Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

# Kernelized Discriminant Analysis for Joint Modeling of Multivariate Categorical Responses

Yisen Jin, Xin Zhang & Aaron J. Molstad

**To cite this article:** Yisen Jin, Xin Zhang & Aaron J. Molstad (26 Aug 2025): Kernelized Discriminant Analysis for Joint Modeling of Multivariate Categorical Responses, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2025.2526412

**To link to this article:** https://doi.org/10.1080/10618600.2025.2526412

View supplementary material

Published online: 26 Aug 2025.

Submit your article to this journal

Article views: 72

View related articles

View Crossmark data

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2025, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/10618600.2025.2526412

Taylor & Francis
Taylor & Francis Group

# Kernelized Discriminant Analysis for Joint Modeling of Multivariate Categorical Responses

Yisen Jin[a], Xin Zhang[b], and Aaron J. Molstad[a,c]

[a]Department of Statistics, University of Florida, Gainesville, FL; [b]Department of Statistics, Florida State University, Tallahassee, FL; [c]School of Statistics, University of Minnesota, Minneapolis, MN

**ABSTRACT**
Modeling the joint probability mass of multiple categorical variables as a function of predictors is a fundamental task in categorical data analysis. When the number of response variables, number of categories per response, and/or the number of predictors is large, existing likelihood-based methods cannot be applied or perform poorly. In this article, we propose a novel approach which assumes a variation of the normal linear discriminant analysis model. In order to estimate unknown parameters in way that exploits dependence amongst the response variables, we propose a new penalized likelihood method based on discrete kernel regression. We propose two estimators, each of which can lead to interpretable and parsimonious fitted models. Theoretically, we establish statistical properties of our method and demonstrate a tradeoff between the statistical error and approximation error. Through simulation studies and an application to genomic data, we demonstrate that our method yields better classification accuracy and more interpretable fitted models than existing methods. Software implementing our method, as well as code for reproducing the results in this article, are available for download at *https://github.com/yjin07/kernelizedDA*. Supplementary materials for this article are available online.

**ARTICLE HISTORY**
Received September 2024
Accepted May 2025

**KEYWORDS**
Categorical data analysis; Kernel methods; Linear discriminant analysis

## 1. Introduction

Consider a regression model where $X \in \mathbb{R}^p$ is the predictor and the response consists of $M$ distinct categorical variables, each with two or more response categories. Specifically, let $Y_1, \dots, Y_M$ be the random responses where $Y_m$ has numerically coded categorical support $\{1, \dots, c_m\} =: [c_m]$ ($c_m \geq 2$) for $m \in \{1, \dots, M\} =: [M]$ ($M \geq 2$). This regression model characterizes the joint conditional probability mass function of interest, namely,

$$\Pr(Y_1 = v_1, \dots, Y_M = v_M \mid X = x), \quad (v_1, \dots, v_M) \in \mathcal{C}, \quad (1)$$

where $\mathcal{C} := [c_1] \times [c_2] \times \cdots \times [c_M]$. Throughout this article, we will use both $(v_1, \dots, v_M)$ and $\boldsymbol{v}$ to denote arbitrary elements of $\mathcal{C}$.

To estimate the probability mass function (1), a simple approach is to fit a separate model for each $(Y_m \mid X)$, for example, using logistic regression or linear discriminant analysis. Then, estimates of (1) are obtained through the product of $M$ estimated marginal probabilities $\prod_{m=1}^{M} \widehat{\Pr}(Y_m = v_m \mid X = x)$. We refer to this approach, which implicitly assumes responses are independent given $X$, as “separate modeling”. In machine learning, many adopt a less restrictive version of the separate modeling approach based on the notion of a “classifier chain” (Read et al. 2009; Zhang and Zhou 2013; Read et al. 2021). A classifier chain is constructed by fitting a particular sequence of conditional models. For example, one would first model $(Y_1 \mid X = x)$, then $(Y_2 \mid X = x, Y_1 = y_1)$, and so on. The classification rule for a new observation with predictor $x_{\text{new}}$ would thus be the argument $(v_1, \dots, v_m) \in \mathcal{C}$ maximizing the product of the successive conditional probabilities. This sequential approach is more flexible than separate modeling, but requires many ad-hoc choices that may have a significant impact on performance (e.g., in what order to fit the chain or whether to condition on predicted or observed response) and yields fitted models whose parameters are difficult to interpret in terms of (1), the mass function of interest.

Beyond separate or sequential approaches, there are many methods for fitting (1) in the classical categorical data analysis literature. In particular, many have proposed link functions with parameters that can be interpreted in terms of their effects on marginal probabilities and higher-order associations (Glonek and McCullagh 1995; Glonek 1996; Lang 1996; Molenberghs and Lesaffre 1999; Ekholm, McDonald, and Smith 2000; Lupparelli and Roverato 2017). While these models are often more flexible than separate models, generally speaking, they are not applicable when $p$ is large, and/or are difficult to compute when $M \geq 3$.

To see why flexibly modeling (1) with large $M$ is challenging, consider representing $(Y_1, \dots, Y_M)$ as a synthetic (univariate) categorical response variable $Y'$ with $c^\star = \prod_{m=1}^{M} c_m$ many categories, and modeling $(Y' \mid X)$ directly. That is, for each $(v_1, \dots, v_M) \in \mathcal{C}$, we define $\Pr\{Y' = f(v_1, \dots, v_M) \mid X = x\} = \Pr(Y_1 = v_1, \dots, Y_M = v_M \mid X = x)$ for

**CONTACT** Aaron J. Molstad *amolstad@umn.edu* School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

bijection $f : \mathcal{C} \to [c^\star]$, and model $(Y' \mid X)$ using standard methods for univariate categorical response regression. We call this approach "aggregate modeling". Aggregate modeling is the regression analog of modeling counts in an $M$-way contingency table as a multinomial random variable (Agresti 2012). In contrast to separate modeling, aggregate modeling allows for arbitrary conditional dependence between components of the response. However, this additional flexibility comes at the cost of model complexity, with the number of parameters growing exponentially in $M$. To make matters worse, if the sample size is small relative to $c^\star$, it is probable that one will not observe a response from at least one of the $c^\star$ many categories of $Y'$. In such a scenario, one cannot use maximum likelihood in a straightforward way. A more efficient modeling approach should exploit the fact that $Y'$ is constructed from $M$ distinct response variables.

Recently, significant efforts have been made to accommodate large $p$. Molstad and Rothman (2023) proposed to fit the multinomial logistic regression aggregate model using penalized maximum likelihood. Their penalty allowed predictors to be interpreted as being either irrelevant, affecting only marginal probabilities, or affecting all higher-order associations. When $M \geq 3$, however, the method from Molstad and Rothman (2023) becomes too computationally burdensome to be useful in practice. Along different lines, Molstad and Zhang (2022) proposed to model (1) using a mixture of regressions model that assumes that the tensor-valued function defined by probabilities $\Pr(Y_1 = v_1, \ldots, Y_M = v_M \mid X = x)$ has a low rank decomposition for all $x \in \mathbb{R}^p$ and performs variable selection under this paradigm. However, their method is not capable of handling arbitrary dependence among all the responses, and is inefficient when the true probability mass function is not low-rank. In the supplementary materials, based on a similar low-rank assumption, we describe a method for estimating the joint probability mass of the responses $(Y_1, \ldots, Y_M)$ unconditional on the predictor.

In this article, we propose a new model-based method for fitting (1) under the normal linear discriminant analysis model. Specifically, our method allows one to model complex conditional dependencies between the response variables indirectly by fitting both the inverse regression of $X$ on $(Y_1, \ldots, Y_M)$ and the joint distribution of $(Y_1, \ldots, Y_M)$, then applying Bayes' theorem. For fitting $X \mid (Y_1, \ldots, Y_M)$, we use a discrete kernel regression technique to straightforwardly handle the case where many combinations of response categories are not observed in the training data. More importantly, our method has parameters which are easily interpretable, and can handle large $M$, $p$, and $c_m$.

We demonstrate these features of our method with an application where we use a patient's $p$-dimensional gene expression profile (taken from a colon tissue sample) to classify the patient in terms of $(Y_1)$ whether they have ulcerative colitis or not, $(Y_2)$ whether the colon tissue was inflamed or not, and $(Y_3)$ the location from which the tissue sample was taken (sigmoid colon, terminal ileum, descending colon, or ascending colon).

## 2. Kernelized Discriminant Analysis

### 2.1. Multivariate Linear Discriminant Analysis Model

Our method assumes a variation of the normal linear discriminant analysis model. Letting $\mathbb{S}^p_+$ denote the set of $p \times p$ symmetric positive definite matrices, we assume

$$X \mid Y_1 = v_1, \ldots, Y_M = v_M \sim \mathrm{N}_p\left(\mu_{*v_1,\ldots,v_M}, \Sigma_*\right), \quad (v_1, \ldots, v_M) \in \mathcal{C}, \tag{2}$$

where $\Sigma_*^{-1} =: \Omega_* \in \mathbb{S}^p_+$ is the unknown precision (inverse covariance) matrix and $\mu_{*v_1,\ldots,v_M} \in \mathbb{R}^p$ is the unknown mean vector corresponding to the response category $M$-tuple $(v_1, \ldots, v_M)$. That is, we assume that given the $M$-dimensional categorical response $(Y_1, \ldots, Y_M)$ the predictor $X$ follows a $p$-dimensional multivariate normal distribution whose mean vector depends on the combination of response categories, but whose covariance is identical across category combinations. Note that (2) is exactly the linear discriminant analysis model under the aggregate model described in the previous section. While this generality provides the flexibility of the aggregate model, we estimate parameters from (2) in a way that exploits the multivariate nature of the response. Recently, Deng, Zhang, and Molstad (2024) proposed a tensor-based approach to parameter estimation under model (2). However, their approach is designed for the classification of a bivariate response, and is not practically applicable with $M \geq 3$. In contrast, our method naturally handles any $M \geq 2$.

Under (2), Bayes' classification rule for a new predictor $x_{\text{new}} \in \mathbb{R}^p$ is given by the $M$-tuple $(v_1, \ldots, v_M)$ maximizing $\Pr(Y_1 = v_1, \ldots, Y_m = v_M \mid X = x_{\text{new}}) \propto f_{*v_1,\ldots,v_M}(x_{\text{new}})\Pr(Y_1 = v_1, \ldots, Y_M = v_M)$ where $f_{*v_1,\ldots,v_M}(x_{\text{new}})$ is the density of $\mathrm{N}_p\left(\mu_{*v_1,\ldots,v_M}, \Sigma_*\right)$ evaluated at $x_{\text{new}}$ and $\Pr(Y_1 = v_1, \ldots, Y_M = v_M) = \pi_{*v_1,\ldots,v_M}$ is the marginal probability that $(Y_1, \ldots, Y_M) = (v_1, \ldots, v_M)$. Naturally, $\pi_{*\boldsymbol{v}} \geq 0$ for all $\boldsymbol{v} \in \mathcal{C}$ and $\sum_{v_1=1}^{c_1} \cdots \sum_{v_M=1}^{c_M} \pi_{*v_1,\ldots,v_M} = 1$. For our regression problem to make sense, however, we require the $\pi_{*\boldsymbol{v}}$ satisfy $\Pr(Y_m = v_m) > 0$ for all $v_m \in [c_m]$ and $m \in [M]$. Thus restated, Bayes' classification rule is

$$\underset{\boldsymbol{v} \in \mathcal{C}}{\arg\max} \left\{\mu_{*\boldsymbol{v}}^\top \Omega_* (2x_{\text{new}} - \mu_{*\boldsymbol{v}}) + 2\log \pi_{*\boldsymbol{v}}\right\}. \tag{3}$$

In practice, one would replace unknown parameters $\pi_{*\boldsymbol{v}}$, $\mu_{*\boldsymbol{v}}$, and $\Omega_*$ appearing in (3) with estimates thereof. Classification with respect to a single response component, say $Y_1$, also requires estimation of $\pi_{*\boldsymbol{v}}$, $\mu_{*\boldsymbol{v}}$, and $\Omega_*$, as Bayes' classification rule for $Y_1$ under (2) is the $v_1 \in [c_1]$ maximizing $\Pr(Y_1 = v_1 \mid X = x_{\text{new}})$, that is,

$$\underset{v_1 \in [c_1]}{\arg\max} \sum_{v_2=1}^{c_2} \cdots \sum_{v_M=1}^{c_M} \Bigg[\pi_{*v_1,\ldots,v_M} \times \exp\left\{\left(x_{\text{new}} - \frac{\mu_{*v_1,\ldots,v_M}}{2}\right)^\top \Omega_* \mu_{*v_1,\ldots,v_M}\right\}\Bigg]. \tag{4}$$

Based on (4), one can see that it is possible for the $v_1$ maximizing the marginal probability to disagree with the $v_1'$ from the $M$-

tuple $(v_1', \dots, v_M')$ maximizing $\Pr(Y_1 = v_1', \dots, Y_m = v_M' \mid X = x_{\text{new}})$. Equation (4) suggests that marginally, the appropriate model for $Y_1 \mid X$ is the mixture discriminant analysis model (Hastie and Tibshirani 1996). If one or more responses from (2) were unobservable, then (2) generates the mixture discriminant analysis model with the same number of mixtures for every class. Later, we show that mixture discriminant analysis model fit to each response separately performs poorly in comparison to methods estimating all parameter from (2) directly.

