3$ generalization of our estimator introduced in

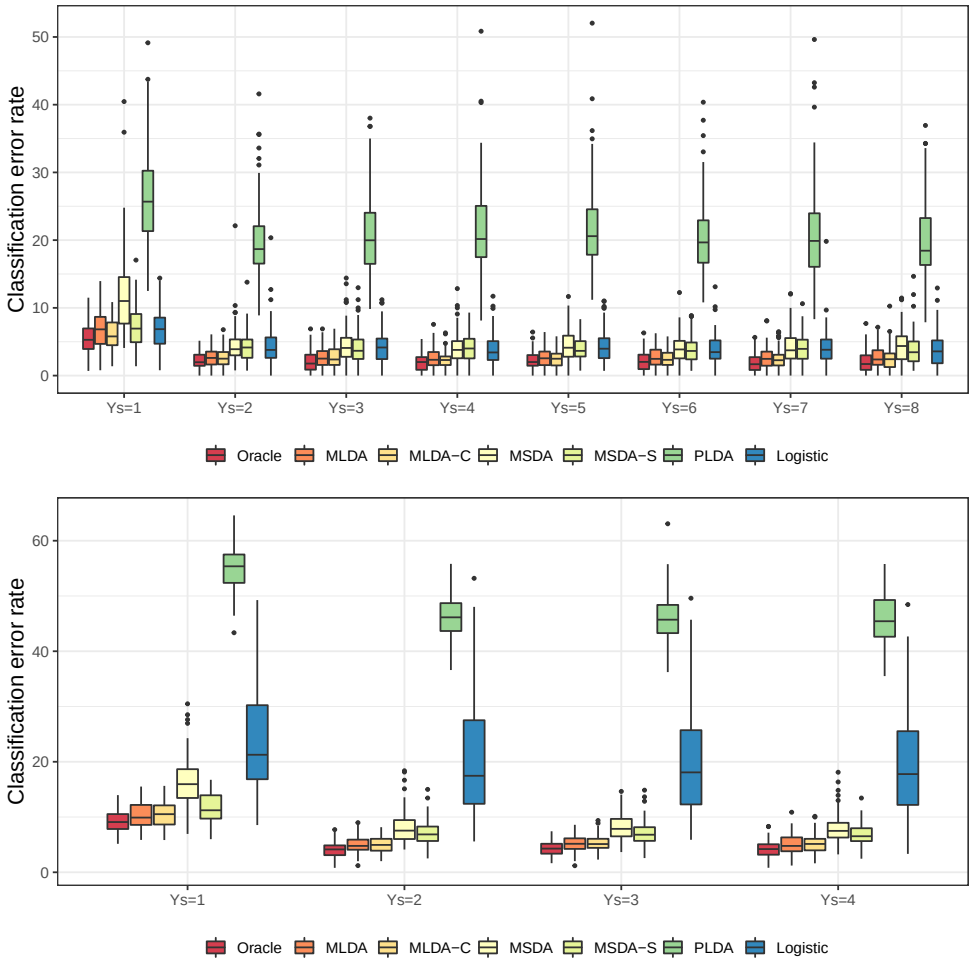

Figure 8: Conditional classification results under **M5**. Top panel: conditional classification of Y_1 from $(\mathbf{X}, Y_2, Y_3, Y_4)$; bottom panel: conditional classification of (Y_1, Y_2) from $(\mathbf{X}, Y_3, Y_4)$.

Section 7. We compare our method with the same competing methods as in the simulation studies, except here, these methods are implemented marginally on each response to obtain joint classification. We also include the pairwise MLDA (MLDA-P) from Algorithm A.4 in the Appendix, and pairwise MSDA (MSDA-P) which transforms each pair of responses to univariate response and then fits MSDA. For pairwise MLDA, instead of maximizing the likelihood of over all $\prod_{m=1}^{M} c_m$ combinations of categories, it is reasonable to only maximize over those labels $\{k_1, \ldots, k_M\}$ observed in the training data. That is, we assume if a category is not observed in the training set, it can not be observed in the testing set. We denote this pairwise MLDA as MLDA$_s$-P.

We demonstrate the joint classification performance of all comparing methods in five datasets with $M > 2$. For each dataset, we randomly split the data 100 times and keep 20% of the samples as a testing set. The descriptions of the dataset, which were obtained from `https://www.uco.es/kdis/mllresources/`, are as follows.

- **Yeast**. The goal is to predict genes' functional classes based on p-dimensional ($p = 103$) gene expression and phylogenetic profiles. In this dataset, there are $n = 2417$ genes and $M = 14$ responses (functional classes) with $c_1 = \cdots = c_{14} = 2$ (i.e., each response indicates whether the gene belongs to a particular functional class or not).

- **Image**. The goal is to classify images into types. The data consists of $n = 2000$ natural scene images that may be classified into $M = 5$ image types (dessert, sunset, trees, mountains, sea) so that $c_1 = \cdots = c_5 = 2$. Each image is transformed into a $p = 49 \times 3 \times 2 = 294$ vector for model fitting.

- **VirusGO**. The goal is to predict the sub-cellular locations of proteins based on their sequences. These data (as well as GpositiveGO and GnegativeGO) consist of binary gene ontology (GO) features derived from protein sequences. In this dataset, there are $n = 207$ protein sequences for virus species, $p = 749$ GO features, and $M = 6$ sub-cellular locations with $c_1 = \cdots = c_6 = 2$.

- **GpositiveGO**. The goal is to predict the sub-cellular locations of proteins based on their sequences. In this dataset, there are $n = 519$ protein sequences for Gram positive species, $p = 912$ GO features, and $M = 4$ sub-cellular locations with $c_1 = \cdots = c_4 = 2$.

- **GnegativeGO**. The goal is to predict the sub-cellular locations of proteins based on their sequencess. In this dataset, there are $n = 1392$ protein sequences for Gram negative bacterial species, $p = 1717$ GO features, and $M = 8$ sub-cellular locations with $c_1 = \cdots = c_8 = 2$.

The classification results of the above five datasets are summarized in Table 1. Each of these data sets contains multiple binary responses, we also separately predict each binary response using the DSDA method from Mai et al. (2012) which is designed for binary classification based on the LDA model. We can see MLDA performs best in terms of joint classification in all datasets. The promising performance of MLDA may be attributable to the fact that we efficiently model dependence between responses. The pairwise likelihood methods MLDA-P and MLDA$_s$-P also achieved competitive classification performance, where MLDA$_s$-P restricted on only observed labels in training data performed better than

	MLDA	MLDA-P	MLDA$_s$-P	MSDA-P	DSDA	PLDA	Logistic	S.E.$\leq$
Yeast	**77.7**	80.3	78.6	80.9	85.0	92.1	85.6	(0.29)
Image	**51.7**	57.0	**52.1**	58.5	64.2	78.1	82.1	(0.24)
VirusGo	**17.8**	20.2	20.1	**18.5**	**17.8**	–	25.5	(0.75)
GpositiveGO	**6.9**	7.4	**7.2**	7.8	9.2	–	10.4	(0.33)
GnegativeGO	**7.2**	8.1	7.8	8.4	8.0	–	11.1	(0.16)

Table 1: Classification of real datasets. Reported are average classification error rate (%). Results are based on 100 random training/testing splits. The maximum standard error among all methods in each dataset (i.e., across each row of the table) is provided in the rightmost column. PLDA failed to be implemented in the gene ontology datasets.

MLDA-P. For joint classification, our method significantly outperforms other classification methods applied marginally on each response.

Besides the joint classification of the above five datasets, in Appendix B, we demonstrate the promising performance of our proposed method for conditional classification on a pan-kidney cancer data with bivariate response describing cancer type and 5-year survival.

Acknowledgements

The authors thank the action editor and two referees for their helpful comments and suggestions. Xin Zhang's research was supported in part by NSF DMS-2053697 and DMS-2113590 and Aaron J. Molstad's research was supported in part by NSF DMS-2113589.

Appendix A. Overview of Appendices

In Appendix B, we provide detailed summaries of classification performance under models **M1**–**M6** and visualizations of the conditional discriminant coefficient tensors under the bivariate LDA models **M1**–**M4**. Appendix C contains the analysis of the pan-kidney cancer data, where we focus on conditional classification. In Appendix D, we provide the ADMM algorithm for conditional classification in high-dimensional bivariate LDA. Appendix E provides more detailed discussion of the $M \geq 3$ extensions of the joint and conditional classifiers introduced in Section 7. Appendices F and G contain proofs of the Propositions and Theorem 1, respectively.

Appendix B. Additional Simulation Results

In this section, we provide additional results for joint and conditional classification under models **M1**–**M6** as described in Section 6 of the paper.

For each bivariate model **M1**–**M4**, we organize the results as follows. First, we visualize the joint discriminant coefficient tensor $\boldsymbol{\beta}$, the conditional discriminant coefficients $\boldsymbol{\theta}^c_{k_1\cdot}$ for $k_1 \in [3]$, and conditional discriminant coefficients $\boldsymbol{\theta}^r_{\cdot k_2}$ for $k_2 \in [3]$. Then, we use three tables to summarize the classification performance of each competing method. Let $\widehat{\mathbf{Y}}(\mathbf{X}) = (\widehat{Y}_1, \widehat{Y}_2)$ be the joint prediction and $\widehat{Y}_1(\mathbf{X}, Y_2)$, $\widehat{Y}_2(\mathbf{X}, Y_1)$ be the conditional predictions. The first table reports joint classification error $\mathrm{Err} = \sum_{i=1}^n I(\widehat{\mathbf{Y}}_i(\mathbf{X}_i) \neq \mathbf{Y}_i)/n$, the classification error of Y_1 defined by $\mathrm{Err}_{Y_1} = \sum_{i=1}^n I(\widehat{Y}_{1i} \neq Y_{1i})/n$, the conditional classification of Y_1 defined by $\mathrm{Err}_{Y_1|Y_2} = \sum_{i=1}^n I(\widehat{Y}_{1i}(\mathbf{X}_i, Y_{2i}) \neq Y_{1i})/n$, and also the classification errors of Y_2 and $Y_2(\mathbf{X}, Y_1)$. The second table reports the classification results for Y_1 conditional on different labels of Y_2, including $\widetilde{\mathrm{Err}}_{Y_1} = \sum_{Y_{2i}=k_2} I(\widehat{Y}_{1i} \neq Y_{1i}) / \sum_{Y_{2i}=k_2} 1$, $\widetilde{\mathrm{Err}}_{Y_1|Y_2} = \sum_{Y_{2i}=k_2} I(\widehat{Y}_{1i}(\mathbf{X}_i, Y_{2i}) \neq Y_{1i}) / \sum_{Y_{2i}=k_2} 1$, true positive rate $\mathrm{TPR}_{1|2}$ and false positive rate $\mathrm{FPR}_{1|2}$ of variables selected by $\widehat{\boldsymbol{\theta}}^r_{\cdot k_2}$ defined as $\mathrm{TPR}_{1|2} = |\widehat{\mathcal{A}}_{\cdot k_2} \cap \mathcal{A}_{\cdot k_2}|/|\mathcal{A}_{\cdot k_2}|$ and $\mathrm{FPR}_{1|2} = |\widehat{\mathcal{A}}_{\cdot k_2} \cap \mathcal{A}^c_{\cdot k_2}|/|\mathcal{A}^c_{\cdot k_2}|$, where recall $\mathcal{A}_{\cdot k_2} = \{j : \boldsymbol{\theta}^r_{[j,k_1,k_2]} \neq 0 \text{ for some } k_1 \in [c_1 - 1]\}$ is the active set of $\boldsymbol{\theta}^r_{\cdot k_2}$. For MLDA and MSDA, $\widehat{\boldsymbol{\theta}}^r_{\cdot k_2}$ is calculated by $\widehat{\boldsymbol{\beta}}_{\cdot k_2}(\mathbf{A}^{c_1})^\top$; for MLDA-C, $\widehat{\boldsymbol{\theta}}^r_{\cdot k_2}$ is the direct output from Algorithm A.1. Finally, in the third table, we report the detailed classification results of Y_2 conditional on different labels of Y_1.

From Tables A.1–A.12, it is clear that MLDA, our proposed method, achieved the lowest classification errors both jointly and conditionally. Also, MLDA obtained the best variable selection performance with a true positive rate close to 100% and a false positive rate close to 0% in most scenarios. Tables A.8–A.9 illustrate that for conditional classification, traditional classification methods such as MSDA-S fitted on partial data can perform much worse than MLDA and its conditional version. This may be due to the fact that MLDA-C accounts for the whole data and only needs to be fitted once, which can improve efficiency. Also, we see that the classification error Err_{Y_1} based on MLDA is greater than or equal to the classification error $\mathrm{Err}_{Y_1|Y_2}$ obtained by MLDA-C, indicating when the information of the other response is available, classification can often be improved compared with the joint classification.

For the $M = 4$ model settings, **M5** and **M6**, we organize the results as follows. Table A.13 summarizes the joint classification performance. We see that MLDA achieved the

best joint classification error—close to the Oracle error—while other competing methods failed to handle large K and a high degree of sparsity. This confirms that MLDA achieves efficiency gain by capturing dependence amongst responses and performing more nuanced variable selection. Thus we expect MLDA to, loosely speaking, have a greater advantage over traditional classification methods (i.e., those that are fit to $\tilde{Y}$ or those that fit a model to each response separately) when number of responses M increases. From Tables A.14 & A.15, we see that MLDA-C achieved lowest conditional classification error and almost perfect variable selection with 100% true positive rate and false positive rate close to 0%. This shows the effectiveness of the extended Algorithm A.3 for conditional classification in the $M \geq 3$ setting.

Additional results under M1

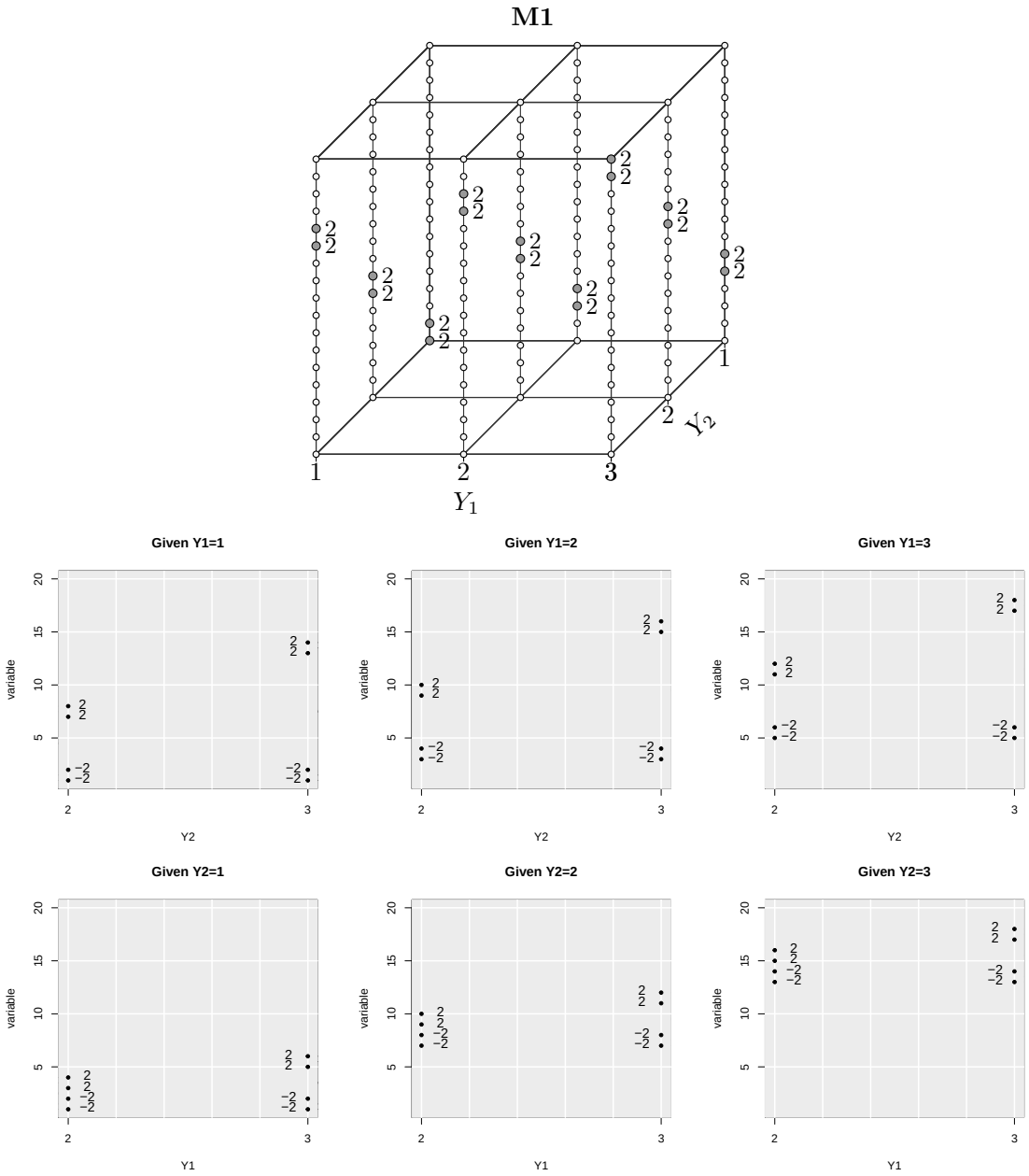

Figure A.1: Illustration of the sparsity pattern of $\boldsymbol{\beta}$ for **M1**. The bottom panels display the sparsity of the conditional discriminant coefficient tensor slices.

	Method	Err	Err_{Y_1}	$\text{Err}_{Y_1\|Y_2}$	Err_{Y_2}	$\text{Err}_{Y_2\|Y_1}$
M1	Oracle	7.4(0.08)	6.0(0.07)	3.3(0.06)	4.8(0.06)	2.1(0.05)
	MLDA	9.6(0.12)	7.8(0.10)	4.5(0.08)	6.0(0.09)	2.9(0.06)
	MSDA	13.5(0.22)	10.9(0.18)	6.6(0.14)	8.4(0.15)	4.2(0.11)
	MSDA-S	–	–	5.2(0.12)	–	3.3(0.10)
	PLDA	59.7(0.21)	50.1(0.18)	53.0(0.26)	34.1(0.25)	49.8(0.27)
	Logistic	43.9(0.27)	29.9(0.30)	8.7(0.24)	20.9(0.23)	6.0(0.24)

Table A.1: Classification results. Average and standard error (in parentheses) of classification errors Err, Err_{Y_1} (%), $\text{Err}_{Y_1|Y_2}$ (%), Err_{Y_2} (%), $\text{Err}_{Y_2|Y_1}$ (%). Results are based on 100 replications.

	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$
M1	$Y_2=1$				$Y_2=2$				$Y_2=3$			
Oracle	5.7 (0.12)	3.2 (0.09)	– –	– –	6.3 (0.13)	3.3 (0.11)	– –	– –	5.9 (0.14)	3.2 (0.11)	– –	– –
MLDA	7.4 (0.18)	4.3 (0.12)	100.0 (0.00)	0.4 (0.05)	8.1 (0.19)	4.5 (0.15)	100.0 (0.00)	0.4 (0.05)	7.8 (0.22)	4.7 (0.15)	99.8 (0.17)	0.4 (0.05)
MSDA	10.5 (0.29)	7.0 (0.22)	98.5 (0.48)	2.7 (0.09)	11.3 (0.28)	6.5 (0.21)	98.0 (0.59)	2.7 (0.08)	11.0 (0.31)	6.5 (0.23)	98.7 (0.45)	2.7 (0.08)
MSDA-S	– –	5.1 (0.16)	97.3 (0.61)	0.4 (0.04)	– –	5.3 (0.18)	98.2 (0.58)	0.4 (0.04)	– –	5.3 (0.24)	98.0 (0.68)	0.3 (0.04)
PLDA	50.2 (0.36)	40.7 (0.30)	100.0 (0.00)	100.0 (0.01)	50.0 (0.43)	40.2 (0.32)	100.0 (0.00)	100.0 (0.01)	50.0 (0.41)	41.0 (0.38)	100.0 (0.00)	100.0 (0.01)
Logistic	29.3 (0.85)	9.0 (0.46)	92.2 (1.04)	0.0 (0.00)	31.1 (0.79)	8.4 (0.32)	92.8 (1.01)	0.0 (0.00)	29.3 (0.87)	8.6 (0.53)	90.8 (0.99)	0.0 (0.00)

Table A.2: Y_1 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_1}$ (%), $\widetilde{\text{Err}}_{Y_1|Y_2}$ (%), $\text{TPR}_{1|2}$ (%), $\text{FPR}_{1|2}$ (%). Results are based on 100 replications.

