A generative approach to modeling data with quantitative and qualitative responses

Xiaoning Kang[a], Lulu Kang[b], Wei Chen[c], Xinwei Deng[*,d]

[a]*Institute of Supply Chain Analytics and International Business College, Dongbei University of Finance and Economics, Dalian, China*
[b]*Department of Applied Mathematics, Illinois Institute of Technology, Chicago, USA*
[c]*Department of Mechanical, Materials & Aerospace Engineering, Illinois Institute of Technology, Chicago, USA*
[d]*Department of Statistics, Virginia Tech, Blacksburg, USA*

Abstract

In many scientific areas, data with mixed quantitative and qualitative (QQ) responses are commonly encountered with a large number of predictors. By exploring the association between QQ responses, existing approaches often consider a joint model of QQ responses given the predictor variables. However, the dependency among predictive variables also provides useful information for fitting QQ responses. Hence in this work, we propose a novel generative approach to jointly model the QQ responses by incorporating the dependency information of predictors. The proposed method is computationally efficient and provides accurate parameter estimation under a penalized likelihood framework. Moreover, because of the generative approach framework, the asymptotically theoretical results of the proposed method are established under some regularity conditions. The performance of the proposed method is examined through simulations and real case studies in material science and genetics.

Key words: Classification, Discriminant analysis, Graphical lasso, Regression, Regularization.
2020 MSC: Primary 62H30, Secondary 62H12

1. Introduction

In supervised learning, analyzing data with heterogeneous types of responses has been an important topic with broad applications. Most often, such heterogeneous data involve both quantitative and qualitative (QQ) responses. For example, Klein et al. [25] described a human health study examining the risk factors of adverse birth outcomes, which contains a qualitative response "presence/absence of low birth weight" and a quantitative response "gestational age". In material science, the properties of a material are often characterized by QQ measures. For example, in the case study of Section 4.3 on Heusler compounds, two metrics, the "mixing enthalpy" (quantitative) and the "global stability based on hull energy" (qualitative) are used to determine the thermodynamic stability of a full Heusler compound. In this paper, we develop a generative approach for joint modeling of two mixed responses: one is a continuous quantitative response and the other is a multi-class qualitative response.

In the literature, it has been recognized that modeling each response separately would overlook the relationship between the responses, hence possibly leading to inaccurate estimation and prediction. In contrast, joint modeling considers and utilizes the information on the association of different responses, enhancing the model interpretation and performance especially in high-dimensional cases [1, 2, 13, 21, 22, 31]. A major difficulty, however, in the joint modeling of such responses is the lack of a natural multivariate distribution. To overcome the difficulty and explore the association between QQ responses, many researches have contributed to this area including early ones such as [14, 16, 18] and recent ones such as [13, 24–26, 34]. Based on the different methodologies, the existing works can be generally grouped into two categories.

The first group of methods considers a factorization of the joint distribution into the product of a marginal and a conditional distribution. Naturally, two possible types of models can be applied, depending on whether the conditional

*Corresponding author. Email address: xdeng@vt.edu

variable is the quantitative or the qualitative response. Note that using which response as the conditional variable leads to different model parameterizations, and consequently has different challenges in parameter estimation and model prediction. In some applications, data are suitable to use the quantitative response as the conditional variable, as in [13] and [21]. In their case study, the data are collected from the lapping stage of the wafer manufacturing process and their primary goal is to build a prediction model for the quantitative response measuring the total thickness variation of the wafer. On the other hand, some applications require the focus to be on the qualitative response. For example, in Kang et al. [23], the quantitative response data are the weights of infants and the qualitative response is whether the birth of an infant is preterm. The latter is a more crucial statistic to the public health expert in their team. But it is more challenging to directly use the qualitative response as the conditional variable, thus Kang et al. [23] introduced latent variable to facilitate such modeling. There have been many other works using the factorization approach. Fitzmaurice and Laird [16] analyzed data from developmental toxicity studies by introducing a conditional regression model of quantitative response conditioned on the qualitative response, and employing the logistic model for the qualitative response. Lin et al. [31] developed a conditional mixed-effects model to analyze clustered data containing QQ responses. Craiu and Sabeti [11] suggested a Bayesian conditional copula model to fit bivariate data with mixed outcomes. These methods are suitable for data with a small number of predictor variables. To handle the high-dimensional data with large number of predictors, Deng and Jin [13] proposed a conditional model that encourages model sparsity through a penalized and constrained likelihood function. However, the corresponding inferences and asymptotic properties of their method cannot be explored due to the complicated constrained likelihood estimation. Kang et al. [21] introduced a Bayesian estimation for the conditional model of Deng and Jin [13] to obtain proper inferences on the model parameters. Nevertheless, their model is not designated for studying the asymptotic properties of the proposed estimator. More related works can be found in [8, 19], among others.

The second group of methods considers a continuous latent variable for the qualitative response, and then assumes a multivariate normal distribution for the latent variable and the quantitative response. For example, Gueorguieva and Agresti [18] studied a probit model for binary response with a latent variable and developed a Monte Carlo expectation-conditional maximization algorithm for parameter estimation. Kürüm et al. [26] used a normal latent variable for characterizing the binary response and suggested a two-stage estimation procedure for analysis of ecological momentary assessment data collected in a smoking cessation study. Klein et al. [25] introduced the idea of latent variable into the framework of copula regressions, constructing a latent continuous representation of binary regression models. However, the use of latent variables often involves considerable computation in the parameter estimation. It also makes the investigation of theoretical properties difficult. Moreover, most of these works focus on the binary qualitative response and their model assumptions may not be easily extended to the multi-class qualitative response cases.

Although various methods have been developed in the literature for joint modeling of QQ responses, they mostly focus on a regression model conditioned on the predictor variables without fully utilizing the dependency information between predictors. Furthermore, to our best knowledge, few works established theoretical results for the QQ responses, since it is a very challenging task under the framework of either conditional models or models characterized by latent variables. In our work, we propose a novel method to jointly model the QQ responses based on the *generative approach*. The proposed generative model considers the joint distribution of the high-dimensional predictor variables, the quantitative response, and the multi-class qualitative response. It is a very unique and different perspective from the existing literature, which enables us to establish the theoretical results for both QQ responses. In addition, the proposed method can easily accommodate multi-class qualitative response and multivariate quantitative responses with attractive theoretical properties. For short, we call the proposed method *GAQQ*, a Generative Approach for QQ responses.

The key contributions of this work are summarized as follows. First, based on the generative model framework, we are able to establish the asymptotic properties of the proposed estimators with respect to both the classification accuracy for the qualitative response and the prediction accuracy for the quantitative response under some regularity conditions. Such conditions are commonly used in the regularized estimation framework [42, 49]. The classification of the qualitative response enjoys the asymptotic optimality of the resulting linear discriminate classification rule. The mean squared error (MSE) of prediction for the quantitative response is as good as the optimal prediction under the Bayes risk. Second, an efficient procedure for parameter estimation is developed via the regularized log-likelihood function of the joint distribution of predictor variables and QQ responses. Specifically, we impose regularization on both the mean differences and the inverse covariance matrix from the joint distribution to achieve sparsity for high-

dimensional predictor variables. Third, the use of the generative approach leads to an effective prediction procedure by inferring the conditional distribution of QQ responses conditioned on the predictor variables. That is, the quantitative response is predicted through the property of conditional multivariate normal distribution, and the linear discriminant analysis (LDA) is employed for classification of the qualitative response. Fourth, the proposed generative model enables the QQ responses to be mutually learned from each other to improve the predictions on both QQ variables, which is different from existing methods in which the prediction of only one type of QQ responses can be benefited from the information of the other type.

The remainder of this paper is organized as follows. Section 2 details the proposed method. The main theoretical results and numerical studies are presented in Sections 3 and 4, respectively. Section 5 concludes this work with some discussion.

2. The proposed methodology

In this section, we lay out the proposed GAQQ model in terms of both estimation and prediction procedures. Section 2.1 focuses on modeling the QQ data with binary qualitative response, and Section 2.2 extends the GAQQ model to deal with the qualitative response with multiple classes.

2.1. The GAQQ model for two-class qualitative response

Suppose the variables of interest are denoted by $(\boldsymbol{X}, y, Z)$ where $\boldsymbol{X} = (X_1, \ldots, X_{p-1})^\top$ is a $(p-1)$ dimensional vector of predictor variables, y is a quantitative response variable and $Z \in \{1, 2\}$ is a qualitative response variable. From a generative modeling perspective, we consider the data generation mechanism as $p(\boldsymbol{X}, y, Z) = p(\boldsymbol{X}, y|Z)p(Z)$, indicating that data are from two classes G_1 and G_2 under $(\boldsymbol{X}, y)|Z$, where $p(\cdot)$ represents a probability density function throughout the paper. Assume that $\boldsymbol{W} = (\boldsymbol{X}^\top, y)^\top$ follows multivariate normal distributions with different means for two classes, but sharing the same covariance matrix as follows

$$G_1 : \boldsymbol{W}|Z = 1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad G_2 : \boldsymbol{W}|Z = 2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}). \tag{1}$$

Suppose that the observed data $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n_1}, \boldsymbol{w}_{n_1+1}, \ldots, \boldsymbol{w}_{n_1+n_2}$ are independent with the first n_1 observations from G_1 and the rest n_2 observations from G_2, where $\boldsymbol{w}_i = (\boldsymbol{x}_i^\top, y_i)^\top, i \in \{1, \ldots, n_1 + n_2\}$. Let $n = n_1 + n_2$, then the log-likelihood function of data can be written as

$$n \ln |\boldsymbol{C}| - \sum_{k=1}^{2} \sum_{i \in G_k} (\boldsymbol{w}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{C} (\boldsymbol{w}_i - \boldsymbol{\mu}_k), \tag{2}$$

up to some constant, where $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix. The parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{C}$ can be estimated by maximizing the log-likelihood function of (2).

For high-dimensional data when $p \geq n$, the regularization is often needed to ensure the proper estimation of inverse covariance matrix $\boldsymbol{C}$ and mean difference $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. We thus propose to penalize $\boldsymbol{C} = (c_{ij})_{1 \leq i,j \leq p}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ simultaneously, resulting in the following optimization

$$\min_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{C})} -n \ln |\boldsymbol{C}| + \sum_{k=1}^{2} \sum_{i \in G_k} (\boldsymbol{w}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{C} (\boldsymbol{w}_i - \boldsymbol{\mu}_k) + \lambda_1 \|\boldsymbol{C}\|_1 + \frac{1}{2} \lambda_2 |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1, \tag{3}$$

where $\|\boldsymbol{C}\|_1 = \sum_{i \neq j} |c_{ij}|$, and $|\boldsymbol{\alpha}|_1 = \sum_i |\alpha_i|$ with α_i being the ith entry of vector $\boldsymbol{\alpha}$. Here $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are two tuning parameters. By applying such regularization, the proposed model can encourage the sparse structures in $\boldsymbol{C}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ at the same time. Note that similar spirits of regularizing both $\boldsymbol{C}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are used in several works on the LDA [7, 42].