In subsequent sections, we will propose new ways to (i) estimate the $\mu_{*\boldsymbol{v}}$ and the precision matrix $\Omega_*$, and (ii) estimate the discriminant vectors $\beta_{*\boldsymbol{v}} = \Omega_* \mu_{*\boldsymbol{v}}$ directly. These discriminant vectors span the same subspace as the classical Fisher-Rao's discriminant vectors, which (without the normality assumption) sequentially maximize the ratio of between-class variance and within-class variance. Therefore, our methodology can be used for off-the-shelf dimension reduction, visualization, and plug-in estimation for Fisher-Rao-type discriminant analysis, which has been used for "multi-label" classification in the literature (Park and Lee 2008; Wang, Ding, and Huang 2010).

As mentioned, a major deficiency of the separate modeling approach is that it implicitly assumes that responses are conditionally independent. In Section S.1 of the supplementary materials, by analyzing the odds ratio, we characterize the manner in which the linear discriminant analysis model (2) induces conditional dependence in $(Y_1, \dots, Y_M) \mid X$. Perhaps surprisingly, under certain restrictions on the precision $\Omega_*$, the mean vectors $\mu_{*\boldsymbol{v}}$, and their products, the model (2) can imply conditional independence.

## 2.2. Considerations for Maximum Likelihood Estimation

In standard applications of the linear discriminant analysis model—when there is only a single (univariate) categorical response—one can expect to observe realizations of $X$ conditional on each response category. As such, one can straightforwardly use standard maximum likelihood estimators for the unknown mean vectors and covariance. Similarly, marginal probabilities for the response categories can be estimated in a straightforward way. For the setting where $p > n$, in which case the maximum likelihood estimator of $\Omega_*$ does not exist, there are many methods for regularized estimation of the parameters from (2) (Rothman et al. 2008; Witten and Tibshirani 2009; Xu et al. 2015; Price, Geyer, and Rothman 2015; Molstad and Rothman 2018), and the discriminant vectors (Cai and Liu 2011; Mai, Zou, and Yuan 2012; Mai, Yang, and Zou 2019).

In the multivariate categorical response context, however, estimation of the parameter from (2) becomes more challenging. For example, in order to estimate the mean vectors from (2) accurately, one must observe sufficiently large sample of $X$ from each of the $\prod_{m=1}^M c_m$ response categories. Moreover, estimating the marginal probabilities $\pi_{*\boldsymbol{v}}$ is challenging: when $n$ is small relative to $\prod_{m=1}^M c_m$, this is essentially the problem of estimating probabilities from a sparse contingency table (Agresti 1992).

One approach for handling this problem is to simply treat response category combinations not appearing in the sample data as occurring with probability zero. That is, for every $\boldsymbol{v} \in \mathcal{C}$ not observed in our training data, we would assume $\pi_{*\boldsymbol{v}} = 0$, which makes estimating $\mu_{*\boldsymbol{v}}$ unnecessary for the task of classification. Of course, if we know certain category combinations occur with nonzero probability, this may be inappropriate. Moreover, the ability to interpret the parameters of (2) is one of the primary reasons for employing the model (2).

Instead, we propose a new approach for estimating all parameters from (2) for the purpose of classification. Loosely speaking, our method exploits the assumption that that if $(v_1, \dots, v_M)$ is similar to $(v_1', \dots, v_M')$ (e.g., many $v_m = v_m'$), then $\mu_{*v_1,\dots,v_M}$ will be similar to $\mu_{*v_1',\dots,v_M'}$. In the case that $p$ is large, we will also consider two (distinct) assumptions which will reduce the number of parameters to be estimated. The first is that many components of the mean vectors $\mu_{*\boldsymbol{v}} \in \mathbb{R}^p$ do not vary across the response category combinations. The second, which is not necessarily implied by the first, is that many of the $p$ variables are irrelevant for classification.

## 2.3. Discrete Kernelized Regression of $X$ on $(Y_1, \dots, Y_M)$

For the remainder of this section, let $g_* : \mathcal{C} \to \mathbb{R}^p$ denote the function $g_*(\boldsymbol{v}) := \mu_{\boldsymbol{v}}$ for each $\boldsymbol{v} \in \mathcal{C}$. Accordingly, (2) can be equivalently characterized $(X \mid Y_1 = v_1, \dots, Y_M = v_M) \sim \mathrm{N}_p\{g_*(v_1, \dots, v_M), \Sigma_*\}$ for each $(v_1, \dots, v_M) \in \mathcal{C}$. That is, $g_*$ is a function whose domain is $\mathcal{C}$ and codomain is $\mathbb{R}^p$. Define the components of $g_*$ at $\boldsymbol{v}$ as $g_*(\boldsymbol{v}) = (g_{*1}(\boldsymbol{v}), \dots, g_{*p}(\boldsymbol{v}))^\top \in \mathbb{R}^p$ where $g_{*\ell} : \mathcal{C} \to \mathbb{R}$ for $\ell \in [p]$.

To exploit the notion that similar combinations of response categories correspond to similar mean vectors, we use a variation of kernelized regression. We assume there exists a transformation $\phi : \mathcal{C} \to \mathbb{R}^d$ $(d \geq 1)$ such that $\|\phi(\boldsymbol{v}) - \phi(\boldsymbol{v}')\|_2$ small implies $\|g_*(\boldsymbol{v}) - g_*(\boldsymbol{v}')\|_2$ is small, loosely speaking. This requires the existence of some transformation from the space of the response, $\mathcal{C}$, to $\mathbb{R}^d$ such that if two response combinations $\boldsymbol{v}$ and $\boldsymbol{v}'$ are close in the transformed space, their corresponding mean vectors are close in $\mathbb{R}^p$. Such transformations $\phi$ are called "feature maps" in nonparametric regression (Schölkopf and Smola 2002).

At a high level, we first apply the transformation $\phi$ to the collection of observed responses, then quantify the similarity between two any response category combinations via the Euclidean inner product in $\mathbb{R}^d$. Specifically, let $k : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ be a symmetric positive-semidefinite kernel function such that for any collection of $n$ responses $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$, where $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iM}) \in \mathcal{C}$, the $n \times n$ matrix with $(i, j)$th entry $k(\boldsymbol{y}_i, \boldsymbol{y}_j) = \langle \phi(\boldsymbol{y}_i), \phi(\boldsymbol{y}_j) \rangle$ is positive semidefinite. Loosely, $k(\boldsymbol{y}_i, \boldsymbol{y}_j)$ will be large if $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are similar, and vice versa.

Formally, we propose to approximate the $\ell$th component of the mean function, $g_{*\ell}$, with a function $g_\ell$ belonging to the hypothesis space of functions

$$\begin{aligned} g_\ell(\cdot) &= \eta_\ell + \widetilde{g}_\ell(\cdot), \quad \eta_\ell \in \mathbb{R}, \\ &\widetilde{g}_\ell(\cdot) \in \operatorname{span}\left\{k(\cdot, \boldsymbol{y}_i) : i \in [n]\right\}, \quad \ell \in [p]. \end{aligned} \tag{5}$$

That is, $g_\ell(\cdot)$ is the set of functions that can be decomposed into a constant plus a function depending on the input element of $\mathcal{C}$. To see how such a function $g_\ell$ satisfies our heuristic, notice that by definition of the hypothesis space, every $g_\ell(\cdot) = \eta_\ell + \sum_{i=1}^n a_{(\ell)i} k(\cdot, \boldsymbol{y}_i)$ for some $a_{(\ell)} \in \mathbb{R}^n$. Therefore, $|g_\ell(\boldsymbol{v}) - g_\ell(\boldsymbol{v}')| = |\sum_{i=1}^n a_{(\ell)i}\{k(\boldsymbol{v}, \boldsymbol{y}_i) - k(\boldsymbol{v}', \boldsymbol{y}_i)\}| \leq \|\phi(\boldsymbol{v}) - \phi(\boldsymbol{v}')\|_2 \left\|\sum_{i=1}^n a_{(\ell)i} \phi(\boldsymbol{y}_i)\right\|_2$ so that for a given $a_{(\ell)}$, $\|\phi(\boldsymbol{v}) - \phi(\boldsymbol{v}')\|_2$ small roughly implies $|g_\ell(\boldsymbol{v}) - g_\ell(\boldsymbol{v}')|$ is small.

We can also justify our hypothesis space of functions (5) more formally. Given positive semidefinite kernel function $k$ with domain $\mathcal{C} \times \mathcal{C}$, we may define a reproducing kernel Hilbert space of functions, $\mathcal{H}$, where for all $\boldsymbol{v} \in \mathcal{C}$, (i) $k(\cdot, \boldsymbol{v}) \in \mathcal{H}$ and (ii) for all $f \in \mathcal{H}$, $\langle f, k(\cdot, \boldsymbol{v}) \rangle_{\mathcal{H}} = f(\boldsymbol{v})$ (Wainwright 2019, chap. 12–13). Suppose, momentarily, $\Omega_*$ were known. It is then natural to consider the (nonparametric) maximum likelihood estimator of $g_*$ given by $\arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \{x_i - h(\boldsymbol{y}_i)\}^\top \Omega_* \{x_i - h(\boldsymbol{y}_i)\}$. By arguments similar to the generalized representer theorem (Schölkopf, Herbrich, and Smola 2001, Theorem 1; see discussion of relaxing strict monotonicity of regularizing function), we can show that one minimizer with respect to $h$ is given by a function of the form $\widehat{h}_\ell(\cdot) = \sum_{i=1}^{n} a_{(\ell)i} k(\cdot, \boldsymbol{y}_i)$ for some $a_{(\ell)} \in \mathbb{R}^n$ for each $\ell \in [p]$. Thus, it is natural to focus our attention to the space of functions (5), the set of all functions having the same linear representation as $\widehat{h}$, a maximum likelihood estimator.

For discrete kernel regression, there are many transformations $\phi : \mathcal{C} \to \mathbb{R}^d$, and corresponding kernel functions $k$, that could be employed. These include the weighted Hamming distance kernel, the weighted pair-agreement kernel, and the weighted triple-agreement kernel, to name a few. The weighted Hamming distance kernel is given by $k^{\mathrm{H}}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum_{m=1}^{M} w_m \mathbf{1}(y_{im} = y_{jm})$, where $w_m \geq 0$ are user specified weights (e.g., $w_m = \sqrt{c_m}$). This kernel computes the (weighted) number of agreements between its two inputs. The weighted pair-agreement kernel is defined as $k^{\mathrm{PA}}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum\sum_{m \neq m'} w_{m,m'} \mathbf{1}\{(y_{im}, y_{im'}) = (y_{jm}, y_{jm'})\}$, which counts the number of agreements between pairs of response components. Here, $w_{m,m'}$ is a user-specified weight. Similarly, the triple-agreement kernel is given by $k^{\mathrm{TA}}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum\sum\sum_{m \neq m' \neq m''} w_{m,m',m''} \mathbf{1}\{(y_{im}, y_{im'}, y_{im''}) = (y_{jm}, y_{jm'}, y_{jm''})\}$, which counts the number of agreements between triplets of response components. We can also define a combined kernel as a weighted sum of the above kernels, allowing for a more flexible similarity measure that captures both individual and higher-order agreements among response components. More flexible versions of these kernels, where weights are category-dependent, are provided in the supplementary material. We compare different choices of kernels numerically in Section S.3.2 in the supplementary materials. In practice, we suggest selecting the kernel via cross-validation.