	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$
M1	$Y_1=1$				$Y_1=2$				$Y_1=3$			
Oracle	5.0 (0.12)	2.1 (0.08)	– –	– –	4.4 (0.11)	2.0 (0.08)	– –	– –	5.1 (0.12)	2.1 (0.08)	– –	– –
MLDA	6.5 (0.17)	3.0 (0.10)	99.8 (0.17)	1.0 (0.04)	5.5 (0.12)	2.8 (0.09)	100.0 (0.00)	0.9 (0.04)	6.1 (0.16)	2.8 (0.10)	100.0 (0.00)	1.0 (0.04)
MSDA	8.4 (0.23)	4.2 (0.16)	98.5 (0.53)	2.7 (0.09)	8.2 (0.24)	4.5 (0.18)	98.2 (0.52)	2.7 (0.08)	8.6 (0.27)	3.9 (0.17)	98.5 (0.58)	2.7 (0.08)
MSDA-S	– –	3.6 (0.23)	98.7 (0.51)	0.4 (0.03)	– –	3.2 (0.12)	99.0 (0.46)	0.4 (0.04)	– –	3.1 (0.11)	98.7 (0.45)	0.5 (0.04)
PLDA	36.8 (0.43)	34.8 (0.37)	100.0 (0.00)	100.0 (0.01)	29.2 (0.35)	34.5 (0.32)	100.0 (0.00)	100.0 (0.01)	36.4 (0.42)	34.2 (0.33)	100.0 (0.00)	100.0 (0.01)
Logistic	25.9 (0.73)	6.1 (0.41)	95.7 (0.84)	0.0 (0.00)	12.6 (0.40)	6.3 (0.46)	93.8 (0.94)	0.0 (0.00)	24.2 (0.63)	5.7 (0.37)	95.7 (0.81)	0.0 (0.00)

Table A.3: Y_2 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_2}$ (%), $\widetilde{\text{Err}}_{Y_2|Y_1}$ (%), $\text{TPR}_{2|1}$ (%), $\text{FPR}_{2|1}$ (%). Results are based on 100 replications.

Additional results under M2

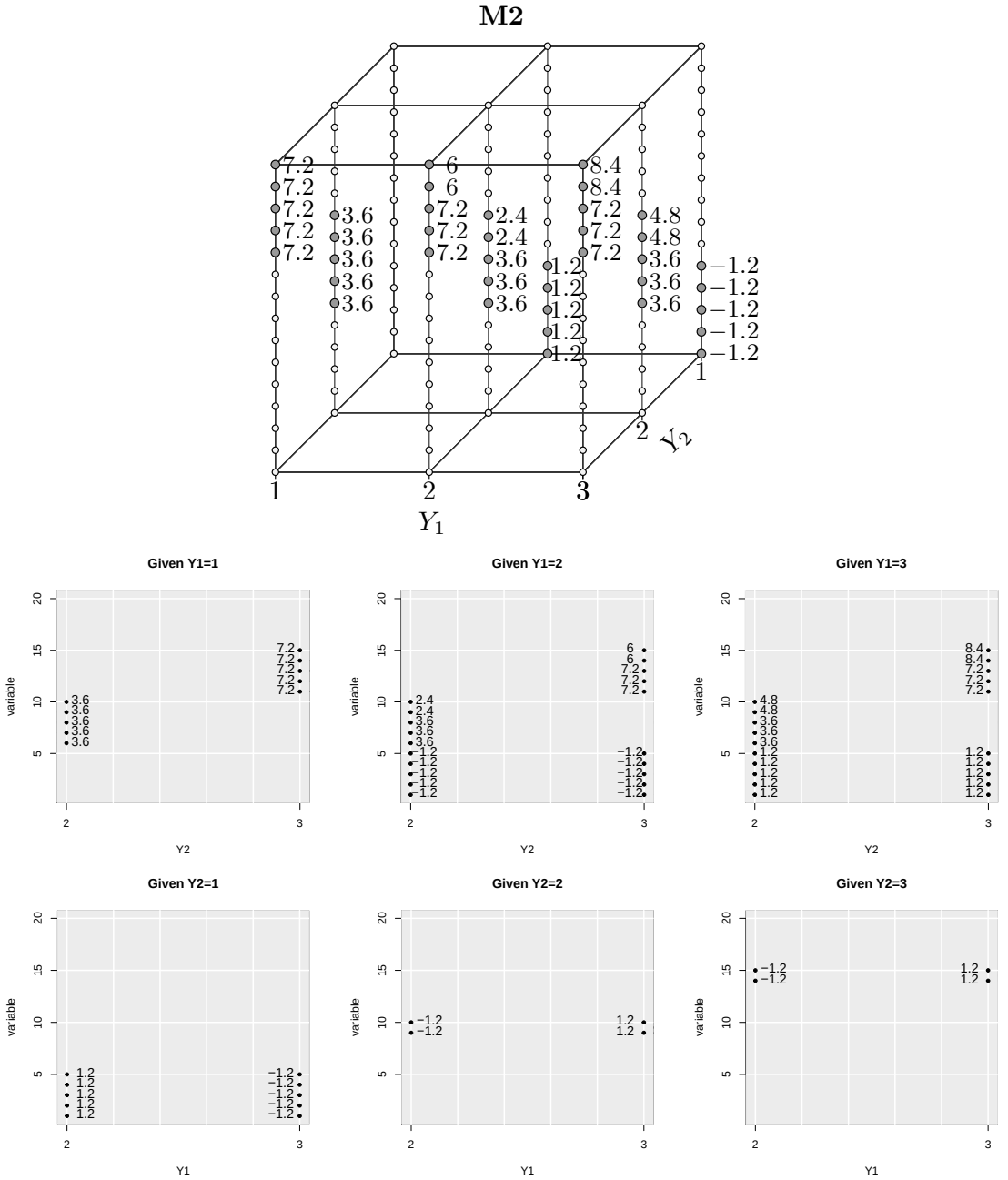

Figure A.2: Illustration of the sparsity pattern of $\boldsymbol{\beta}$ for **M2**. The bottom panels display the sparsity of the conditional discriminant coefficient tensor slices.

	Method	Err	Err_{Y_1}	$\text{Err}_{Y_1\|Y_2}$	Err_{Y_2}	$\text{Err}_{Y_2\|Y_1}$
M2	Oracle	12.1(0.10)	12.1(0.10)	12.1(0.10)	0.0(0.00)	0.0(0.00)
	MLDA	15.0(0.14)	15.0(0.14)	15.0(0.14)	0.0(0.00)	0.0(0.00)
	MLDA-C	–	–	12.9(0.12)	–	0.0(0.00)
	MSDA	20.0(0.27)	20.0(0.27)	20.0(0.27)	0.0(0.00)	0.0(0.00)
	MSDA-S	–	–	13.7(0.18)	–	0.0(0.00)
	PLDA	31.7(0.14)	30.6(0.14)	27.8(0.18)	2.6(0.02)	20.5(0.17)
	Logistic	52.1(0.32)	42.3(0.22)	28.9(0.37)	23.5(0.59)	26.6(0.49)

Table A.4: Classification results. Average and standard error (in parentheses) of classification errors Err, Err_{Y_1} (%), $\text{Err}_{Y_1|Y_2}$ (%), Err_{Y_2} (%), $\text{Err}_{Y_2|Y_1}$ (%). Results are based on 100 replications.

	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$
M2	$Y_2=1$				$Y_2=2$				$Y_2=3$			
Oracle	5.8 (0.10)	5.8 (0.10)	– –	– –	21.5 (0.31)	21.5 (0.31)	– –	– –	21.7 (0.27)	21.7 (0.27)	– –	– –
MLDA	7.9 (0.14)	7.9 (0.14)	95.2 (0.99)	0.0 (0.00)	23.1 (0.32)	23.1 (0.32)	100.0 (0.00)	0.6 (0.02)	27.9 (0.34)	27.9 (0.34)	100.0 (0.00)	17.8 (0.32)
MLDA-C	– –	6.3 (0.13)	95.6 (0.92)	0.0 (0.00)	– –	22.4 (0.30)	100.0 (0.00)	0.0 (0.00)	– –	22.8 (0.30)	100.0 (0.00)	0.0 (0.00)
MSDA	16.3 (0.43)	16.3 (0.43)	73.4 (1.91)	8.5 (0.37)	26.0 (0.38)	26.0 (0.38)	100.0 (0.00)	8.7 (0.38)	25.1 (0.27)	25.1 (0.27)	100.0 (0.00)	8.7 (0.38)
MSDA-S	– –	6.2 (0.10)	100.0 (0.00)	0.3 (0.04)	– –	24.8 (0.53)	99.5 (0.50)	0.1 (0.02)	– –	25.1 (0.60)	99.0 (0.70)	0.0 (0.02)
PLDA	22.1 (0.11)	11.1 (0.14)	99.6 (0.40)	96.9 (1.71)	45.3 (0.34)	50.4 (0.42)	99.0 (1.00)	99.0 (1.00)	41.7 (0.37)	50.2 (0.45)	97.0 (1.71)	97.0 (1.71)
Logistic	28.7 (0.33)	27.5 (0.54)	74.8 (1.35)	0.0 (0.00)	62.3 (0.34)	31.3 (0.50)	100.0 (0.00)	0.0 (0.00)	63.0 (0.37)	30.9 (0.44)	99.5 (0.50)	0.0 (0.00)

Table A.5: Y_1 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_1}$ (%), $\widetilde{\text{Err}}_{Y_1|Y_2}$ (%), $\text{TPR}_{1|2}$ (%), $\text{FPR}_{1|2}$ (%). Results are based on 100 replications.

	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$
M2	$Y_1=1$				$Y_1=2$				$Y_1=3$			
Oracle	0.0 (0.00)	0.0 (0.00)	– –	– –	0.0 (0.00)	0.0 (0.00)	– –	– –	0.0 (0.00)	0.0 (0.00)	– –	– –
MLDA	0.0 (0.00)	0.0 (0.00)	100.0 (0.00)	18.2 (0.32)	0.0 (0.00)	0.0 (0.00)	98.5 (0.31)	17.9 (0.32)	0.0 (0.00)	0.0 (0.00)	98.5 (0.31)	17.9 (0.32)
MLDA-C	– –	0.0 (0.00)	100.0 (0.00)	18.1 (0.32)	– –	0.0 (0.01)	97.9 (0.39)	17.7 (0.32)	– –	0.0 (0.00)	98.4 (0.32)	17.7 (0.32)
MSDA	0.0 (0.00)	0.0 (0.00)	100.0 (0.00)	8.0 (0.38)	0.0 (0.00)	0.0 (0.00)	91.1 (0.64)	7.7 (0.38)	0.0 (0.00)	0.0 (0.00)	91.1 (0.64)	7.7 (0.38)
MSDA-S	– –	0.0 (0.00)	100.0 (0.00)	1.5 (0.04)	– –	0.0 (0.00)	81.1 (0.74)	1.6 (0.05)	– –	0.0 (0.00)	77.4 (0.75)	1.3 (0.04)
PLDA	2.6 (0.03)	10.4 (0.13)	100.0 (0.00)	95.7 (1.96)	2.6 (0.03)	10.3 (0.12)	100.0 (0.00)	99.6 (0.02)	2.6 (0.03)	10.0 (0.20)	100.0 (0.00)	99.6 (0.02)
Logistic	23.9 (0.66)	26.8 (0.85)	53.1 (0.99)	0.0 (0.00)	27.4 (0.59)	27.6 (0.84)	36.3 (0.59)	0.0 (0.00)	17.9 (0.69)	25.1 (0.96)	32.3 (0.66)	0.0 (0.00)

Table A.6: Y_2 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_2}$ (%), $\widetilde{\text{Err}}_{Y_2|Y_1}$ (%), $\text{TPR}_{2|1}$ (%), $\text{FPR}_{2|1}$ (%). Results are based on 100 replications.

Additional results under M3

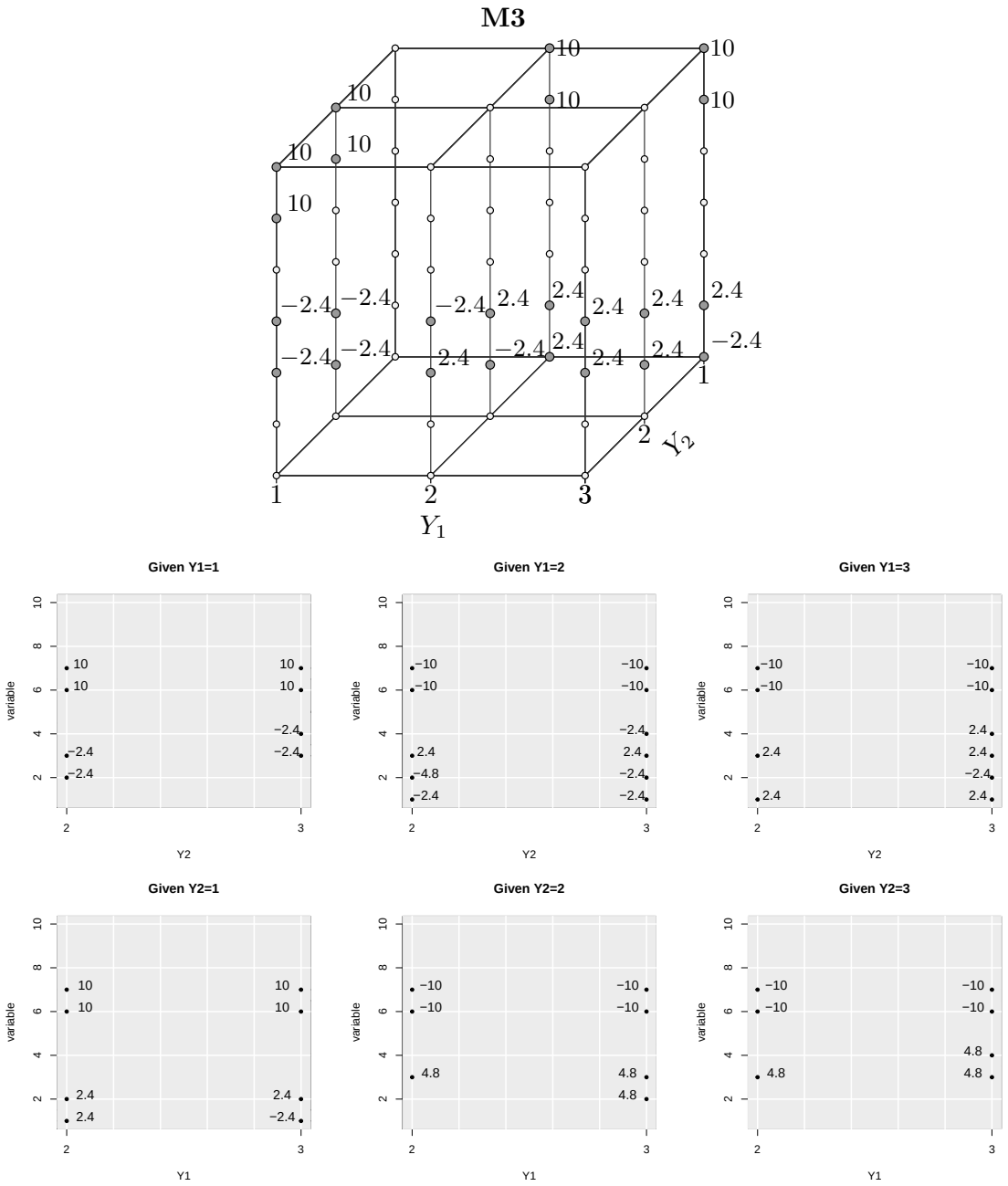

Figure A.3: Illustration of the sparsity pattern of $\boldsymbol{\beta}$ for **M3**. The bottom panels display the sparsity of the conditional discriminant coefficient tensor slices.

Table A.7: Classification results. Average and standard error (in parentheses) of classification errors Err, Err_{Y_1} (%), $\text{Err}_{Y_1|Y_2}$ (%), Err_{Y_2} (%), $\text{Err}_{Y_2|Y_1}$ (%). Results are based on 100 replications.

	Method	Err	Err_{Y_1}	$\text{Err}_{Y_1\|Y_2}$	Err_{Y_2}	$\text{Err}_{Y_2\|Y_1}$
M3	Oracle	9.9(0.09)	5.0(0.06)	0.6(0.02)	9.7(0.09)	5.3(0.06)
	MLDA	13.6(0.18)	7.4(0.15)	1.1(0.04)	12.9(0.18)	6.7(0.14)
	MSDA	15.3(0.24)	8.6(0.16)	2.8(0.12)	13.2(0.18)	7.2(0.13)
	MSDA-S	–	–	4.3(0.13)	–	25.1(0.46)
	PLDA	47.8(0.23)	23.9(0.34)	6.7(0.08)	43.9(0.22)	33.4(0.17)
	Logistic	69.4(0.54)	43.9(0.54)	15.4(0.53)	48.9(0.45)	32.5(0.31)

	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$
M3	$Y_2=1$				$Y_2=2$				$Y_2=3$			
Oracle	6.4 (0.14)	0.6 (0.04)	– –	– –	4.3 (0.12)	0.5 (0.04)	– –	– –	4.4 (0.11)	0.6 (0.05)	– –	– –
MLDA	11.0 (0.34)	1.9 (0.10)	97.2 (0.79)	2.1 (0.06)	4.9 (0.23)	0.6 (0.04)	100.0 (0.00)	0.8 (0.05)	6.3 (0.26)	0.9 (0.06)	100.0 (0.00)	0.9 (0.05)
MSDA	14.3 (0.45)	5.0 (0.26)	94.8 (1.02)	2.1 (0.10)	4.8 (0.15)	1.0 (0.06)	100.0 (0.00)	2.1 (0.10)	6.8 (0.22)	2.6 (0.15)	100.0 (0.00)	2.1 (0.10)
MSDA-S	– –	8.0 (0.24)	62.8 (1.35)	1.8 (0.07)	– –	1.9 (0.15)	99.2 (0.43)	1.1 (0.08)	– –	3.0 (0.25)	98.5 (0.60)	1.1 (0.08)
PLDA	29.8 (0.53)	6.4 (0.13)	100.0 (0.00)	100.0 (0.00)	20.7 (0.69)	6.7 (0.14)	100.0 (0.00)	100.0 (0.00)	21.2 (0.74)	6.8 (0.12)	100.0 (0.00)	100.0 (0.00)
Logistic	46.9 (1.04)	24.3 (0.85)	68.5 (1.10)	0.0 (0.00)	42.7 (0.75)	11.7 (1.01)	73.2 (0.64)	0.0 (0.00)	42.2 (0.72)	10.5 (0.85)	73.8 (0.55)	0.0 (0.00)

Table A.8: Y_1 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_1}$ (%), $\widetilde{\text{Err}}_{Y_1|Y_2}$ (%), $\text{TPR}_{1|2}$ (%), $\text{FPR}_{1|2}$ (%). Results are based on 100 replications.