To estimate the parameters, we develop an iterative procedure to solve the sub-optimization problems with respect to $\boldsymbol{C}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ respectively. Define $\boldsymbol{\delta}_2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2$ as well as $\boldsymbol{\gamma} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, then accordingly we have $\boldsymbol{\mu}_1 = \boldsymbol{\delta}_2 + \boldsymbol{\gamma}$ and $\boldsymbol{\mu}_2 = \boldsymbol{\gamma} - \boldsymbol{\delta}_2$. As a result, the optimization problem (3) is re-written as

$$\min_{(\boldsymbol{\delta}_2, \boldsymbol{\gamma}, \boldsymbol{C})} - n \ln |\boldsymbol{C}| + \sum_{i \in G_1} (\boldsymbol{w}_i - \boldsymbol{\delta}_2 - \boldsymbol{\gamma})^\top \boldsymbol{C} (\boldsymbol{w}_i - \boldsymbol{\delta}_2 - \boldsymbol{\gamma}) + \sum_{i \in G_2} (\boldsymbol{w}_i + \boldsymbol{\delta}_2 - \boldsymbol{\gamma})^\top \boldsymbol{C} (\boldsymbol{w}_i + \boldsymbol{\delta}_2 - \boldsymbol{\gamma}) + \lambda_1 \|\boldsymbol{C}\|_1 + \lambda_2 |\boldsymbol{\delta}_2|_1. \tag{4}$$

It is thus easy to obtain the maximum likelihood estimate of $\boldsymbol{\gamma}$ from (4) as

$$\hat{\boldsymbol{\gamma}} = \bar{\boldsymbol{w}} + \frac{n_2 - n_1}{n}\boldsymbol{\delta}_2, \tag{5}$$

where $\bar{\boldsymbol{w}} = (\sum_{i=1}^{n} \boldsymbol{w}_i)/n$ is the overall mean of data. Subsequently, plugging $\hat{\boldsymbol{\gamma}}$ back into (4) yields

$$\begin{aligned}(\hat{\boldsymbol{\delta}}_2, \hat{\boldsymbol{C}}) = \arg\min_{\boldsymbol{\delta}_2, \boldsymbol{C}} & - n \ln |\boldsymbol{C}| + \sum_{i \in G_1} (\boldsymbol{w}_i - \frac{2n_2}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}})^\top \boldsymbol{C} (\boldsymbol{w}_i - \frac{2n_2}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}}) \\ & + \sum_{i \in G_2} (\boldsymbol{w}_i + \frac{2n_1}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}})^\top \boldsymbol{C} (\boldsymbol{w}_i + \frac{2n_1}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}}) + \lambda_1 \|\boldsymbol{C}\|_1 + \lambda_2 |\boldsymbol{\delta}_2|_1. \end{aligned} \tag{6}$$

In this manner, solving the optimization problem (3) is equivalent to solving the optimization (6). Next, we show that (6) can be decomposed as a graphical lasso model (Glasso) [47] in terms of $\boldsymbol{C}$ and a Lasso regression [43] in terms of $\boldsymbol{\delta}_2$ with the other parameter fixed, such that these two parameters can be estimated iteratively. To be more precise, for a given value of $\boldsymbol{\delta}_2$, the minimization problem (6) with respect to $\boldsymbol{C}$ is

$$\min_{\boldsymbol{C}} -n \ln |\boldsymbol{C}| + \mathrm{tr}(\boldsymbol{C}\boldsymbol{S}) + \lambda_1 \|\boldsymbol{C}\|_1, \tag{7}$$

where $\boldsymbol{S} = \sum_{i \in G_1} (\boldsymbol{w}_i - 2n_2\boldsymbol{\delta}_2/n - \bar{\boldsymbol{w}})(\boldsymbol{w}_i - 2n_2\boldsymbol{\delta}_2/n - \bar{\boldsymbol{w}})^\top + \sum_{i \in G_2} (\boldsymbol{w}_i + 2n_1\boldsymbol{\delta}_2/n - \bar{\boldsymbol{w}})(\boldsymbol{w}_i + 2n_1\boldsymbol{\delta}_2/n - \bar{\boldsymbol{w}})^\top$. It has the same form as the Glasso, which has been extensively studied in the literature [17, 27, 32, 39, 47]. On the other hand, when the inverse covariance matrix $\boldsymbol{C}$ is fixed, the minimization problem (6) regarding $\boldsymbol{\delta}_2$ becomes

$$\min_{\boldsymbol{\delta}_2} \sum_{i \in G_1} (\boldsymbol{w}_i - \frac{2n_2}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}})^\top \boldsymbol{C} (\boldsymbol{w}_i - \frac{2n_2}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}}) + \sum_{i \in G_2} (\boldsymbol{w}_i + \frac{2n_1}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}})^\top \boldsymbol{C} (\boldsymbol{w}_i + \frac{2n_1}{n}\boldsymbol{\delta}_2 - \bar{\boldsymbol{w}}) + \lambda_2 |\boldsymbol{\delta}_2|_1, \tag{8}$$

which is equivalent to

$$\min_{\boldsymbol{\delta}_2} (\tilde{\boldsymbol{y}} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_2)^\top (\tilde{\boldsymbol{y}} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_2) + \lambda_2 |\boldsymbol{\delta}_2|_1, \tag{9}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{C}^{1/2}(n_2 \sum_{i \in G_1} \boldsymbol{w}_i - n_1 \sum_{i \in G_2} \boldsymbol{w}_i)/(2n_1 n_2)$. A detailed derivation of (9) from (8) is provided in the Appendix. We solve the minimization problem (9) by the Lasso technique. Consequently, solving the complicated optimization problem (6) is decomposed to the simple tasks of iteratively solving a Glasso estimate for $\boldsymbol{C}$ and a Lasso estimate for $\boldsymbol{\delta}_2$ until both of them are converged. We summarize the above estimation procedure for the proposed model in Algorithm 1.

Algorithm 1 (Estimation Procedure)

Step 0: Set an initial value of $\boldsymbol{\delta}_2$.

Step 1: Given $\boldsymbol{\delta}_2 = \hat{\boldsymbol{\delta}}_{2,t}$, solve $\boldsymbol{C}$ in (7) by the Glasso technique.

Step 2: Given $\boldsymbol{C} = \hat{\boldsymbol{C}}_t$, solve $\boldsymbol{\delta}_2$ in (9) by the Lasso technique.

Step 3: Repeat Step 1 and 2 till both $\hat{\boldsymbol{C}}_t$ and $\hat{\boldsymbol{\delta}}_{2,t}$ converge.

Here $\hat{\boldsymbol{C}}_t$ and $\hat{\boldsymbol{\delta}}_{2,t}$ represent the estimates of $\boldsymbol{C}$ and $\boldsymbol{\delta}_2$ in the tth iteration. The convergence criteria are $\|\hat{\boldsymbol{C}}_t - \hat{\boldsymbol{C}}_{t-1}\|_F^2 < \tau_1$ and $\|\hat{\boldsymbol{\delta}}_{2,t} - \hat{\boldsymbol{\delta}}_{2,t-1}\|_2^2 < \tau_2$, where τ_1 and τ_2 are two pre-selected small quantities, $\|\cdot\|_F$ stands for the Frobenius norm, and $\|\boldsymbol{\alpha}\|_2^2 = \sum_i \alpha_i^2$ with α_i being the ith entry of vector $\boldsymbol{\alpha}$. We set the initial value of $\boldsymbol{\delta}_2$ as $(\bar{\boldsymbol{w}}_1 - \bar{\boldsymbol{w}}_2)/2$, where $\bar{\boldsymbol{w}}_k$ is the sample mean for the kth class, $k \in \{1, 2\}$. With value of $\hat{\boldsymbol{\delta}}_2$, the estimate $\hat{\boldsymbol{\gamma}}$ is calculated by (5), and then we have $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\delta}}_2 + \hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\delta}}_2$. Therefore, Algorithm 1 provides the estimates of three parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{C}$ in (3).

Note that there are two tuning parameters λ_1 and λ_2 in the optimization problem (6). To choose their optimal values, we minimize a modified Bayesian information criterion (BIC) proposed by Wang et al. [44] as

$$\mathrm{BIC}(\lambda_1, \lambda_2) = -n \ln |\hat{\boldsymbol{C}}| + \mathrm{tr}(\hat{\boldsymbol{C}}\boldsymbol{S}) + \{v(\hat{\boldsymbol{\delta}}_2) + v(\hat{\boldsymbol{C}}) + 1\} \ln(n), \tag{10}$$

where $v(\hat{\boldsymbol{\delta}}_2)$ and $v(\hat{\boldsymbol{C}})$ stand for the number of nonzero entries in the estimates $\hat{\boldsymbol{\delta}}_2$ and $\hat{\boldsymbol{C}}$, respectively. This criterion enjoys consistency properties and has been commonly used in the literature [35, 46].

The standard errors for $\hat{\boldsymbol{\delta}}_2$ can be obtained as it is estimated from penalized likelihood function. Several works have established the asymptotic properties of the maximum penalized likelihood estimates under some regularity conditions and provided the formula of estimated standard errors, see [10, 15, 29]. Let $L = (\tilde{\boldsymbol{y}} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_2)^\top(\tilde{\boldsymbol{y}} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_2)$ and

$$\nabla L(\hat{\boldsymbol{\delta}}_2) = \frac{\partial L}{\partial \boldsymbol{\delta}_2}|_{\boldsymbol{\delta}_2=\hat{\boldsymbol{\delta}}_2}, \quad \nabla^2 L(\hat{\boldsymbol{\delta}}_2) = \frac{\partial^2 L}{\partial \boldsymbol{\delta}_2 \partial \boldsymbol{\delta}_2^\top}|_{\boldsymbol{\delta}_2=\hat{\boldsymbol{\delta}}_2}, \quad \Sigma_{\lambda_2}(\hat{\boldsymbol{\delta}}_2) = \mathrm{diag}\left\{\frac{\lambda_2 \mathrm{sign}(\hat{\boldsymbol{\delta}}_{21})}{|\hat{\boldsymbol{\delta}}_{21}|}, \ldots, \frac{\lambda_2 \mathrm{sign}(\hat{\boldsymbol{\delta}}_{2p})}{|\hat{\boldsymbol{\delta}}_{2p}|}\right\},$$

where $\hat{\boldsymbol{\delta}}_{2j}$ represents the jth entry of $\hat{\boldsymbol{\delta}}_2$. Denote by $\hat{\boldsymbol{\delta}}_{2,1}$ the nonzero components of $\hat{\boldsymbol{\delta}}_2$, then based on the conventional technique in the likelihood setting, one can employ the sandwich formula to estimate the covariance of $\hat{\boldsymbol{\delta}}_{2,1}$ as $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\delta}}_{2,1}) = \{\nabla^2 L(\hat{\boldsymbol{\delta}}_{2,1}) + n\Sigma_{\lambda_2}(\hat{\boldsymbol{\delta}}_{2,1})\}^{-1}\widehat{\mathrm{cov}}\{\nabla L(\hat{\boldsymbol{\delta}}_{2,1})\}\{\nabla^2 L(\hat{\boldsymbol{\delta}}_{2,1}) + n\Sigma_{\lambda_2}(\hat{\boldsymbol{\delta}}_{2,1})\}^{-1}$, where $\widehat{\mathrm{cov}}\{\nabla L(\hat{\boldsymbol{\delta}}_{2,1})\}$ is the sample covariance of $\nabla L(\hat{\boldsymbol{\delta}}_{2,1})$. Such estimated standard errors can be used for subsequent statistical inferences.

Next, we demonstrate how to conduct model prediction by the proposed method. For convenience, write

$$\boldsymbol{\mu}_1 = \begin{bmatrix} \boldsymbol{\mu}_{1X} \\ \mu_{1y} \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} \boldsymbol{\mu}_{2X} \\ \mu_{2y} \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_X & \boldsymbol{C}_{Xy} \\ \boldsymbol{C}_{Xy}^\top & c_y^2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{Xy} \\ \boldsymbol{\Sigma}_{Xy}^\top & \sigma_y^2 \end{bmatrix},$$

where $\boldsymbol{\mu}_{1X}$ and $\boldsymbol{\mu}_{2X}$ are $p-1$ dimensional vectors representing the means of variable $\boldsymbol{X}$ for two classes, and $\boldsymbol{\Sigma}_X$ is the $(p-1)\times(p-1)$ covariance matrix of $\boldsymbol{X}$. Then one can partition their estimates correspondingly as

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} \hat{\boldsymbol{\mu}}_{1X} \\ \hat{\mu}_{1y} \end{bmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} \hat{\boldsymbol{\mu}}_{2X} \\ \hat{\mu}_{2y} \end{bmatrix}, \quad \hat{\boldsymbol{C}} = \begin{bmatrix} \hat{\boldsymbol{C}}_X & \hat{\boldsymbol{C}}_{Xy} \\ \hat{\boldsymbol{C}}_{Xy}^\top & \hat{c}_y^2 \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_X & \hat{\boldsymbol{\Sigma}}_{Xy} \\ \hat{\boldsymbol{\Sigma}}_{Xy}^\top & \hat{\sigma}_y^2 \end{bmatrix}.$$

From model assumption (1) as well as the property of multivariate normal distribution, the prediction for the quantitative response y from a new observation $\boldsymbol{x}$ is

$$\hat{y} = \begin{cases} \hat{\mu}_{1y} + \hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{1X}), \text{ if } \hat{Z} = 1 \\ \hat{\mu}_{2y} + \hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{2X}), \text{ if } \hat{Z} = 2. \end{cases} \tag{11}$$

Note that $\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1} = -\hat{\boldsymbol{C}}_{Xy}^\top / \hat{c}_y^2$ where $\hat{c}_y^2$ is a scalar, implying that the sparsity of $\hat{\boldsymbol{C}}_{Xy}$ will lead to a sparse model for the prediction of y.