Before we formally describe how we will estimate $g_*$, we note that the span of $\{k(\cdot, \boldsymbol{y}_i), i \in [n]\}$, the space from which we will estimate $\widetilde{g}$, is determined by the set of unique responses $\boldsymbol{y}_i$. For example, if $\boldsymbol{y}_n = \boldsymbol{y}_{n-1}$, then $\mathrm{span}\left\{k(\cdot, \boldsymbol{y}_i) : i \in [n]\right\} = \mathrm{span}\left\{k(\cdot, \boldsymbol{y}_i) : i \in [n-1]\right\}$. Consequently, we define $\widetilde{\boldsymbol{Y}} = \{\widetilde{\boldsymbol{y}}_i\}_{i=1}^{\widetilde{n}}$ as the set of distinct response category combinations observed in $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{n}$ (i.e., $\widetilde{n} \leq n$ and $\widetilde{\boldsymbol{Y}} \subseteq \boldsymbol{Y}$) where $\widetilde{\boldsymbol{y}}_i \neq \widetilde{\boldsymbol{y}}_j$ for all $i \neq j$. We then define $k_{\widetilde{\boldsymbol{Y}}}(\cdot) : \mathcal{C} \to \mathbb{R}^{\widetilde{n}}$ as $k_{\widetilde{\boldsymbol{Y}}}(\cdot) = (k(\cdot, \widetilde{\boldsymbol{y}}_1), \dots, k(\cdot, \widetilde{\boldsymbol{y}}_{\widetilde{n}}))^\top$, define $K_{\widetilde{\boldsymbol{Y}}} \in \mathbb{R}^{n \times \widetilde{n}}$ as the matrix with $(i,j)$th entry $k(\boldsymbol{y}_i, \widetilde{\boldsymbol{y}}_j)$, and define $K_{\widetilde{\boldsymbol{Y}}}^{\dagger} \in \mathbb{R}^{\widetilde{n} \times \widetilde{n}}$ as the matrix with $(i,j)$th entry $k(\widetilde{\boldsymbol{y}}_i, \widetilde{\boldsymbol{y}}_j)$. It is easy to see that $\mathrm{span}\{k(\cdot, \widetilde{\boldsymbol{y}}_i) : i \in [\widetilde{n}]\} = \mathrm{span}\{k(\cdot, \boldsymbol{y}_i) : i \in [n]\}$. The implication of this fact is that any $\widetilde{g}(\cdot) = (\widetilde{g}_1(\cdot), \dots, \widetilde{g}_p(\cdot))^\top$ can be represented as $\alpha^\top k_{\widetilde{\boldsymbol{Y}}}(\cdot)$ where $\alpha \in \mathbb{R}^{\widetilde{n} \times p}$. Thus, in contrast to applications of the representer theorem with predictors drawn from a continuous distribution (wherein the coefficient dimension is $n$), we need only estimate $\widetilde{n}$ coefficients per component of $\widetilde{g}$. It is this reduction in the number of coefficients that makes our method particularly scalable to large $M$, since in general, we expect $\widetilde{n} \ll n \ll \prod_{m=1}^{M} c_m$ for large $M$. In the supplementary materials Section S.7.3, we explain that this feature of our method can be understood as exploiting an exact version of the Nyström approximation (Williams and Seeger 2000). Thus, for a given set of $n$ independent observations $\{(\boldsymbol{y}_1, x_1), \dots, \{(\boldsymbol{y}_n, x_n)\}$, we approximate the function $g_*(\cdot)$ with a function of the form $\eta + \alpha^\top k_{\widetilde{\boldsymbol{Y}}}(\cdot)$. To fit the model (2), $\eta \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}^{\widetilde{n} \times p}$ will be estimated using penalized maximum likelihood.

To establish some of the results in this article, we will often require an assumption about the user-chosen kernel function $k$. Let $\varphi_{\min}(A)$ be the smallest singular value of a matrix $A$.

*Assumption 1.* The kernel function $k$ is chosen so that for any $\widetilde{\boldsymbol{Y}}$, there exists a constant $c_0 > 0$ such that $\varphi_{\min}(K_{\widetilde{\boldsymbol{Y}}}^{\dagger}) \geq c_0 > 0$.

Assumption 1 would be satisfied by a kernel function $k'$ for any sample $\boldsymbol{Y}$, if, for example, we defined $k'(\boldsymbol{y}_i, \boldsymbol{y}_j) = k(\boldsymbol{y}_i, \boldsymbol{y}_j) + c_0 \mathbf{1}(\boldsymbol{y}_i = \boldsymbol{y}_j)$ with $k$ being any of the three example kernels. This amounts to upweighting an exact match between the two arguments of the kernel function by positive constant $c_0$.

## 3. Estimation

### 3.1. Regularized Estimation of $g_*$

In order to estimate the coefficients $(\eta, \alpha)$ and the precision matrix $\Omega_*$, we propose to minimize the negative log-likelihood with $g_*(\cdot)$ approximated by $\eta + \alpha^\top k_{\widetilde{\boldsymbol{Y}}}(\cdot)$, which is given by

$$\frac{1}{n} \sum_{i=1}^{n} \{x_i - \eta - \alpha^\top k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{y}_i)\}^\top \Omega \{x_i - \eta - \alpha^\top k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{y}_i)\} - \log\det(\Omega), \tag{6}$$

where we have ignored some constants and det is the determinant of a matrix. In low-dimensional settings, minimizing (6) with respect to $\alpha$, $\eta$, and $\Omega$ may work well. Specifically, we have the following result.

*Proposition 1.* Define $g_{\mathrm{MLE}}(\cdot) = \eta_{\mathrm{MLE}} + \alpha_{\mathrm{MLE}}^\top k_{\widetilde{\boldsymbol{Y}}}(\cdot)$ where $\eta_{\mathrm{MLE}}$ and $\alpha_{\mathrm{MLE}}$ are minimizers of (6) with respect to $\eta$ and $\alpha$. Let $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. For each $\{\widetilde{\boldsymbol{y}}_i\}_{i=1}^{\widetilde{n}}$, define $\bar{x}_{\widetilde{\boldsymbol{y}}_i} = \sum_{k=1}^{n} \mathbf{1}\{\boldsymbol{y}_k = \widetilde{\boldsymbol{y}}_i\} x_k / \sum_{\ell=1}^{n} \mathbf{1}\{\boldsymbol{y}_\ell = \widetilde{\boldsymbol{y}}_i\}$ as the sample mean for the observed response category combination $\widetilde{\boldsymbol{y}}_i \in \mathcal{C}$, and define $\bar{x}_{\widetilde{\boldsymbol{y}}_i}^0 = \sum_{k=1}^{n} \mathbf{1}\{\boldsymbol{y}_k = \widetilde{\boldsymbol{y}}_i\}(x_k - \bar{x}) / \sum_{\ell=1}^{n} \mathbf{1}\{\boldsymbol{y}_\ell = \widetilde{\boldsymbol{y}}_i\}$. If Assumption 1 holds, then

$$g_{\mathrm{MLE}}(\boldsymbol{v}) = \begin{cases} \bar{x}_{\boldsymbol{v}} & : \text{ if } \boldsymbol{y}_i = \boldsymbol{v} \text{ for any } i \in [n] \\ \sum_{i=1}^{\widetilde{n}} w_i(\boldsymbol{v}) \bar{x}_{\widetilde{\boldsymbol{y}}_i}^0 + \bar{x} & : \text{otherwise} \end{cases},$$

where $w(\boldsymbol{v}) = (w_1(\boldsymbol{v}), \dots, w_{\widetilde{n}}(\boldsymbol{v}))^\top = K_{\widetilde{\boldsymbol{Y}}}^{0\dagger^{-1}} k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{v})$and $K_{\widetilde{\boldsymbol{Y}}}^{0\dagger} \in \mathbb{R}^{\widetilde{n} \times \widetilde{n}}$ is a matrix with $(i,j)$th entry $k(\widetilde{\boldsymbol{y}}_i, \widetilde{\boldsymbol{y}}_j) - n^{-1} \sum_{i'=1}^{n} k(\boldsymbol{y}_{i'}, \widetilde{\boldsymbol{y}}_j)$. The minimizer with respect to $\Omega$, if it exists, is $(n^{-1} \sum_{i=1}^{n} \{x_i - g_{\mathrm{MLE}}(\boldsymbol{y}_i)\}\{x_i - g_{\mathrm{MLE}}(\boldsymbol{y}_i)\}^\top)^{-1}$.

Proposition 1 establishes that when we minimize (6) with respect to $\eta$ and $\alpha$, our estimate of $g_*$, $g_{\mathrm{MLE}}$, is equivalent to the conditional sample mean for all response category combinations $\boldsymbol{v} \in \mathcal{C}$ that are observed in the training data. For category combinations that are not observed, $g_{\mathrm{MLE}}$ is a weighted sum of the overall sample mean and the conditional sample means for observed response category combinations. The weights are determined by the choice of kernel function $k$ and the collection of observed responses $\{\boldsymbol{y}_i\}_{i=1}^{n}$. Thus, if $p$ and $c^\star$ were fixed, and the $\pi_{*\boldsymbol{v}}$ are bounded away from zero, as $n \to \infty$, our method will perform identically to standard maximum likelihood.

However, in finite samples when $p$ and $c^\star$ are large relative to $n$, this may be problematic. First, the $\bar{x}_{\boldsymbol{v}}$ may be computed from a small number of observations, and moreover, it is well known that with $p$ diverging quickly relative to $n$, linear discriminant analysis will eventually perform no better than random guessing due to noise accumulation in the mean estimates (Fan and Fan 2008; Elman et al. 2020). Second, when $p \geq n$, the maximum likelihood estimator of $\Omega$ will not exist because $n^{-1}\sum_{i=1}^{n}\{x_i - g_{\mathrm{MLE}}(\boldsymbol{y}_i)\}\{x_i - g_{\mathrm{MLE}}(\boldsymbol{y}_i)\}^\top$ will not be invertible. Even when $p < n$, this matrix will be singular when $c^\star$ is large relative to $n$.

As mentioned, in this article we consider two schemes for shrinkage estimation of parameters in (6) which exploit different assumptions about the model (2). The first assumption we consider is that many components of $\mu_{*\boldsymbol{v}}$ do not vary across all $\boldsymbol{v} \in \mathcal{C}$. That is, there are many components $j$ such that $[\mu_{*\boldsymbol{v}}]_j = \eta_j \in \mathbb{R}$ for all $\boldsymbol{v} \in \mathcal{C}$ where $[a]_j$ denotes the $j$th component of a vector $a$. To encourage fitted models with this property, we need a way in which to estimate $g$ such that many of the $\widetilde{g}_\ell = 0$ for all inputs.

Notice that for any fixed $\alpha$, $n^{-1}\sum_{i=1}^{n} x_i - n^{-1}\sum_{i=1}^{n}\alpha^\top k_{\widetilde{\mathbf{Y}}}(\boldsymbol{y}_i)$ minimizes the negative log-likelihood with respect to $\eta$. Thus, we need only focus on minimizing

$$\mathcal{L}(\alpha,\Omega) = \frac{1}{n}\mathrm{tr}\big\{(X^0 - K^0_{\widetilde{\mathbf{Y}}}\alpha)\Omega(X^0 - K^0_{\widetilde{\mathbf{Y}}}\alpha)^\top\big\} - \log\det(\Omega), \quad (7)$$

where $X^0 = (x_1 - n^{-1}\sum_{i=1}^{n} x_i, \ldots, x_n - n^{-1}\sum_{i=1}^{n} x_i)^\top \in \mathbb{R}^{n\times p}$ and $K^0_{\widetilde{\mathbf{Y}}} \in \mathbb{R}^{n\times\widetilde{n}}$ is a matrix with $(j,k)$th entry $k(\boldsymbol{y}_j, \widetilde{\boldsymbol{y}}_k) - n^{-1}\sum_{i=1}^{n} k(\boldsymbol{y}_i, \widetilde{\boldsymbol{y}}_k)$.

If, for example, the $j$th column of $\alpha$ is entirely zero, then we are ensured $[\alpha^\top k_{\widetilde{\mathbf{Y}}}(\boldsymbol{v})]_j = 0$ for all $\boldsymbol{v} \in \mathcal{C}$, which would imply that $[\widehat{g}(\boldsymbol{v})]_j$ is constant. Thus, to encourage estimates of $g$ such that $\widetilde{g}_\ell(\boldsymbol{v}) = 0$ for all $\boldsymbol{v}$, we apply a group lasso penalty on the columns of $\alpha$. Specifically, we propose to estimate the pair $(\alpha, \Omega)$ with $(\widehat{\alpha}, \widehat{\Omega})$, defined as

$$\operatorname*{arg\,min}_{\alpha\in\mathbb{R}^{\widetilde{n}\times p},\,\Omega\in\mathbb{S}^{p}_{+}} \Big\{\mathcal{L}(\alpha,\Omega) + \lambda\|\alpha\|_{1,2} + \frac{\gamma}{2}\|\Omega\|_F^2\Big\}, \quad (8)$$

where $\|\alpha\|_{1,2} := \sum_{j=1}^{p}(\sum_{i=1}^{\widetilde{n}}\alpha_{i,j}^2)^{1/2}$, $\|\Omega\|_F := \{\mathrm{tr}(\Omega^\top\Omega)\}^{1/2}$, and $(\lambda,\gamma) \in (0,\infty)\times(0,\infty)$ are user-specified tuning parameters. The ridge penalty on $\Omega$ serves to shrink the sum of squared elements of $\Omega$ and ensures that with $\alpha$ fixed, a minimizer with respect to $\Omega$ exists. Though the optimization problem in (8) is nonconvex, it is biconvex. That is, with $\alpha$ held fixed, the optimization is convex with respect to $\Omega$ and vice versa.