	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$
M3	$Y_1=1$				$Y_1=2$				$Y_1=3$			
Oracle	11.6 (0.19)	5.2 (0.13)	– –	– –	9.3 (0.15)	5.4 (0.11)	– –	– –	8.2 (0.15)	5.4 (0.12)	– –	– –
MLDA	21.1 (0.55)	9.0 (0.39)	96.6 (0.86)	1.9 (0.05)	9.8 (0.20)	5.8 (0.13)	99.0 (0.40)	1.9 (0.05)	8.8 (0.18)	5.8 (0.12)	97.7 (0.58)	1.8 (0.05)
MSDA	17.6 (0.43)	7.4 (0.21)	100.0 (0.00)	2.0 (0.10)	11.1 (0.22)	7.0 (0.19)	96.5 (0.68)	1.9 (0.10)	11.7 (0.21)	7.4 (0.19)	96.5 (0.68)	1.9 (0.10)
MSDA-S	– –	30.6 (0.65)	49.4 (1.25)	1.0 (0.08)	– –	17.4 (0.82)	57.8 (1.35)	2.0 (0.10)	– –	29.9 (0.75)	43.8 (1.33)	1.0 (0.08)
PLDA	56.7 (0.59)	32.1 (0.33)	100.0 (0.00)	100.0 (0.00)	40.7 (0.56)	31.4 (0.30)	100.0 (0.00)	100.0 (0.00)	35.3 (0.45)	31.7 (0.34)	100.0 (0.00)	100.0 (0.00)
Logistic	82.7 (0.90)	31.7 (0.67)	44.4 (1.05)	0.0 (0.00)	25.3 (1.01)	33.1 (0.45)	42.3 (1.02)	0.0 (0.00)	46.9 (0.92)	32.6 (0.40)	38.8 (1.01)	0.0 (0.00)

Table A.9: Y_2 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_2}$ (%), $\widetilde{\text{Err}}_{Y_2|Y_1}$ (%), $\text{TPR}_{2|1}$ (%), $\text{FPR}_{2|1}$ (%). Results are based on 100 replications.

Additional results under M4

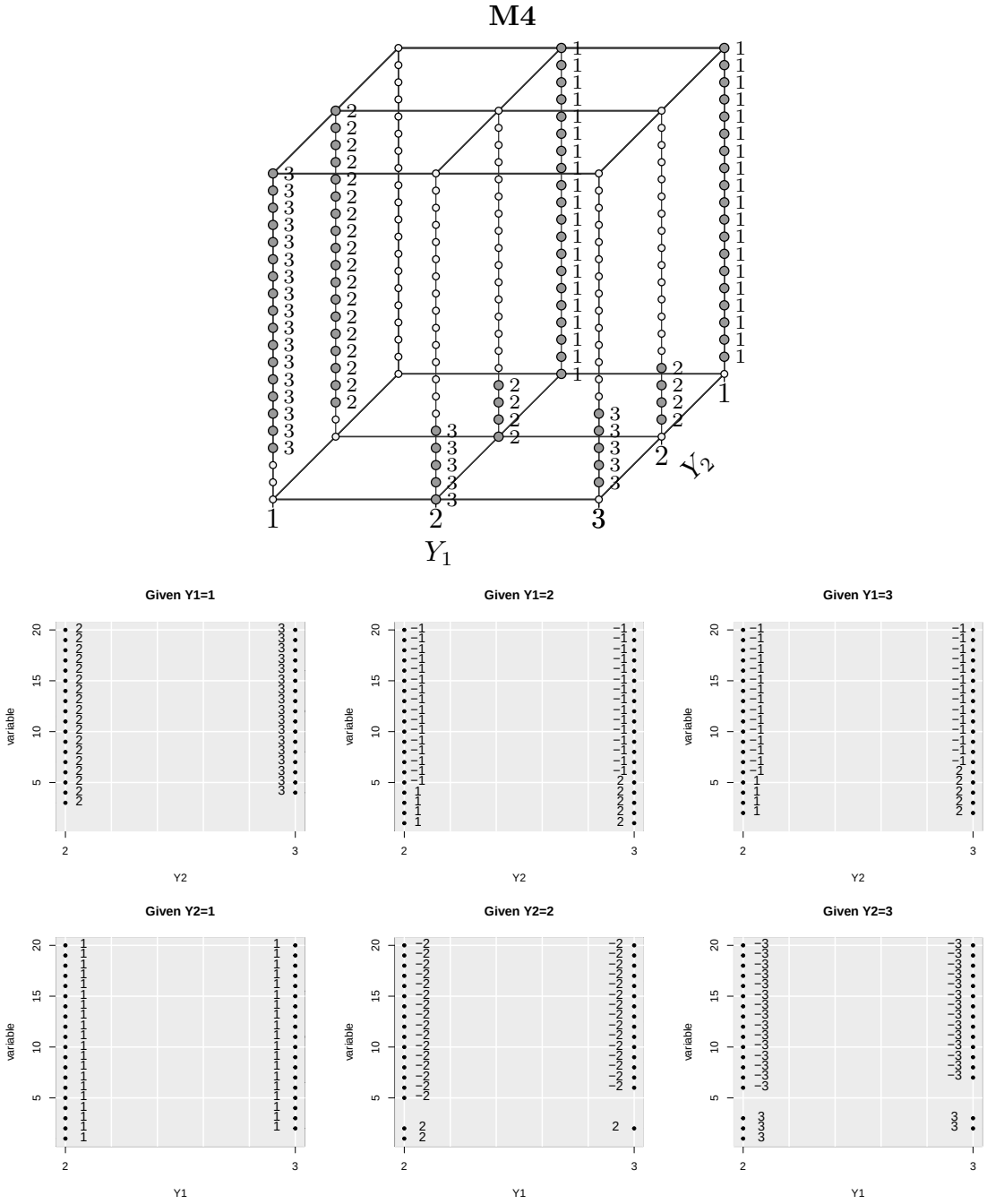

Figure A.4: Illustration of the sparsity pattern of $\boldsymbol{\beta}$ for **M4**. The bottom panels display the sparsity of the conditional discriminant coefficient tensor slices.

	Method	Err	Err_{Y_1}	$\text{Err}_{Y_1\|Y_2}$	Err_{Y_2}	$\text{Err}_{Y_2\|Y_1}$
M4	Oracle	11.2(0.11)	10.9(0.11)	10.7(0.10)	0.7(0.02)	0.5(0.02)
	15.3(0.15)	13.3(0.14)	12.7(0.14)	3.4(0.11)	2.8(0.10)	
	MSDA	27.3(0.50)	25.6(0.42)	24.6(0.41)	4.5(0.35)	3.6(0.34)
	MSDA-S	–	–	29.1(0.32)	–	4.0(0.09)
	PLDA	34.5(0.19)	31.4(0.19)	32.5(0.15)	6.8(0.07)	6.3(0.07)
	Logistic	79.4(0.38)	40.8(0.26)	30.1(0.28)	60.9(0.66)	21.3(0.58)

Table A.10: Classification results. Average and standard error (in parentheses) of classification errors Err, Err_{Y_1} (%), $\text{Err}_{Y_1|Y_2}$ (%), Err_{Y_2} (%), $\text{Err}_{Y_2|Y_1}$ (%). Results are based on 100 replications.

	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$	$\widetilde{\text{Err}}_{Y_1}$	$\widetilde{\text{Err}}_{Y_1\|Y_2}$	$\text{TPR}_{1\|2}$	$\text{FPR}_{1\|2}$
M4	$Y_2=1$				$Y_2=2$				$Y_2=3$			
Oracle	21.4 (0.24)	21.3 (0.24)	– –	– –	8.5 (0.16)	8.2 (0.15)	– –	– –	2.8 (0.10)	2.7 (0.10)	– –	– –
MLDA	24.6 (0.26)	23.7 (0.27)	79.0 (0.80)	10.9 (0.17)	11.4 (0.22)	10.5 (0.21)	75.6 (0.91)	11.0 (0.18)	3.9 (0.15)	3.9 (0.15)	75.6 (0.91)	11.3 (0.17)
MSDA	29.8 (0.32)	28.7 (0.29)	78.1 (1.51)	0.3 (0.04)	24.6 (0.53)	23.8 (0.55)	77.1 (1.59)	0.5 (0.05)	22.6 (0.83)	21.5 (0.84)	75.7 (1.68)	0.5 (0.04)
MSDA-S	– –	28.9 (0.49)	50.4 (1.96)	1.1 (0.11)	– –	30.1 (0.43)	53.7 (2.30)	0.6 (0.09)	– –	28.5 (0.77)	58.2 (2.06)	0.6 (0.07)
PLDA	28.9 (0.28)	27.4 (0.24)	100.0 (0.00)	100.0 (0.00)	33.3 (0.31)	33.3 (0.28)	100.0 (0.00)	100.0 (0.00)	31.9 (0.41)	32.5 (0.36)	100.0 (0.00)	100.0 (0.00)
Logistic	60.7 (0.38)	30.8 (0.40)	16.9 (0.71)	0.4 (0.02)	31.8 (0.50)	30.3 (0.38)	22.6 (0.64)	0.1 (0.01)	30.3 (0.31)	29.0 (0.59)	13.3 (0.59)	0.0 (0.00)

Table A.11: Y_1 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_1}$ (%), $\widetilde{\text{Err}}_{Y_1|Y_2}$ (%), $\text{TPR}_{1|2}$ (%), $\text{FPR}_{1|2}$ (%). Results are based on 100 replications.

	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$	$\widetilde{\text{Err}}_{Y_2}$	$\widetilde{\text{Err}}_{Y_2\|Y_1}$	$\text{TPR}_{2\|1}$	$\text{FPR}_{2\|1}$
M4	$Y_1=1$				$Y_1=2$				$Y_1=3$			
Oracle	0.1 (0.02)	0.0 (0.00)	– –	– –	0.8 (0.05)	0.6 (0.04)	– –	– –	1.0 (0.06)	0.8 (0.05)	– –	– –
MLDA	0.2 (0.02)	0.0 (0.01)	76.2 (0.88)	11.3 (0.16)	4.2 (0.16)	3.4 (0.15)	73.2 (0.87)	5.5 (0.08)	5.7 (0.19)	4.8 (0.19)	72.5 (0.90)	5.4 (0.08)
MSDA	0.9 (0.11)	0.0 (0.00)	83.8 (1.53)	0.4 (0.05)	4.4 (0.38)	4.3 (0.37)	78.1 (1.51)	0.3 (0.04)	8.1 (0.65)	6.1 (0.69)	81.5 (1.55)	0.3 (0.05)
MSDA-S	– –	0.0 (0.00)	56.7 (1.15)	2.5 (0.06)	– –	4.8 (0.15)	61.0 (0.71)	1.0 (0.08)	– –	6.8 (0.19)	56.4 (0.90)	1.2 (0.07)
PLDA	0.5 (0.05)	0.0 (0.00)	100.0 (0.00)	100.0 (0.00)	8.3 (0.13)	7.4 (0.13)	100.0 (0.00)	100.0 (0.00)	11.3 (0.15)	11.0 (0.15)	100.0 (0.00)	100.0 (0.00)
Logistic	33.8 (0.47)	32.9 (1.16)	16.3 (0.49)	0.1 (0.01)	73.5 (0.97)	15.0 (0.99)	30.1 (0.68)	0.0 (0.00)	71.1 (0.98)	18.2 (0.87)	31.4 (0.62)	0.0 (0.01)

Table A.12: Y_2 prediction results. Average and standard error (in parentheses) of classification errors $\widetilde{\text{Err}}_{Y_2}$ (%), $\widetilde{\text{Err}}_{Y_2|Y_1}$ (%), $\text{TPR}_{2|1}$ (%), $\text{FPR}_{2|1}$ (%). Results are based on 100 replications.

Additional results under M5–M6

Method	Oracle	MLDA	MSDA	PLDA	Logistic
M5	14.3(0.11)	15.6(0.13)	22.5(0.24)	73.8(0.18)	76.0(0.20)
M6	17.6(0.13)	21.5(0.15)	37.6(0.50)	76.3(0.15)	68.8(0.30)

Table A.13: Joint classification results of $\mathbf{Y}$ under **M5**–**M6**. Average and standard error (in parentheses) of joint classification error rate. Results are based on 20 replications.

Method	Err_{Y_1} (%)	$\text{Err}_{Y_1\|\mathbf{Y}_{-1}}$	$\text{TPR}_{Y_1\|\mathbf{Y}_{-1}}$	$\text{FPR}_{Y_1\|\mathbf{Y}_{-1}}$
Oracle	8.2(0.09)	2.5(0.05)	–	–
MLDA	9.5(0.10)	3.1(0.06)	100.0(0.00)	3.3(0.04)
MLDA-C	–	2.9(0.06)	100.0(0.03)	0.1(0.00)
MSDA	13.3(0.17)	5.3(0.12)	99.0(0.23)	4.2(0.11)
MSDA-S	–	4.4(0.07)	69.3(0.41)	2.1(0.06)
PLDA	43.2(0.13)	21.5(0.19)	99.5(0.25)	96.6(0.24)
Logistic	40.9(0.31)	4.4(0.08)	92.6(0.48)	0.0(0.00)

Table A.14: Conditional classification results of $Y_1|(Y_2, Y_3, Y_4)$ under **M5**. Average and standard error (in parentheses) of classification errors Err_{Y_1} (%), $\text{Err}_{Y_1|\mathbf{Y}_{-1}}$ (%), $\text{TPR}_{Y_1|\mathbf{Y}_{-1}}$ (%), $\text{FPR}_{Y_1|\mathbf{Y}_{-1}}$ (%). Results are based on 100 replications.

Method	$\text{Err}_{\mathbf{Y}_s}$ (%)	$\text{Err}_{\mathbf{Y}_s\|\mathbf{Y}_{-s}}$	$\text{TPR}_{\mathbf{Y}_s\|\mathbf{Y}_{-s}}$	$\text{FPR}_{\mathbf{Y}_s\|\mathbf{Y}_{-s}}$
Oracle	11.7(0.10)	5.5(0.08)	–	–
MLDA	13.2(0.12)	6.4(0.09)	100.0(0.00)	2.1(0.02)
MLDA-C	–	6.5(0.09)	100.0(0.00)	0.2(0.01)
MSDA	18.6(0.23)	10.1(0.19)	99.0(0.23)	4.0(0.11)
MSDA-S	–	8.1(0.10)	98.7(0.20)	0.4(0.02)
PLDA	61.4(0.16)	62.3(0.18)	100.0(0.00)	100.0(0.00)
Logistic	47.7(0.41)	21.0(0.51)	85.1(0.61)	0.0(0.00)

Table A.15: Conditional classification results of $(Y_1, Y_2)|(Y_3, Y_4)$ under **M5**. Average and standard error (in parentheses) of classification errors $\text{Err}_{\mathbf{Y}_s}$ (%), $\text{Err}_{\mathbf{Y}_s|\mathbf{Y}_{-s}}$ (%), $\text{TPR}_{\mathbf{Y}_s|\mathbf{Y}_{-s}}$ (%), $\text{FPR}_{\mathbf{Y}_s|\mathbf{Y}_{-s}}$ (%). Results are based on 100 replications.

Appendix C. Pan-kidney Cancer Data Analysis

We analyze the pan-kidney cancer data described below to demonstrate our method for conditional classification.

The TCGA pan-kidney cancer data consists of the gene expressions of $n = 420$ patients. The goal is to predict the bivariate categorical response of 5-year survival (binary; survived or failed within five years) and types of cancer: kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), or kidney chromophobe (KICH). Following Molstad and Rothman (2023), we normalize the the j-th gene as $\log\{(v_{i,j} + 1)/q_{i,0.75}\}$, where $v_{i,j}$ is the sequencing count for the i-th subject's j-th gene and $q_{i,0.75}$ is the 75-th percentile of counts for the i-th subject. We then order genes by median absolute deviation decreasingly. Next, we perform a pruning step on the ordered genes such that no two genes

Method	MLDA	MLDA-C	MSDA	MSDA-S	PLDA	Logistic
$\text{Err}_{Y_1\|Y_2}$	5.6(0.24)	5.0(0.22)	9.8(0.39)	19.8(0.37)	10.2(0.33)	24.5(0.47)
$\widehat{s}_1$	21.0(0.21)	17.4(0.20)	3.9(0.12)	10.2(0.34)	98.7(0.03)	1.3(0.02)

Table A.16: Conditional classification results of pan-cancer kidney data. Y_1 denotes cancer type, Y_2 denotes 5-year survival. Reported are average and standard error (in parentheses) of conditional classification error $\text{Err}_{Y_1|Y_2}$ (%) and estimated sparsity level $\widehat{s}_1$ (%), where $\widehat{s}_1 = \sum_{k_2} |\widehat{\mathcal{A}}_{\cdot k_2}|/(pc_2)$. Results are based on 100 replications.

have absolute correlation greater than 0.75. After pruning, we keep the first $p = 500$ genes that remain. We use 4/5 of the samples for training and the remaining 1/5 for testing.

We focus on the conditional classification of cancer type given 5-year survival. The classification results are summarized in Table A.16. We can see that MLDA-C achieved the lowest classification error among all methods considered. Also, MLDA-C provided more sparse estimates than MLDA and improved conditional classification error. Figure A.5 summarizes the conditional classification of cancer type condition on each label of 5-year survival. It is evident that MLDA-C achieved best classification error rate no matter the label of 5-year survival, which slightly improved the classification performance based on MLDA while providing more sparse discriminant coefficient estimates. On the other hand, MSDA-S and Logistic fitted on the partial data performed worst in the conditional classification of cancer type.

Appendix D. ADMM Algorithm for Bivariate LDA

Algorithm A.1 is used to estimate conditional discriminant coefficient tensors under the bivariate LDA model. Details are provided in Section 3.3 of the main text.

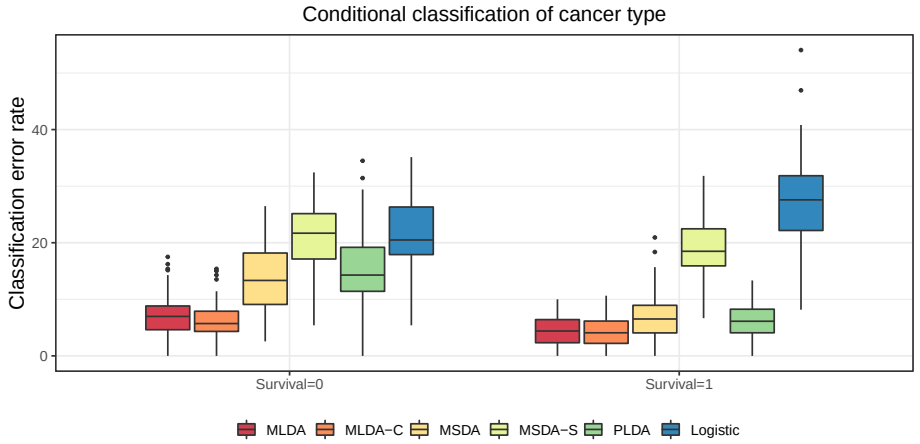

Figure A.5: Conditional classification of cancer type given 5-year survival to be 0 or 1.

Algorithm A.1 ADMM algorithm update for conditional classification (13).