On the other hand, denote by π_1 and π_2 the prior probability of $\boldsymbol{w}$ belonging to classes G_1 and G_2, respectively. The prediction for the qualitative response Z by the proposed model is naturally based on the estimated LDA classification rule as

$$\ln \frac{\Pr(G_1|\boldsymbol{W} = (\boldsymbol{x}^\top, \hat{y})^\top)}{\Pr(G_2|\boldsymbol{W} = (\boldsymbol{x}^\top, \hat{y})^\top)} = \ln \frac{\hat{\pi}_1}{\hat{\pi}_2} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)^\top \hat{\boldsymbol{C}}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) + (\boldsymbol{x}^\top, \hat{y})\hat{\boldsymbol{C}}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2), \tag{12}$$

where Pr stands for probability, and the estimates $\hat{\pi}_1$ and $\hat{\pi}_2$ are the empirical proportions of data from each class. From (11) and (12), however, we note that the prediction of one response variable depends on the information of the other. To address this issue, we propose to calculate two candidate values of y for a new observation $\boldsymbol{x}$ by (11) for two different classes, denoted by $\hat{y}_1$ and $\hat{y}_2$. Then the conditional probability densities $p(\boldsymbol{W} = (\boldsymbol{x}^\top, \hat{y}_1)^\top|G_1)$ and $p(\boldsymbol{W} = (\boldsymbol{x}^\top, \hat{y}_2)^\top|G_2)$ can be estimated via the density functions of $N(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}})$ and $N(\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}})$. Denote such two values as $\hat{p}_1$ and $\hat{p}_2$. The prediction of y at this new observation is then obtained as $\hat{y}_k$ corresponding to the larger value of $\hat{\pi}_k\hat{p}_k, k \in \{1, 2\}$. To express it clearly, we describe the above steps of the model prediction in Algorithm 2 for a new observation $\boldsymbol{x}$.

Algorithm 2 (Prediction Procedure)

Step 1: For $k \in \{1, 2\}$, $\hat{y}_k = \hat{\mu}_{ky} + \hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{kX})$, and consequently obtain the probability densities $\hat{p}_k$ by plugging $(\boldsymbol{x}^\top, \hat{y}_k)^\top$ into the density functions of $N(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}})$.

Step 2a: If $\hat{\pi}_1\hat{p}_1 > \hat{\pi}_2\hat{p}_2$, let $\hat{y} = \hat{y}_1$; otherwise let $\hat{y} = \hat{y}_2$.

Step 2b: Apply the LDA classification rule (12) to predict Z by $\boldsymbol{w} = (\boldsymbol{x}^\top, \hat{y})^\top$.

It is seen that in Algorithm 2, we obtain the prediction of y first, and then predict Z with the value of $\hat{y}$. One would argue that it is not a unique way of making predictions on QQ responses, as we may also predict Z first and then variable y. The following Proposition 1 provides an interesting insight into this issue.

Proposition 1. *For the prediction of Z by the proposed model, the class label k obtained from Step 2b of Algorithm 2 maximizes $\hat{\pi}_k\hat{p}_k$.*

We first state Lemma 1, which is used to prove Proposition 1.

Lemma 1. *Suppose a random vector $(\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top \sim N(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are multivariate variables. For a given value of $\boldsymbol{a}$, then $\boldsymbol{b} = \boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ab}^\top\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_a)$ maximizes $exp\{-(1/2)[(\boldsymbol{a}^\top, \boldsymbol{b}^\top) - \boldsymbol{\mu}_*^\top]\boldsymbol{\Sigma}_*^{-1}[(\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top - \boldsymbol{\mu}_*]\}$, where $\boldsymbol{\mu}_* = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$ and $\boldsymbol{\Sigma}_* = \begin{bmatrix} \boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ab}^\top & \boldsymbol{\Sigma}_b \end{bmatrix}$.*

Proof: Let C denote a generic constant thereafter in the paper for convenience. We need to search for $\boldsymbol{b}$ to minimize $\{(\boldsymbol{a}^\top, \boldsymbol{b}^\top) - \boldsymbol{\mu}^\top\}\boldsymbol{\Omega}\{(\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top - \boldsymbol{\mu}\}$, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_*^{-1} = \begin{bmatrix} \boldsymbol{\Omega}_a & \boldsymbol{\Omega}_{ab} \\ \boldsymbol{\Omega}_{ab}^\top & \boldsymbol{\Omega}_b \end{bmatrix}$. That is, we minimize

$$\begin{aligned} L(\boldsymbol{b}) &= (\boldsymbol{a}^\top, \boldsymbol{b}^\top)\boldsymbol{\Omega}(\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top - 2\boldsymbol{\mu}^\top\boldsymbol{\Omega}(\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top = (\boldsymbol{a}^\top, \boldsymbol{b}^\top)\begin{bmatrix} \boldsymbol{\Omega}_a & \boldsymbol{\Omega}_{ab} \\ \boldsymbol{\Omega}_{ab}^\top & \boldsymbol{\Omega}_b \end{bmatrix}\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} - 2(\boldsymbol{\mu}_a^\top, \boldsymbol{\mu}_b^\top)\begin{bmatrix} \boldsymbol{\Omega}_a & \boldsymbol{\Omega}_{ab} \\ \boldsymbol{\Omega}_{ab}^\top & \boldsymbol{\Omega}_b \end{bmatrix}\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \\ &= 2\boldsymbol{a}^\top\boldsymbol{\Omega}_{ab}\boldsymbol{b} + \boldsymbol{b}^\top\boldsymbol{\Omega}_b\boldsymbol{b} - 2(\boldsymbol{\mu}_a^\top\boldsymbol{\Omega}_{ab} + \boldsymbol{\mu}_b^\top\boldsymbol{\Omega}_b)\boldsymbol{b} + C. \end{aligned}$$

Taking derivative of $L(\boldsymbol{b})$ and setting to zero yields

$$\begin{aligned} \frac{\partial L(\boldsymbol{b})}{\partial \boldsymbol{b}} &= 2\boldsymbol{\Omega}_{ab}^\top\boldsymbol{a} + 2\boldsymbol{\Omega}_b\boldsymbol{b} - 2(\boldsymbol{\Omega}_{ab}^\top\boldsymbol{\mu}_a + \boldsymbol{\Omega}_b\boldsymbol{\mu}_b) = \boldsymbol{0} \\ \boldsymbol{b} &= \boldsymbol{\mu}_b - \boldsymbol{\Omega}_b^{-1}\boldsymbol{\Omega}_{ab}^\top(\boldsymbol{a} - \boldsymbol{\mu}_a). \end{aligned}$$

This, together with a property of block matrix that $\boldsymbol{\Omega}_{ab}^\top = -\boldsymbol{\Omega}_b\boldsymbol{\Sigma}_{ab}^\top\boldsymbol{\Sigma}_a^{-1}$, completes the proof. □

For a new observation $\boldsymbol{x}$, let $y_1 = \mu_{1y} + \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X})$ and $y_2 = \mu_{2y} + \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{2X})$. Denote by $p_1 = p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top|G_1)$ and $p_2 = p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_2)^\top|G_2)$. Now we prove Proposition 1.

Proof of Proposition 1: Without loss of generality, we suppose $\pi_1 p_1 > \pi_2 p_2$, then we show below that the LDA classification rule would assign $(\boldsymbol{x}^\top, y_1)^\top$ to G_1. In order to achieve this, we only need to prove

$$p_2 \geq p_3 = p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_3)^\top|G_2) \tag{13}$$

for any value of y_3. That is, we need to prove $\boldsymbol{W} = (\boldsymbol{x}^\top, y_2)^\top$ will maximize the density function of $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, which is the conclusion of Lemma 1. As a result, $\pi_1 p_1 = \pi_1 p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top|G_1) > \pi_2 p_2 \geq \pi_2 p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top|G_2)$ by taking $y_3 = y_1$ in (13). Hence,

$$p(\boldsymbol{x} \in G_1|\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top) = \frac{\pi_1 p_1}{p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top)} > \frac{\pi_2 p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top|\boldsymbol{x} \in G_2)}{p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top)} = p(\boldsymbol{x} \in G_2|\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top),$$

implying that the LDA assigns $(\boldsymbol{x}^\top, y_1)^\top$ to G_1. Here we also use π_i to represent $p(\boldsymbol{x} \in G_i)$ to express the probability of this observation belonging to class $G_i, i \in \{1, 2\}$. □

Proposition 1 implies that we can predict the response variable Z by simply comparing values of $\hat{\pi}_k\hat{p}_k$ instead of employing LDA. Therefore, the order of which response variable to be predicted first is not a concern. Actually, the Step 2a and Step 2b are equivalent to the following Step 2 as

Step 2: If $\hat{\pi}_1\hat{p}_1 > \hat{\pi}_2\hat{p}_2$, let $\hat{y} = \hat{y}_1$ and $\hat{Z} = 1$; otherwise let $\hat{y} = \hat{y}_2$ and $\hat{Z} = 2$.

2.2. The GAQQ model for multi-class qualitative response

We now extend the GAQQ model to handle the QQ data with multi-class qualitative response, i.e., the qualitative variable $Z \in \{1, \ldots, K\}$. In such cases, the GAQQ method is expressed as

$$G_k : \boldsymbol{W}|Z = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}),\ k \in \{1, \ldots, K\}.$$

Based on a baseline class G_1, we regularize on the differences between means through $\boldsymbol{\mu}_k - \boldsymbol{\mu}_1$ for $k \in \{2, \ldots, K\}$. The objective function is thus formulated as

$$\min_{(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\boldsymbol{C})} -n \ln |\boldsymbol{C}| + \sum_{k=1}^{K} \sum_{i \in G_k} (\boldsymbol{w}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{C} (\boldsymbol{w}_i - \boldsymbol{\mu}_k) + \lambda_1 \|\boldsymbol{C}\|_1 + \lambda_2 \sum_{k=2}^{K} |\boldsymbol{\mu}_k - \boldsymbol{\mu}_1|_1. \tag{14}$$

The subsequent derivation follows similar steps as that described in Section 2.1. Let $K\boldsymbol{\delta}_k = \boldsymbol{\mu}_k - \boldsymbol{\mu}_1$ and $K\boldsymbol{\gamma} = \sum_{k=1}^{K} \boldsymbol{\mu}_k$, then we have $\boldsymbol{\mu}_k = \boldsymbol{\gamma} - \sum_{g=2}^{K} \boldsymbol{\delta}_g + K\boldsymbol{\delta}_k,\ k \in \{1, \ldots, K\}$. As a result, the optimization problem (14) can be re-written as

$$\min_{(\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_K,\boldsymbol{\gamma},\boldsymbol{C})} -n \ln |\boldsymbol{C}| + \sum_{k=1}^{K} \sum_{i \in G_k} (\boldsymbol{w}_i - \boldsymbol{\gamma} + \sum_{g=2}^{K} \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k)^\top \boldsymbol{C} (\boldsymbol{w}_i - \boldsymbol{\gamma} + \sum_{g=2}^{K} \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k) + \lambda_1 \|\boldsymbol{C}\|_1 + \lambda_2 \sum_{k=2}^{K} |\boldsymbol{\delta}_k|_1. \tag{15}$$

Let n_k represent the number of observations belonging to class G_k. The maximum likelihood estimator of $\boldsymbol{\gamma}$ from (15) is $\hat{\boldsymbol{\gamma}} = \bar{\boldsymbol{w}} + \sum_{g=2}^{K} \boldsymbol{\delta}_g - (K/n) \sum_{g=2}^{K} n_g \boldsymbol{\delta}_g$. Consequently, the optimization problem (15) becomes

$$\min_{(\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_K,\boldsymbol{C})} -n \ln |\boldsymbol{C}| + \sum_{k=1}^{K} \sum_{i \in G_k} (\boldsymbol{w}_i - \bar{\boldsymbol{w}} + \frac{K}{n} \sum_{g=2}^{K} n_g \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k)^\top \boldsymbol{C} (\boldsymbol{w}_i - \bar{\boldsymbol{w}} + \frac{K}{n} \sum_{g=2}^{K} n_g \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k) + \lambda_1 \|\boldsymbol{C}\|_1 + \lambda_2 \sum_{k=2}^{K} |\boldsymbol{\delta}_k|_1. \tag{16}$$

Let $\tilde{\boldsymbol{S}} = \sum_{k=1}^{K} \sum_{i \in G_k} \{\boldsymbol{w}_i - \bar{\boldsymbol{w}} + (K/n) \sum_{g=2}^{K} n_g \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k\} \{\boldsymbol{w}_i - \bar{\boldsymbol{w}} + (K/n) \sum_{g=2}^{K} n_g \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k\}^\top$, then the formula (16) can be decomposed as one Glasso problem

$$\min_{\boldsymbol{C}} -n \ln |\boldsymbol{C}| + \operatorname{tr}(\boldsymbol{C}\tilde{\boldsymbol{S}}) + \lambda_1 \|\boldsymbol{C}\|_1, \tag{17}$$

and

$$\min_{(\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_K)} \sum_{k=1}^{K} \sum_{i \in G_k} (\boldsymbol{w}_i - \bar{\boldsymbol{w}} + \frac{K}{n} \sum_{g=2}^{K} n_g \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k)^\top \boldsymbol{C} (\boldsymbol{w}_i - \bar{\boldsymbol{w}} + \frac{K}{n} \sum_{g=2}^{K} n_g \boldsymbol{\delta}_g - K\boldsymbol{\delta}_k) + \lambda_2 \sum_{k=2}^{K} |\boldsymbol{\delta}_k|_1. \tag{18}$$