Our choice of ridge penalty on $\Omega$ is primarily for computational convenience. With $\alpha$ fixed, the minimizer of (8) with respect to $\Omega$ has a closed form. One could instead penalize the sum of the absolute values of the off-diagonals of $\Omega$ if it is reasonable to assume that $\Omega_*$ is sparse (Rothman et al. 2008). This approach, however, would in general require an iterative algorithm to minimize (8) with respect to $\Omega$.

With $(\widehat{\alpha}, \widehat{\Omega})$ (and consequently, $\widehat{\eta}$) in hand, we classify a subject with predictors $x_{\mathrm{new}}$ into response category set given by $\arg\max_{\boldsymbol{v}\in\mathcal{C}}[\{\widehat{\eta} + \widehat{\alpha}^\top k_{\widetilde{\mathbf{Y}}}(\boldsymbol{v})\}^\top\widehat{\Omega}\{2x_{\mathrm{new}} - \widehat{\eta} - \widehat{\alpha}^\top k_{\widetilde{\mathbf{Y}}}(\boldsymbol{v})\} + 2\log\widehat{\pi}_{\boldsymbol{v}}]$, where $\widehat{\pi}_{\boldsymbol{v}}$ is an estimate of $\pi_{*\boldsymbol{v}}$, which we discuss in the supplementary materials.

In the supplementary materials, we perform numerical studies also including the penalty term $\mathrm{tr}(\alpha^\top K^{\dagger}_{\widetilde{\mathbf{Y}}}\alpha)$ in (8), as is often seen in kernel-based nonparametric regression. We also discuss briefly why, in our context, we think penalizing $\alpha$'s $(1,2)$-norm alone is preferable.

### 3.2. Convexifying Reparameterization for Direct Variable Selection

The estimator in (8) is motivated by the assumption that many elements of the mean vectors $\mu_{*\boldsymbol{v}}$ do not differ across all $\boldsymbol{v} \in \mathcal{C}$. While this affords interpretability in terms of how the parameters from (2) depend on $\boldsymbol{v}$, this does not lead to variable selection (i.e., removal of irrelevant variables) without additional constraints on $\Omega_*$. Recall that $\beta_{*\boldsymbol{v}} := \Omega_*\mu_{*\boldsymbol{v}}$ is the discriminant vector for category combination $\boldsymbol{v} \in \mathcal{C}$. As mentioned in Remark 2 of the supplementary material, for the $j$th variable to be irrelevant for distinguishing between all combinations of response categories, it must be that $[\beta_{*\boldsymbol{v}}]_j = \nu_j \in \mathbb{R}$ for all $\boldsymbol{v} \in \mathcal{C}$. Intuitively, it is insufficient that $[\mu_{*\boldsymbol{v}}]_j = [\mu_{*\boldsymbol{v}'}]_j$ for all $\boldsymbol{v}, \boldsymbol{v}'$ because if the $j$th variable is conditionally correlated with a variable whose means are unequal, the $j$th variable will affect the decision rule (Xu et al. 2015).

When we approximate $g_*(\cdot)$ with $\eta + \alpha^\top k_{\widetilde{\mathbf{Y}}}(\cdot)$, we approximate $\beta_{*\boldsymbol{v}} = \Omega_* g_*(\boldsymbol{v})$ with $\Omega_*\{\eta + \alpha^\top k_{\widetilde{\mathbf{Y}}}(\boldsymbol{v})\}$ for some $\eta \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}^{\widetilde{n}\times p}$. Therefore, if we want to estimate $\Omega_*$ and $(\eta, \alpha)$ so that our fitted model can be interpreted directly in terms of which variables are irrelevant for discriminating between response categories, we need to encourage fitted models such that many elements of $\widehat{\Omega}\widehat{\alpha}^\top k_{\widetilde{\mathbf{Y}}}(\boldsymbol{v})$ will be zero for all $\boldsymbol{v} \in \mathcal{C}$ for estimates $(\widehat{\Omega}, \widehat{\alpha})$. If the $j$th row of $\widehat{\Omega}\widehat{\alpha}^\top$ were entirely zero, then $[\widehat{\Omega}\widehat{g}(\boldsymbol{v})]_j = [\widehat{\Omega}\widehat{\eta}]_j = c \in \mathbb{R}$ for all $\boldsymbol{v}$, that is, the $j$th predictor has no effect the estimated decision rule. Therefore, if we let $\Theta := \alpha\Omega$, imposing sparsity on the columns of $\Theta$ would correspond to componentwise equality cross the discriminant vectors. Under this parameterization, we can write the negative log-likelihood

$$\begin{aligned}\mathcal{L}_c(\Theta,\Omega) = {}& \frac{1}{n}\mathrm{tr}\big\{(X^0\Omega - K^0_{\widetilde{\mathbf{Y}}}\Theta)\Omega^{-1}(X^0\Omega - K^0_{\widetilde{\mathbf{Y}}}\Theta)^\top\big\} \\ & - \log\det(\Omega), \quad (9)\end{aligned}$$

where the subscript $c$ denotes that this is the negative log-likelihood under the parameterization $\Theta = \alpha\Omega$. Analogous to (8), our estimator for direct variable selection is

$$\begin{aligned}&\operatorname*{arg\,min}_{\Theta\in\mathbb{R}^{\widetilde{n}\times p},\,\Omega=\Omega^\top} \Big\{\mathcal{L}_c(\Theta,\Omega) + \lambda\|\Theta\|_{1,2} + \frac{\eta}{2}\|\Omega\|_F^2\Big\} \\ &\text{subject to } \ \Omega \succeq \epsilon I_p, \quad (10)\end{aligned}$$

where $\epsilon > 0$ is a lower bound on the smallest eigenvalue of $\Omega_*$. Remarkably, the optimization problem in (10) is jointly convex in $(\Theta, \Omega)$ (Zhu 2020, Theorem 1). This implies that (10) is also biconvex as both $\Theta \mapsto \mathcal{L}_c(\Theta, \Omega)$ and $\Omega \mapsto \mathcal{L}_c(\Theta, \Omega)$ are convex. Though this convexifying reparameterization has been studied in regression (Yu and Bien 2019; Zhu 2020), to the best of our knowledge, it has not been used for variable selection in the linear discriminant analysis model.

In practice, we impose a lower bound on the smallest eigenvalue of $\Omega$, $\epsilon$, making the feasible set closed (as opposed to $\mathbb{S}_+^p$, which is open). Though $\epsilon$ is a tuning parameter, we find that simply setting $\epsilon$ equal to some reasonably small constant (e.g., $\epsilon = 10^{-4}$) seems to work well across a variety of settings.

With a solution to (10) in hand, say $(\ddot{\Theta}, \ddot{\Omega})$, we use classification rule

$$\operatorname*{arg\,max}_{\boldsymbol{v} \in \mathcal{C}} \Big[ \{ \ddot{\eta}^\top \ddot{\Omega} + k_{\ddot{\boldsymbol{Y}}}(\boldsymbol{v})^\top \ddot{\Theta} \} \\ \{ 2x_{\text{new}} - \ddot{\eta} + \ddot{\Omega}^{-1} \ddot{\Theta}^\top k_{\ddot{\boldsymbol{Y}}}(\boldsymbol{v}) \} + 2 \log \widehat{\pi}_{\boldsymbol{v}} \Big], \qquad (11)$$

where $\ddot{\eta} = n^{-1} \sum_{i=1}^n x_i - n^{-1} \sum_{i=1}^n k_{\ddot{\boldsymbol{Y}}}(\boldsymbol{y}_i) \ddot{\Theta} \ddot{\Omega}^{-1}$. Examining (11), it is immediate to see that if the $j$th column of $\ddot{\Theta}$ is entirely zero, then the $j$th component of $x_{\text{new}}$ has no effect on the decision rule (since $\ddot{\eta}^\top \ddot{\Omega} x_{\text{new}}$ is constant with respect to $\boldsymbol{v}$).

## 4. Computation

To exploit the biconvexity of the objective function from (8), we use a blockwise coordinate descent algorithm. That is, we iteratively update $\alpha$ with $\Omega$ held fixed and vice versa. With $\alpha$ fixed at its $(t)$th iterate, $\alpha^{(t)}$, obtaining the $(t)$th iterate for $\Omega$ requires solving a ridge penalized normal precision matrix estimation problem, $\Omega^{(t)} = \arg\min_{\Omega \in \mathbb{S}_+^p} [\operatorname{tr}\{S(\alpha)\Omega\} - \log\det(\Omega) + \frac{\eta}{2} \|\Omega\|_F^2]$, with $S(\alpha) = n^{-1}(X^0 - K^0_{\widetilde{\boldsymbol{Y}}}\alpha)^\top (X^0 - K^0_{\widetilde{\boldsymbol{Y}}}\alpha)$. It can be shown that $\Omega^{(t+1)} = \frac{1}{2\eta} V\{-D + (D^2 + 4\eta I_p)^{1/2}\} V^\top$ where $S(\alpha) = VDV^\top$ is the eigendecomposition of $S(\alpha)$ where $V \in \mathbb{R}^{p \times p}$ is orthogonal and $D \in \mathbb{R}^{p \times p}$ diagonal (Witten and Tibshirani 2009; Price, Geyer, and Rothman 2015).

With $\Omega$ fixed at $\Omega^{(t)}$, the $(t+1)$th iterate for $\alpha$ is $\alpha^{(t+1)} \in \arg\min_{\alpha \in \mathbb{R}^{\widetilde{n} \times p}} \left\{ \mathcal{L}(\alpha, \Omega) + \gamma \|\alpha\|_{1,2} \right\}$. We use a variation of the proximal gradient descent algorithm to compute $\alpha^{(t+1)}$ (Beck and Teboulle 2009; Polson, Scott, and Willard 2015). For this subalgorithm, we will use $r$ as an iteration counter. Specifically, given step size $s > 0$ sufficiently small and $(r)$th iterate of $\alpha$, $\alpha^{(r)}$, the $(r+1)$th iterate of the proximal gradient descent subalgorithm is defined as

$$\alpha^{(r+1)} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^{\widetilde{n} \times p}} \left\{ \frac{1}{2} \|\alpha - \alpha^{(r)} + s\nabla_\alpha \mathcal{L}(\alpha^{(r)}, \Omega)\|_F^2 + s\gamma \|\alpha\|_{1,2} \right\} \qquad (12)$$

where $\nabla_\alpha \mathcal{L}(\alpha, \Omega) = -\frac{2}{n}\{K^{0\,\top}_{\widetilde{\boldsymbol{Y}}} X^0 \Omega - K^{0\,\top}_{\widetilde{\boldsymbol{Y}}} K^0_{\widetilde{\boldsymbol{Y}}} \alpha \Omega\}$. One can use subgradient calculus to show that (12) can be solved column-by-column in closed form. Namely,

$$\alpha_{\cdot,j}^{(r+1)} = \max \left( 1 - \frac{s\gamma}{\|\alpha_{\cdot,j}^{(r)} - s[\nabla_\alpha \mathcal{L}(\alpha^{(r)}, \Omega)]_{\cdot,j}\|_2}, 0 \right) \\ \times \left( \alpha_{\cdot,j}^{(r)} - s[\nabla_\alpha \mathcal{L}(\alpha^{(r)}, \Omega)]_{\cdot,j} \right) \qquad (13)$$

where $\alpha_{\cdot,j}^{(r)}$ is the $j$th column of $\alpha^{(r)}$ and $[\nabla_\alpha \mathcal{L}(\alpha^{(r)}, \Omega)]_{\cdot,j}$ is the $j$th column of $\nabla_\alpha \mathcal{L}(\alpha^{(r)}, \Omega)$ (Yuan and Lin 2006; Simon et al. 2013). We repeat (13) for $r = 1, 2, 3, \ldots$ in sequence until the objective function value converges.