1. **Input**: Sample estimates $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\delta}}$, convergence tolerances $\epsilon_s > 0$ and $\epsilon_r > 0$
2. **Initialize**: $(\boldsymbol{\theta}^r)^{(0)} = 0, (\boldsymbol{\theta}^c)^{(0)} = 0, (\boldsymbol{w}^r)^{(0)} = 0, (\boldsymbol{w}^c)^{(0)} = 0$.
3. **Iterate**: For steps $t = 1, 2, \ldots,$ do the following until convergence.
   (a) Update vectorized $\boldsymbol{\beta}^{(t)}$ as $\mathrm{vec}(\boldsymbol{\beta}^{(t)}) = (\mathbf{0}_p^\top, (\boldsymbol{\beta}_{-1}^{(t)})^\top)^\top$, where $\boldsymbol{\beta}_{-1}^{(t)} \in \mathbb{R}^{(c_1c_2-1)p}$ is updated with the closed-form solution of (22),

$$\boldsymbol{\beta}_{-1}^{(t)} = \big(\mathbf{I}_{c_1c_2-1} \otimes \widehat{\boldsymbol{\Sigma}} + \rho \mathbf{A}_{-p}^\top \mathbf{A}_{-p}\big)^{-1} \big(\rho \mathbf{A}_{-p}^\top (\boldsymbol{\theta}^{(t-1)} - \boldsymbol{w}^{(t-1)}) + \widehat{\boldsymbol{\delta}}_{-1}\big).$$

   (b) Update $\boldsymbol{\theta}^{(r)}$ and $\boldsymbol{\theta}^{(c)}$:
      i. For $j = 1, \ldots, p$ and $k_2 = 1, \ldots, c_2$:

$$(\boldsymbol{\theta}^r_{[j,:,k_2]})^{(t)} = \left(1 - \frac{\lambda_1/\rho}{\|\mathbf{A}^{c_1}\boldsymbol{\beta}^{(t)}_{[j,:,k_2]} + (\boldsymbol{w}^r_{[j,:,k_2]})^{(t-1)}\|_2}\right)_+ \left\{\mathbf{A}^{c_1}\boldsymbol{\beta}^{(t)}_{[j,:,k_2]} + (\boldsymbol{w}^r_{[j,:,k_2]})^{(t-1)}\right\}.$$

      ii. For $j = 1, \ldots, p$ and $k_1 = 1, \ldots, c_1$:

$$(\boldsymbol{\theta}^c_{[j,k_1,:]})^{(t)} = \left(1 - \frac{\lambda_2/\rho}{\|\mathbf{A}^{c_2}\boldsymbol{\beta}^{(t)}_{[j,k_1,:]} + (\boldsymbol{w}^c_{[j,k_1,:]})^{(t-1)}\|_2}\right)_+ \left\{\mathbf{A}^{c_2}\boldsymbol{\beta}^{(t)}_{[j,k_1,:]} + (\boldsymbol{w}^c_{[j,k_1,:]})^{(t-1)}\right\}.$$

   (c) Calculate:

$$\text{dual residual: } s^{(t)} = \rho \mathbf{A}^\top (\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)});$$
$$\text{primal residual: } r^{(t)} = \mathbf{A}\mathrm{vec}(\boldsymbol{\beta}^{(t)}) - \boldsymbol{\theta}^{(t)}.$$

   (d) Update $\boldsymbol{w}^{(r)}$ and $\boldsymbol{w}^{(c)}$:

$$(\boldsymbol{w}^r)^{(t)} = (\boldsymbol{w}^r)^{(t-1)} + \boldsymbol{\beta}^{(t)} \times_2 \mathbf{A}^{c_1} - (\boldsymbol{\theta}^r)^{(t)},$$
$$(\boldsymbol{w}^c)^{(t)} = (\boldsymbol{w}^c)^{(t-1)} + \boldsymbol{\beta}^{(t)} \times_3 \mathbf{A}^{c_2} - (\boldsymbol{\theta}^c)^{(t)}.$$

4. **Output**: $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}^r, \widehat{\boldsymbol{\theta}}^c)$, the iterates $(\boldsymbol{\beta}^{(t)}, (\boldsymbol{\theta}^r)^{(t)}, (\boldsymbol{\theta}^c)^{(t)})$ after both $\|s^{(t)}\|_2 \le \epsilon_s$ and $\|r^{(t)}\|_2 \le \epsilon_r$.

Appendix E. Extension to Arbitrary Multivariate Response ($M \ge 3$)

E.1 Bayes' rules

We first derive the Bayes' rules for both joint and conditional classification under the multivariate LDA model (25) and characterize the corresponding discriminant coefficient tensors.

For joint classification, the Bayes' rule $\phi_{\mathbf{Y}} : \mathbb{R}^p \to [c_1] \times \cdots \times [c_M]$ achieves the lowest joint classification error rate. Under model (25), $\phi_{\mathbf{Y}}(\mathbf{X})$ can be written as

$$\phi_{\mathbf{Y}}(\mathbf{X}) = \underset{k_m \in [c_m], m \in [M]}{\operatorname{argmax}} \left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_{k_1 \ldots k_M} + \boldsymbol{\mu}_{1\ldots1}}{2} \right)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{k_1 \ldots k_M} - \boldsymbol{\mu}_{1\ldots1}) + \log \pi_{k_1 \ldots k_M} \right\}. \tag{A.1}$$

Let $\boldsymbol{\delta} = \boldsymbol{\mu} - \boldsymbol{\mu}_{1\ldots1} \circ \mathbf{1}_{c_1} \cdots \circ \mathbf{1}_{c_M} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$ and $\boldsymbol{\beta} = \boldsymbol{\delta} \times_1 \boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$. Then, we have $\boldsymbol{\beta}_{k_1 \ldots k_M} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{k_1 \ldots k_M} - \boldsymbol{\mu}_{1\ldots1})$ for $k_m \in [c_m]$ and $m \in [M]$. Thus, $\boldsymbol{\beta}$ is the joint discriminant coefficient tensor for joint classification that contains all the discriminant directions.

To describe our method for conditional classification, we must define a number of quantities. Let Z_s be the univariate response transformed from $\mathbf{Y}_{-s}$ with total $d_s := \prod_{m \neq s} c_m$ categories. And let $\psi : [d_s] \to [c_1] \times \cdots \times [c_{s-1}] \times [c_{s+1}] \times \cdots \times [c_M]$ be the category mapping from the univariate Z_s to multivariate $\mathbf{Y}_{-s}$ such that for $u \in [d_s]$, $\psi(u) = (k_1, \ldots, k_{s-1}, k_{s+1}, \ldots, k_M)$ for some $k_m \in [c_m]$, $m \neq s$. We then use $\boldsymbol{\beta}_{k_s \cup \psi(u)}$ to denote $\boldsymbol{\beta}_{k_1,\ldots,k_s,\ldots,k_M}$ and $\boldsymbol{\mu}_{k_s \cup \psi(u)}$ to denote $\boldsymbol{\mu}_{k_1,\ldots,k_s,\ldots,k_M}$.

For conditional classification of a univariate response Y_s, the conditional Bayes' rule $\phi_{Y_s} : \mathbb{R}^p \times [c_1] \times \cdots \times [c_{s-1}] \times [c_{s+1}] \times \cdots \times [c_M] \to [c_s]$ that achieves the lowest conditional classification error rate under model (25) can be written as

$$\begin{aligned} \phi_{Y_s}(\mathbf{X}, \mathbf{Y}_{-s}) = \underset{k_s \in [c_s]}{\operatorname{argmax}} \Bigg\{ & \left(\mathbf{X} - \frac{\boldsymbol{\mu}_{k_s \cup \psi(u)} + \boldsymbol{\mu}_{1 \cup \psi(u)}}{2} \right)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{k_s \cup \psi(u)} - \boldsymbol{\mu}_{1 \cup \psi(u)}) \\ & + \log \pi_{k_s \cup \psi(u)} \Bigg\}. \end{aligned} \tag{A.2}$$

Therefore, when given $\mathbf{Y}_{-s} = u$ for $u \in [d_s]$, the parameter of interest becomes $(\boldsymbol{\beta}_{2 \cup \psi(u)} - \boldsymbol{\beta}_{1 \cup \psi(u)}, \ldots, \boldsymbol{\beta}_{c_s \cup \psi(u)} - \boldsymbol{\beta}_{1 \cup \psi(u)}) \in \mathbb{R}^{p \times (c_s - 1)}$.

In what follows, we extend Algorithms 1 and A.1 to estimate the discriminant coefficient tensors efficiently under both joint and conditional classification settings for multivariate LDA (25).

E.2 Extending Algorithm 1

For joint classification, we propose to estimate the discriminant coefficient tensor using

$$\widehat{\boldsymbol{\beta}} \in \underset{\substack{\boldsymbol{\beta} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M} \\ \boldsymbol{\beta}_{1\ldots1} = 0}}{\operatorname{argmin}} \left\{ \widetilde{g}_n(\boldsymbol{\beta}) + \widetilde{\mathcal{P}}_\lambda(\boldsymbol{\beta}) \right\}. \tag{A.3}$$

As discussed in Section 7, to classify multivariate response $\mathbf{Y} = (Y_1, \ldots, Y_M)$, the target parameter to estimate now is a $(M+1)$-way tensor $\boldsymbol{\beta} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$. In the multivariate case, it is reasonable to consider $\lambda_s = \lambda$ for all $s \in [M]$ so that the tuning parameter can be easily tuned.

By introducing the latent overlapping group lasso analog to $\widetilde{\mathcal{P}}_\lambda$, $\widetilde{\mathcal{P}}_{\mathcal{V},\lambda}$, we estimate $\boldsymbol{\beta}$ by estimating $p \sum_{s=1}^M d_s$ overlapped groups of parameters in $\boldsymbol{\beta}$. Let G be the set of $p \sum_{s=1}^M d_s$ groups of $p \prod_{s=1}^M c_s$ parameters in $\boldsymbol{\beta}$ to be estimated. For each $g \in G$, let $\boldsymbol{\nu}^{(g)} \in \mathbb{R}^{p \times c_1 \times \cdots \times c_M}$

be a tensor whose nonzero entries corresponds to one group parameters (e.g., as in Figure 3) and let $\mathcal{V}^{(g)} \subseteq \mathbb{R}^{p\times c_1\times\cdots\times c_M}$ be the subspace of such tensors. Similar to the bivariate response case, we use the latent overlapping group lasso penalty to estimate $\boldsymbol{\nu}^{(g)}$ such that $\boldsymbol{\beta} = \sum_{g\in G}\boldsymbol{\nu}^{(g)}$ by solving

$$\underset{\substack{\boldsymbol{\nu}^{(g)}\in\mathcal{V}^{(g)}\\ g\in G}}{\operatorname{argmin}} \left[\sum_{k_1,\ldots,k_M} \left\{ \frac{1}{2}\big(\sum_{g\in G}\boldsymbol{\nu}^{(g)}_{k_1\ldots k_M}\big)^\top \widehat{\boldsymbol{\Sigma}}\big(\sum_{g\in G}\boldsymbol{\nu}^{(g)}_{k_1\ldots k_M}\big) - \widehat{\boldsymbol{\delta}}^\top_{k_1\ldots k_M}\big(\sum_{g\in G}\boldsymbol{\nu}^{(g)}_{k_1\ldots k_M}\big) \right\} + \sum_{g\in G}\lambda_g\|\boldsymbol{\nu}^{(g)}_g\|_2 \right], \tag{A.4}$$

where $\boldsymbol{\nu}^{(g)}_g \in \mathbb{R}^{c_s}$ and $\lambda_g = \lambda_s$ if group g is for the classification and variable selection of response Y_s. With the latent overlapping group lasso penalty defined in terms of $\boldsymbol{\beta}$ analogously to (15), it can be shown that for estimating $\boldsymbol{\beta}$, solving the optimization (A.4) is equivalent to solving

$$\underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^{p\times c_1\times\cdots\times c_M}\\ \boldsymbol{\beta}_{1\ldots1}=0}}{\operatorname{argmin}} \left\{ \widetilde{g}_n(\boldsymbol{\beta}) + \widetilde{\mathcal{P}}_{\mathcal{V},\lambda}(\boldsymbol{\beta}) \right\}. \tag{A.5}$$

We summarize the estimation procedure for high-dimensional multivariate LDA in Algorithm A.2.

E.3 Extending Algorithm A.1

For conditional classification of Y_s given $\mathbf{Y}_{-s}$ ($s \in [M]$), we estimate the conditional discriminant coefficients $\boldsymbol{\theta}^s$ by solving

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}^s} \in \underset{\substack{\boldsymbol{\beta}\in\mathbb{R}^{p\times c_1\times\cdots\times c_M}\\ \boldsymbol{\beta}_{1\ldots1}=0}}{\operatorname{argmin}} \left\{ \widetilde{g}_n(\boldsymbol{\beta}) + \widetilde{\mathcal{H}}^s_\lambda(\boldsymbol{\beta}) \right\}, \tag{A.6}$$

where $\widetilde{\mathcal{H}}^s_\lambda(\boldsymbol{\beta}) = \lambda\sum_{u=1}^{d_s}\|\boldsymbol{\theta}^s_u\|_{2,1}$ where $\boldsymbol{\theta}^s_u \in \mathbb{R}^{p\times(c_s-1)}$ is the matrix of discriminant coefficients for $(Y_s \mid \mathbf{X}, \mathbf{Y}_{-s})$ and $\boldsymbol{\theta}^s \in \mathbb{R}^{p\times(c_s-1)\times d_s}$ Note that this approach is somewhat distinct from the bivariate case since here, we target only predictors important for $Y_s \mid (\mathbf{X}, \mathbf{Y}_{-s})$, whereas in the bivariate case, we can target both $Y_1 \mid (\mathbf{X}, Y_2)$ and $Y_2 \mid (\mathbf{X}, Y_1)$ simultaneously. The optimization (A.6) can be solved similarly to Algorithm A.1 with some modification on the ADMM constraints. Define $\boldsymbol{w}^s \in \mathbb{R}^{p\times(c_s-1)\times d_s}$ as the Lagrangian multipliers associated with the two constraints in (A.6), and let $\boldsymbol{\theta} = \operatorname{vec}(\boldsymbol{\theta}^s)^\top \in \mathbb{R}^{p(c_s-1)d_s}$, $\boldsymbol{w} = \operatorname{vec}(\boldsymbol{w}^s)^\top \in \mathbb{R}^{p(c_s-1)d_s}$. Then for $\rho > 0$, we write the augmented Lagrangian of the objective function (13) as

$$\begin{aligned} L_\rho(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{w}) = &\sum_{k_1,\ldots,k_M}\left\{\frac{1}{2}\boldsymbol{\beta}^\top_{k_1\ldots k_M}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1\ldots k_M} - \widehat{\boldsymbol{\delta}}^\top_{k_1\ldots k_M}\boldsymbol{\beta}_{k_1\ldots k_M}\right\} + \lambda\sum_{u=1}^{d_s}\|\boldsymbol{\theta}^s_u\|_{2,1} \\ &+ \frac{\rho}{2}\|\mathbf{A}\operatorname{vec}(\boldsymbol{\beta}) - \boldsymbol{\theta} + \boldsymbol{w}\|_2^2 - \frac{\rho}{2}\|\boldsymbol{w}\|_2^2, \end{aligned} \tag{A.8}$$

where $\mathbf{A} = \mathbf{I}_{d_s} \otimes (\mathbf{A}^{c_s})^\top \otimes \mathbf{I}_s \in \mathbb{R}^{p(c_s-1)d_s \times pc_s d_s}$. Then we apply the alternating direction method of multipliers (ADMM) algorithm to solve the optimization (A.6). We summarize the estimation procedure in Algorithm A.3.

Similar to the bivariate response scenario, we consider conditional variable selection as an additional step after joint variable selection. After the joint discriminant coefficient $\widehat{\boldsymbol{\beta}}$ and the set of important variables for joint classification is obtained from Algorithm A.1, we further apply Algorithm A.2 on the reduced data for conditional classification variable selection and estimation.

Moreover, if the problem of interest is to predict multivariate response $\mathbf{Y}_S = (Y_{s_1}, \ldots, Y_{s_N})$, $S = \{s_k\}_{k=1}^N \subset [M]$, conditionally from $(\mathbf{X}, \mathbf{Y}_{-S})$, where $\mathbf{Y}_{-S}$ is the multivariate response excluding $\mathbf{Y}_S$ from $\mathbf{Y}$, we transform $\mathbf{Y}_S$ to a univariate response $\widetilde{Y}_S$ so that it becomes the same problem as (A.6). Here we remark that it is preferable to consider the responses in $\widetilde{Y}_S$ distinctly instead of combining them. However, this extension is nontrivial and would require a substantially different algorithm from Algorithm 1 and A.1. Therefore, we leave this extension as future work.

E.4 Pairwise Likelihood Approach

E.4.1 Overview

In many applications when the number of responses M is large, some combinations of the M categories may not be observed in a random sample. The pairwise likelihood is a special case of composite likelihood (Cox and Reid, 2004; Lindsay, 1988). The idea is to

---

Algorithm A.2 Blockwise coordinate descent algorithm for joint classification (A.5).

1. **Input**: Sample estimates $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\delta}}$, parameter groups G.
2. **Initialize**: $(\widehat{\boldsymbol{\nu}}^{(g)})^{(0)} = 0$ for all $g \in G$.
3. **Iterate**: For steps $t = 1, 2, \ldots,$ do the following until convergence.
   For $g \in G$:
   (a) Compute
   $$(\mathbf{Z}^{(g)})^{(t-1)} = \frac{\widehat{\boldsymbol{\delta}}_g - (\boldsymbol{\beta}^{(t-1)}_{\cdot\psi(u)})^\top \widehat{\boldsymbol{\Sigma}}_{\cdot j}}{\widehat{\sigma}_{jj}} + \boldsymbol{\beta}^{(t-1)}_g,$$
   where group g is for classification of response Y_s, $j = j(g) \in \{1, \ldots, p\}$, $s = s(g) \in [M]$ and $u = u(g) \in [d_s]$.
   (b) Update $(\widehat{\boldsymbol{\nu}}^{(g)})^{(t)}$ by
   $$(\widehat{\boldsymbol{\nu}}^{(g)}_g)^{(t)} \longleftarrow \left(1 - \frac{\lambda_g/\widehat{\sigma}_{jj}}{\|(\mathbf{Z}^{(g)})^{(t-1)} - \sum_{\substack{\ell \in G,\\ \ell \neq g}} (\widehat{\boldsymbol{\nu}}^{(\ell)}_g)^{(t-1)}\|_2}\right)_+ \left\{(\mathbf{Z}^{(g)})^{(t-1)} - \sum_{\substack{\ell \in G,\\ \ell \neq g}} (\widehat{\boldsymbol{\nu}}^{(\ell)}_g)^{(t-1)}\right\}.$$
   (c) Update $\boldsymbol{\beta}^{(t)} = \sum_{g \in G} (\widehat{\boldsymbol{\nu}}^{(g)})^{(t)}$.
4. **Output**: $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$.

---

construct pseudolikelihoods by compounding lower-dimensional margins to approximate the full likelihood that is difficult to evaluate or estimate. Following the definition in Varin et al. (2011), suppose we have a M-dimensional vector random variable Y with density function $f(y;\theta)$ for unknown $\theta \in \mathbb{R}^p$ and a set of marginal or conditional events $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ with associated likelihoods $\mathcal{L}_k(\theta; y) \propto f(y \in \mathcal{A}_k; \theta)$. Then the composite likelihood is defined as

$$\mathcal{L}_C(\theta; y) = \prod_{k=1}^{K} \mathcal{L}_k(\theta; y). \tag{A.9}$$

The pairwise likelihood considers the dependence between each pair of the M variables $Y_1, \ldots, Y_M$ in $\mathbf{Y}$ and is particularly popular for modeling correlated categorical responses (le Cessie and Van Houwelingen, 1994; Varin, 2008). The pairwise likelihood is defined as

$$\mathcal{L}_{pair}(\theta; y) = \prod_{s>r} f(y_r, y_s; \theta) = \prod_{r=1}^{M-1} \prod_{s=r+1}^{M} f(y_r, y_s; \theta). \tag{A.10}$$

In the case of multivariate categorical response regression, the pairwise likelihood essentially models every pair of the M responses to form the pseudolikelihood to obtain parameter estimates and perform classification. The first example is from le Cessie and Van Houwelingen (1994), where they model each pair by logistic marginals.

Algorithm A.3 ADMM algorithm update for conditional classification (A.6).