The optimization (18) is equivalent to the following $K - 1$ Lasso regressions separately

$$\min_{\boldsymbol{\delta}_k} (\tilde{\boldsymbol{y}} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_k)^\top (\tilde{\boldsymbol{y}} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_k) + \lambda_2 |\boldsymbol{\delta}_k|_1, \quad k \in \{2, \ldots, K\}, \tag{19}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{C}^{1/2} \{(n - n_k) \sum_{i \in G_k} \boldsymbol{w}_i - n_k \sum_{i \notin G_k} \boldsymbol{w}_i + Kn_k \sum_{g=2, g \neq k}^{K} n_g \boldsymbol{\delta}_g\} / (Knn_k)$. The detailed derivation from (18) to (19) is provided in the Appendix. Therefore, the parameters $\boldsymbol{\delta}_k$ and $\boldsymbol{C}$ can be solved iteratively until convergence following the spirit of Algorithm 1. The optimal values of tuning parameters are chosen by the modified BIC extended for the multi-class problem as

$$\mathrm{BIC}(\lambda_1, \lambda_2) = -n \ln |\hat{\boldsymbol{C}}| + \operatorname{tr}(\hat{\boldsymbol{C}}\tilde{\boldsymbol{S}}) + \{v(\hat{\boldsymbol{\delta}}) + v(\hat{\boldsymbol{C}}) + K - 1\} \ln(n), \tag{20}$$

where $v(\hat{\boldsymbol{\delta}})$ represents the number of nonzero entries in all the estimates $\hat{\boldsymbol{\delta}}_k$. The estimated covariances for each $\hat{\boldsymbol{\delta}}_k$ can be obtain by following the similar techniques for estimating the covariance of $\hat{\boldsymbol{\delta}}_2$ in two-class setting in Section 2.1.

For a new observation $\boldsymbol{x}$, the quantitative response y is predicted, similarly as in Algorithm 2, to be $\hat{y}_k = \hat{\mu}_{ky} + \hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{kX})$, where k maximizes $\hat{\pi}_k \hat{p}_k$ with $\hat{p}_k = p(\boldsymbol{W} = (\boldsymbol{x}^\top, \hat{y}_k)^\top | G_k)$, computed by plugging $(\boldsymbol{x}^\top, \hat{y}_k)^\top$ into the density functions of $N(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}})$. The class label is estimated as $\hat{Z} = \arg\max_k \hat{\pi}_k \hat{p}_k$, or equivalently by the LDA rule as

$$\hat{Z} = \arg\max_k \ \ln \frac{\hat{\pi}_k}{\hat{\pi}_1} + K\{(\boldsymbol{x}^\top, \hat{y}_k)^\top - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_k}{2}\}^\top \hat{\boldsymbol{C}} \hat{\boldsymbol{\delta}}_k.$$

3. Theoretical properties

In this section, we study the theoretical properties of the proposed GAQQ model. The asymptotic optimality of the classification rule is investigated via Theorems 1 - 3 in Section 3.1. The asymptotic consistency properties of the prediction of y are established in Theorem 4 in Section 3.2.

3.1. Asymptotic optimality of the classification rule

For the proposed classification rule, we first derived the theoretical results for the multi-class problem and then provide a thorough discussion of the two-class case. We use the same definition of asymptotic optimality for a classification rule as defined in Shao et al. [42]. Denote by R_{Bayes} and $R_{\text{PROP}}(\mathcal{T})$ the Bayes error and the conditional misclassification rate of the proposed rule, where $\mathcal{T}$ denotes the training samples. The asymptotic optimality for a classification rule is defined as follows.

Definition 1. Let T be a classification rule with conditional misclassification rate $R_T(\mathcal{T})$, given the training samples $\mathcal{T}$.

(i) T is asymptotically optimal if $R_T(\mathcal{T})/R_{\text{Bayes}} \xrightarrow{P} 1$.

(ii) T is asymptotically sub-optimal if $R_T(\mathcal{T}) - R_{\text{Bayes}} \xrightarrow{P} 0$.

Note that if $\lim_{n\to\infty} R_{\text{Bayes}} > 0$, then the asymptotically sub-optimality is the same as the asymptotically optimality. To facilitate the construction of theoretical results, we need to introduce some notation and make assumptions on the true model. Define the true values of $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$, $\boldsymbol{C}$ and $\boldsymbol{\delta}_k$ as $\boldsymbol{\mu}_k^0$, $\boldsymbol{\Sigma}^0$, $\boldsymbol{C}^0$ and $\boldsymbol{\delta}_k^0 = (\boldsymbol{\mu}_k^0 - \boldsymbol{\mu}_1^0)/K = ((\boldsymbol{\delta}_{kX}^0)^\top, \delta_{ky}^0)^\top$, where $\boldsymbol{\delta}_{kX}^0$ is a $(p-1)$ dimensional vector representing the true mean difference of variable $\boldsymbol{X}$ between classes G_1 and G_k. Denote the true inverse covariance matrix of variable $\boldsymbol{X}$ by $\boldsymbol{C}_X^0$. Also define $\Delta_k = \sqrt{(\boldsymbol{\delta}_{kX}^0)^\top \boldsymbol{C}_X^0 \boldsymbol{\delta}_{kX}^0}$ and $\Delta = \max\{\Delta_k\}_{k=1}^K$. Denote $\mathbb{S}_{\delta_k} = \{j; (\boldsymbol{\delta}_k^0)_j \neq 0\}$, which is the set containing location indices of the nonzero entries in $\boldsymbol{\delta}_k^0$. Let $\tilde{s}_k$ be the cardinalities of set $\mathbb{S}_{\delta_k}$. Define $s_k = \tilde{s}_k$ if $\delta_{ky}^0 = 0$; otherwise $s_k = \tilde{s}_k - 1$. That is, s_k is the number of nonzero entries of $\boldsymbol{\delta}_{kX}^0$. Additionally, we use the same sparsity measure $S_{h;p} = \max_{i\le p} \sum_{j=1}^p |\sigma_{ij}^0|^h$ on $\boldsymbol{\Sigma}^0 = (\sigma_{ij}^0)_{1\le i,j\le p}$ as in Bickel and Levina [3], where $0 \le h < 1$ and 0^0 is defined to be 0. Hence firstly, $S_{0;p}$ equals the maximum of the numbers of nonzero entries in each row of the matrix $\boldsymbol{\Sigma}^0$. In this case, a smaller value of $S_{0;p}$ compared with p implies a sparse structure in matrix $\boldsymbol{\Sigma}^0$. Secondly, if $S_{h;p}$ is smaller than p for $0 < h < 1$, it indicates that many entries of matrix $\boldsymbol{\Sigma}^0$ are very small. Moreover, we assume the following regularity conditions:

C1: There exists a constant θ such that $0 < \theta^{-1} < \lambda_{\min}(\boldsymbol{C}^0) \le \lambda_{\max}(\boldsymbol{C}^0) < \theta < \infty$, where $\lambda_{\min}(\boldsymbol{C}^0)$ and $\lambda_{\max}(\boldsymbol{C}^0)$ are the minimum and maximum eigenvalues of matrix $\boldsymbol{C}^0$;

C2: $\lambda_1 = O(\sqrt{\ln p/n})$, $\lambda_2 = O(\sqrt{\ln p/n})$;

C3: Restricted eigenvalue condition: for some constant $\varphi_k > 0$, assume $\boldsymbol{C}^0$ satisfies $\|(\boldsymbol{C}^0)^{1/2}\boldsymbol{\delta}_k^0\|_2^2 \ge n\varphi_k\|\boldsymbol{\delta}_k^0\|_2^2$ for all subsets $J \subseteq \{1,\dots,p\}$ such that the cardinality of J equals $\tilde{s}_k$, and $|(\boldsymbol{\delta}_k^0)_{J^c}|_1 \le 3|(\boldsymbol{\delta}_k^0)_J|_1$. Here $(\boldsymbol{\delta}_k^0)_J = ((\boldsymbol{\delta}_k^0)_j \cdot I\{j \in J\})_{1\le j\le p}$, and J^c represents the complement set of J;

C4: Irrepresentable condition: without loss of generality, write $\boldsymbol{\delta}_k^0 = ((\boldsymbol{\delta}_k^0)_{\mathbb{S}_\delta}^\top, (\boldsymbol{\delta}_k^0)_{\mathbb{S}_\delta^c}^\top)^\top$, and correspondingly let $\boldsymbol{C}^0 = \begin{bmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{bmatrix}$, where $\boldsymbol{\Psi}_{11}$ is an $\tilde{s}_k \times \tilde{s}_k$ matrix. Then there exists a positive constant vector $\boldsymbol{\zeta}$ such that $|\boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\text{sign}((\boldsymbol{\delta}_k^0)_{\mathbb{S}_\delta}^\top)| \le \mathbf{1} - \boldsymbol{\zeta}$, where $\mathbf{1}$ is a $p - \tilde{s}_k$ dimensional unit vector, and the inequality holds element-wise;

C5: There exist $0 \le c_1 < c_2 \le 1$ and $M > 0$, such that $n^{(1-c_2)/2}\min_{1\le i\le \tilde{s}_k}|(\boldsymbol{\delta}_k^0)_i| \ge M$, $\tilde{s}_k = O(n^{c_1})$. $\lambda_2 = o(n^{(c_2-c_1+1)/2})$, $p = o(\lambda_2^2 n)$;

C6: There exists a constant $c_3 > 0$ such that $(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{\ell X}^0)^\top \boldsymbol{C}_X^0 (\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{\ell X}^0) > c_3 > 0, k \neq \ell$;

C7: There exists a constant c_4 such that $c_4^{-1} \le K\pi_k \le c_4, k \in \{1,\dots,K\}$.

By Conditions C1 and C2, Rothman et al. [40] and Lam and Fan [27] derived the convergence rate of Glasso estimate. We thus have

$$\|\hat{\boldsymbol{C}}_X - \boldsymbol{C}_X^0\| = O_p(d_n), \tag{21}$$

where $d_n = S_{h;p}(\ln p/n)^{(1-h)/2}$, and $\|\boldsymbol{A}\|$ is the matrix spectral norm defined as the squared root of the maximum eigenvalue of matrix $\boldsymbol{A}^\top\boldsymbol{A}$. The Conditions C2 and C3 are used in Bühlmann and Van De Geer [5] to study the theoretical property of Lasso estimate, and we have

$$\|\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{\delta}_{kX}^0\|_2 = O_p(b_k^{(n)}), \tag{22}$$

where $b_k^{(n)} = \sqrt{\tilde{s}_k \ln p/(n\varphi_k^2)}$. Under Conditions C4 and C5, Zhao and Yu [49] showed that the Lasso estimate is model selection consistency. Condition C6 requires that all the classes should be separated from each other. Also note that

Condition C6 is equivalent to that Δ_k is bounded away from 0. The Condition C7 guarantees a balanced sample size for each class, which is commonly used in the literature to bound the term $\ln(\pi_k/\pi_1)$ in the LDA rule for establishing the properties of classification rules. Based on the above results, we present the following theories on the consistency of the classification rule by the proposed method.

Theorem 1. *Assume that Conditions C1 - C7 hold, and*

$$\xi_{n;k} = \max\{d_n, \frac{b_k^{(n)}}{\Delta_k}, \frac{\sqrt{s_k S_{h;p}}}{\sqrt{n}\Delta_k} \text{ for any } k\} \to 0.$$

Then the proposed rule for the multi-class problem is asymptotically sub-optimal if either one of the following two conditions are satisfied
(i) $\Delta = \max\{\Delta_k\}_{k=1}^K$ *is bounded;*
(ii) if $\Delta \to \infty$*, then there exists a constant* $\alpha \in (0, 1/2)$ *such that* $\Delta^2 \xi_{n;k}^{1-2\alpha} \to 0$.

Before the proof of Theorem 1, we need Proposition 2 and Lemmas 2 - 4.

Proposition 2. *For an observation* $\boldsymbol{x}$*, recall that* $y_1 = \mu_{1y} + \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X})$*,* $y_2 = \mu_{2y} + \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{2X})$*,* $p_1 = p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_1)^\top | G_1)$ *and* $p_2 = p(\boldsymbol{W} = (\boldsymbol{x}^\top, y_2)^\top | G_2)$*. Then* $p(\boldsymbol{x} \in G_1 | \boldsymbol{X} = \boldsymbol{x}) > p(\boldsymbol{x} \in G_2 | \boldsymbol{X} = \boldsymbol{x})$ *is equivalent to* $\pi_1 p_1 > \pi_2 p_2$.