In our implementation, we use an accelerated variation of this algorithm (Parikh and Boyd 2014, Chapter 4.3). Briefly, the accelerated version replaces search point $\alpha^{(r)} - s\nabla_\alpha \mathcal{L}(\alpha^{(r)}, \Omega)$ with $\alpha^{(r,r-1)} - s\nabla_\alpha \mathcal{L}(\alpha^{(r,r-1)}, \Omega)$ where $\alpha^{(r,r-1)} = \alpha^{(r)} + \frac{r-1}{r+2}(\alpha^{(r)} - \alpha^{(r-1)})$. It is well known that if $s$ is fixed and chosen sufficiently small—or if $s$ is chosen by backtracking line search—the objective function value converges at a quadratic rate (Beck and Teboulle 2009; Parikh and Boyd 2014). We provide an outline of the algorithm to solve (8), Algorithm S.1 in the supplementary material. To solve (10), we use a blockwise coordinate descent scheme similar to that described in the previous section. The complete algorithm we use for computing (10) can be found in the supplementary material Algorithm S.2. Different from the nonconvex case, with $\Theta$ fixed at $(t)$th iterate $\Theta^{(t)}$, the update for $\Omega$ does not have a closed form solution and we use a projected gradient descent algorithm (Algorithm S.3) to solve the optimization.

We recommend selecting all tuning parameters using cross-validation.

## 5. Theoretical Properties

In this section, we study the finite sample properties of our estimator of $g_*$ based on kernelized regression in (8). Let $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$ be the observed responses, and let $\widetilde{\boldsymbol{Y}} = \{\widetilde{\boldsymbol{y}}_i\}_{i=1}^{\widetilde{n}}$ be the set of distinct response category combinations. The following results will apply conditional on the set of observed responses $\boldsymbol{Y}$. In our context, the mean vectors and discriminant vectors are of primary interest. The precision matrix $\Omega_*$ plays a crucial role in estimation and classification, but mainly serves to bridge the mean vectors and the discriminant vectors. As such, we assume $\Omega_*$ is known in order to focus our attention specifically on recovery of $g_*$. Without loss of generality, we also assume that the predictors are centralized in the sense that their componentwise marginal expectation is zero. In this setting, if for a particular $\ell \in [p]$, $g_{*\ell}$ is constant across all response category combinations, then $g_{*\ell}(\boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in \mathcal{C}$. Hence, we consider estimating $g_*(\boldsymbol{v}) = \mathrm{E}\{X \mid (Y_1, \ldots, Y_M) = \boldsymbol{v}\}$ with $\widehat{g}(\boldsymbol{v}) = \sum_{j=1}^{\widetilde{n}} \widehat{\alpha}_j k(\boldsymbol{v}, \widetilde{\boldsymbol{y}}_j)$ where we define $\widehat{\alpha}$ as

$$\operatorname*{arg\,min}_{\alpha \in \mathbb{R}^{\widetilde{n} \times p}} \left[ \frac{1}{n} \sum_{i=1}^n \{x_i - \alpha^\top k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{y}_i)\}^\top \Omega_* \{x_i - \alpha^\top k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{y}_i)\} \\ + \lambda \|\alpha\|_{1,2} \right]. \qquad (14)$$

Notice that we can write the random predictor (conditional on $\boldsymbol{y}_i$) as $X_i = \Omega_*^{-1/2} Z_i + g_*(\boldsymbol{y}_i)$, where entries of $Z_i \in \mathbb{R}^p$ are independent standard normals for $i \in [n]$. Let $M_* \in \mathbb{R}^{n \times p}$ have $i$th row $g_*(\boldsymbol{y}_i)$. Because many rows of $M_*$ will be duplicated if we observed $\boldsymbol{y}_i = \boldsymbol{y}_j$ for many $i \neq j$, it is convenient to define $M_*^\dagger \in \mathbb{R}^{\widetilde{n} \times p}$ as the matrix that contains one mean vector corresponding to each element of $\widetilde{\boldsymbol{Y}}$ and define $Q \in \mathbb{R}^{n \times \widetilde{n}}$ as a matrix with $(i,j)$th entry $Q_{i,j} = \mathbf{1}(\boldsymbol{y}_i = \widetilde{\boldsymbol{y}}_j)$, so that $M_* = QM_*^\dagger$.

Now, let us define $\alpha_* \in \arg\min_{\alpha \in \mathbb{R}^{\widetilde{n} \times p}} \|M_* - K_{\widetilde{\boldsymbol{Y}}}\alpha\|_F^2$. If Assumption 1 holds, then $\alpha_*$ is unique and given by $\alpha_* = (K_{\widetilde{\boldsymbol{Y}}}^\top K_{\widetilde{\boldsymbol{Y}}})^{-1} K_{\widetilde{\boldsymbol{Y}}}^\top M_*$. Moreover, it can be easily verified that under Assumption 1, $\text{minimum}_{\alpha \in \mathbb{R}^{\widetilde{n} \times p}} \|M_* - K_{\widetilde{\boldsymbol{Y}}}\alpha\|_F^2 = \|M_* - K_{\widetilde{\boldsymbol{Y}}}\alpha_*\|_F^2 = 0$. For completeness, we provide a proof of this fact in the supplementary material. The equality $\|M_* - K_{\widetilde{\boldsymbol{Y}}}\alpha_*\|_F^2 = 0$ implies the existence of an $\alpha$ such that $K_{\widetilde{\boldsymbol{Y}}}\alpha_*$ perfectly recovers $M_*$. This suggests that we may treat $\alpha_*$ as an estimand, and $\widehat{\alpha}$ as our estimator thereof.

Next, let us state our second assumption.

*Assumption 2.* There exists a constant $c_1$ such that $0 < c_1 \leq \varphi_{\min}(\Omega_*) \leq \varphi_{\max}(\Omega_*) \leq 1/c_1 < \infty$, where $\varphi_{\max}$ and $\varphi_{\min}$ denote the largest and smallest eigenvalue of their argument, respectively.

Recall from Section 2.3 that we assume few components of $g_*$ differ as a function of its argument. In terms of $M_*$, this would imply that $M_{*\cdot,j} = 0$ for many $j \in [p]$, where $M_{*\cdot,j}$ is the $j$th column of $M_*$. If $M_{*\cdot,j} = 0$, this implies that $\alpha_{*\cdot,j} = 0$ (since $M_*$ is the rightmost term in the product defining $\alpha_*$). Hence, define $\mathcal{S} = \{j : \alpha_{*\cdot,j} \neq 0\}$, define $\mathcal{S}^c = [p] \setminus \mathcal{S}$, and let $s$ be the cardinality of the set $\mathcal{S}$. Recall that $K_{\widetilde{\boldsymbol{Y}}}^\dagger \in \mathbb{R}^{\widetilde{n} \times \widetilde{n}}$ is the matrix with $(i,j)$th entry $k(\widetilde{\boldsymbol{y}}_i, \widetilde{\boldsymbol{y}}_j)$ (i.e., $K_{\widetilde{\boldsymbol{Y}}} = QK_{\widetilde{\boldsymbol{Y}}}^\dagger$). Define $\kappa_{\mathcal{S}} = \inf_{\Delta \in \mathbb{C}(\mathcal{S}), \|\Delta\|_F = 1} \|K_{\widetilde{\boldsymbol{Y}}}^\dagger \Delta \Omega_*^{1/2}\|_F^2/\widetilde{n}$ where $\mathbb{C}(\mathcal{S}) = \{\nu \in \mathbb{R}^{\widetilde{n} \times p} : \|\nu_{\cdot\mathcal{S}^c}\|_{1,2} \leq 3\|\nu_{\cdot\mathcal{S}}\|_{1,2}\}$. Under Assumptions 1 and 2, $\kappa_{\mathcal{S}} \geq c_0^2 c_1/\widetilde{n} > 0$, but for $\mathcal{S}$ with small cardinality, $\kappa_{\mathcal{S}}$ may be larger. Notice that $\kappa_{\mathcal{S}}$ is effectively a restricted eigenvalue of the matrix $\widetilde{n}^{-1}(\Omega_* \otimes K_{\widetilde{\boldsymbol{Y}}}^\dagger K_{\widetilde{\boldsymbol{Y}}}^\dagger)$ (Wainwright 2019, chap. 7.3.1), hence, the stated lower bound.

We are now prepared to present our first result. For the remainder of the article, let $\|\cdot\|$ denote the spectral norm of matrix.

*Theorem 1* (Average in-sample mean estimation error). Suppose Assumptions 1 and 2 hold. Let $c_2 > 1$ be a fixed constant and let $\omega_* = \max_{j \in [p]} \|\Omega_{*j,\cdot}^{1/2}\|_2$. If $\lambda = 4\omega_*(\|K_{\widetilde{\boldsymbol{Y}}}\|_F + \|K_{\widetilde{\boldsymbol{Y}}}\|\sqrt{2c_2 \log p})/n$, then with probability at least $1 - p^{1-c_2}$,

$$\frac{1}{\widetilde{n}} \sum_{i=1}^{\widetilde{n}} \|g_*(\widetilde{\boldsymbol{y}}_i) - \widehat{g}(\widetilde{\boldsymbol{y}}_i)\|_2^2 \leq \frac{36\omega_*^2 s}{\varphi_{\min}(\Omega_*)\kappa_{\mathcal{S}}} \left(\frac{\widetilde{n}_{\max}}{\widetilde{n}_{\min}^2}\right) \times \left\{\frac{\|K_{\widetilde{\boldsymbol{Y}}}^\dagger\|_F}{\widetilde{n}} + \frac{\|K_{\widetilde{\boldsymbol{Y}}}^\dagger\|}{\widetilde{n}}\sqrt{2c_2 \log p}\right\}^2,$$

where $\widetilde{n}_{\max} := \max_{j \in [\widetilde{n}]} \sum_{i=1}^n \mathbf{1}(\boldsymbol{y}_i = \widetilde{\boldsymbol{y}}_j)$ and $\widetilde{n}_{\min} := \min_{j \in [\widetilde{n}]} \sum_{i=1}^n \mathbf{1}(\boldsymbol{y}_i = \widetilde{\boldsymbol{y}}_j)$.

Theorem 1 demonstrates how well we can recover the mean vectors, on average, corresponding to the response category combinations observed in the training data. If $\widetilde{n}_{\min}$ is small relative to $\widetilde{n}_{\max}$, the bound would be worse than, say, in the best-case scenario when $\widetilde{n}_{\min} = \widetilde{n}_{\max}$. Note that our result in Theorem 1 is distinct from standard results for kernel regression estimators. Our proof technique, which focuses on estimation error for the "optimal" coefficients $\alpha_*$, allows us to account for sparsity in $M_*$ in a direct way using the proof strategy from Negahban et al. (2012).

Next, we illustrate how judicious choice of kernel can improve estimation of means for all category combinations—including those not observed in the training data.

*Lemma 1* (Statistical versus approximation error). For any $\boldsymbol{v} \in \mathcal{C}$, we have $\|\widehat{g}(\boldsymbol{v}) - g_*(\boldsymbol{v})\|_2 \leq \inf_{w \in \mathbb{R}^{\widetilde{n}}}\{h_w^{g_*}(\boldsymbol{v}) + h_w^{\phi}(\boldsymbol{v})\} + \|(\widehat{\alpha} - \alpha_*)^\top k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{v})\|_2$, where $h_w^{g_*}(\boldsymbol{v}) := \|g_*(\boldsymbol{v}) - \sum_{i=1}^{\widetilde{n}} w_i g_*(\widetilde{\boldsymbol{y}}_i)\|_2$ and $h_w^{\phi}(\boldsymbol{v}) := \|\alpha_*^\top\{\sum_{i=1}^{\widetilde{n}} w_i k_{\widetilde{\boldsymbol{Y}}}(\widetilde{\boldsymbol{y}}_i) - k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{v})\}\|_2$. If $\boldsymbol{v} \in \widetilde{\boldsymbol{Y}}$, then $\inf_{w \in \mathbb{R}^{\widetilde{n}}}\{h_w^{g_*}(\boldsymbol{v}) + h_w^{\phi}(\boldsymbol{v})\} = 0$.