1. **Input**: Sample estimates $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\delta}}$, convergence tolerances $\epsilon_s > 0$ and $\epsilon_r > 0$
2. **Initialize**: $(\boldsymbol{\theta}^s)^{(0)} = 0, (\boldsymbol{w}^s)^{(0)} = 0$.
3. **Iterate**: For steps $t = 1, 2, \ldots,$ do the following until convergence.
   (a) Update vectorized $\boldsymbol{\beta}$ as $\mathrm{vec}(\boldsymbol{\beta}^{(t)}) = (\mathbf{0}_p^\top, (\boldsymbol{\beta}_{-1}^{(t)})^\top)^\top$, where $\boldsymbol{\beta}_{-1}^{(t)} \in \mathbb{R}^{(c_s d_s - 1)p}$ is updated by

$$\boldsymbol{\beta}_{-1}^{(t)} = \big(\mathbf{I}_{K-1} \otimes \widehat{\boldsymbol{\Sigma}} + \rho \mathbf{A}_{-p}^\top \mathbf{A}_{-p}\big)^{-1} \big(\rho \mathbf{A}_{-p}^\top (\boldsymbol{\theta}^{(t-1)} - \boldsymbol{w}^{(t-1)}) + \widehat{\boldsymbol{\delta}}_{-1}\big). \tag{A.7}$$

   (b) Update $\boldsymbol{\theta}^s$: For $j = 1, \ldots, p$ and $u = 1, \ldots, d_s$,

$$(\boldsymbol{\theta}^s_{[j,:,u]})^{(t)} = \left(1 - \frac{\lambda/\rho}{\|\mathbf{A}^{c_s}(\boldsymbol{\beta}^{(t)}_{(1,1+s)})_{[j,:,u]} + (\boldsymbol{w}^s_{[j,:,u]})^{(t-1)}\|_2}\right)_+ \left\{\mathbf{A}^{c_s}(\boldsymbol{\beta}^{(t)}_{(1,1+s)})_{[j,:,u]} + (\boldsymbol{w}^s_{[j,:,u]})^{(t-1)}\right\}.$$

   (c) Calculate:

$$\text{dual residual: } s^{(t)} = \rho \mathbf{A}^\top (\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)});$$
$$\text{primal residual: } r^{(t)} = \mathbf{A}\mathrm{vec}(\boldsymbol{\beta}^{(t)}) - \boldsymbol{\theta}^{(t)}.$$

   (d) Update $\boldsymbol{w}^{(s)}$:

$$(\boldsymbol{w}^s)^{(t)} = (\boldsymbol{w}^s)^{(t-1)} + \boldsymbol{\beta}^{(t)}_{(1,s+1)} \times_2 \mathbf{A}^{c_s} - (\boldsymbol{\theta}^s)^{(t)}.$$

4. **Output**: $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}^s}, \widehat{\boldsymbol{\theta}}^s)$, the iterates $(\boldsymbol{\beta}^{(t)}, (\boldsymbol{\theta}^s)^{(t)})$ after both $\|s^{(t)}\|_2 \le \epsilon_s$ and $\|r^{(t)}\|_2 \le \epsilon_r$.

E.4.2 Joint Classification

Here we adopt the idea of pairwise likelihood and combine it with the proposed bivariate LDA model to achieve joint classification of a multivariate ($M \geq 3$) response. The motivation of pairwise likelihood formulation in (A.10) is to obtain pseudolikelihood estimation of the unknown parameters $\boldsymbol{\Theta}$, while we use each model fitted on each pair of responses to approximate the joint distribution of all responses $\mathbf{Y} = (Y_1, \ldots, Y_M)$ by $f(Y_1, \ldots, Y_M; \theta) = \prod_{r=1}^{M-1} \prod_{s=r+1}^{M} f(Y_r, Y_s; \theta^{rs})$, where $\boldsymbol{\Theta} = \{\theta^{rs}\}_{s>r}^{M}$. Specifically, under the bivariate LDA model (1), we assume $\mathbf{X}|(Y_r = k_r, Y_s = k_s) \sim \mathrm{N}(\boldsymbol{\mu}_{k_r k_s}^{rs}, \boldsymbol{\Sigma}^{rs})$ and $\Pr(Y_r = k_r, Y_s = k_s) = \pi_{k_r k_s}^{rs} > 0$. Thus the likelihood function $f(Y_r, Y_s; \theta^{rs})$ is given by

$$\begin{aligned} f(Y_r, Y_s; \theta^{rs}) &= f(\mathbf{X}|Y_r, Y_s; \theta^{rs}) f(Y_r, Y_s)/f(\mathbf{X}) \\ &\propto \pi_{Y_r Y_s}^{rs} \exp\left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_{Y_r Y_s}^{rs} + \boldsymbol{\mu}_{11}^{rs}}{2} \right)^{\top} \boldsymbol{\beta}_{Y_r Y_s}^{rs} \right\}, \end{aligned} \tag{A.11}$$

where $\theta^{rs} = \{\boldsymbol{\mu}^{rs}, \boldsymbol{\beta}^{rs}, \pi^{rs}\}$, $\boldsymbol{\mu}^{rs} \in \mathbb{R}^{p \times c_r \times c_s}$, $\boldsymbol{\beta}^{rs} \in \mathbb{R}^{p \times c_r \times c_s}$ satisfies $\boldsymbol{\beta}_{Y_r Y_s}^{rs} = (\boldsymbol{\Sigma}^{rs})^{-1}(\boldsymbol{\mu}_{Y_r Y_s}^{rs} - \boldsymbol{\mu}_{11}^{rs})$ and $\sum_{Y_r, Y_s} \pi_{Y_r Y_s}^{rs} = 1$. Then we approximate the joint likelihood of $\mathbf{Y} \mid \mathbf{X}$ using

$$f(Y_1, \ldots, Y_M; \boldsymbol{\Theta}) \propto \prod_{r=1}^{M-1} \prod_{s=r+1}^{M} \pi_{Y_r Y_s}^{rs} \exp\left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_{Y_r Y_s}^{rs} + \boldsymbol{\mu}_{11}^{rs}}{2} \right)^{\top} \boldsymbol{\beta}_{Y_r Y_s}^{rs} \right\}. \tag{A.12}$$

Based on the approximate joint likelihood (A.12), we can obtain the estimated response $(\widehat{Y}_1, \ldots, \widehat{Y}_M)$ by finding which response category combination maximizes $f(Y_1, \ldots, Y_M; \boldsymbol{\Theta})$.

Besides the pairwise approximation, a common approach in multivariate response classification problem is to use a separate model for each response. This ignores responses' dependence by implicitly assuming $\Pr(Y_1 = k_1, \ldots, Y_M = k_M \mid \mathbf{X}) = \prod_{m=1}^{M} \Pr(Y_m = k_m \mid \mathbf{X})$. This separate approximation can be viewed as an application of the simplest composite likelihood referred as the independence likelihood, defined as

$$\mathcal{L}_{ind}(\theta; Y) = \prod_{r=1}^{M} f(Y^r; \boldsymbol{\Theta}). \tag{A.16}$$

Our pairwise likelihood classification procedure is summarized in Algorithm A.4.

It may also be worthwhile to consider the weighted version of the pairwise likelihood:

$$\mathcal{L}_{pair}^{w}(\theta; y) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r, y_s; \boldsymbol{\Theta})^{w_{rs}}, \tag{A.17}$$

where w_{rs} are nonnegative weights to be chosen. In our implementation of bivariate LDA classification, we can choose bigger weights on pairs with higher training classification accuracy.

E.4.3 Conditional Classification

We can also apply the composite likelihood for conditional classification. Suppose we want to predict multivariate response $\mathbf{Y}_S = (Y_{s_1}, \ldots, Y_{s_M})$ conditionally given $\mathbf{Y}_{-S}$, we can

approximate the conditional likelihood with independence composite likelihood (A.16) as

$$f(\mathbf{Y}_S|\mathbf{Y}_{-S};\widehat{\mathbf{\Theta}}) = \prod_{r=1}^{m} f(Y_{s_r}|\mathbf{Y}_{-S};\widehat{\boldsymbol{\theta}}^{r,-S}) \\ \propto \prod_{r=1}^{m} \widehat{\pi}^{r,-S}_{Y_{s_r}\mathbf{Y}_{-S}} \exp\left\{\left(\mathbf{X} - \frac{\boldsymbol{\mu}^{r,-S}_{Y_{s_r}\mathbf{Y}_{-S}} + \boldsymbol{\mu}^{r,-S}_{1\mathbf{Y}_{-S}}}{2}\right)^{\top} \widehat{\boldsymbol{\theta}}^{r,-S}_{Y_r\mathbf{Y}_{-S}}\right\}, \tag{A.18}$$

where $\widehat{\boldsymbol{\theta}}^{r,-S}$ is the output after solving (A.6).

Another way of approximating $f(\mathbf{Y}_s|\mathbf{Y}_{-s};\mathbf{\Theta})$ is to directly use the pairwise likelihood just introduced. Note that $f(\mathbf{Y}_s|\mathbf{Y}_{-s},\mathbf{\Theta}) \propto f(\mathbf{Y}_s,\mathbf{Y}_{-s};\mathbf{\Theta})$, we can compute the conditional likelihood by first approximating the joint likelihood $f(\mathbf{Y}_s,\mathbf{Y}_{-s};\mathbf{\Theta})$ with (A.12), then the conditional classification can be carried out by maximizing $f(\mathbf{Y}_s|\mathbf{Y}_{-s},\mathbf{\Theta})$ over all possible categories of $\mathbf{Y}_s$.

Appendix F. Proof of Propositions

F.1 Proof of Proposition 1

Following the proof technique from Min et al. (2023), Proposition 1follows directly from arguments used to prove their Theorem 3.3. We provide a brief sketch of how their arguments apply here. In their first step, Min et al. (2023) show that under the conditions of Lemma A.8, if we compare any two response category combinations, say (k_1,k_2) to (k_1',k_2'), there

Algorithm A.4 Pairwise classification with bivariate LDA.

1. **Input**: Training data $\{X_i,Y_{1i},\ldots,Y_{iM}\}_{i=1}^n$; new data $\mathbf{X}\in\mathbb{R}^p$ to be classified.
2. For all $M(M-1)/2$ pairs of (r,s) where $1\le r<s\le M$:
   (a) Fit bivariate MLDA on $\{X_i,(Y_{ri},Y_{si})\}_{i=1}^n$ to obtain estimates $\widehat{\theta}^{rs}=\{\widehat{\boldsymbol{\mu}}^{rs},\widehat{\boldsymbol{\beta}}^{rs},\widehat{\pi}^{rs}\}$.
   (b) Obtain likelihood $f(Y_r=k_r,Y_s=k_s|\mathbf{X},\widehat{\theta}^{rs})$ by

$$f(Y_r=k_r,Y_s=k_s|\mathbf{X},\widehat{\theta}^{rs}) = \frac{\widehat{\pi}^{rs}_{k_rk_s}\exp\left\{(\mathbf{X}-\frac{\widehat{\boldsymbol{\mu}}^{rs}_{k_rk_s}+\widehat{\boldsymbol{\mu}}^{rs}_{11}}{2})^{\top}\widehat{\boldsymbol{\beta}}^{rs}_{k_rk_s}\right\}}{\sum_{k_r=1}^{c_r}\sum_{k_s=1}^{c_s}\widehat{\pi}^{rs}_{k_rk_s}\exp\left\{(\mathbf{X}-\frac{\widehat{\boldsymbol{\mu}}^{rs}_{k_rk_s}+\widehat{\boldsymbol{\mu}}^{rs}_{11}}{2})^{\top}\widehat{\boldsymbol{\beta}}^{rs}_{k_rk_s}\right\}}. \tag{A.13}$$

3. Obtain log-likelihood of $(Y_1=k_1,\ldots,Y_M=k_M)$ for $j_m\in\{1,\ldots,c_m\}$, $m=1,\ldots,M$ as

$$\log f(Y_1=k_1,\ldots,Y_M=k_M;\widehat{\mathbf{\Theta}}) = \sum_{r=1}^{M-1}\sum_{s=r+1}^{M}\log f(Y_r=k_r,Y_s=k_s|\mathbf{X},\widehat{\theta}^{rs}). \tag{A.14}$$

4. **Output**: The predicted label $\widehat{\mathbf{Y}}=(\widehat{Y}_1,\ldots,\widehat{Y}_M)$ is given by

$$(\widehat{Y}_1,\ldots,\widehat{Y}_M) = \underset{k_1,\ldots,k_M}{\operatorname{argmax}}\log f(Y_1=k_1,\ldots,Y_M=k_M;\widehat{\mathbf{\Theta}}). \tag{A.15}$$

exists a constant $c_1 \in (0, \infty)$ such that with probability at least $1 - O(p^{-1})$,

$$\begin{aligned}\widehat{R}^{\text{err}}((k_1,k_2),(k_1',k_2')) &:= \pi_{k_1,k_2}\text{Pr}_{\boldsymbol{\Theta}}\{\widehat{D}_{k_1,k_2}(\mathbf{X}) < \widehat{D}_{k_1',k_2'}(\mathbf{X}) \mid \text{labels}(\mathbf{X}) = (k_1,k_2)\} \\ &- \pi_{k_1,k_2}\text{Pr}_{\boldsymbol{\Theta}}\{D_{k_1,k_2}(\mathbf{X}) < D_{k_1',k_2'}(\mathbf{X}) \mid \text{labels}(\mathbf{X}) = (k_1,k_2)\} \\ &+ \pi_{k_1',k_2'}\text{Pr}_{\boldsymbol{\Theta}}\{\widehat{D}_{k_1',k_2'}(\mathbf{X}) < \widehat{D}_{k_1,k_2}(\mathbf{X}) \mid \text{labels}(\mathbf{X}) = (k_1',k_2')\} \\ &- \pi_{k_1',k_2'}\text{Pr}_{\boldsymbol{\Theta}}\{D_{k_1',k_2'}(\mathbf{X}) < D_{k_1,k_2}(\mathbf{X}) \mid \text{labels}(\mathbf{X}) = (k_1',k_2')\} \leq c_1 \max(s_1^*, s_2^*)\frac{\log(p)}{n}\end{aligned}$$

for n sufficiently large. This is proven in their Theorem 3.1. To see how their proof applies, take their $\widehat{\boldsymbol{\beta}}$ equal to our $\widehat{\boldsymbol{\beta}}_{k_1k_2} - \widehat{\boldsymbol{\beta}}_{k_1'k_2'}$, their $\widehat{\mathbf{v}}_1$ (resp. $\widehat{\pi}_1$) equal to our $\widehat{\boldsymbol{\mu}}_{k_1k_2}$ (resp. $\widehat{\pi}_{k_1k_2}$), their $\widehat{\mathbf{v}}_2$ (resp. $\widehat{\pi}_2$) equal to our $\widehat{\boldsymbol{\mu}}_{k_1'k_2'}$ (resp. $\widehat{\pi}_{k_1'k_2'}$). Define all population versions similarly. Letting

$$\|\mathbf{a}\|_{2,d} := \sup_{\mathbf{v}\in\Gamma(d)} \frac{|\mathbf{a}^\top\mathbf{v}|}{\|\mathbf{v}\|_2} \text{ where } \Gamma(d) = \{\mathbf{v} : \mathbf{v} \neq 0, \|\mathbf{v}_{\mathcal{S}^c}\|_1 \leq 2\|\mathbf{v}_{\mathcal{S}}\|_1 \text{ for some } \mathcal{S} \subset [p] \text{ with } |\mathcal{S}| = d\},$$

Min et al. (2023) show that when defining

$$\delta_n = \max(\|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \widehat{\boldsymbol{\beta}}_{k_1'k_2'} - \boldsymbol{\beta}_{k_1k_2} + \boldsymbol{\beta}_{k_1'k_2'}\|_2, \|\widehat{\boldsymbol{\mu}}_{k_1k_2} - \boldsymbol{\mu}_{k_1k_2}\|_{2,s^*}, \|\widehat{\boldsymbol{\mu}}_{k_1'k_2'} - \boldsymbol{\mu}_{k_1'k_2'}\|_{2,s^*})$$

with probability at least $1 - O(p^{-1})$, with $\boldsymbol{\beta} = \boldsymbol{\beta}_{k_1k_2} - \boldsymbol{\beta}_{k_1'k_2'}$,

$$\widehat{R}^{\text{err}}((k_1,k_2),(k_1',k_2')) \lesssim \delta_n^2 \exp(-\boldsymbol{\beta}^\top\Sigma\boldsymbol{\beta}/8)\sqrt{\boldsymbol{\beta}^\top\Sigma\boldsymbol{\beta}}.$$

For the latter two terms in δ_n, their Lemma A.7 applies, and for the first term, our Theorem 1 applies, so $\delta_n \lesssim \sqrt{s^*\log(p)/n}$ with probability at least $1 - O(p^{-1})$. Thus, because $\boldsymbol{\beta}^\top\Sigma\boldsymbol{\beta}$ is bounded above and below by **A3**, it follows that $\widehat{R}^{\text{err}}((k_1,k_2),(k_1',k_2')) \lesssim s^*\frac{\log(p)}{n}$ with probability at least $1 - O(p^{-1})$. Finally, since the strong misclassification rate sums $\widehat{R}^{\text{err}}$ over all possible combinations of response categories (which is a constant, since c_1 and c_2 are constant), this implies our result.

F.2 Proof of Proposition 2

Proof of Proposition 2. We begin by simplifying the the objective function in (14). Let $\boldsymbol{\beta}_{k_1k_2} = \sum_{g\in G}\boldsymbol{\nu}_{k_1k_2}^{(g)}$, and recall that the summation in the first part is $\boldsymbol{\beta}_{k_1k_2}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2}/2 - \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top\boldsymbol{\beta}_{k_1k_2}$. Thus, we have

$$\begin{aligned}\boldsymbol{\beta}_{k_1k_2}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} &= \sum_{i,m=1}^{p}\boldsymbol{\beta}_{[i,k_1k_2]}\boldsymbol{\beta}_{[m,k_1k_2]}\widehat{\sigma}_{im} \\ &= \boldsymbol{\beta}_{[j,k_1k_2]}^2\widehat{\sigma}_{jj} + 2\sum_{i\neq j}\boldsymbol{\beta}_{[i,k_1k_2]}\boldsymbol{\beta}_{[j,k_1k_2]}\widehat{\sigma}_{ij} + \sum_{i,m\neq j}\boldsymbol{\beta}_{[i,k_1k_2]}\boldsymbol{\beta}_{[m,k_1k_2]}\widehat{\sigma}_{im},\end{aligned}$$

and for the second part,

$$\widehat{\boldsymbol{\delta}}_{k_1k_2}^\top\boldsymbol{\beta}_{k_1k_2} = \widehat{\boldsymbol{\delta}}_{[j,k_1k_2]}\boldsymbol{\beta}_{[j,k_1k_2]} + \sum_{i\neq j}\widehat{\boldsymbol{\delta}}_{[i,k_1k_2]}\boldsymbol{\beta}_{[i,k_1k_2]}.$$

Combining the above two parts, we obtain

$$\begin{aligned}\frac{1}{2}\boldsymbol{\beta}_{k_1k_2}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2}-\widehat{\boldsymbol{\delta}}_{k_1k_2}^\top\boldsymbol{\beta}_{k_1k_2}=&\frac{1}{2}\boldsymbol{\beta}_{[j,k_1k_2]}^2\widehat{\sigma}_{jj}-(\widehat{\boldsymbol{\delta}}_{[j,k_1k_2]}-\sum_{i\neq j}\widehat{\sigma}_{ij}\boldsymbol{\beta}_{[i,k_1k_2]})\boldsymbol{\beta}_{[j,k_1k_2]}+C_1\\=&\frac{1}{2}\widehat{\sigma}_{jj}\left(\frac{\widehat{\boldsymbol{\delta}}_{[j,k_1k_2]}-\sum_{i\neq j}\widehat{\sigma}_{ij}\boldsymbol{\beta}_{[i,k_1k_2]}}{\widehat{\sigma}_{jj}}-\boldsymbol{\beta}_{[j,k_1k_2]}\right)^2+C_2,\end{aligned}$$

where C_1 and C_2 are terms that do not involve $\boldsymbol{\beta}_{[j,k_1k_2]}$ and thus do not involve $\boldsymbol{\nu}^{(g)}$. Therefore, when given $\boldsymbol{\nu}^{(\ell)}$, $\ell\neq g$ and $\ell\in G$, the optimization in (14) can be written as

$$\widehat{\boldsymbol{\nu}}^{(g)}=\underset{\boldsymbol{\nu}^{(g)}\in\mathcal{V}^{(g)}}{\operatorname{argmin}}\frac{1}{2}\sum_{k_1,k_2}\left\{\frac{\widehat{\boldsymbol{\delta}}_{[j,k_1k_2]}-\sum_{i\neq j}\widehat{\sigma}_{ij}\sum_{\ell\in G}\boldsymbol{\nu}_{[i,k_1k_2]}^{(\ell)}}{\widehat{\sigma}_{jj}}-\sum_{\ell\in G}\boldsymbol{\nu}_{[j,k_1k_2]}^{(\ell)}\right\}^2+\frac{\lambda_g}{\widehat{\sigma}_{jj}}\|\boldsymbol{\nu}_g^{(g)}\|_2. \tag{A.19}$$