Proof: Since $p(\boldsymbol{x} \in G_1 | \boldsymbol{X} = \boldsymbol{x}) > p(\boldsymbol{x} \in G_2 | \boldsymbol{X} = \boldsymbol{x})$, we have

$$\begin{aligned}
&\pi_1 p(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{x} \in G_1) > \pi_2 p(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{x} \in G_2), \\
&\pi_1 \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{1X})^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X})\} > \pi_2 \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{2X})^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{2X})\}, \\
&\ln \pi_1 - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{1X})^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X}) > \ln \pi_2 - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{2X})^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{2X}).
\end{aligned} \tag{23}$$

On the other hand, $\pi_1 p_1 > \pi_2 p_2$ yields

$$\begin{aligned}
&\ln \pi_1 - \frac{1}{2}\left\{\begin{pmatrix} \boldsymbol{x} \\ y_1 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_{1X} \\ \mu_{1y} \end{pmatrix}\right\}^\top \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{Xy} \\ \boldsymbol{\Sigma}_{Xy}^\top & \sigma_y^2 \end{bmatrix}^{-1} \left\{\begin{pmatrix} \boldsymbol{x} \\ y_1 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_{1X} \\ \mu_{1y} \end{pmatrix}\right\} \\
&\quad > \ln \pi_2 - \frac{1}{2}\left\{\begin{pmatrix} \boldsymbol{x} \\ y_2 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_{2X} \\ \mu_{2y} \end{pmatrix}\right\}^\top \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{Xy} \\ \boldsymbol{\Sigma}_{Xy}^\top & \sigma_y^2 \end{bmatrix}^{-1} \left\{\begin{pmatrix} \boldsymbol{x} \\ y_2 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_{2X} \\ \mu_{2y} \end{pmatrix}\right\}.
\end{aligned} \tag{24}$$

Now we prove (23) and (24) are equivalent. Since

$$\begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{Xy} \\ \boldsymbol{\Sigma}_{Xy}^\top & \sigma_y^2 \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_X^{-1} + \frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}\boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}}{\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}} & -\frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}}{\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}} \\ -\frac{\boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}}{\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}} & \frac{1}{\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_X^{-1} + \frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}\boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}}{\mathrm{Var}(y|\boldsymbol{X})} & -\frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}}{\mathrm{Var}(y|\boldsymbol{X})} \\ -\frac{\boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}}{\mathrm{Var}(y|\boldsymbol{X})} & \frac{1}{\mathrm{Var}(y|\boldsymbol{X})} \end{bmatrix},$$

the left side of (24) equals

$$\begin{aligned}
&\ln \pi_1 - \frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_{1X})^\top, y_1 - \mu_{1y}\right]^\top \begin{bmatrix} \boldsymbol{\Sigma}_X^{-1} + \frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}\boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}}{\mathrm{Var}(y|\boldsymbol{X})} & -\frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}}{\mathrm{Var}(y|\boldsymbol{X})} \\ -\frac{\boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}}{\mathrm{Var}(y|\boldsymbol{X})} & \frac{1}{\mathrm{Var}(y|\boldsymbol{X})} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} - \boldsymbol{\mu}_{1X} \\ y_1 - \mu_{1y} \end{bmatrix} \\
&= \ln \pi_1 - \frac{1}{2}\{(\boldsymbol{x} - \boldsymbol{\mu}_{1X})^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X}) + (\boldsymbol{x} - \boldsymbol{\mu}_{1X})^\top \frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}\boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}}{\mathrm{Var}(y|\boldsymbol{X})}(\boldsymbol{x} - \boldsymbol{\mu}_{1X}) \\
&\quad - \frac{y_1 - \mu_{1y}}{\mathrm{Var}(y|\boldsymbol{X})}\boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X}) + \frac{(y_1 - \mu_{1y})^2}{\mathrm{Var}(y|\boldsymbol{X})} - (\boldsymbol{x} - \boldsymbol{\mu}_{1X})^\top \frac{\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}}{\mathrm{Var}(y|\boldsymbol{X})}(y_1 - \mu_{1y})\} \\
&= \ln \pi_1 - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{1X})^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X}),
\end{aligned}$$

where the last equality applies $y_1 - \mu_{1y} = \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{1X})$. Similarly, the right side of (24) equals $\ln \pi_2 - (\boldsymbol{x} - \boldsymbol{\mu}_{2X})^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{2X})/2$. This completes the proof. □

The inequality $p(\boldsymbol{x} \in G_1|\boldsymbol{X} = \boldsymbol{x}) > p(\boldsymbol{x} \in G_2|\boldsymbol{X} = \boldsymbol{x})$ in Proposition 2 indicates that the LDA rule assigns $\boldsymbol{x}$ to G_1. Therefore, Proposition 2 implies that Step 2b of Algorithm 2 is equivalent to applying the LDA classification rule directly on $\boldsymbol{x}$ instead of $\boldsymbol{w} = (\boldsymbol{x}^\top, \hat{y})^\top$. This fact enables us to give theoretical proof of Theorem 1 for the consistency properties of the proposed classification rule based on variable $\boldsymbol{X}$ rather than $\boldsymbol{W} = (\boldsymbol{X}^\top, y)^\top$.

Lemma 2. *For any $k \in \{2, \ldots, K\}$, we have*

$$(\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X - (\boldsymbol{\delta}_{kX}^0)^\top \boldsymbol{C}_X^0)\boldsymbol{\Sigma}_X^0(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0) = \Delta_k^2\left[O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)\right]$$

for the multi-class problem.

Proof: Decompose

$$(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{\Sigma}_X^0(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0) = \hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\boldsymbol{\Sigma}_X^0\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - 2\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\boldsymbol{\delta}_{kX}^0 + (\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0. \tag{25}$$

On one hand, by the result (21) we have

$$\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\boldsymbol{\Sigma}_X^0\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} = \hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}[1 + O_p(d_n)] = \hat{\boldsymbol{\delta}}_{kX}^\top\boldsymbol{C}_X^0\hat{\boldsymbol{\delta}}_{kX}[1 + O_p(d_n)].$$

Since $E[(\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{\delta}_{kX}^0)]^2 \le \Delta_k^2 E[(\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{\delta}_{kX}^0)]$ and by (22), we obtain

$$\begin{aligned}\hat{\boldsymbol{\delta}}_{kX}^\top\boldsymbol{C}_X^0\hat{\boldsymbol{\delta}}_{kX} &= (\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0 + 2(\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{\delta}_{kX}^0) + (\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{\delta}_{kX}^0)\\ &= \Delta_k^2 + O_p(b_k^{(n)}\Delta_k) + O_p((b_k^{(n)})^2) = \Delta_k^2[1 + O_p(\frac{b_k^{(n)}}{\Delta_k})].\end{aligned}$$

As a result,

$$\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} = \hat{\boldsymbol{\delta}}_{kX}^\top\boldsymbol{C}_X^0\hat{\boldsymbol{\delta}}_{kX}[1 + O_p(d_n)] = \Delta_k^2[1 + O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)]. \tag{26}$$

On the other hand, since $\|\boldsymbol{\delta}_{kX}^0\|_2^2 = O(\Delta_k^2)$, we have

$$(\boldsymbol{\delta}_{kX}^0)^\top\hat{\boldsymbol{C}}_X\boldsymbol{\delta}_{kX}^0 = (\boldsymbol{\delta}_{kX}^0)^\top(\hat{\boldsymbol{C}}_X - \boldsymbol{C}_X^0)\boldsymbol{\delta}_{kX}^0 + (\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0 = O_p(\Delta_k^2 d_n) + \Delta_k^2 = \Delta_k^2[1 + O_p(d_n)]. \tag{27}$$

Consequently,

$$\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\boldsymbol{\delta}_{kX}^0 = \Delta_k\sqrt{1 + O_p(d_n)}\,\Delta_k\sqrt{1 + O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)} = \Delta_k^2\sqrt{1 + O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)}. \tag{28}$$

Combing (25), (26) and (28) yields

$$\begin{aligned}(\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X - (\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0)\boldsymbol{\Sigma}_X^0(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0) =& \Delta_k^2[1 + O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)] - 2\Delta_k^2\sqrt{1 + O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)} + \Delta_k^2\\ =& \Delta_k^2\left[O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)\right],\end{aligned}$$

where the last equality uses the Taylor expansion of $\sqrt{1+x} = 1 + x/2 + o(x)$. □

Write $\boldsymbol{\mu}_k^0 = ((\boldsymbol{\mu}_{kX}^0)^\top, \mu_{ky}^0)^\top$, where $\boldsymbol{\mu}_{kX}^0$ is the true mean value of variable $\boldsymbol{X}$ for class G_k. Correspondingly, write $\hat{\boldsymbol{\mu}}_k = (\hat{\boldsymbol{\mu}}_{kX}^\top, \hat{\mu}_{ky})^\top$. Let $a_n \asymp b_n$ represent that two sequences a_n and b_n are the same order. Now we state Lemma 3.

Lemma 3. *Let $q_k^{(n)}$ be the number of nonzero entries of $\hat{\boldsymbol{\delta}}_{kX}$. For $k \in \{2, \ldots, K\}$, we have*

$$\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{1X} - \boldsymbol{\mu}_{1X}^0) \asymp \hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{kX} - \boldsymbol{\mu}_{kX}^0) = O_p(\sqrt{\frac{q_k^{(n)}}{n}})\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}} - O_p(\sqrt{\frac{S_{h;p}q_k^{(n)}}{n}})\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}}.$$

Proof: Without loss of generality, we assume that $\hat{\boldsymbol{\delta}}_{kX} = (\hat{\boldsymbol{\delta}}_{k,1}^\top, \mathbf{0}^\top)^\top$, where $\hat{\boldsymbol{\delta}}_{k,1}^\top$ is a $q_k^{(n)}$-dimensional vector containing all the nonzero entries of $\hat{\boldsymbol{\delta}}_{kX}$. Note that $\lim_{n\to\infty} q_k^{(n)} = s_k$. Conformally, we write

$$\boldsymbol{\Sigma}_X^0 = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^0 & \boldsymbol{\Sigma}_{12}^0 \\ (\boldsymbol{\Sigma}_{12}^0)^\top & \boldsymbol{\Sigma}_{22}^0 \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}}_X = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ (\hat{\boldsymbol{\Sigma}}_{12})^\top & \hat{\boldsymbol{\Sigma}}_{22} \end{bmatrix}, \quad \boldsymbol{C}_X^0 = \begin{bmatrix} \boldsymbol{C}_{11}^0 & \boldsymbol{C}_{12}^0 \\ (\boldsymbol{C}_{12}^0)^\top & \boldsymbol{C}_{22}^0 \end{bmatrix}, \quad \hat{\boldsymbol{C}}_X = \begin{bmatrix} \hat{\boldsymbol{C}}_{11} & \hat{\boldsymbol{C}}_{12} \\ (\hat{\boldsymbol{C}}_{12})^\top & \hat{\boldsymbol{C}}_{22} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{11}^0, \hat{\boldsymbol{\Sigma}}_{11}, \boldsymbol{C}_{11}^0$ and $\hat{\boldsymbol{C}}_{11}$ are $q_k^{(n)} \times q_k^{(n)}$ matrices. Let $\hat{\boldsymbol{\mu}}_{1X} - \boldsymbol{\mu}_{1X}^0 = (\boldsymbol{\eta}_1^\top, \boldsymbol{\eta}_2^\top)^\top$ with $\boldsymbol{\eta}_1$ a $q_k^{(n)}$-dimensional vector. Hence,

$$\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{1X} - \boldsymbol{\mu}_{1X}^0) = \hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{C}}_{11}\boldsymbol{\eta}_1 + \hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{C}}_{12}\boldsymbol{\eta}_2 = \hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{C}}_{11}\boldsymbol{\eta}_1 - \hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{C}}_{22}\boldsymbol{\eta}_2.$$

On one hand,

$$(\hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{C}}_{11}\boldsymbol{\eta}_1)^2 \le (\hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{C}}_{11}\hat{\boldsymbol{\delta}}_{k,1})(\boldsymbol{\eta}_1^\top \hat{\boldsymbol{C}}_{11}\boldsymbol{\eta}_1) = (\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX})(\boldsymbol{\eta}_1^\top \hat{\boldsymbol{C}}_{11}\boldsymbol{\eta}_1) = O_p(\frac{q_k^{(n)}}{n})(\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}).$$