The generic error bound from Lemma 1 can be decomposed into two parts: approximation error, $\inf_{w \in \mathbb{R}^{\widetilde{n}}}\{h_w^{g_*}(\cdot) + h_w^{\phi}(\cdot)\}$, and statistical error, $\|(\widehat{\alpha} - \alpha_*)^\top k_{\widetilde{\boldsymbol{Y}}}(\cdot)\|_F$. The approximation error can be further decomposed into two pieces, represented by $h_w^{g_*}$ and $h_w^{\phi}$. The magnitude of $h_w^{g_*}(\boldsymbol{v})$ quantifies how well we can approximate $g_*(\boldsymbol{v})$ with any linear combination of the $\{g_*(\widetilde{\boldsymbol{y}}_i)\}_{i=1}^{\widetilde{n}}$. If we observe a sufficiently large number of response category combinations in our training data, we could expect there to exist $w$ such that this term is small. However, $h_w^{g_*}$ cannot be disentangled from $h_w^{\phi}$. The term $h_w^{\phi}$ reflects the quality of our choice of kernel function $k$. In particular, we can write $h_w^{\phi}(\boldsymbol{v}) = \|\sum_{\ell=1}^{\widetilde{n}} \alpha_{\ell,\cdot}^* \langle \sum_{i=1}^{\widetilde{n}} w_i \phi(\widetilde{\boldsymbol{y}}_i) - \phi(\boldsymbol{v}), \phi(\widetilde{\boldsymbol{y}}_\ell)\rangle\|_2$, which will be small if $\sum_{i=1}^{\widetilde{n}} w_i \phi(\widetilde{\boldsymbol{y}}_i) - \phi(\boldsymbol{v})$ is small. Ideally, we could select a kernel $k$ (and consequently, $\phi$) for which there exists a $w$ such that $\|\sum_{i=1}^{\widetilde{n}} w_i \phi(\widetilde{\boldsymbol{y}}_i) - \phi(\boldsymbol{v})\|_2$ and $\|\sum_{i=1}^{\widetilde{n}} w_i g_*(\widetilde{\boldsymbol{y}}_i) - g_*(\boldsymbol{v})\|_2$ are both small. The optimal choice of kernel, then, is one in which both $\sum_{i=1}^{\widetilde{n}} w_i \phi(\widetilde{\boldsymbol{y}}_i) = \phi(\boldsymbol{v})$ and $\sum_{i=1}^{\widetilde{n}} w_i g_*(\widetilde{\boldsymbol{y}}_i) = g_*(\boldsymbol{v})$ for a single vector $w$ for all $\boldsymbol{v} \in \mathcal{C}$.

Finally, we apply Lemma 1 to establish the following.

*Theorem 2* (Out-of-sample mean estimation). Suppose Assumption 1 and 2 hold. If $\lambda = 4\omega_*(\|K_{\widetilde{\boldsymbol{Y}}}\|_F + \|K_{\widetilde{\boldsymbol{Y}}}\|\sqrt{2c_2 \log p})/n$, then for any $\boldsymbol{v} \in \mathcal{C}$, with probability at least $1 - p^{1-c_2}$

$$\|\widehat{g}(\boldsymbol{v}) - g_*(\boldsymbol{v})\|_2 \leq \|g_*(\boldsymbol{v}) - M_*^{\dagger\top} K_{\widetilde{\boldsymbol{Y}}}^{\dagger\,-1} k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{v})\|_2 + \left[\frac{6\|k_{\widetilde{\boldsymbol{Y}}}(\boldsymbol{v})\|_2 \omega_* \sqrt{s}}{\kappa_{\mathcal{S}}} \left(\frac{\sqrt{\widetilde{n}_{\max}}}{\widetilde{n}_{\min}}\right) \times \left\{\frac{\|K_{\widetilde{\boldsymbol{Y}}}^\dagger\|_F}{\widetilde{n}} + \left(\frac{\|K_{\widetilde{\boldsymbol{Y}}}^\dagger\|}{\widetilde{n}}\right)\sqrt{2c_2 \log p}\right\}\right].$$

The result of Theorem 2 follows from the fact that under Assumption 1, there always exists a $w$ such that $h_w^{\phi}(\boldsymbol{v}) = 0$. Proofs can be found in the supplementary material.

## 6. Simulation Studies

### 6.1. Data Generating Models and Competing Method

In this section, we illustrate the performance of our method through simulation studies. We compare our estimators to competitors under a variety of data generating models. For 100 independent replications under each setting, we first generate $n = 200$ independent responses $(Y_1, \ldots, Y_M)$. To do so, we generate a $c^\star$-variate vector which has independent Uniform(0,1) components, say $u \in (0,1) \times \cdots \times (0,1)$, then divide by its sum so that $\pi_* = u/\sum_{j=1}^{c^\star} u_j$ belongs to the $(c^\star - 1)$-dimensional probability simplex. Then, we generate realizations of $(Y_1, \dots, Y_M)$, denoted $\boldsymbol{y}$, from the categorical distribution with probabilities $\pi_*$. Thus, the components of the response are marginally dependent with arbitrary dependencies.

Given $\boldsymbol{y}$, we then generate $x$ from the $p$-dimensional multivariate normal distribution (2). In each scenario, we set $\Omega^{-1}_{*s,t} = 0.7^{|s-t|}$ for all $(s,t) \in [p] \times [p]$. We consider two different models (Models A and B) for the mean vectors determined by $g_*$. Specifically, the two models differ in terms of how the mean vectors from (2) depend on the response categories. For both models, we vary $p$ and a parameter controlling the difficulty of classification.

- **Model A.** We randomly select 10 distinct elements of $[p]$, say $\{k_1, \dots, k_{10}\}$ and set, for $k \in [p]$, $g_{*k}(\boldsymbol{y}) = b_\ell^\top \boldsymbol{y}$ if $k = k_\ell$ for some $\ell \in \{1, \dots, 10\}$ and 0 otherwise. where $b_\ell \in \{0, 2\}^M$ with the collection $\{b_\ell\}_{\ell=1}^{10}$ having $10M/2$ components equal to two and $10M/2$ equal to zero in randomly chosen positions.

Note that here (and in Model B), we use the numeric form $\boldsymbol{y} \in [c_1] \times \cdots [c_M]$ so that $b_\ell^\top \boldsymbol{y} \in \mathbb{R}$.

Model A is ideal for the nonconvex estimator: only ten elements of the $\mu_{*\boldsymbol{v}}$ differ as a function of $\boldsymbol{v}$, which the nonconvex estimator is designed to exploit. The convex estimator, on the other hand, exploits sparsity in the collection of discriminant vectors $\Omega_*(\mu_{*\boldsymbol{v}} - \mu_{*\boldsymbol{v}'})$ for $\boldsymbol{v} \neq \boldsymbol{v}'$. Under Model A, $\Omega_*(\mu_{*\boldsymbol{v}} - \mu_{*\boldsymbol{v}'})$ can have as many as 30 nonzero elements because $\Omega_*$ is tridiagonal. Model B, in contrast, imposes sparsity on the discriminant vectors directly.

- **Model B.** We randomly select 10 distinct elements of $[p]$, say $\{k_1, \dots, k_{10}\}$ and set, for $k \in [p]$, $[\beta_{*\boldsymbol{y}}]_k = b_\ell^\top \boldsymbol{y}/\nu$ if $k = k_\ell$ for some $\ell \in \{1, \dots, 10\}$ and 0 otherwise, where $b_\ell \in \{0, 2\}^M$ with the collection $\{b_\ell\}_{\ell=1}^{10}$ having $10M/2$ components equal to two and $10M/2$ equal to zero in randomly chosen positions. Then, we set $g_*(\boldsymbol{v}) = \Omega_*^{-1}\beta_{*\boldsymbol{v}}$ so that $\beta_{*\boldsymbol{v}} - \beta_{*\boldsymbol{v}'} = \Omega_*(\mu_{*\boldsymbol{v}} - \mu_{*\boldsymbol{v}'})$.

Under Model B, all components of $g_*$ can differ as a function of the response categories. However, only ten variables are relevant for classification. Model B is thus ideal for (10).

We consider two versions of each model. In **Model A-4** and **Model B-4**, we set $M = 4$ and $c_1 = \cdots = c_4 = 3$; in **Model A-6** and **Model B-6**, we set $M = 6$ and $c_1 = \cdots = c_6 = 2$. Throughout our simulations, we will consider $p \in \{50, 100, 150\}$ and $\nu \in \{0.8, 1.0, 1.2, 1.4\}$. Under both models, $\nu$ controls the difficulty of the classification problem. If $\nu$ is small, differences between category combination means are large, so the problem is easier.

We compare our two estimators: (8) (KLDA-M) and (10) (KLDA-D)—both using Hamming distance kernel—to competitors that either fit separate models for each response, or formulate a synthetic (univariate) categorical response and fit a singular model (i.e., the aggregate model). The first competitor is the separate multinomial logistic regression estimator (S-Logistic), which fits a separate group-lasso penalized multinomial logistic regression model for each response. We also consider an aggregate version, A-Logistic, which fits a single group-lasso penalized multinomial logistic regression model to the synthetic $c^\star = \prod_{\ell=1}^{M} c_m$-category response. Note that if a category combination is not observed in the training data, this method sets its conditional probability to zero. In addition, we consider separate multiclass sparse discriminant analysis models (Mai, Yang, and Zou 2019, S-MSDA), and an aggregate multiclass sparse discriminant analysis model A-MSDA. The latter correctly specifies the LDA model from which we generate data. Finally, we also compared to oracle, which uses the true parameters from (2) plugged into Bayes' classification rule. This serves as an upper bound for the performance of any method, but is not available in practice.

The first performance metric we considered is the mean estimation error. Since none of the competitors are capable of estimating the means corresponding different categories, we compare the mean estimation error of the sample mean (i.e., the MLE) with the mean estimates obtained using our methods. Specifically we display mean estimation error, defined as $\frac{1}{\widetilde{n}} \sum_{i=1}^{\widetilde{n}} \|g_*(\widetilde{\boldsymbol{y}}_i) - \bar{x}_{\widetilde{\boldsymbol{y}}_i}\|_2^2$ for the MLE, and $\frac{1}{\widetilde{n}} \sum_{i=1}^{\widetilde{n}} \|g_*(\widetilde{\boldsymbol{y}}_i) - \widehat{g}(\widetilde{\boldsymbol{y}}_i)\|_2^2$ for the KLDA variants, where $\bar{x}_{\widetilde{\boldsymbol{y}}_i} = \frac{1}{\sum_{i=1}^{n} \mathbf{1}(\boldsymbol{y}_i = \widetilde{\boldsymbol{y}}_i)} \sum_{i=1}^{n} x_i \mathbf{1}(\boldsymbol{y}_i = \widetilde{\boldsymbol{y}}_i)$ is the sample mean corresponding to $\widetilde{\boldsymbol{y}}_i$. The other two performance metrics we considered are prediction accuracy and Hamming distance, which are defined as $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\boldsymbol{y}_i = \widehat{\boldsymbol{y}}_i)$, and $\frac{1}{nM} \sum_{i=1}^{n} \sum_{\ell=1}^{M} \mathbf{1}(\boldsymbol{y}_{i\ell} = \widehat{\boldsymbol{y}}_{i\ell})$, respectively.

## 6.2. Results

In Table 1, we present the average mean estimation error results with varying parameters $p$, $\nu$ and different models. Notably, our proposed estimators result in considerably smaller errors in mean estimation compared to the sample mean estimates across all considered scenarios. The nonconvex estimator KLDA-M, which directly estimates the mean vectors, does not invariably outperform the convex estimator KLDA-D, which approximates the discriminant vectors directly. Under Model A, the mean vectors are sparse, and KLDA-M is able to exploit this, whereas KLDA-D is not. Conversely, under Model B, it is the discriminant vectors that are sparse, whereas the mean vectors are nonsparse. Naturally, this favors KLDA-D in the mean estimation.

In Figures 1 and 2, we display the prediction accuracy on testing set across various models. These figures clearly demonstrate our proposed estimators, KLDA-M and KLDA-D, have superior prediction accuracy relative to the competitors, particularly noticeable with a larger number of responses as in Model A-6 and Model B-6.