Then, without loss of generality, suppose $g\in G$ is from row selection groups, then $k_2=k_2(g)$ is fixed and the vector we are trying to update here is essentially $\boldsymbol{\nu}_g^{(g)}=\boldsymbol{\nu}_{[j,:,k_2]}\in\mathbb{R}^{c_1}$. Next, we further remove the terms in the summation of (A.19) that do not involve g and obtain

$$\begin{aligned}&\sum_{k_1}\left\{\frac{\widehat{\boldsymbol{\delta}}_{[j,k_1k_2]}-\sum_{i\neq j}\widehat{\sigma}_{ij}\sum_{\ell\in G}\boldsymbol{\nu}_{[i,k_1k_2]}^{(\ell)}}{\widehat{\sigma}_{jj}}-\sum_{\ell\in G}\boldsymbol{\nu}_{[j,k_1k_2]}^{(\ell)}\right\}^2\\&=\sum_{k_1}\left\{\frac{\widehat{\boldsymbol{\delta}}_{[j,k_1k_2]}-\sum_{i}\widehat{\sigma}_{ij}\sum_{\ell\in G}\boldsymbol{\nu}_{[i,k_1k_2]}^{(\ell)}+\widehat{\sigma}_{jj}\sum_{\ell\in G}\boldsymbol{\nu}_{[j,k_1k_2]}^{(\ell)}}{\widehat{\sigma}_{jj}}-\sum_{\ell\in G}\boldsymbol{\nu}_{[j,k_1k_2]}^{(\ell)}\right\}^2\\&=\sum_{k_1}\left\{\frac{\widehat{\boldsymbol{\delta}}_{[j,k_1k_2]}-\boldsymbol{\beta}_{[:,k_1k_2]}^\top\widehat{\boldsymbol{\Sigma}}_{\cdot j}}{\widehat{\sigma}_{jj}}+\boldsymbol{\beta}_{[j,k_1k_2]}-\sum_{\ell\in G}\boldsymbol{\nu}_{[j,k_1k_2]}^{(\ell)}\right\}^2\\&=\left\|\frac{\widehat{\boldsymbol{\delta}}_g-\boldsymbol{\beta}_{[:,:,k_2]}^\top\widehat{\boldsymbol{\Sigma}}_{\cdot j}}{\widehat{\sigma}_{jj}}+\boldsymbol{\beta}_g-\sum_{\ell\in G}\boldsymbol{\nu}_g^{(\ell)}\right\|_2^2.\end{aligned}$$

Finally, we can see that $\boldsymbol{\nu}^{(g)}$ can be updated by solving

$$\widehat{\boldsymbol{\nu}}_g^{(g)}=\underset{\boldsymbol{\nu}_g^{(g)}}{\operatorname{argmin}}\left\{\frac{1}{2}\|\mathbf{Z}^{(g)}-\sum_{\ell\in G}\boldsymbol{\nu}_g^{(\ell)}\|_2^2+\frac{\lambda_g}{\widehat{\sigma}_{jj}}\|\boldsymbol{\nu}_g^{(g)}\|_2\right\},$$

where $\mathbf{Z}^{(g)}=(\widehat{\boldsymbol{\delta}}_g-\boldsymbol{\beta}_{[:,:,k_2]}^\top\widehat{\boldsymbol{\Sigma}}_{\cdot j})/\widehat{\sigma}_{jj}+\boldsymbol{\beta}_g$. ■

F.3 Proof of Proposition 3

The proof of Proposition 3 will use the following definitions and propositions from (Tseng, 2001). These results concern a generic objective function

$$f(\mathbf{x}_1,\ldots,\mathbf{x}_N)=f_0(\mathbf{x}_1,\ldots,\mathbf{x}_N)+\sum_{k=1}^N f_k(\mathbf{x}_k),$$

where $f_0:\mathbb{R}^{n_1+\cdots+n_N}\mapsto\mathbb{R}\cup\{\infty\}$ and $f_k:\mathbb{R}^{n_k}\mapsto\mathbb{R}\cup\{\infty\}$.

Definition A.1 *(Gâteaux-differentiable, Bertsekas, 1997). For a function $F : \mathbb{R}^p \mapsto \mathbb{R}$, its Gâteaux derivative is defined as*

$$F'(\mathbf{x};\mathbf{y}) = \lim_{\lambda \searrow 0} \frac{F(\mathbf{x}+\lambda\mathbf{y}) - F(\mathbf{x})}{\lambda}.$$

If $F'(\mathbf{x};\mathbf{y})$ exists for all $\mathbf{y}$ at $\mathbf{x}$, then F is Gâteaux-differentiable at $\mathbf{x}$.

Definition A.2 *(Stationary points, Tseng, 2001). A point $\mathbf{z}$ is stationary for function h if $g(\mathbf{z};\mathbf{d}) \geq 0$ for any $\mathbf{d}$, where*

$$g(\mathbf{z};\mathbf{d}) = \liminf_{\lambda \searrow 0} \frac{h(\mathbf{z}+\lambda\mathbf{d}) - h(\mathbf{z})}{\lambda}$$

Proposition A.3 *(Regular function, Tseng, 2001). A function f is regular at $\mathbf{z}$ if $f'(\mathbf{z};\mathbf{d}) \geq 0$ for any $\mathbf{d} = (d_1, \ldots, d_N)$ such that $f'(\mathbf{z}; (0, \ldots, d_k, \ldots, 0)) \geq 0$, $k = 1, \ldots, N$.*

Proposition A.4 *(Tseng, 2001) If f_0 is Gâteaux-differentiable, f is regular at each $\mathbf{z}$.*

Proposition A.5 *(Tseng, 2001) Assume that the level set $\mathbf{X}^0 = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$ is compact, where $\mathbf{x}^0$ is the initial value of the algorithm, and that f is continuous on $\mathbf{X}^0$. Further assume that $f(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is pseudoconvex in $(\mathbf{x}_k, \mathbf{x}_i)$ for all $i, k \in \{1, \ldots, N\}$, and if f is regular at every $\mathbf{x}$, then the solution generated by the cyclic coordinate descent method converges to a stationary point of f.*

To prove Proposition 3, we first prove the following Lemma A.6.

Lemma A.6 *The set $\mathcal{B}^0 = \{\boldsymbol{\beta} : L(\boldsymbol{\beta}) \leq L(\boldsymbol{\beta}^0)\}$ is compact for any $\boldsymbol{\beta}^0$, where*

$$L(\boldsymbol{\beta}) = \sum_{k_1=1}^{c_1} \sum_{k_2=1}^{c_2} \left\{ \frac{1}{2} \boldsymbol{\beta}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \widehat{\boldsymbol{\beta}}_{k_1k_2} \right\} + \mathcal{P}_{\mathcal{V},\lambda}(\boldsymbol{\beta}).$$

Proof of Lemma A.6. Since $L(\boldsymbol{\beta})$ is continuous, $\mathcal{B}^0$ must be closed. It suffices to prove that $\mathcal{B}^0$ is bounded. When $\widehat{\boldsymbol{\Sigma}}$ is invertible, it is straightforward to see that, for any $\boldsymbol{\beta}_k \in \mathbb{R}^p$,

$$\boldsymbol{\beta}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{k_1k_2} - 2\widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \boldsymbol{\beta}_{k_1k_2} \geq -\widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\delta}}_{k_1k_2}.$$

It follows that, for any $\boldsymbol{\beta} \in \mathcal{B}^0$,

$$-\sum_{k_1=1}^{c_1} \sum_{k_2=1}^{c_2} \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\delta}}_{k_1k_2} + \mathcal{P}_{\mathcal{V},\lambda}(\boldsymbol{\beta}) \leq L(\boldsymbol{\beta}) \leq L(\boldsymbol{\beta}^0),$$

which implies that for any $\boldsymbol{\beta} \in \mathcal{B}^0$, we have

$$\mathcal{P}_{\mathcal{V},\lambda}(\boldsymbol{\beta}) \leq \sum_{k_1=1}^{c_1} \sum_{k_2=1}^{c_2} \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\delta}}_{k_1k_2} + L(\boldsymbol{\beta}^0).$$

Therefore, $\mathcal{B}^0$ is bounded and hence compact. When $\mathbf{\Sigma}$ is positive semidefinite, we further assume $\|\boldsymbol{\beta}^{(t)}\|_2 \leq C$ for some constant $C > 0$ within each block update in Algorithm 1. This directly guarantees $\mathcal{B}^0$ is compact. ■

Finally, we prove the result stated in Proposition 3 in what follows.

Proof of Proposition 3. Let $f_0(\boldsymbol{\beta}) = \sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\left\{\frac{1}{2}\boldsymbol{\beta}_{k_1k_2}^\top \widehat{\mathbf{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \widehat{\boldsymbol{\beta}}_{k_1k_2}\right\}$, $f_1(\boldsymbol{\beta}) = \mathcal{P}_{\mathcal{V},\lambda}(\boldsymbol{\beta})$. Then we have $L(\boldsymbol{\beta}) = f_0(\boldsymbol{\beta}) + f_1(\boldsymbol{\beta})$. Because $f_0(\boldsymbol{\beta})$ is differentiable and convex, and f_1 is convex, by Proposition A.4, we have $L(\boldsymbol{\beta})$ is regular at each $\boldsymbol{\beta}$. Then if $\widehat{\mathbf{\Sigma}}$ is positive definite, by Proposition A.6, $\mathbf{X}^0 = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$ is compact. Further note that $L(\boldsymbol{\beta})$ is both continuous and convex, thus by Proposition A.5, the blockwise coordinate descent algorithm summarized in Algorithm 1 converges to a stationary point of $L(\boldsymbol{\beta})$. Finally, because $L(\boldsymbol{\beta})$ is strict convex, the stationary point is the global minimizer. ■

F.4 Proof of Proposition 5

The proof of Proposition 5 relies on the results from (Deng and Yin, 2016). Consider a generic constrainted convex optimization problem with separable objective function

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{y}} \ f(\mathbf{x}) + g(\mathbf{y}) \\ &\text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b}, \end{aligned} \tag{A.20}$$

where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ are unknown variables, $\mathbf{A} \in \mathbb{R}^{p\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times m}$ are given matrices, $\mathbf{b} \in \mathbb{R}^p$, and $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \mapsto \mathbb{R} \cup \{+\infty\}$ are closed proper convex functions. The augmented Lagrangian function of (A.20) is

$$L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = f(\mathbf{x}) + g(\mathbf{y}) - \boldsymbol{\lambda}^\top(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{b}) + \frac{\beta}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{b}\|_2^2, \tag{A.21}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^p$ is the Lagrangian multiplier vector and $\beta > 0$ is a penalty parameter. Let $\mathbf{u}^* = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ be the point satisfying the KKT condition of (A.20), and let $\mathbf{u}^{(t)} = (\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \boldsymbol{\lambda}^{(t)})$ denote the iterate from ADMM algorithm after t iterations. We will use the following proposition from (Deng and Yin, 2016).

Proposition A.7 *(Theorem 2.3 (Deng and Yin, 2016)) Assume the KKT point $\mathbf{u}^* = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ exists and the functions f and g are convex. Further assume that $\{\mathbf{u}^{(t)}\}_{t\geq 0}$ of the ADMM algorithm is bounded. We have (i) $\boldsymbol{\lambda}^{(t)} \to \boldsymbol{\lambda}^*$; (ii) $\mathbf{A}\mathbf{x}^{(t)} \to \mathbf{A}\mathbf{x}^*$; (iii) $\mathbf{B}\mathbf{y}^{(t)} \to \mathbf{B}\mathbf{y}^*$.*

Proof of Proposition 5. We first write the optimization problem (13) in the form of (A.20) as the following:

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \ f(\boldsymbol{\beta}) + g(\boldsymbol{\theta}) \\ &\text{s.t. } \mathbf{A}\text{vec}(\boldsymbol{\beta}) - \boldsymbol{\theta} = 0, \end{aligned} \tag{A.22}$$

where $f(\boldsymbol{\beta}) = \sum_{k_1=1}^{c_1}\sum_{k_2=1}^{c_2}\left\{\frac{1}{2}\boldsymbol{\beta}_{k_1k_2}^\top \widehat{\mathbf{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \boldsymbol{\beta}_{k_1k_2}\right\}$ and $g(\boldsymbol{\theta}) = \lambda_1\sum_{k_2=1}^{c_2}\|\boldsymbol{\theta}^r_{\cdot k_2}\|_{2,1} + \lambda_2\sum_{k_1=1}^{c_1}\|\boldsymbol{\theta}^c_{k_1\cdot}\|_{2,1}$. The functions f and g are proper closed convex functions. Let $\widehat{\boldsymbol{w}}$ be the solution of the scaled Lagrangian multiplier of (18). Since (13) is a convex optimization satisfying Slater's condition, the solution $(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ is the KKT point of (13). Moreover, by Remark 3 in Deng and Yin (2016), because $\mathbf{A}$ has full column rank, the iterates

$(\boldsymbol{w}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}(t))$ are bounded. Finally, by Proposition A.7, we have $\boldsymbol{\beta}^{(t)} \to \widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta}^{(t)} \to \widehat{\boldsymbol{\theta}}$. Which gives $\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{\beta}}\|_F \to 0$, $\|(\boldsymbol{\theta}^r)^{(t)} - \widehat{\boldsymbol{\theta}}^r\|_F \to 0$ and $\|(\boldsymbol{\theta}^c)^{(t)} - \widehat{\boldsymbol{\theta}}^c\|_F \to 0$. ■

Appendix G. Proof of Theorem 1

G.1 Preliminaries

Our proof of Theorem 1 follows similar arguments as the proof of Theorem 3.1 (which actually verifies the first part of Theorem 3.3) from Min et al. (2023). For completeness, we restate the assumptions and define a number of important quantities that will be used throughout our proof.

- **A1. [Covariance]** There exists constant v_φ such that

$$0 < v_\varphi^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}) \leq \varphi_{\max}(\boldsymbol{\Sigma}) \leq v_\varphi < \infty.$$

- **A2. [Marginal probabilities]** There exists a constant $v_\pi > 0$ such that $\pi_{k_1k_2} \geq v_\pi > 0$ for all (k_1, k_2), $k_1 \in [c_1], k_2 \in [c_2]$.

- **A3. [Signal strength]** There exists a constant v_κ such that

$$0 < v_\kappa \leq (\boldsymbol{\mu}_{k_1k_2} - \boldsymbol{\mu}_{11})^\top \boldsymbol{\Sigma}^{-1} (\mu_{k_1k_2} - \mu_{11}) \leq 3v_\kappa < \infty$$

  for all $k_1 \in [c_1], k_2 \in [c_2]$ with $(k_1, k_2) \neq (1, 1)$.

For a matrix $\mathbf{A}$, let $\|\mathbf{A}\|_{2,1} = \sum_j \|\mathbf{A}_{[j,:]}\|_2$ where $\|\mathbf{a}\|_2$ is the Euclidean norm of the vector $\mathbf{a}$. Let $\|\mathbf{A}\|_\infty = \max_{i,j} |\mathbf{A}_{[i,j]}|$ and let $\|\mathbf{a}\|_1 = \sum_i |\mathbf{a}_i|$. For any positive integer n, let $[n] = \{1, 2, \ldots, n\}$. Recall that $\widehat{\boldsymbol{\delta}}_{k_1k_2} = \widehat{\boldsymbol{\mu}}_{k_1k_2} - \widehat{\boldsymbol{\mu}}_{11}$. Finally, let $\mathcal{T} = [c_1] \times [c_2] \setminus \{1, 1\}$, so that $\mathcal{T}$ has cardinality $c_1c_2 - 1$.

Recall that the set of important predictors for each response are defined as

$$\mathcal{S}_{k_1\cdot} =: \mathcal{S}_{k_1}^{(1)} = \{j : \boldsymbol{\beta}_{[j,k_1,k]} \neq 0 \text{ for any } k \in [c_2]\}, \quad k_1 \in [c_1]$$

and

$$\mathcal{S}_{\cdot k_2} =: \mathcal{S}_{k_2}^{(2)} = \{j : \boldsymbol{\beta}_{[j,k,k_2]} \neq 0 \text{ for any } k \in [c_1]\}, \quad k_2 \in [c_2].$$

Define $s_{\ell,k} := \text{card}(\mathcal{S}_k^{(\ell)})$ for $k \in [c_\ell]$, and define $s_\ell^* = \max_{k \in [c_\ell]} s_{\ell,k}$ for $\ell \in [2]$.

G.2 Proof of Theorem 1

To prove Theorem 1, we will rely on the following lemma. This and all subsequent lemmas are proven in a later section.

Lemma A.8 *Assume data are generated from the linear discriminant analysis model, and define $\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}}$ as the maximum likelihood estimators of $\boldsymbol{\delta}, \boldsymbol{\Sigma}$. Define $E_b(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$ as the event that*

(i) $\|\widehat{\boldsymbol{\delta}}_{k_1k_2} - \boldsymbol{\delta}_{k_1k_2}\|_\infty \leq b_0 \sqrt{\frac{\log(p)}{n}}$

(ii) $\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_{k_1k_2}\|_\infty \leq b_1 \sqrt{\frac{\log(p)}{n}}$

(iii) $\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \leq b_2 \sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})$

for positive finite constants b_0, b_1, *and* b_2. *If* $\lambda_1 = M\phi\{c_2 \log(p)/n\}^{1/2}$ *and* $\lambda_2 = M(1-\phi)\{c_1 \log(p)/n\}^{1/2}$ *for fixed constant* $\phi \in [0,1]$ *and* M *sufficiently large, then there exists a constant* $b_3 \in (0,\infty)$ *such that*

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \leq b_3 \left(\phi^2 \max_{k\in[c_1]} s_{1,k} + (1-\phi)^2 \max_{k\in[c_2]} s_{2,k} \right) \frac{\log(p)}{n},$$

on event $E_b(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$, *where* b_3 *depends only on constants* (b_0, b_1, b_2) *and* (c_1, c_2), *the number of response categories.*

Therefore, to prove Theorem 1 as stated, we need only establish conditions under which $E_b(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$ occurs with probability at least $1 - O(p^{-1})$. For this, we have the following two lemmas. The first is taken directly from Min et al. (2023) (setting their $M = 1$ and their $K = c_1c_2$).

Lemma A.9 *(Lemmas A.2 and A.3, Min et al., 2023) Under* ***A1*** *and* ***A2****, with probability at least* $1 - O(p^{-1})$, *there exists constants* $b_0 > 0$ *and* $b_0' > 0$ *such that*

$$\|\widehat{\boldsymbol{\delta}}_{k_1k_2} - \boldsymbol{\delta}_{k_1k_2}\|_\infty \leq b_0 \sqrt{\frac{\log(p)}{n}}$$

and

$$\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_{k_1k_2}\|_\infty \leq b_0' \|\boldsymbol{\beta}_{k_1k_2}\|_2 \sqrt{\frac{\log(p)}{n}}.$$

If ***A3*** *also holds, then* $\|\boldsymbol{\beta}_{k_1k_2}\|_2$ *is bounded, so the latter implies that there exists a* $b_1 > 0$ *such that*

$$\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_{k_1k_2}\|_\infty \leq b_1 \sqrt{\frac{\log(p)}{n}}.$$

Finally, we need only deal with (iii) from Lemma A.8.