On the other hand,

$$\begin{aligned}(\hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{C}}_{22}\boldsymbol{\eta}_2)^2 &\le (\hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\delta}}_{k,1})(\boldsymbol{\eta}_2^\top \hat{\boldsymbol{C}}_{22}\hat{\boldsymbol{\Sigma}}_{12}^\top \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{C}}_{22}\boldsymbol{\eta}_2) \le (\hat{\boldsymbol{\delta}}_{k,1}^\top \hat{\boldsymbol{C}}_{11}\hat{\boldsymbol{\delta}}_{k,1})(\boldsymbol{\eta}_2^\top \hat{\boldsymbol{C}}_{22}\hat{\boldsymbol{\Sigma}}_{12}^\top \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{C}}_{22}\boldsymbol{\eta}_2) \\ &= (\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX})(\boldsymbol{\eta}_2^\top \hat{\boldsymbol{C}}_{22}\hat{\boldsymbol{\Sigma}}_{12}^\top \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{C}}_{22}\boldsymbol{\eta}_2) = (\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX})(\boldsymbol{\eta}_2^\top \boldsymbol{C}_{22}^0(\boldsymbol{\Sigma}_{12}^0)^\top(\boldsymbol{\Sigma}_{11}^0)^{-1}\boldsymbol{\Sigma}_{12}^0\boldsymbol{C}_{22}^0\boldsymbol{\eta}_2[1 + O_p(d_n)]) \\ &= (\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX})(\omega_n[1 + O_p(d_n)]),\end{aligned}$$

where the forth equation is obtained from (21), and $\omega_n = (\boldsymbol{\eta}_2^\top \boldsymbol{C}_{22}^0(\boldsymbol{\Sigma}_{12}^0)^\top(\boldsymbol{\Sigma}_{11}^0)^{-1}\boldsymbol{\Sigma}_{12}^0\boldsymbol{C}_{22}^0\boldsymbol{\eta}_2$. Hence, we have

$$\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{1X} - \boldsymbol{\mu}_{1X}^0) = O_p(\sqrt{\frac{q_k^{(n)}}{n}})\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}} - \sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}}\sqrt{\omega_n[1 + O_p(d_n)]}.$$

Under Condition C1,

$$E(\omega_n) \le \theta E(\boldsymbol{\eta}_2^\top \boldsymbol{C}_{22}^0(\boldsymbol{\Sigma}_{12}^0)^\top \boldsymbol{\Sigma}_{12}^0\boldsymbol{C}_{22}^0\boldsymbol{\eta}_2) = \frac{\theta}{n}\mathrm{tr}[\boldsymbol{\Sigma}_{12}^0\boldsymbol{C}_{22}^0\boldsymbol{\Sigma}_{22}^0\boldsymbol{C}_{22}^0(\boldsymbol{\Sigma}_{12}^0)^\top] \le \frac{\theta^4}{n}\mathrm{tr}[\boldsymbol{\Sigma}_{12}^0(\boldsymbol{\Sigma}_{12}^0)^\top].$$

Recall that $\boldsymbol{\Sigma}^0 = (\sigma_{ij}^0)_{1\le i,j\le p}$, then

$$E(\omega_n) \le \frac{\theta^4}{n}\sum_{i=1}^{q_k^{(n)}}\sum_{j=q_k^{(n)}+1}^{p}(\sigma_{ij}^0)^2 \le \frac{\theta^4}{n}q_k^{(n)}\max_i\sum_{j=q_k^{(n)}+1}^{p}(\sigma_{ij}^0)^2 \le \frac{\theta^{6-h}}{n}q_k^{(n)}\max_{j\le p}\sum_{i=1}^{p}|\sigma_{ij}^0|^h = O(\frac{S_{h;p}q_k^{(n)}}{n}).$$

Consequently,

$$\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{1X} - \boldsymbol{\mu}_{1X}^0) = O_p(\sqrt{\frac{q_k^{(n)}}{n}})\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}} - O_p(\sqrt{\frac{S_{h;p}q_k^{(n)}}{n}})\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}}.$$

Similarly, we have

$$\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{kX} - \boldsymbol{\mu}_{kX}^0) = O_p(\sqrt{\frac{q_k^{(n)}}{n}})\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}} - O_p(\sqrt{\frac{S_{h;p}q_k^{(n)}}{n}})\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}}.$$

□

Lemma 4. *For $t \in \{1, \ldots, K\}$ and $k \in \{2, \ldots, K\}$, we have*

$$\begin{aligned}&(\boldsymbol{\mu}_{tX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0) - (\frac{\hat{\boldsymbol{\mu}}_{1X} + \hat{\boldsymbol{\mu}}_{kX}}{2})^\top \hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} + (\frac{\boldsymbol{\mu}_{1X}^0 + \boldsymbol{\mu}_{kX}^0}{2})^\top \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0 \\ =&\Delta^2\left[O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n) + O_p(\frac{\sqrt{S_{h;p}q_k^{(n)}}}{\sqrt{n}\Delta_k})\right] + \frac{1}{2}(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)^\top \boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)\end{aligned}$$

for the multi-class problem.

Proof: It is not difficult to derive

$$
\begin{aligned}
&(\boldsymbol{\mu}_{tX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)-(\frac{\hat{\boldsymbol{\mu}}_{1X}+\hat{\boldsymbol{\mu}}_{kX}}{2})^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}+(\frac{\boldsymbol{\mu}_{1X}^0+\boldsymbol{\mu}_{kX}^0}{2})^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0\\
=&(\boldsymbol{\mu}_{tX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)-\frac{1}{2}\{(\hat{\boldsymbol{\mu}}_{1X}+\hat{\boldsymbol{\mu}}_{kX})-(\boldsymbol{\mu}_{1X}^0+\boldsymbol{\mu}_{kX}^0)\}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\frac{1}{2}(\boldsymbol{\mu}_{1X}^0+\boldsymbol{\mu}_{kX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)\\
=&\{\frac{(\boldsymbol{\mu}_{tX}^0-\boldsymbol{\mu}_{1X}^0)+(\boldsymbol{\mu}_{tX}^0-\boldsymbol{\mu}_{kX}^0)}{2}\}^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)-\frac{1}{2}\{(\hat{\boldsymbol{\mu}}_{1X}-\boldsymbol{\mu}_{1X}^0)+(\hat{\boldsymbol{\mu}}_{kX}-\boldsymbol{\mu}_{kX}^0)\}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}\\
=&K(\boldsymbol{\delta}_{tX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)-\frac{K}{2}(\boldsymbol{\delta}_{kX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)-\frac{1}{2}\{(\hat{\boldsymbol{\mu}}_{1X}-\boldsymbol{\mu}_{1X}^0)+(\hat{\boldsymbol{\mu}}_{kX}-\boldsymbol{\mu}_{kX}^0)\}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}.
\end{aligned}
$$

Because

$$
(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)=\Delta_k^2+\Delta_t^2-2(\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0,
$$

we hence have

$$
-(\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0=\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)-\frac{1}{2}(\Delta_k^2+\Delta_t^2)\le\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)-\Delta_k\Delta_t.
$$

Consequently, applying the Cauchy-Schwarz inequality together with (26) and (27), we can obtain

$$
\begin{aligned}
(\boldsymbol{\delta}_{tX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)\le&\Delta_t\sqrt{1+O_p(d_n)}\,\Delta_k\sqrt{1+O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)}+\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)-\Delta_t\Delta_k\\
\le&\Delta_t\Delta_k[1+O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)]-\Delta_t\Delta_k+\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)\\
\le&\Delta^2[O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)]+\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0).
\end{aligned}
$$

Similarly,

$$
(\boldsymbol{\delta}_{kX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)=\Delta_k^2[O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)]\le\Delta^2[O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)].
$$

As a result, according to Lemma 3, we have

$$
\begin{aligned}
&(\boldsymbol{\mu}_{tX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}-\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)-(\frac{\hat{\boldsymbol{\mu}}_{1X}+\hat{\boldsymbol{\mu}}_{kX}}{2})^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}+(\frac{\boldsymbol{\mu}_{1X}^0+\boldsymbol{\mu}_{kX}^0}{2})^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0\\
\le&\Delta^2[O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)]+\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)+\left\{O_p(\sqrt{\frac{S_{h;p}q_k^{(n)}}{n}})-O_p(\sqrt{\frac{q_k^{(n)}}{n}})\right\}\sqrt{\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}}\\
\le&\Delta^2[O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)]+\Delta_kO_p(\sqrt{\frac{S_{h;p}q_k^{(n)}}{n}})\sqrt{1+O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)}+\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)\\
=&\Delta^2[O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)]+\Delta_k^2[O_p(\frac{b_k^{(n)}}{\Delta_k^2})+O_p(\frac{d_n}{\Delta_k})+O_p(\frac{\sqrt{S_{h;p}q_k^{(n)}}}{\sqrt{n}\Delta_k})]+\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)\\
\le&\Delta^2\left[O_p(\frac{b_k^{(n)}}{\Delta_k})+O_p(d_n)+O_p(\frac{\sqrt{S_{h;p}q_k^{(n)}}}{\sqrt{n}\Delta_k})\right]+\frac{1}{2}(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0-\boldsymbol{\delta}_{tX}^0).
\end{aligned}
$$

□

From Lemma 4, note that when $t = k$, we have $(\boldsymbol{\mu}_{kX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0) - (\hat{\boldsymbol{\mu}}_{1X} + \hat{\boldsymbol{\mu}}_{kX})^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}/2 + (\boldsymbol{\mu}_{1X}^0 + \boldsymbol{\mu}_{kX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0/2 = \Delta^2\left[O_p(b_k^{(n)}/\Delta_k) + O_p(d_n) + O_p(\sqrt{S_{h;p}q_k^{(n)}}/(\sqrt{n}\Delta_k))\right]$. With Lemmas 2 - 4, we are ready to complete the proof of Theorem 1.

Proof of Theorem 1: Let $\hat{Z}_{\text{PROP}}$ and $\hat{Z}_{\text{Bayes}}$ denote the predicted class labels obtained by the proposed model and the Bayes rule, respectively. For simplicity, we assume $\pi_1 = \ldots = \pi_K$ instead of Condition C7 in the proofs of Theorems 1 - 3 with no influence on the theoretical results, since Condition C7 is only used to bound the term $\ln(\pi_k/\pi_1)$ in the LDA rule. Define $\vartheta_k = (\boldsymbol{x} - (\boldsymbol{\mu}_{1X}^0 + \boldsymbol{\mu}_{kX}^0)/2)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0$ and $\hat{\vartheta}_k = (\boldsymbol{x} - (\hat{\boldsymbol{\mu}}_{1X} + \hat{\boldsymbol{\mu}}_{kX})/2)^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}$ for a new sample $\boldsymbol{x}$. Then for any $\epsilon > 0$,

$$\begin{aligned} R_{\text{PROP}}(\mathcal{T}) - R_{\text{Bayes}} &\le \Pr(\hat{Z}_{\text{PROP}} \neq \hat{Z}_{\text{Bayes}}) \le 1 - \Pr(|\hat{\vartheta}_k - \vartheta_k| < \frac{\epsilon}{2}, |\vartheta_k - \vartheta_\ell| > \epsilon \text{ for any } k, \ell) \\ &\le \Pr(|\hat{\vartheta}_k - \vartheta_k| \ge \frac{\epsilon}{2} \text{ for some } k) + \Pr(|\vartheta_k - \vartheta_\ell| \le \epsilon \text{ for some } k, \ell). \end{aligned}$$

Firstly, we bound the probability $\Pr(|\vartheta_k - \vartheta_\ell| \le \epsilon \text{ for some } k, \ell)$. Since $\vartheta_k - \vartheta_\ell = \boldsymbol{x}^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{\ell X}^0) - (\boldsymbol{\mu}_{1X}^0 + \boldsymbol{\mu}_{kX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0/2 + (\boldsymbol{\mu}_{1X}^0 + \boldsymbol{\mu}_{\ell X}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{\ell X}^0/2$, the variance of $\vartheta_k - \vartheta_\ell$ is $(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{\ell X}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{\ell X}^0)$. Hence,

$$\Pr(|\vartheta_k - \vartheta_\ell| \le \epsilon \text{ for some } k, \ell) = \sum_{t=1}^K \Pr(|\vartheta_k - \vartheta_\ell| \le \epsilon | Z = t)\pi_t \le \sum_{k,\ell,t}\pi_t\frac{C\epsilon}{\sqrt{(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{\ell X}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{\ell X}^0)}} \le CK^2\epsilon,$$

where the last inequality is obtained by Condition C6. Secondly, we bound the term $\Pr(|\hat{\vartheta}_k - \vartheta_k| \ge \epsilon/2 \text{ for some } k)$. As $(\hat{\vartheta}_k - \vartheta_k | Z = t) = \boldsymbol{x}^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0) - (\hat{\boldsymbol{\mu}}_{1X} + \hat{\boldsymbol{\mu}}_{kX})^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX}/2 + (\boldsymbol{\mu}_{1X}^0 + \boldsymbol{\mu}_{kX}^0)^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0/2$, the conditional difference term $(\hat{\vartheta}_k - \vartheta_k | Z = t)$ is from normal distribution $N(\mu_\vartheta, \sigma_\vartheta^2)$ with

$$\mu_\vartheta^{(t)} = (\boldsymbol{\mu}_{tX}^0)^\top(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0) - (\frac{\hat{\boldsymbol{\mu}}_{1X} + \hat{\boldsymbol{\mu}}_{kX}}{2})^\top\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} + (\frac{\boldsymbol{\mu}_{1X}^0 + \boldsymbol{\mu}_{kX}^0}{2})^\top\boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0$$

and

$$\sigma_\vartheta^2 = (\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X - (\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0)\boldsymbol{\Sigma}_X^0(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0).$$