For Model A-4, KLDA-M and KLDA-D have comparable performance when $p$ is 50 or 100. However, when $p$ is increased to 150, KLDA-M clearly outperforms KLDA-D. This coheres with expectations, given that KLDA-M leverages the sparsity of the mean vectors. In contrast, the discriminant vectors, onto which KLDA-D imposes sparsity, are nonsparse in this this scenario, so the regularization scheme of KLDA-D may impose unhelpful bias. Consequently, S-MSDA mirrors the performance of KLDA-D under these conditions. Under Model B, KLDA-D

**Table 1.** Average mean estimation errors for the MLE versus KLDA averaged over 100 independent replications under Model A and Model B.

| $p$ | $\nu$ | Model A-4 | | | Model A-6 | | | Model B-4 | | | Model B-6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MLE | KLDA-M | KDLA-D | MLE | KLDA-M | KLDA-D | MLE | KLDA-M | KLDA-D | MLE | KLDA-M | KLDA-D |
| 50 | 0.8 | 24.475 | 1.149 | 2.176 | 21.072 | 1.022 | 1.747 | 24.475 | 2.945 | 2.174 | 21.072 | 2.636 | 1.864 |
| | 1.0 | 24.346 | 0.945 | 2.262 | 20.737 | 0.787 | 1.703 | 24.346 | 2.815 | 2.044 | 20.737 | 2.153 | 1.759 |
| | 1.2 | 24.361 | 0.854 | 2.314 | 20.950 | 0.716 | 1.731 | 24.361 | 2.833 | 2.130 | 20.950 | 2.161 | 1.920 |
| | 1.4 | 24.364 | 0.813 | 2.359 | 20.970 | 0.657 | 1.741 | 24.364 | 2.626 | 2.131 | 20.970 | 2.192 | 1.851 |
| 100 | 0.8 | 48.267 | 1.435 | 4.301 | 41.685 | 1.299 | 3.247 | 48.267 | 5.228 | 3.646 | 41.685 | 4.607 | 3.456 |
| | 1.0 | 48.955 | 1.255 | 4.460 | 42.290 | 1.028 | 3.252 | 48.955 | 4.944 | 3.541 | 42.290 | 4.259 | 3.358 |
| | 1.2 | 48.551 | 1.179 | 4.557 | 42.178 | 0.993 | 3.282 | 48.551 | 5.073 | 3.486 | 42.178 | 4.101 | 3.379 |
| | 1.4 | 48.670 | 1.088 | 4.498 | 41.878 | 0.983 | 3.156 | 48.670 | 4.828 | 3.316 | 41.878 | 4.299 | 3.165 |
| 150 | 0.8 | 72.536 | 1.657 | 6.310 | 62.511 | 1.519 | 4.637 | 72.536 | 7.714 | 5.119 | 62.511 | 6.011 | 4.777 |
| | 1.0 | 72.256 | 1.537 | 6.575 | 62.976 | 1.378 | 4.616 | 72.256 | 6.811 | 4.855 | 62.976 | 5.726 | 4.465 |
| | 1.2 | 72.193 | 1.465 | 6.498 | 63.267 | 1.309 | 4.547 | 72.193 | 6.726 | 4.639 | 63.267 | 5.658 | 4.302 |
| | 1.4 | 72.222 | 1.425 | 6.266 | 63.715 | 1.226 | 4.330 | 72.222 | 6.974 | 4.563 | 63.715 | 5.707 | 4.057 |

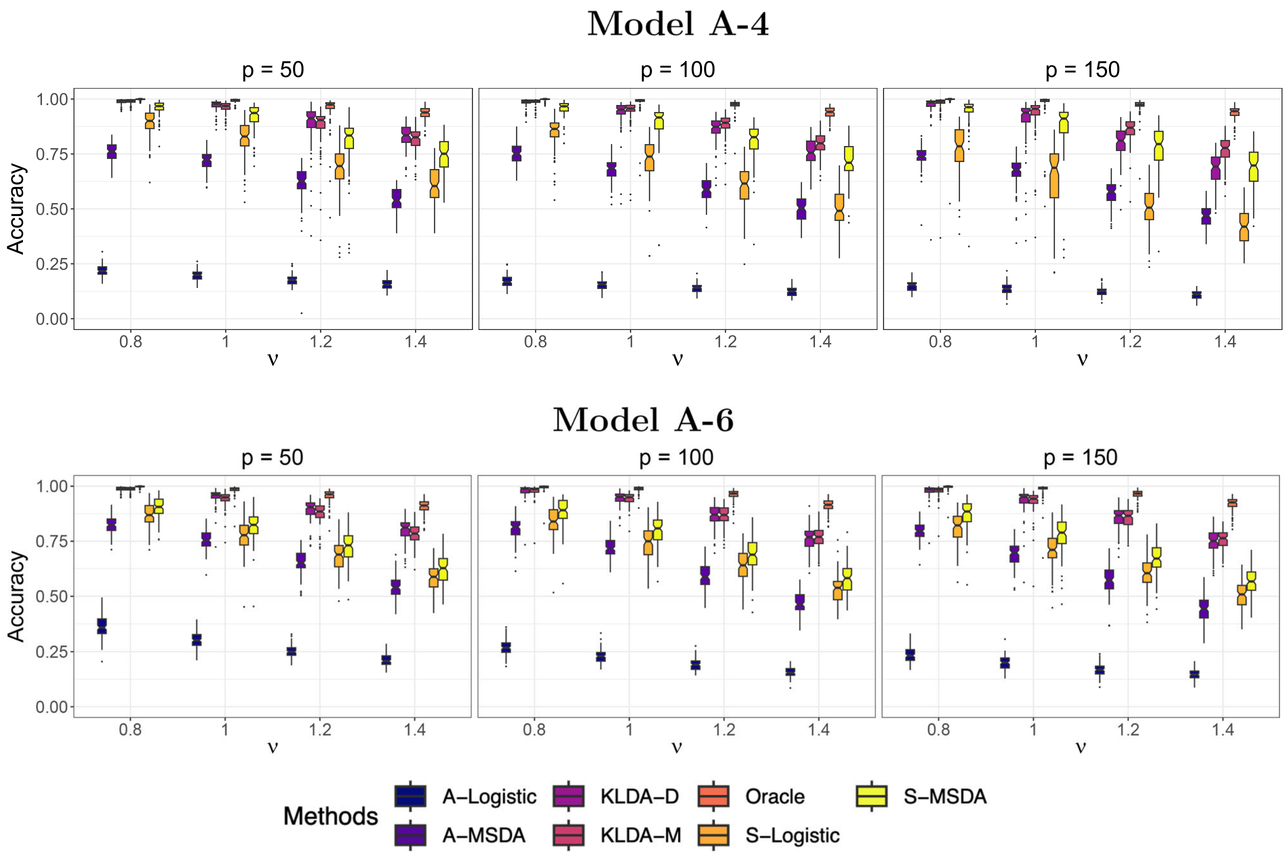


**Figure 1.** Prediction accuracy over 100 independent replications under **Model A-4** and **Model A-6** with $(p, \nu) \in \{50, 100, 150\} \times \{0.8, 1.0, 1.2, 1.4\}$..

consistently outperforms KLDA-M. This can be attributed to the fact that the mean vectors, which KLDA-M regularizes, are nonsparse. As expected, KLDA-D outperforms the other competitors as it exploits the sparsity of the discriminant vectors.

In the supplementary materials, we also include Hamming distance and variable selection results under all simulation settings. To summarize briefly, in terms of Hamming distance, we observe similar general trends as in Figures 1 and 2, except that S-MSDA can, at times, outperform our proposed methods. This can be understood from the fact that Hamming distance is inherently measuring quality of estimation of marginal probabilities, whereas our proposal is focused on estimation of the joint probability mass of $(Y_1, \dots, Y_M \mid X)$.

### 6.3. Additional Simulation Study Results

In the supplementary material, we include many additional numerical studies. In Section S.3.2, we compare various choices of kernel functions, and include results under simulation settings more amendable to the pair-agreement kernel. In Section S.3.3, we study the effect of including a squared Hilbert-norm regularization term to the objective function for MLDA-M. In Section S.3.4, we compare our method to fitting mixture

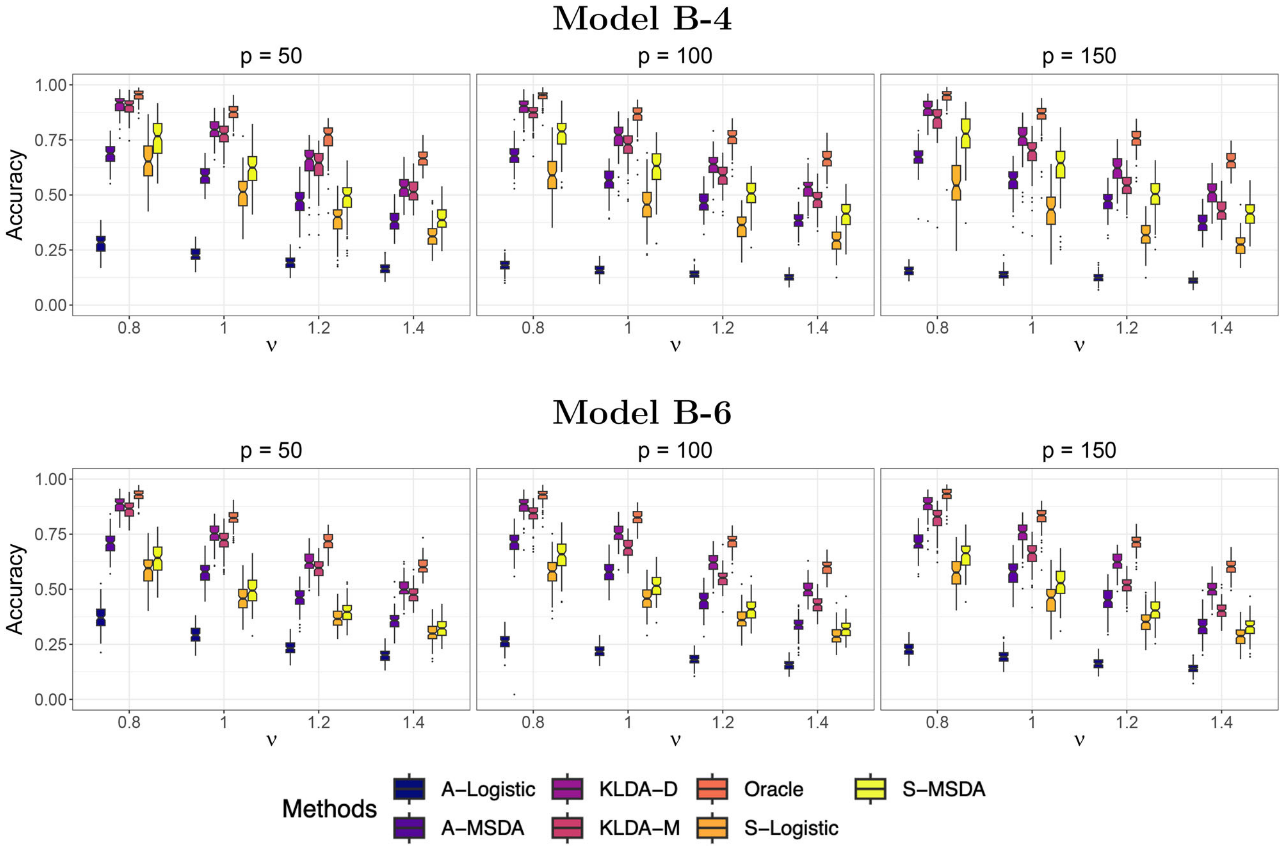


**Figure 2.** Prediction accuracy over 100 independent replications under **Model B-4** and **Model B-6** with $(p, \nu) \in \{50, 100, 150\} \times \{0.8, 1.0, 1.2, 1.4\}$.

discriminant analysis models to each response (Hastie and Tibshirani 1996).

## 7. Classification of Colon Tissue Samples

In this section, we demonstrate the application of our method on a dataset consisting of gene expression profiles from colon biopsies (Noble et al. 2008). This dataset, which can be downloaded from the Gene Expression Omnibus (GDS3268), contains 44,290 gene expression levels from 202 tissue samples. There are three labels for each sample: patient state (normal/ulcerative colitis), tissue state (inflamed/uninflamed) and anatomical locations (sigmoid colon/terminal ileum/descending colon/ascending colon).

Following our simulation studies, we analyze the data using the proposed `KLDA-M` and `KLDA-D`—with Hamming distance kernel—in conjunction with `S-Logistic`, `S-MSDA`, `A-Logistic` and `A-MDSA`. Results for our method with alternative kernels are provided in Section S.6 of the supplementary material. We partitioned the data by randomly selecting $n$ samples for the training set, allocating 50 samples for the validation set, and designating the remaining $152 - n$ samples for the testing set. To mitigate computational demands, we undertook a screening process on the genes. This involved ranking gene expression levels based on their median absolute deviation and subsequently selecting the top $p$ genes. To avoid issues of collinearity, genes exhibiting high correlation were pruned. For our analysis, we consider $p \in \{100, 200, 300, 400, 500\}$ and $n \in \{50, 100\}$.

The results based on 100 replicates are listed in Table 2. It can be seen that our proposed methods achieve the highest accuracy under most choices of $n$ and $p$, and `KLDA-M` outperforms `KLDA-D` in all but two settings. This underscores the potential advantage of regularizing mean vectors as opposed to discriminant vectors for this particular problem.