Lemma A.10 *Under the conditions of Lemma A.8, if* $\max(s_1^*, s_2^*)\log(p)/n \to 0$, *then there exists a constant* $b_2 \in (0,\infty)$ *such that*

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \leq b_2 \sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})$$

with probability at least $1 - O(p^{-1})$ *for* n *sufficiently large.*

The proof of Theorem 1 is thus an immediate application of Lemma A.8, the conditions of which are implied by Lemmas A.9 and A.10. In the next section, will prove Lemma A.8 and A.10.

G.3 Proofs of Lemmas

We begin this section with the proof of Lemma A.8. In all subsequent proofs, we use b_l and d_l ($l \in \mathbb{N}$) to denote generic positive constants that may differ from place to place.

Proof of Lemma A.8. Let $\lambda_1 = M\phi\{c_2 \log(p)/n\}^{1/2}$ and $\lambda_2 = M(1-\phi)\{c_1 \log(p)/n\}^{1/2}$. Suppose that (i), (ii), and (iii) of $E_b(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$ hold. By definition of $\widehat{\boldsymbol{\beta}}$, the global minimizer of our estimation criterion,

$$2\lambda_1 \sum_{k_1=1}^{c_1} (\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}) + 2\lambda_2 \sum_{k_2=1}^{c_2} (\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1})$$
$$\leq \sum_{(k_1,k_2)\in\mathcal{T}} \{-(\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}) - 2(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2})^\top (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})\}.$$

Thus, rearranging terms,

$$\sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})$$
$$\leq 2 \sum_{(k_1,k_2)\in\mathcal{T}} |(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2})^\top (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})| + 2\lambda_1 \sum_{k_1=1}^{c_1} |\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}|$$
$$+ 2\lambda_2 \sum_{k_2=1}^{c_2} |\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}|$$
$$\leq 2 \sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 + 2\lambda_1 \sum_{k_1=1}^{c_1} |\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}|$$
$$+ 2\lambda_2 \sum_{k_2=1}^{c_2} |\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}|$$

by Hölder's inequality. On $E_b(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$, there exists a constant $d_0 \in (0,\infty)$ such that $\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty \leq \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_{k_1k_2}\|_\infty + \|\boldsymbol{\delta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty \leq d_0\sqrt{\frac{\log(p)}{n}}$, which implies

$$\sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})$$
$$\leq 2d_0\sqrt{\frac{\log(p)}{n}} \sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 + 2M\phi\sqrt{\frac{c_2\log(p)}{n}} \sum_{k_1=1}^{c_1} |\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}|$$
$$+ 2M(1-\phi)\sqrt{\frac{c_1\log(p)}{n}} \sum_{k_2=1}^{c_2} |\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}|$$
$$\leq d_1\sqrt{\frac{\log(p)}{n}} \left(\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 + \phi\sqrt{c_2} \sum_{k_1=1}^{c_1} \|\widehat{\boldsymbol{\beta}}_{k_1\cdot} - \boldsymbol{\beta}_{k_1\cdot}\|_{2,1} \right. \tag{A.23}$$
$$\left. + (1-\phi)\sqrt{c_1} \sum_{k_2=1}^{c_2} \|\widehat{\boldsymbol{\beta}}_{\cdot k_2} - \boldsymbol{\beta}_{\cdot k_2}\|_{2,1} \right)$$

$$= d_1\sqrt{\frac{\log(p)}{n}}\Big(\underbrace{\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\Delta}}_{k_1k_2}\|_1}_{T_1} + \underbrace{\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_{2,1} + (1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_{2,1}}_{T_2}\Big)$$

for constant $d_1 \in (0,\infty)$ and $\widehat{\boldsymbol{\Delta}} := \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, $\widehat{\boldsymbol{\Delta}}_{k_1k_2} := \widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}$, $\widehat{\boldsymbol{\Delta}}_{k_1\cdot} := \widehat{\boldsymbol{\beta}}_{k_1\cdot} - \boldsymbol{\beta}_{k_1\cdot}$, and $\widehat{\boldsymbol{\Delta}}_{\cdot k_2} := \widehat{\boldsymbol{\beta}}_{\cdot k_2} - \boldsymbol{\beta}_{\cdot k_2}$. Note that (A.23) follows in part from $|\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}| = |\|\widehat{\boldsymbol{\beta}}_{k_1\cdot} - \boldsymbol{\beta}_{k_1\cdot} + \boldsymbol{\beta}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}| \le \|\widehat{\boldsymbol{\beta}}_{k_1\cdot} - \boldsymbol{\beta}_{k_1\cdot}\|_{2,1}$, and similarly for the final term. Next, we bound T_1 and T_2. Starting with T_2,

$$\begin{aligned}
T_2 &= \phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_{2,1} + (1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_{2,1} \\
&= \phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\left(\sum_{j\in\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + \sum_{j\in[p]\setminus\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2\right) \\
&\qquad + (1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\left(\sum_{j\in\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 + \sum_{j\in[p]\setminus\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2\right)
\end{aligned}$$

so that by (A.28), the above implies

$$\begin{aligned}
T_2 &\le 3\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\sum_{j\in\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + 3(1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\sum_{j\in\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 \\
&\le 3\sum_{k_1=1}^{c_1}\phi\sqrt{c_2 s_{1,k_1}}\|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_F + 3\sum_{k_2=1}^{c_2}(1-\phi)\sqrt{c_1 s_{2,k_2}}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_F \\
&\le 3\phi\sqrt{c_2 s_1^*}\sum_{k_1=1}^{c_1}\|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_F + 3(1-\phi)\sqrt{c_1 s_2^*}\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_F \\
&\le 3\left(\phi\sqrt{c_1c_2 s_1^*} + (1-\phi)\sqrt{c_1c_2 s_2^*}\right)\|\widehat{\boldsymbol{\Delta}}\|_F,
\end{aligned}$$

where the final inequality follows from Cauchy-Schwarz (i.e., $\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_F \le \sqrt{c_2}\|\widehat{\boldsymbol{\Delta}}\|_F$). Following the same arguments,

$$\begin{aligned}
T_1 = \sum_{(k_1,k_2)\in\mathcal{T}}\|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 &= \phi\sum_{(k_1,k_2)\in\mathcal{T}}\|\widehat{\boldsymbol{\Delta}}_{k_1k_2}\|_1 + (1-\phi)\sum_{(k_1,k_2)\in\mathcal{T}}\|\widehat{\boldsymbol{\Delta}}_{k_1k_2}\|_1 \\
&\le \phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_{2,1} + (1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_{2,1} \\
&\le 3\left(\phi\sqrt{c_1c_2 s_1^*} + (1-\phi)\sqrt{c_1c_2 s_2^*}\right)\|\widehat{\boldsymbol{\Delta}}\|_F.
\end{aligned}$$

Hence, we have shown that there exists a constant $d_2 \in (0, \infty)$ such that

$$\sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}) \tag{A.24}$$
$$\leq d_1 \sqrt{\frac{\log(p)}{n}} (T_1 + T_2) \leq d_2 \sqrt{\frac{\log(p)}{n}} \left(\phi \sqrt{c_1 c_2 s_1^*} + (1-\phi)\sqrt{c_1 c_2 s_2^*} \right) \|\widehat{\boldsymbol{\Delta}}\|_F.$$

Recall that on event $E_b(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$,

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \leq b_2 \sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}), \tag{A.25}$$

so that (A.24) and (A.25) together imply that there exists a constant $d_3 \in (0, \infty)$ such that

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \leq d_3 \sqrt{\frac{\log(p)}{n}} \left(\sqrt{\phi^2 c_1 c_2 s_1^*} + \sqrt{(1-\phi)^2 c_1 c_2 s_2^*} \right) \|\widehat{\boldsymbol{\Delta}}\|_F$$

which finally yields

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2 \leq 2 d_3 c_1 c_2 \left\{ \phi^2 s_1^* + (1-\phi)^2 s_2^* \right\} \frac{\log(p)}{n}.$$

The result as written follows from the fact that c_1 and c_2 are fixed. ■

We next prove Lemma A.10. For a matrix $\mathbf{A}$ and $s \in [p]$, define

$$\phi_{\min}^{\mathbf{A}}(s) = \inf_{\|\mathbf{u}\|_0 \leq s, \mathbf{u}\neq 0} \frac{\mathbf{u}^\top \mathbf{A} \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}, \quad \phi_{\max}^{\mathbf{A}}(s) = \sup_{\|\mathbf{u}\|_0 \leq s, \mathbf{u}\neq 0} \frac{\mathbf{u}^\top \mathbf{A} \mathbf{u}}{\mathbf{u}^\top \mathbf{u}},$$

where $\|\mathbf{u}\|_0 = \sum_i \mathbf{1}(\mathbf{u}_i \neq 0)$. The following lemma establishes conditions under which Lemma A.10 holds.

Lemma A.11 *Let $\widehat{\boldsymbol{\Delta}}_{[j,k_1,k_2]} = \widehat{\boldsymbol{\beta}}_{[j,k_1,k_2]} - \boldsymbol{\beta}_{[j,k_1,k_2]}$ for each $(j, k_1, k_2) \in [p] \times [c_1] \times [c_2]$. If both*

(iv) $\phi\sqrt{c_2} \sum_{k_1=1}^{c_1} \sum_{j\in[p]\setminus\mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + (1-\phi)\sqrt{c_1} \sum_{k_2=1}^{c_2} \sum_{j\in[p]\setminus\mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2$
$\leq 2\phi\sqrt{c_2} \sum_{k_1=1}^{c_1} \sum_{j\in\mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + 2(1-\phi)\sqrt{c_1} \sum_{k_2=1}^{c_2} \sum_{j\in\mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2,$

(v) there exists constants $b_3 > 0$ and $b_4 > 0$ such that

$$b_3^{-1} - b_4 \epsilon_{s^*} \leq \phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s^*) \leq \phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(s^*) \leq b_3 + b_4 \epsilon_{s^*}$$

for some $\epsilon_{s^*} \to 0$,

then, there exists a constant $b_2 > 0$ such that $\sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}) \geq b_2 \sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_2^2$ with probability at least $1 - O(p^{-1})$.

Let us now prove Lemma A.11, then we will certify (iv) and (v).

Proof of Lemma A.11. Recall that $\widehat{\boldsymbol{\Delta}}_{k_1k_2} = \widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}$, $\widehat{\boldsymbol{\Delta}}_{[j,k_1,k_2]} = \widehat{\boldsymbol{\beta}}_{[j,k_1,k_2]} - \boldsymbol{\beta}_{[j,k_1,k_2]}$, and

$$\sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}) \\ = \phi \sum_{k_1=1}^{c_1} \mathrm{tr}(\widehat{\boldsymbol{\Delta}}_{k_1\cdot} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Delta}}_{k_1\cdot}^\top) + (1-\phi) \sum_{k_2=1}^{c_2} \mathrm{tr}(\widehat{\boldsymbol{\Delta}}_{\cdot k_2} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Delta}}_{\cdot k_2}^\top)$$

We will deal with the first term (scaled by ϕ) first. Let $J_{0,1,k_1} = \mathcal{S}_{k_1}^{(1)}$. Let $J_{1,1,k_1} \subset [p] \setminus \mathcal{S}_{k_1}^{(1)}$ be an index set denoting the t largest Euclidean norms $\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2$ over all $j \in [p] \setminus \mathcal{S}_{k_1}^{(1)}$. Then, partition $[p] \setminus \mathcal{S}_{k_1}^{(1)}$ into $J_{2,1,k_1}, \ldots, J_{L,1,k_1}$ where each set $J_{l,1,k_1}$ corresponds to the t largest Euclidean norms of $\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2$ over all $j \notin \cup_{k=0}^{l-1} J_{k,1,k_1}$. In this way, $\|\widehat{\boldsymbol{\Delta}}_{[\ell,k_1,:]}\|_2 \geq \|\widehat{\boldsymbol{\Delta}}_{[\ell',k_1,:]}\|_2$ for all $\ell \in J_{v,1,k_1}$ and $\ell' \in J_{v+m,1,k_1}$ for $v \geq 1$ and $m \geq 1$.

For ease of display, we momentarily omit the second and third subscripts on the $J_{v,1,k_1}$: take these to be $1, k_1$. Similarly, define $\boldsymbol{\theta} := \widehat{\boldsymbol{\Delta}}_{k_1\cdot} \in \mathbb{R}^{p\times c_2}$. By the triangle inequality, we have

$$\sqrt{\mathrm{tr}(\widehat{\boldsymbol{\Delta}}_{k_1\cdot} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Delta}}_{k_1\cdot}^\top)} \geq \|\widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\theta}_{J_0\cup J_1}\|_F - \sum_{l\geq 2} \|\widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\theta}_{J_l}\|_F,$$

where $\boldsymbol{\theta}_{J_0\cup J_1} \in \mathbb{R}^{p\times c_1}$ is the matrix which is equal to $\boldsymbol{\theta}$ in rows indexed by $J_0 \cup J_1$ and has zeros elsewhere. Define $\boldsymbol{\theta}_{J_l}$ similarly for $l \geq 2$. Then, by Lemma A.13, since $J_0 \cup J_1$ has cardinality $s_{1,k_1} + t$,

$$\|\widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\theta}_{J_0\cup J_1}\|_F = \sqrt{\sum_{\ell=1}^{c_1} \mathrm{tr}(\mathbf{v}_\ell^\top \widehat{\boldsymbol{\Sigma}} \mathbf{v}_\ell)} \geq \sqrt{\varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s_{1,k_1}+t) \sum_{\ell=1}^{c_1} \|\mathbf{v}_\ell\|_2^2} \\ \geq \sqrt{\varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s_{1,k_1}+t)} \|\boldsymbol{\theta}_{J_0\cup J_1}\|_F,$$

where $\mathbf{v}_\ell \in \mathbb{R}^p$ is the ℓth column of $\boldsymbol{\theta}_{J_0\cup J_1}$. Similarly, $\|\widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\theta}_{J_l}\|_F \leq \sqrt{\varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)} \|\boldsymbol{\theta}_{J_l}\|_F$ so that together,

$$\sqrt{\mathrm{tr}(\widehat{\boldsymbol{\Delta}}_{k_1\cdot} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Delta}}_{k_1\cdot}^\top)} \geq \sqrt{\varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s_{1,k_1}+t)} \|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{0,1,k_1}\cup J_{1,1,k_1},:]}\|_F - \sqrt{\varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)} \sum_{l\geq 2} \|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{l,1,k_1},:]}\|_F$$

for each k_1, and by identical arguments, for each k_2,

$$\sqrt{\mathrm{tr}(\widehat{\boldsymbol{\Delta}}_{\cdot k_2} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Delta}}_{\cdot k_2}^\top)} \geq \sqrt{\varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s_{2,k_2}+t)} \|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{0,2,k_2}\cup J_{1,2,k_2},:]}\|_F - \sqrt{\varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)} \sum_{l\geq 2} \|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{l,2,k_2},:]}\|_F.$$

Thus far, we have shown

$$
\begin{aligned}
&\sqrt{\sum_{(k_1,k_2)\in\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})} \\
&\quad= \phi\left\{\sum_{k_1=1}^{c_1}\operatorname{tr}(\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Delta}}_{k_1\cdot}^\top)\right\}^{1/2} + (1-\phi)\left\{\sum_{k_2=1}^{c_2}\operatorname{tr}(\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Delta}}_{\cdot k_2}^\top)\right\}^{1/2} \\
&\quad\geq \phi\sum_{k_1=1}^{c_1}\left\{\frac{1}{c_1}\operatorname{tr}(\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Delta}}_{k_1\cdot}^\top)\right\}^{1/2} + (1-\phi)\sum_{k_2=1}^{c_2}\left\{\frac{1}{c_2}\operatorname{tr}(\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Delta}}_{\cdot k_2}^\top)\right\}^{1/2} \\
&\quad\geq \frac{\phi}{\sqrt{c_1}}\sum_{k_1=1}^{c_1}\left\{\sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s_{1,k_1}+t)}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{0,1,k_1}\cup J_{1,1,k_1},:]}\|_F - \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(t)}\sum_{l\geq 2}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{l,1,k_1},:]}\|_F\right\} \\
&+ \frac{(1-\phi)}{\sqrt{c_2}}\sum_{k_2=1}^{c_2}\left\{\sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s_{2,k_2}+t)}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{0,2,k_2}\cup J_{1,2,k_2},:]}\|_F - \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(t)}\sum_{l\geq 2}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{l,2,k_2},:]}\|_F\right\}.
\end{aligned}
$$

Then, to deal with the terms summing over $\ell \geq 2$ in the previous inequality, notice

$$
\begin{aligned}
&\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\sum_{l\geq 2}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{l,1,k_1},:]}\|_F + (1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\sum_{l\geq 2}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{l,2,k_2},:]}\|_F \\
&\quad\leq \phi\sqrt{tc_2}\sum_{k_1=1}^{c_1}\sum_{l\geq 2}\max_{j\in J_{l,1,k_1}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[j,:]}\|_2 + (1-\phi)\sqrt{tc_1}\sum_{k_2=1}^{c_2}\sum_{l\geq 2}\max_{j\in J_{l,2,k_2}}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[j,:]}\|_2 \\
&\quad\leq \phi\sqrt{tc_2}\sum_{k_1=1}^{c_1}\sum_{l\geq 2}\frac{1}{t}\sum_{j\in J_{l-1,1,k_1}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[j,:]}\|_2 + (1-\phi)\sqrt{tc_1}\sum_{k_2=1}^{c_2}\sum_{l\geq 2}\frac{1}{t}\sum_{j\in J_{l-1,2,k_2}}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[j,:]}\|_2 \\
&\quad\leq \phi\sqrt{\frac{c_2}{t}}\sum_{k_1=1}^{c_1}\sum_{j\in[p]\setminus\mathcal{S}^{(1)}_{k_1}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[j,:]}\|_2 + (1-\phi)\sqrt{\frac{c_1}{t}}\sum_{k_2=1}^{c_2}\sum_{j\in[p]\setminus\mathcal{S}^{(2)}_{k_2}}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[j,:]}\|_2 \\
&\quad\leq 2\phi\sqrt{\frac{c_2}{t}}\sum_{k_1=1}^{c_1}\sum_{j\in\mathcal{S}^{(1)}_{k_1}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[j,:]}\|_2 + 2(1-\phi)\sqrt{\frac{c_1}{t}}\sum_{k_2=1}^{c_2}\sum_{j\in\mathcal{S}^{(2)}_{k_2}}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[j,:]}\|_2 \\
&\quad\leq 2\phi\sum_{k_1=1}^{c_1}\sqrt{\frac{c_2 s_{1,k_1}}{t}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[\mathcal{S}^{(1)}_{k_1},:]}\|_F + 2(1-\phi)\sum_{k_2=1}^{c_2}\sqrt{\frac{c_1 s_{2,k_2}}{t}}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[\mathcal{S}^{(2)}_{k_2},:]}\|_F
\end{aligned}
$$

where the fourth inequality follows from (iv), and the fifth from Cauchy-Schwarz. In the third inequality, we used that for $l \geq 2$, $\max_{j\in J_{l,1,k_1}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{j,:}\|_2 \leq \frac{1}{t}\sum_{j\in J_{l-1,1,k_1}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{j,:}\|_2$ as the $J_{l,1,k_1}$ sorted rows in descending order in terms of their Euclidean norms. We have

therefore shown

$$\sqrt{c_1c_2\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})}$$
$$\geq \phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\left(\sqrt{\varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s_{1,k_1}+t)}-2\sqrt{\frac{s_{1,k_1}}{t}\varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)}\right)\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{0,1,k_1}\cup J_{1,1,k_1},:]}\|_F$$
$$+(1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\left(\sqrt{\varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s_{2,k_2}+t)}-2\sqrt{\frac{s_{2,k_2}}{t}\varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)}\right)\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{0,2,k_2}\cup J_{1,2,k_2},:]}\|_F.$$