By Markov's inequality, together with Lemmas 2 and 4, we have

$$\begin{aligned} \Pr(|\hat{\vartheta}_k - \vartheta_k| \ge \frac{\epsilon}{2} \text{ for some } k) &= \sum_{t\neq k}^K \pi_t \Pr(|\hat{\vartheta}_k - \vartheta_k| \ge \frac{\epsilon}{2} | Z = t) + \pi_k \Pr(|\hat{\vartheta}_k - \vartheta_k| \ge \frac{\epsilon}{2} | Z = k) \\ \le& \frac{C\max_k(\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X - (\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0)\boldsymbol{\Sigma}_X^0(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)}{(\epsilon - \mu_\vartheta^{(t\neq k)})^2} + \frac{(\hat{\boldsymbol{\delta}}_{kX}^\top\hat{\boldsymbol{C}}_X - (\boldsymbol{\delta}_{kX}^0)^\top\boldsymbol{C}_X^0)\boldsymbol{\Sigma}_X^0(\hat{\boldsymbol{C}}_X\hat{\boldsymbol{\delta}}_{kX} - \boldsymbol{C}_X^0\boldsymbol{\delta}_{kX}^0)}{(\epsilon - \mu_\vartheta^{(k)})^2} \\ \le& \frac{C\max_k \Delta_k^2[O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)]}{\left[\epsilon - \Delta^2[O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n) + O_p(\frac{\sqrt{S_{h;p}q_k^{(n)}}}{\sqrt{n}\Delta_k})] - \frac{1}{2}(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)\right]^2} \\ &+ \frac{\Delta_k^2[O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n)]}{\left[\epsilon - \Delta^2[O_p(\frac{b_k^{(n)}}{\Delta_k}) + O_p(d_n) + O_p(\frac{\sqrt{S_{h;p}q_k^{(n)}}}{\sqrt{n}\Delta_k})]\right]^2} \\ \le& \frac{\Delta^2 O_p(\xi_{n;k})}{[\epsilon - \Delta^2 O_p(\xi_{n;k}) - \frac{1}{2}(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)^\top\boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)]^2} + \frac{\Delta^2 O_p(\xi_{n;k})}{[\epsilon - \Delta^2 O_p(\xi_{n;k})]^2}. \end{aligned} \tag{29}$$

By Condition C6, the first term of (29) converges to 0 in probability. Choose $\epsilon = C\xi_{n;k}^{\alpha}$, where $0 < \alpha < 1/2$ with a positive constant C, then

$$\Pr(|\hat{\vartheta}_k - \vartheta_k| \geq \frac{\epsilon}{2} \text{ for some } k) \leq \frac{\Delta^2 O_p(\xi_{n;k})}{[\epsilon - \Delta^2 O_p(\xi_{n;k}) - \frac{1}{2}(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)^\top \boldsymbol{C}_X^0(\boldsymbol{\delta}_{kX}^0 - \boldsymbol{\delta}_{tX}^0)]^2} + \frac{\Delta^2 O_p(\xi_{n;k}^{1-2\alpha})}{[C - \Delta^2 O_p(\xi_{n;k}^{1-\alpha})]^2} \xrightarrow{P} 0.$$

□

Theorem 1 establishes the sub-optimality property of the proposed classification rule for the multi-class problem. In the case of two-class problem, the Bayes error rate can be expressed in a closed form of

$$R_{\text{Bayes}} = \Phi(-\Delta_2/2)$$

when the data are from normal distribution, where Φ represents the cumulative distribution function of $N(0, 1)$, and $\Delta_2 = \sqrt{(\boldsymbol{\delta}_{2X}^0)^\top \boldsymbol{C}_X^0 \boldsymbol{\delta}_{2X}^0} = \sqrt{(\boldsymbol{\mu}_{2X}^0 - \boldsymbol{\mu}_{1X}^0)^\top \boldsymbol{C}_X^0 (\boldsymbol{\mu}_{2X}^0 - \boldsymbol{\mu}_{1X}^0)}$. Accordingly, in Theorems 2 and 3, we can compute the convergence rate of the proposed rule for the two-class problem, and subsequently investigate its properties.

Theorem 2. *Assume that Conditions C1 - C7 hold with $K = 2$, and*

$$\xi_n = \max\{d_n, \frac{b_2^{(n)}}{\Delta_2}, \frac{\sqrt{s_2 S_{h;p}}}{\sqrt{n}\Delta_2}\} \to 0.$$

Then we have $R_{PROP}(\mathcal{T}) = \Phi(-(\Delta_2/2)[1 + O_p(\xi_n)])$.

Theorem 3. *Assuming that all the conditions in Theorem 2 are satisfied, we have*
(i) if Δ_2 is bounded, then the proposed rule is asymptotically optimal and $R_{PROP}(\mathcal{T})/R_{Bayes} - 1 = O_p(\xi_n)$;
(ii) if $\Delta_2 \to \infty$, then the proposed rule is asymptotically sub-optimal;
(iii) if $\Delta_2 \to \infty$ and $\xi_n \Delta_2^2 \to 0$, then the proposed rule is asymptotically optimal.

Proof of Theorem 2: The conditional misclassification rate is

$$\begin{aligned} R_{\text{PROP}}(\mathcal{T}) &= \frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{(-1)^k \hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X(\boldsymbol{\mu}_{kX}^0 - \hat{\boldsymbol{\mu}}_{kX}) - \hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{1X} - \hat{\boldsymbol{\mu}}_{2X})/2}{\sqrt{\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \boldsymbol{\Sigma}_X^0 \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}}\right) \\ &= \frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{(-1)^k \hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X(\boldsymbol{\mu}_{kX}^0 - \hat{\boldsymbol{\mu}}_{kX}) - \hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}{\sqrt{\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \boldsymbol{\Sigma}_X^0 \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}}\right). \end{aligned}$$

By the result (21), we have

$$\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \boldsymbol{\Sigma}_X^0 \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X} = \hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}[1 + O_p(d_n)] = \hat{\boldsymbol{\delta}}_{2X}^\top \boldsymbol{C}_X^0 \hat{\boldsymbol{\delta}}_{2X}[1 + O_p(d_n)].$$

From the result (22), together with $E[(\boldsymbol{\delta}_{2X}^0)^\top \boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{2X} - \boldsymbol{\delta}_{2X}^0)]^2 \leq \Delta_2^2 E[(\hat{\boldsymbol{\delta}}_{2X} - \boldsymbol{\delta}_{2X}^0)^\top \boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{2X} - \boldsymbol{\delta}_{2X}^0)]$, it is easy to derive

$$\begin{aligned} \hat{\boldsymbol{\delta}}_{2X}^\top \boldsymbol{C}_X^0 \hat{\boldsymbol{\delta}}_{2X} &= (\boldsymbol{\delta}_{2X}^0)^\top \boldsymbol{C}_X^0 \boldsymbol{\delta}_{2X}^0 + (\boldsymbol{\delta}_{2X}^0)^\top \boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{2X} - \boldsymbol{\delta}_{2X}^0) + (\hat{\boldsymbol{\delta}}_{2X} - \boldsymbol{\delta}_{2X}^0)^\top \boldsymbol{C}_X^0(\hat{\boldsymbol{\delta}}_{2X} - \boldsymbol{\delta}_{2X}^0) \\ &= \Delta_2^2 + O_p(b_2^{(n)}\Delta_2) + O_p((b_2^{(n)})^2) = \Delta_2^2[1 + O_p(\frac{b_2^{(n)}}{\Delta_2})]. \end{aligned}$$

Hence we have

$$\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X} = \hat{\boldsymbol{\delta}}_{2X}^\top \boldsymbol{C}_X^0 \hat{\boldsymbol{\delta}}_{2X}[1 + O_p(d_n)] = \Delta_2^2[1 + O_p(\frac{b_2^{(n)}}{\Delta_2}) + O_p(d_n)].$$

Therefore, by Lemma 3, we obtain

$$\frac{\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X(\hat{\boldsymbol{\mu}}_{1X} - \boldsymbol{\mu}_{1X}^0) - \hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}{\sqrt{\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \boldsymbol{\Sigma}_X^0 \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}} = \frac{O_p(\sqrt{\frac{q_2^{(n)}}{n}}) + O_p(\sqrt{\frac{S_{h;p} q_2^{(n)}}{n}}) - \sqrt{\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}}{\sqrt{1 + O_p(d_n)}}$$

$$= -\frac{\Delta_2}{2} \frac{\sqrt{1 + O_p(\frac{b_2^{(n)}}{\Delta_2}) + O_p(d_n)}}{\sqrt{1 + O_p(d_n)}} + \frac{O_p(\sqrt{\frac{S_{h;p} q_2^{(n)}}{n}})}{\sqrt{1 + O_p(d_n)}} = -\frac{\Delta_2}{2}[1 + O_p(\frac{b_2^{(n)}}{\Delta_2}) + O_p(d_n)] + O_p(\sqrt{\frac{S_{h;p} q_2^{(n)}}{n}})$$

$$= -\frac{\Delta_2}{2}[1 + O_p(\xi_n)].$$

Similarly, we have

$$\frac{\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X(\boldsymbol{\mu}_{2X}^0 - \hat{\boldsymbol{\mu}}_{2X}) - \hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}{\sqrt{\hat{\boldsymbol{\delta}}_{2X}^\top \hat{\boldsymbol{C}}_X \boldsymbol{\Sigma}_X^0 \hat{\boldsymbol{C}}_X \hat{\boldsymbol{\delta}}_{2X}}} = -\frac{\Delta_2}{2}[1 + O_p(\xi_n)],$$

which proves the theorem. □

To establish the theoretical results in Theorem 3, we need a lemma from Shao et al. [42]. We state it here for completeness, and then prove Theorem 3.

Lemma 5. *Let $a_n^{(1)}$ and $a_n^{(2)}$ be two sequences of positive numbers such that $a_n^{(1)} \to \infty$ and $a_n^{(2)} \to 0$ as $n \to \infty$. If $\lim_{n\to\infty} a_n^{(1)} a_n^{(2)} = \rho$, where ρ may be 0, positive, or ∞, then*

$$\lim_{n\to\infty} \frac{\Phi(-\sqrt{a_n^{(1)}}(1 - a_n^{(2)}))}{\Phi(-\sqrt{a_n^{(1)}})} = e^\rho.$$

Proof: See the proof of Lemma 1 in Shao et al. [42]. □

Proof of Theorem 3: (i) Let ϕ be the density function of $N(0, 1)$. By the mean value theorem,

$$R_{\text{PROP}}(\mathcal{T}) - R_{\text{Bayes}} = \Phi(-\frac{\Delta_2}{2}[1 + O_p(\xi_n)]) - \Phi(-\frac{\Delta_2}{2}) = -\phi(\tau_n)\frac{\Delta_2}{2} O_p(\xi_n),$$

where τ_n is between $-\Delta_2/2$ and $-(\Delta_2/2)[1 + O_p(\xi_n)]$. Since Δ_2 is bounded, then R_{Bayes} is bounded away from 0. Hence,

$$\frac{R_{\text{PROP}}(\mathcal{T})}{R_{\text{Bayes}}} - 1 = -\frac{\Delta_2}{2}\frac{\phi(\tau_n)}{R_{\text{Bayes}}} O_p(\xi_n) = O_p(\xi_n).$$

(ii) When $\Delta_2 \to \infty$, we have $R_{\text{PROP}}(\mathcal{T}) \xrightarrow{P} 0$. This, together with $\lim_{\Delta_2\to\infty} R_{\text{Bayes}} = 0$, proves (2).