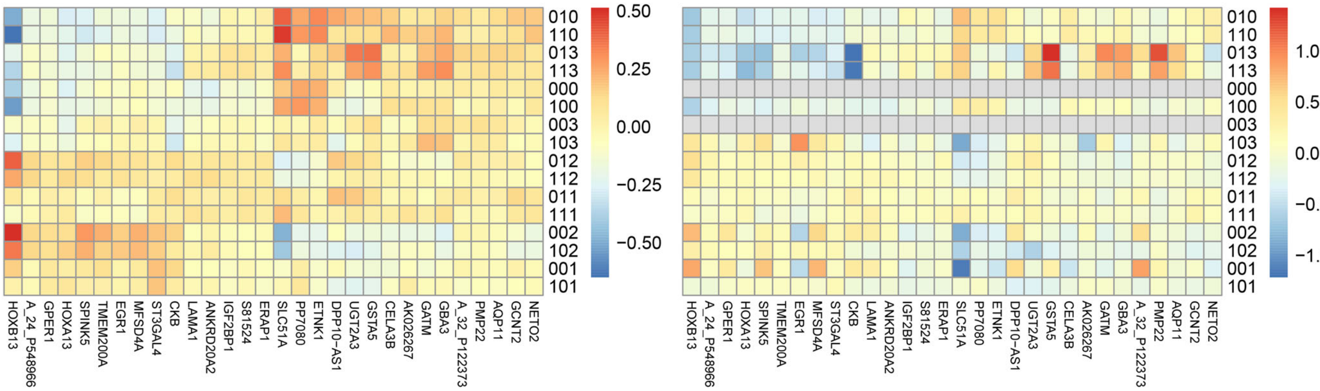
Next, we turn our attention to the mean estimation using `KLDA-M`. For this analysis, we set $p = 200$, allocate 152 samples for training, and select tuning parameters with the remaining 50 samples. The results are shown in Figure 3, where we only include 30 genes to save space. Our fitted model estimated 111 genes' means varied as a function of the response category combinations. The three-digit numbers along the rows denote distinct combinations of response labels. The first digit represents the patient state: 0 for normal and 1 for ulcerative colitis. The second digit represents the tissue's state, with 0 indicating inflamed and 1 denoting uninflamed. The final digit represents anatomical locations: 0 for the ascending colon, 1 for the descending colon, 2 for the sigmoid colon, and 3 for the terminal ileum. It is important to note that both the estimated and sample mean vectors have been centralized; we've subtracted the global mean from each. Note that the combinations of response categories 000 (normal patient, inflamed tissue, ascending colon) and 003 (normal patient, inflamed tissue, terminal ileum) are not observed in the dataset, so we don't have corresponding sample mean estimates.

Our estimates reveal a more distinct pattern of gene expression levels across various response category combinations compared to the sample mean estimates. Notably, genes such as

**Table 2.** Prediction accuracy on GDS3268 dataset over 100 independent replications with $p \in \{100, 200, 300, 400, 500\}$ and $n \in \{50, 100\}$.

| $n$ | $p$ | KLDA-M | KLDA-D | S-Logistic | S-MSDA | A-Logistic | A-MSDA |
|---|---|---|---|---|---|---|---|
| | 100 | **0.404** | 0.397 | 0.389 | 0.392 | 0.321 | 0.322 |
| | 200 | **0.425** | 0.422 | 0.408 | 0.405 | 0.348 | 0.335 |
| 50 | 300 | **0.431** | 0.429 | 0.394 | 0.397 | 0.333 | 0.332 |
| | 400 | 0.434 | **0.438** | 0.400 | 0.408 | 0.323 | 0.324 |
| | 500 | 0.446 | **0.450** | 0.390 | 0.397 | 0.332 | 0.322 |
| | 100 | 0.446 | 0.436 | **0.479** | 0.472 | 0.388 | 0.387 |
| | 200 | 0.503 | 0.491 | **0.505** | 0.486 | 0.445 | 0.404 |
| 100 | 300 | **0.543** | 0.529 | 0.488 | 0.491 | 0.424 | 0.408 |
| | 400 | **0.535** | 0.525 | 0.487 | 0.505 | 0.403 | 0.410 |
| | 500 | **0.547** | 0.537 | 0.488 | 0.479 | 0.412 | 0.406 |

NOTE: When $n = 50$, standard errors were never larger than 0.007; when $n = 100$ standard errors were never larger than 0.009. Bolded values indicate the highest average prediction accuracy within a row.



(a) Estimated Mean

(b) Sample Mean

**Figure 3.** Mean estimates (minus the columnwise global average) using `KLDA-M` and the MLE with $p = 200$. Each column corresponds to a gene and we include 30 genes. Each row corresponds to a combination of response categories.

HOXB13, HOXA13, and CKB exhibit lower expression levels in ascending colon and terminal ileum and heightened levels in descending colon and sigmoid colon. Conversely, genes like SLC51A, ETNK1, and UGT2A3 demonstrate an inverse pattern. Regarding HOXB13, normal patients have a higher expression level relative to those with ulcerative colitis.

## 8. Discussion

Due to space limitations, there are many aspects of our method that we could not discuss in this manuscript. In the supplementary materials Section S.7.1, we contrast our method to using a multiway ANOVA model to estimate the mean function $g_*$. In brief, though this approach can provide flexible and interpretable estimates of $g_*$, there are numerous practical issues that our method avoids. We also argue that our method is more scalable with large $M$ and $c_m$. In Section S.7.2, we discuss how our work relates to methods proposed for "dynamic linear discriminant analysis" (Jiang, Chen, and Leng 2020; Jiang et al. 2021). In particular, the method of Jiang et al. (2021), though motivated from a different perspective from our own, essentially estimates $g_*$ and $\Sigma_*$ using kernel smoothing. The discriminant vectors $\beta_{*v}$ are then estimated in a second step by solving penalized quadratic programs.

## Supplementary Materials

The supplementary materials contains extended discussions, computational details, proofs, additional results from the colon tissue data analysis, and extended simulation study results. Also included in the Supplementary Material is code to reproduce many of the results from this article.

## Acknowledgments

The authors would like to thank the Associate Editor and reviewers for helpful comments.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Funding

X. Zhang's and A. J. Molstad's contributions were supported by NSF DMS-2113590 and DMS-2415067, respectively.

## References

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–153. [3]

——— (2012), *Categorical Data Analysis* (Vol. 792), Hoboken, NJ: Wiley. [2]

Beck, A., and Teboulle, M. (2009), "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, 2, 183–202. [6]

Cai, T., and Liu, W. (2011), "A Direct Estimation Approach to Sparse Linear Discriminant Analysis," *Journal of the American Statistical Association*, 106, 1566–1577. [3]

Deng, K., Zhang, X., and Molstad, A. J. (2024), "Multi-Response Linear Discriminant Analysis in High Dimensions," *Journal of Machine Learning Research*, 25, 1–66. [2]

Ekholm, A., McDonald, J. W., and Smith, P. W. (2000), "Association Models for a Multivariate Binary Response," *Biometrics*, 56, 712–718. [1]

Elman, M. R., Minnier, J., Chang, X., and Choi, D. (2020), "Noise Accumulation in High Dimensional Classification and Total Signal Index," *The Journal of Machine Learning Research*, 21, 1383–1405. [5]

Fan, J., and Fan, Y. (2008), "High Dimensional Classification Using Features Annealed Independence Rules," *Annals of Statistics*, 36, 2605. [5]

Glonek, G. F. (1996), "A Class of Regression Models for Multivariate Categorical Responses," *Biometrika*, 83, 15–28. [1]

Glonek, G. F., and McCullagh, P. (1995), "Multivariate Logistic Models," *Journal of the Royal Statistical Society*, Series B, 57, 533–546. [1]

Hastie, T., and Tibshirani, R. (1996), "Discriminant Analysis by Gaussian Mixtures," *Journal of the Royal Statistical Society*, Series B, 58, 155–176. [3,10]

Jiang, B., Chen, Z., and Leng, C. (2020), "Dynamic Linear Discriminant Analysis in High Dimensional Space," *Bernoulli*, 26, 1234–1268. [11]

Jiang, B., Leng, C., Wang, C., Yang, Z., and Yu, X. (2021), "Linear Discriminant Analysis with High-Dimensional Mixed Variables," arXiv preprint arXiv:2112.07145. [11]

Lang, J. B. (1996), "Maximum Likelihood Methods for a Generalized Class of Log-Linear Models," *The Annals of Statistics*, 24, 726–752. [1]

Lupparelli, M., and Roverato, A. (2017), "Log-Mean Linear Regression Models for Binary Responses with an Application to Multimorbidity," *Journal of the Royal Statistical Society*, Series C, 66, 227–252. [1]

Mai, Q., Yang, Y., and Zou, H. (2019), "Multiclass Sparse Discriminant Analysis," *Statistica Sinica*, 29, 97–111. [3,8]

Mai, Q., Zou, H., and Yuan, M. (2012), "A Direct Approach to Sparse Discriminant Analysis in Ultra-High Dimensions," *Biometrika*, 99, 29–42. [3]

Molenberghs, G., and Lesaffre, E. (1999), "Marginal Modelling of Multivariate Categorical Data," *Statistics in Medicine*, 18, 2237–2255. [1]

Molstad, A. J., and Rothman, A. J. (2018), "Shrinking Characteristics of Precision Matrix Estimators," *Biometrika*, 105, 563–574. [3]

——— (2023), "A Likelihood-based Approach for Multivariate Categorical Response Regression in High Dimensions," *Journal of the American Statistical Association*, 118, 1402–1414. [2]

Molstad, A. J., and Zhang, X. (2022), "Conditional Probability Tensor Decompositions for Multivariate Categorical Response Regression," arXiv:2206.10676. [2]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of m-estimators with Decomposable Regularizers," *Statistical Science*, 27, 538–557. [7]

Noble, C. L., Abbas, A. R., Cornelius, J., Lees, C. W., Ho, G.-T., Toy, K., Modrusan, Z., Pal, N., Zhong, F., Chalasani, S., et al. (2008), "Regional Variation in Gene Expression in the Healthy Colon is Dysregulated in Ulcerative Colitis," *Gut*, 57, 1398–1405. [10]

Parikh, N., and Boyd, S. (2014), "Proximal Algorithms," *Foundations and Trends in Optimization*, 1, 127–239. [6]

Park, C. H., and Lee, M. (2008), "On Applying Linear Discriminant Analysis for Multi-Labeled Problems," *Pattern Recognition Letters*, 29, 878–887. [3]

Polson, N., Scott, J. G., and Willard, B. T. (2015), "Proximal Algorithms in Statistics and Machine Learning," *Statistical Science*, 30, 559–581. [6]

Price, B. S., Geyer, C. J., and Rothman, A. J. (2015), "Ridge Fusion in Statistical Learning," *Journal of Computational and Graphical Statistics*, 24, 439–454. [3,6]

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2009), "Classifier Chains for Multi-Label Classification," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, pp. 254–269, Springer. [1]

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2021), "Classifier Chains: A Review and Perspectives," *Journal of Artificial Intelligence Research*, 70, 683–718. [1]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [3,5]

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001), "A Generalized Representer Theorem," in *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, pp. 416–426, Springer. [4]

Schölkopf, B., and Smola, A. J. (2002), *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge, MA: MIT Press. [3]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [6]

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Vol. 48), Cambridge: Cambridge University Press. [4,7]

Wang, H., Ding, C., and Huang, H. (2010), "Multi-Label Linear Discriminant Analysis," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*, pp. 126–139, Springer. [3]

Williams, C., and Seeger, M. (2000), "Using the Nyström Method to Speed Up Kernel Machines," in *Advances in Neural Information Processing Systems* (Vol. 13). [4]

Witten, D. M., and Tibshirani, R. (2009), "Covariance-Regularized Regression and Classification for High Dimensional Problems," *Journal of the Royal Statistical Society*, Series B, 71, 615–636. [3,6]

Xu, P., Zhu, J., Zhu, L., and Li, Y. (2015), "Covariance-Enhanced Discriminant Analysis," *Biometrika*, 102, 33–45. [3,5]

Yu, G., and Bien, J. (2019), "Estimating the Error Variance in a High-Dimensional Linear Model," *Biometrika*, 106, 533–546. [6]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [6]

Zhang, M.-L., and Zhou, Z.-H. (2013), "A Review on Multi-Label Learning Algorithms," *IEEE Transactions on Knowledge and Data Engineering*, 26, 1819–1837. [1]

Zhu, Y. (2020), "A Convex Optimization Formulation for Multivariate Regression," *Advances in Neural Information Processing Systems*, 33, 17652–17661. [6]