Let $t=d_0s^*$ for some constant $d_0\in(0,\infty)$ where $s^*=\max(s_1^*,s_2^*)$. When (v) holds, there exists positive constants b_3 and b_4 such that $b_3^{-1}-b_4\epsilon_s\leq\phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s)\leq\phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(s)\leq b_3+b_4\epsilon_s$ for any s such that $s\log(p)/n$ is bounded above. Thus, we have

$$\left(\sqrt{\phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s_{2,k_2}+t)}-2\sqrt{\frac{s_{2,k_2}}{t}\phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)}\right)\geq\sqrt{b_3^{-1}-b_4\epsilon_{s^*(1+d_0)}}-2\sqrt{\frac{b_3+b_4\epsilon_{s^*}}{d_0}}$$

By assumption $\epsilon_{s^*}=o(1)$, so by taking d_0 sufficiently large, there exists a constant $d_1\in(0,\infty)$

$$\sqrt{c_1c_2\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})}\tag{A.26}$$
$$\geq d_1\left\{\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{0,1,k_1}\cup J_{1,1,k_1},:]}\|_F+(1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{0,2,k_2}\cup J_{1,2,k_2},:]}\|_F\right\}$$

for n sufficiently large. Finally, to complete the proof, notice

$$\sqrt{\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})}$$
$$=\phi\Big\{\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})\Big\}^{1/2}$$
$$+(1-\phi)\Big\{\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})\Big\}^{1/2}$$
$$\leq\phi\sum_{k_1=1}^{c_1}\{\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1}\cup J_{0,1,k_1},:]}\|_F+\sum_{l\geq2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1},:]}\|_F\}$$
$$+(1-\phi)\sum_{k_2=1}^{c_2}\{\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2}\cup J_{0,2,k_2},:]}\|_F+\sum_{l\geq2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2},:]}\|_F\}$$
$$\leq\phi\sum_{k_1=1}^{c_1}\{\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1}\cup J_{0,1,k_1},:]}\|_F+\sqrt{c_2}\sum_{l\geq2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1},:]}\|_F\}$$
$$+(1-\phi)\sum_{k_2=1}^{c_2}\{\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2}\cup J_{0,2,k_2},:]}\|_F+\sqrt{c_1}\sum_{l\geq2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2},:]}\|_F\}$$

which implies

$$\sqrt{\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})}$$
$$\leq \phi\sum_{k_1=1}^{c_1}\{\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1}\cup J_{0,1,k_1},:]}\|_F+\sum_{l\geq 2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1},:]}\|_F\}$$
$$+(1-\phi)\sum_{k_2=1}^{c_2}\{\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2}\cup J_{0,2,k_2},:]}\|_F+\sum_{l\geq 2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2},:]}\|_F\}.$$

Recall, we just showed that

$$\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\sum_{l\geq 2}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[J_{l,1,k_1},:]}\|_F+(1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\sum_{l\geq 2}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[J_{l,2,k_2},:]}\|_F$$
$$\leq 2\phi\sum_{k_1=1}^{c_1}\sqrt{\frac{c_2 s_{1,k_1}}{t}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[\mathcal{S}_{k_1}^{(1)},:]}\|_F+2(1-\phi)\sum_{k_2=1}^{c_2}\sqrt{\frac{c_1 s_{2,k_2}}{t}}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[\mathcal{S}_{k_2}^{(2)},:]}\|_F$$
$$\leq 2\phi\sum_{k_1=1}^{c_1}\sqrt{\frac{c_2}{d_0}}\|[\widehat{\boldsymbol{\Delta}}_{k_1\cdot}]_{[\mathcal{S}_{k_1}^{(1)},:]}\|_F+2(1-\phi)\sum_{k_2=1}^{c_2}\sqrt{\frac{c_1}{d_0}}\|[\widehat{\boldsymbol{\Delta}}_{\cdot k_2}]_{[\mathcal{S}_{k_2}^{(2)},:]}\|_F$$

so that we have

$$\sqrt{\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})}$$
$$\leq \phi\left(1+2\sqrt{\frac{c_2}{d_0}}\right)\sum_{k_1=1}^{c_1}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1}\cup J_{0,1,k_1},:]}\|_F+(1-\phi)\left(1+2\sqrt{\frac{c_1}{d_0}}\right)\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2}\cup J_{0,2,k_2},:]}\|_F.$$

Finally, taking d_0 sufficiently large, this implies

$$\sqrt{\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})}$$
$$\leq \phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,1,k_1}\cup J_{0,1,k_1},:]}\|_F+(1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{[J_{1,2,k_2}\cup J_{0,2,k_2},:]}\|_F$$
$$\leq \frac{\sqrt{c_1c_2}}{d_1}\left\{\sum_{(k_1,k_2)\in\mathcal{T}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})^\top\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_{k_1k_2}-\boldsymbol{\beta}_{k_1k_2})\right\}^{1/2}$$

where the final inequality follows from (A.26), which completes the proof. ■

Therefore, it remains only to show that under the conditions of our theorem, (iv) and (v) hold. We start with (iv).

Lemma A.12 *Under conditions (i) and (ii) of Lemma A.8, we have*

$$2\lambda_1 \sum_{k_1=1}^{c_1} (\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}) + 2\lambda_2 \sum_{k_2=1}^{c_2} (\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}) \tag{A.27}$$
$$\leq \sum_{(k_1,k_2)\in\mathcal{T}} \{-(\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}) - 2(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2})^\top (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})\}$$

and

$$\phi\sqrt{c_2} \sum_{k_1=1}^{c_1} \sum_{j\in[p]\setminus\mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + (1-\phi)\sqrt{c_1} \sum_{k_2=1}^{c_2} \sum_{j\in[p]\setminus\mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 \tag{A.28}$$
$$\leq 2\phi\sqrt{c_2} \sum_{k_1=1}^{c_1} \sum_{j\in\mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + 2(1-\phi)\sqrt{c_1} \sum_{k_2=1}^{c_2} \sum_{j\in\mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2.$$

Proof of Lemma A.12. The first inequality, (A.27), follows immediately from the fact that $\widehat{\beta}$ is the global minimizer, i.e.,

$$\sum_{(k_1,k_2)\in\mathcal{T}} \left\{ \frac{1}{2} \widehat{\boldsymbol{\beta}}_{k_1k_2}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}^\top \widehat{\boldsymbol{\beta}}_{k_1k_2} \right\} + \lambda_1 \sum_{k_1=1}^{c_1} \|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} + \lambda_2 \sum_{k_2=1}^{c_2} \|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1}$$
$$\leq \sum_{(k_1,k_2)\in\mathcal{T}} \left\{ \frac{1}{2} \boldsymbol{\beta}_{k_1k_2}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_{k_1k_2} - \boldsymbol{\delta}_{k_1k_2}^\top \boldsymbol{\beta}_{k_1k_2} \right\} + \lambda_1 \sum_{k_1=1}^{c_1} \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1} + \lambda_2 \sum_{k_2=1}^{c_2} \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}$$

from which (A.27) follows from straightforward algebra.

For (A.28), we follow the same arguments as in the proof of Lemma A.4 from Min et al. (2023). In particular,

$$\begin{aligned}
&2\lambda_1 \sum_{k_1=1}^{c_1} (\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}) + 2\lambda_2 \sum_{k_2=1}^{c_2} (\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}) \\
&\quad\leq \sum_{(k_1,k_2)\in\mathcal{T}} \{-(\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})^\top \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}) - 2(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2})^\top (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})\} \\
&\quad\leq \sum_{(k_1,k_2)\in\mathcal{T}} |(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2})^\top (\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2})| \\
&\quad\leq \sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1
\end{aligned}$$

On (i) and (ii) of $E_b(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$, $\|\widehat{\boldsymbol{\delta}}_{k_1k_2} - \boldsymbol{\delta}_{k_1k_2}\|_\infty \leq b_0\sqrt{\frac{\log(p)}{n}}$ and $\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_{k_1k_2}\|_\infty \leq b_1\sqrt{\frac{\log(p)}{n}}$, so that there exists a constant $d_0 \in (0,\infty)$ such that

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty &= \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \boldsymbol{\Sigma}\boldsymbol{\beta}_{k_1k_2} + \boldsymbol{\Sigma}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty \\
&\leq \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_{k_1k_2}\|_\infty + \|\boldsymbol{\delta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty \leq d_0\sqrt{\frac{\log(p)}{n}}
\end{aligned}$$

where the inequality follows from the triangle inequality and fact that $\boldsymbol{\beta}_{k_1k_2} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_{k_1k_2}$. Next, considering the difference in penalties,

$$2\lambda_1 \sum_{k_1=1}^{c_1} (\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}) + 2\lambda_2 \sum_{k_2=1}^{c_2} (\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1})$$
$$\leq \sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{k_1k_2} - \widehat{\boldsymbol{\delta}}_{k_1k_2}\|_\infty \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 \leq d_0\sqrt{\frac{\log(p)}{n}} \sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1. \tag{A.29}$$

Then, notice that

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 \leq \sqrt{c_2} \sum_{k_1=1}^{c_1} \| \underbrace{\widehat{\boldsymbol{\beta}}_{k_1\cdot} - \boldsymbol{\beta}_{k_1\cdot}}_{=\widehat{\boldsymbol{\Delta}}_{k_1\cdot}} \|_{2,1}$$

and similarly,

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 \leq \sqrt{c_1} \sum_{k_2=1}^{c_2} \| \underbrace{\widehat{\boldsymbol{\beta}}_{\cdot k_2} - \boldsymbol{\beta}_{\cdot k_2}}_{=\widehat{\boldsymbol{\Delta}}_{\cdot k_2}} \|_{2,1}$$

so that for any $\phi \in [0,1]$,

$$\sum_{(k_1,k_2)\in\mathcal{T}} \|\widehat{\boldsymbol{\beta}}_{k_1k_2} - \boldsymbol{\beta}_{k_1k_2}\|_1 \leq \left\{ \phi\sqrt{c_2} \sum_{k_1=1}^{c_1} \|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_{2,1} + (1-\phi)\sqrt{c_1} \sum_{k_2=1}^{c_2} \|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_{2,1} \right\}.$$

Therefore, from (A.29) it follows that

$$2\lambda_1 \sum_{k_1=1}^{c_1} (\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}) + 2\lambda_2 \sum_{k_2=1}^{c_2} (\|\widehat{\boldsymbol{\beta}}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1})$$
$$\leq d_0\sqrt{\frac{\log(p)}{n}} \left\{ \phi\sqrt{c_2} \sum_{k_1=1}^{c_1} \|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_{2,1} + (1-\phi)\sqrt{c_1} \sum_{k_2=1}^{c_2} \|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_{2,1} \right\}. \tag{A.30}$$

Now, considering the difference

$$\begin{aligned}
\lambda_1 \sum_{k_1=1}^{c_1} (\|\widehat{\boldsymbol{\beta}}_{k_1\cdot}\|_{2,1} - \|\boldsymbol{\beta}_{k_1\cdot}\|_{2,1}) &= \lambda_1 \sum_{k_1=1}^{c_1} \sum_{j=1}^{p} (\|\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2 - \|\boldsymbol{\beta}_{[j,k_1,:]}\|_2) \\
&= \lambda_1 \sum_{k_1=1}^{c_1} \left\{ \sum_{j \in \mathcal{S}_{k_1}^{(1)}} (\|\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2 - \|\boldsymbol{\beta}_{[j,k_1,:]}\|_2) + \sum_{j \in [p] \setminus \mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2 \right\} \\
&= \lambda_1 \sum_{k_1=1}^{c_1} \left\{ \sum_{j \in \mathcal{S}_{k_1}^{(1)}} (\|\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2 - \|\boldsymbol{\beta}_{[j,k_1,:]} - \widehat{\boldsymbol{\beta}}_{[j,k_1,:]} + \widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2) + \sum_{j \in [p] \setminus \mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2 \right\} \\
&\geq \lambda_1 \sum_{k_1=1}^{c_1} \left\{ \sum_{j \in \mathcal{S}_{k_1}^{(1)}} (\|\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2 - \| \underbrace{\boldsymbol{\beta}_{[j,k_1,:]} - \widehat{\boldsymbol{\beta}}_{[j,k_1,:]}}_{=:-\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}} \|_2 - \|\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}\|_2) + \sum_{j \in [p] \setminus \mathcal{S}_{k_1}^{(1)}} \| \underbrace{\widehat{\boldsymbol{\beta}}_{[j,k_1,:]}}_{=:\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}} \|_2 \right\} \\
&= \lambda_1 \sum_{k_1=1}^{c_1} \left\{ \sum_{j \in [p] \setminus \mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 - \sum_{j \in \mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 \right\}
\end{aligned}$$

and by nearly identical arguments,

$$\lambda_2 \sum_{k_2=1}^{c_2} (\|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1} - \|\boldsymbol{\beta}_{\cdot k_2}\|_{2,1}) \geq \lambda_2 \sum_{k_2=1}^{c_2} \left\{ \sum_{j \in [p] \setminus \mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 - \sum_{j \in \mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 \right\}.$$

We have so far shown that

$$\begin{aligned}
&\lambda_1 \sum_{k_1=1}^{c_1} \left\{ \sum_{j \in [p] \setminus \mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 - \sum_{j \in \mathcal{S}_{k_1}^{(1)}} \|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 \right\} + \lambda_2 \sum_{k_2=1}^{c_2} \left\{ \sum_{j \in [p] \setminus \mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 - \sum_{j \in \mathcal{S}_{k_2}^{(2)}} \|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 \right\} \\
&\qquad \leq \frac{d_0}{2} \sqrt{\frac{\log(p)}{n}} \left\{ \phi \sqrt{c_2} \sum_{k_1=1}^{c_1} \|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_{2,1} + (1-\phi) \sqrt{c_1} \sum_{k_2=1}^{c_2} \|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_{2,1} \right\}
\end{aligned}$$

so that with $\lambda_1 = M\phi\sqrt{\frac{c_2\log(p)}{n}}$ and $\lambda_2 = M(1-\phi)\sqrt{\frac{c_1\log(p)}{n}}$ for $M \geq 3d_0/2$, the previous inequality implies

$$\begin{aligned}
&\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\left\{\sum_{j\in[p]\setminus\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 - \sum_{j\in\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2\right\} \\
&\qquad\qquad + (1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\left\{\sum_{j\in[p]\setminus\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 - \sum_{j\in\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2\right\} \\
&\leq \left\{\frac{\phi\sqrt{c_2}}{3}\sum_{k_1=1}^{c_1}\|\widehat{\boldsymbol{\Delta}}_{k_1\cdot}\|_{2,1} + \frac{(1-\phi)\sqrt{c_1}}{3}\sum_{k_2=1}^{c_2}\|\widehat{\boldsymbol{\Delta}}_{\cdot k_2}\|_{2,1}\right\} \\
&= \frac{\phi\sqrt{c_2}}{3}\sum_{k_1=1}^{c_1}\left(\sum_{j\in\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + \sum_{j\in[p]\setminus\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2\right) \\
&\qquad\qquad + \frac{(1-\phi)\sqrt{c_1}}{3}\sum_{k_2=1}^{c_2}\left(\sum_{j\in\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 + \sum_{j\in[p]\setminus\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2\right)
\end{aligned}$$

which finally implies

$$\begin{aligned}
&\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\sum_{j\in[p]\setminus\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + (1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\sum_{j\in[p]\setminus\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2 \\
&\qquad\leq 2\phi\sqrt{c_2}\sum_{k_1=1}^{c_1}\sum_{j\in\mathcal{S}_{k_1}^{(1)}}\|\widehat{\boldsymbol{\Delta}}_{[j,k_1,:]}\|_2 + 2(1-\phi)\sqrt{c_1}\sum_{k_2=1}^{c_2}\sum_{j\in\mathcal{S}_{k_2}^{(2)}}\|\widehat{\boldsymbol{\Delta}}_{[j,:,k_2]}\|_2. \qquad \blacksquare
\end{aligned}$$

Finally, for (v), we have the next lemma from Min et al. (2023). Note that this exactly their Lemma A.5 with their $M = 1$.

Lemma A.13 *(Lemma A.5, Min et al., 2023) Let $s \in [p]$ and suppose $s\log(p)/n < b_5$ for constant b_5. Suppose there exists a constant $b_6 \in (0,\infty)$ such that $0 < b_6^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}) \leq \varphi_{\max}(\boldsymbol{\Sigma}) \leq b_6 < \infty$. For sample estimator $\widehat{\boldsymbol{\Sigma}}$, if $\epsilon_s = \sqrt{s\log(p)/n}$, then*

$$b_6^{-1} - b_7\epsilon_s \leq \phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s) \leq \phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(s) + \leq b_6 + b_7\epsilon_s$$

with probability at least $1 - O(p^{-1})$ for some constant $b_7 \in (0,\infty)$.

Hence, when $s^*\log(p)/n \to 0$, (v) holds; whereas under (i) and (ii) of Lemma A.8, (iv) holds. Therefore, Lemma A.11 and A.13 imply Lemma A.10 holds under the conditions of our theorem.

References

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

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

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.

Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

David R Cox and Nancy Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66:889–916, 2016.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605, 2008.

Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771, 2012.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

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

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

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, volume 2. Springer, 2009.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

Binyan Jiang and Chenlei Leng. High dimensional discrimination analysis via a semiparametric model. *Statistics and Probability Letters*, 110:103–110, 2016.

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

Saskia le Cessie and JC Van Houwelingen. Logistic regression for correlated binary data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):95–108, 1994.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

Y Lin and Y Jeon. Discriminant analysis through a semiparametric model. *Biometrika*, 90 (2):379–392, 2003.

Bruce G Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239, 1988.

Chongliang Luo, Jian Liang, Gen Li, Fei Wang, Changshui Zhang, Dipak K Dey, and Kun Chen. Leveraging mixed and incomplete outcomes via reduced-rank modeling. *Journal of Multivariate Analysis*, 167:378–394, 2018.

Qing Mai and Hui Zou. Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188, 2015.

Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

Qing Mai, Yi Yang, and Hui Zou. Multiclass sparse discriminant analysis. *Statistica Sinica*, 29(1):97–111, 2019.

Qing Mai, Di He, and Hui Zou. Coordinatewise gaussianization: Theories and applications. *Journal of the American Statistical Association*, 118(544):2329–2343, 2023.

Peter McCullagh and John A Nelder. Generalized linear models, volume 37 of. *Monographs on Statistics and Applied Probability*, 1989.

Keqian Min, Qing Mai, and Junge Li. Optimality in high-dimensional tensor discriminant analysis. *Pattern Recognition*, 143:109803, 2023.

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

Aaron J Molstad and Xin Zhang. Conditional probability tensor decompositions for multivariate categorical response regression. *arXiv preprint arXiv:2206.10676*, 2022.

Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

Jing Ouyang and Gongjun Xu. Identifiability of latent class models with covariates. *psychometrika*, 87(4):1343–1360, 2022.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Cheong Hee Park and Moonhwi Lee. On applying linear discriminant analysis for multi-labeled problems. *Pattern Recognition Letters*, 29(7):878–887, 2008.

Seyoung Park, Eun Ryung Lee, and Hongyu Zhao. Low-rank regression models for multiple binary responses and their applications to cancer cell-line encyclopedia data. *Journal of the American Statistical Association*, 119(545):202–216, 2024.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer, 2009.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

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

Cristiano Varin. On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1–28, 2008.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

Hua Wang, Chris Ding, and Heng Huang. Multi-label linear discriminant analysis. In *European Conference on Computer Vision*, pages 126–139. Springer, 2010.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. CNN-RNN: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294. IEEE, 2016.

Wei Weng, Da-Han Wang, Chin-Ling Chen, Juan Wen, and Shun-Xiang Wu. Label specific features-based classifier chains for multi-label classification. *IEEE Access*, 8:51265–51275, 2020.

Daniela M Witten and Robert Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

Xiaohan Yan and Jacob Bien. Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531–560, 2017.

Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. *Advances in Neural Information Processing Systems*, 24, 2011.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.