(iii) The conditions $\Delta_2 \to \infty$, $\lim_{n\to\infty} \xi_n \Delta_2^2 = 0$, together with Lemma 5, prove that $R_{\text{PROP}}(\mathcal{T})/R_{\text{Bayes}} \xrightarrow{P} 1$. □

Theorem 2 provides the convergence rate of the proposed classification rule for the two-class problem with respect to ξ_n. Base on such a result, Theorem 3 demonstrates that the property of the proposed classification rule (optimality or sub-optimality) depends on the scenarios of the true model's Δ_2. Specifically, (1) when Δ_2 is bounded, i.e., $\lim_{n\to\infty} R_{\text{Bayes}} > 0$, then $R_{\text{PROP}}(\mathcal{T})$ converges in probability to the same limit as R_{Bayes}. (2) When $\Delta_2 \to \infty$, i.e., $R_{\text{Bayes}} \to 0$, then $R_{\text{PROP}}(\mathcal{T}) \xrightarrow{P} 0$; in this case, if we further have $\xi_n \Delta_2^2 \to 0$, then $R_{\text{PROP}}(\mathcal{T})$ and R_{Bayes} have the same convergence rate.

3.2. Consistency property of the prediction of y

Now, we derive the consistency property for the proposed estimate of y. Denote by $\hat{y}^P$ the predicted value of y obtained from the proposed model. Define $\hat{y}^B$ to be the predicted value of y for $\boldsymbol{x}$ when all parameters are known. Specifically, first obtain the class label k via the Bayes LDA rule, then $\hat{y}^B = y_k = \mu_{ky} + \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{kX})$. Hence, the mean squared errors (MSE) of estimates $\hat{y}^B$ and $\hat{y}^P$ are $\text{MSE}_{\text{Bayes}} = E[(\hat{y}^B - y)^2|\mathcal{T}]$ and $\text{MSE}_{\text{PROP}} = E[(\hat{y}^P - y)^2|\mathcal{T}]$. Now we establish the theoretical results of $\hat{y}^P$ in Theorem 4.

Theorem 4. *Assume that Conditions C1 - C7 hold and the conditions in Theorem 1 are satisfied. Then we have*

$$MSE_{PROP} - MSE_{Bayes} \xrightarrow{P} 0,$$

for the multi-class qualitative response.

Proof of Theorem 4: Define $r_{ik} = \Pr(\hat{Z} = i|Z = k)$ for $i, k \in \{1, \ldots, K\}$. Let R be the misclassification error for a classifier, it is then calculated via

$$R = \sum_{k=1}^{K} \Pr(Z = k) \Pr(\hat{Z} \neq k|Z = k) = \sum_{k=1}^{K} \left(\pi_k \sum_{i \neq k}^{K} r_{ik} \right). \tag{30}$$

Now we derive an upper bound of $(\hat{y} - y)^2$. Since it is random, we focus on the average, i.e.,

$$E[(\hat{y} - y)^2|\boldsymbol{x}, \mathcal{T}] = E_y E_{\hat{y}|\mathcal{T}}[(\hat{y} - y)^2|\boldsymbol{x}, \mathcal{T}]. \tag{31}$$

To simplify the notation, we omit $\boldsymbol{x}$ and $\mathcal{T}$ and write it as $E_y E_{\hat{y}|\mathcal{T}}[(\hat{y} - y)^2]$. Then (31) becomes

$$E[(\hat{y} - y)^2] = E_y E_{\hat{y}|\mathcal{T}}[(\hat{y} - y)^2] = E_Z[E_{y|Z} E_{\hat{y}|\mathcal{T}}[(\hat{y} - y)^2]|Z] = \sum_{k=1}^{K} \pi_k E_{y|Z=k} \left[\sum_{i=1}^{K} r_{ik} (\hat{y}_i - y)^2 | Z = k \right].$$

Next, we derive

$$\begin{aligned} &E_{y|Z=1}\left[a_1(\hat{y}_1 - y)^2|Z = 1\right] = E_{y|Z=1}\left[a_1\{\hat{y}_1 - E(y|Z = 1) + E(y|Z = 1) - y\}^2|Z = 1\right] \\ &= E_{y|Z=1}\left[a_1\{\hat{y}_1 - E(y|Z = 1)\}^2|Z = 1\right] + E_{y|Z=1}\left[a_1\{y - E(y|Z = 1)\}^2|Z = 1\right] = a_1\{\hat{y}_1 - E(y|Z = 1)\}^2 + a_1 \text{Var}(y|Z = 1). \end{aligned}$$

Similarly, we have for $c > 0$ and $i, k \in \{1, \ldots, K\}$

$$E_{y|Z=k}\left[c(\hat{y}_i - y)^2|Z = k\right] = c\{\hat{y}_i - E(y|Z = k)\}^2 + c\text{Var}(y|Z = k).$$

As a result, (31) is decomposed as

$$\begin{aligned} E[(\hat{y} - y)^2|\boldsymbol{x}, \mathcal{T}] &= \sum_{k=1}^{K} \pi_k \left[\sum_{i=1}^{K} r_{ik}\{\hat{y}_i - E(y|Z = k)\}^2 + \sum_{i=1}^{K} r_{ik} \text{Var}(y|Z = k) \right] \\ &= \sum_{k=1}^{K} \sum_{i=1}^{K} \pi_k r_{ik}\{\hat{y}_i - E(y|Z = k)\}^2 + (\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{Xy}) \sum_{k=1}^{K} \sum_{i=1}^{K} \pi_k r_{ik} \\ &= \sum_{k=1}^{K} \sum_{i=1}^{K} \pi_k r_{ik}\{\hat{y}_i - E(y|Z = k)\}^2 + (\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{Xy}), \end{aligned}$$

where in the second equality $\text{Var}(y|Z = k) = \sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{Xy}$ is used, and the third equality uses $\sum_{k=1}^{K} \sum_{i=1}^{K} \pi_k r_{ik} = 1$. Now we tackle with each term of

$$\{\hat{y}_k - E(y|Z = k)\}^2 = \left\{ (\hat{\mu}_{ky} - \mu_{ky}) + \left(\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1} \right) \boldsymbol{x} - \left(\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1} \hat{\boldsymbol{\mu}}_{kX} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_{kX} \right) \right\}^2$$

and

$$\begin{aligned}\{\hat{y}_k - E(y|Z=k')\}^2 &= \left\{(\hat{\mu}_{ky} - \mu_{k'y}) + \left(\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\right)\boldsymbol{x} - \left(\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\mu}}_{kX} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_{k'X}\right)\right\}^2 \\ &= \left\{(\hat{\mu}_{ky} - \mu_{ky}) + \left(\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\right)\boldsymbol{x} - \left(\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\mu}}_{kX} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_{kX}\right)\right. \\ &\quad \left. + (\mu_{ky} - \mu_{k'y}) - \left(\boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_{kX} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_{k'X}\right)\right\}^2, \quad k \neq k'.\end{aligned}$$

For $k \in \{1, \ldots, K\}$ and $k \neq k'$, define the following terms

$$\begin{aligned}&b_{ky} = \hat{\mu}_{ky} - \mu_{ky}, \quad D_{kk'} = E(y|Z=k) - E(y|Z=k') = (\mu_{ky} - \mu_{k'y}) - \left(\boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_{kX} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_{k'X}\right), \\ &\boldsymbol{h} = \left(\hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\right)^\top, \quad d_k = \hat{\boldsymbol{\Sigma}}_{Xy}^\top \hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\mu}}_{kX} - \boldsymbol{\Sigma}_{Xy}^\top \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_{kX}.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}E[(\hat{y}-y)^2|\boldsymbol{x},\mathcal{T}] &= \sum_{i=1}^K \pi_i r_{ii}(b_{iy} + \boldsymbol{h}^\top\boldsymbol{x} - d_i)^2 + \sum_{k=1}^K\sum_{i\neq k}^K \pi_k r_{ik}(b_{iy} + \boldsymbol{h}^\top\boldsymbol{x} - d_i + D_{ik})^2 + (\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}) \\ &= \sum_{i=1}^K \pi_i r_{ii}(b_{iy} + \boldsymbol{h}^\top\boldsymbol{x} - d_i)^2 + \sum_{k=1}^K\sum_{i\neq k}^K \pi_k r_{ik}(b_{iy} + \boldsymbol{h}^\top\boldsymbol{x} - d_i)^2 + \sum_{k=1}^K\sum_{i\neq k}^K \pi_k r_{ik} D_{ik}^2 \\ &\quad + \sum_{k=1}^K\sum_{i\neq k}^K 2\pi_k r_{ik}(b_{iy} + \boldsymbol{h}^\top\boldsymbol{x} - d_i)D_{ik} + (\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}) \\ &= \mathbb{M} + \sum_{k=1}^K\sum_{i\neq k}^K \pi_k r_{ik} D_{ik}^2 + (\sigma_y^2 - \boldsymbol{\Sigma}_{Xy}^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{Xy}),\end{aligned}$$

where

$$\mathbb{M} = \sum_{k=1}^K\sum_{i=1}^K \pi_k r_{ik}(b_{iy} + \boldsymbol{h}^\top\boldsymbol{x} - d_i)^2 + \sum_{k=1}^K\sum_{i\neq k}^K 2\pi_k r_{ik}(b_{iy} + \boldsymbol{h}^\top\boldsymbol{x} - d_i)D_{ik}.$$

Now if the classification of Z is based on the known distribution, the misclassification rate R is R_{Bayes}. For $i, k \in \{1, \ldots, K\}$, let $r_{ik}^B = \Pr(\hat{Z} = i|Z = k)$ represent the corresponding r_{ik} with $\hat{Z}$ obtained from Bayes rule. Similarly, let symbol r_{ik}^P be the corresponding r_{ik} with $\hat{Z}$ from the proposed model. Denote by $\mathbb{M}_{\text{PROP}}$ the value of $\mathbb{M}$ computed from the proposed model. Note that the value of $\mathbb{M}$ computed from the Bayes rule is equal to 0. Then we have

$$E[(\hat{y}^P - y)^2|\boldsymbol{x},\mathcal{T}] - E[(\hat{y}^B - y)^2|\boldsymbol{x},\mathcal{T}] = \mathbb{M}_{\text{PROP}} + \sum_{k=1}^K\sum_{i\neq k}^K (\pi_k r_{ik}^P - \pi_k r_{ik}^B)D_{ik}^2 \leq \mathbb{M}_{\text{PROP}} + [R_{\text{PROP}}(\mathcal{T}) - R_{\text{Bayes}}]D_{\max}^2,$$

where $D_{\max}^2 = \max\{D_{kk'}^2\}$, and the last inequality uses (30). By conditions in Theorem 1, $E_{\boldsymbol{x}}(\mathbb{M}_{\text{PROP}}) \xrightarrow{P} 0$ as $n \to \infty$. Consequently, we have

$$\begin{aligned}\text{MSE}_{\text{PROP}} - \text{MSE}_{\text{Bayes}} &= E[(\hat{y}^P - y)^2|\mathcal{T}] - E[(\hat{y}^B - y)^2|\mathcal{T}] = E_{\boldsymbol{x}}E[(\hat{y}^P - y)^2|\boldsymbol{x},\mathcal{T}] - E_{\boldsymbol{x}}E[(\hat{y}^B - y)^2|\boldsymbol{x},\mathcal{T}] \\ &\leq E_{\boldsymbol{x}}(\mathbb{M}_{\text{PROP}}) + [R_{\text{PROP}}(\mathcal{T}) - R_{\text{Bayes}}]D_{\max}^2 \xrightarrow{P} 0.\end{aligned}$$

□

Theorem 4 compares the MSE of the proposed estimate of y with that from the optimal Bayes rule (under the assumption that all parameters are known). Since the classification errors from a classification rule might be larger than 0, the MSE of $\hat{y}$ may not converge to 0 even though the sample size n is sufficiently large. Here we adopt the $\text{MSE}_{\text{Bayes}}$ as a reasonable performance benchmark to evaluate the property of the proposed model with respect to y. Theorem 4 states that the difference of MSE between the proposed and the Bayes methods converges to 0 in probability.

4. Numerical results

In this section, we study the empirical performance of the proposed GAQQ model from multiple perspectives. In Sections 4.1 and 4.2, we examine the performance of our method via simulated QQ data with two-class and multi-class responses, respectively. In Section 4.3, two real-world data examples from material science and genetics are used to illustrate the merits of the proposed approach.

4.1. Two-class settings of the qualitative response

We evaluate the performance of the proposed GAQQ method for a binary response Z under different inverse covariance matrices $\boldsymbol{C}$ and mean differences $\boldsymbol{\delta}_2$. The proposed GAQQ model is compared with several benchmark methods, denoted as GLDA, CL, and ENET, which use the predictor variables $\boldsymbol{X}$ to predict Z and y. The GLDA employs the LDA classification rule for Z using the Moore-Penrose generalized inverse of the sample covariance matrix of $\boldsymbol{X}$ when $p > n$. The CL method applies the LPD technique [7] to predict the response Z based on $\boldsymbol{X}$. With their estimated class label of Z, the GLDA and CL predict y by (11). The ENET method uses the elastic-net logistic model [50] on predictor variables $\boldsymbol{X}$ to fit the response Z and hence predicts Z for the testing data. For the quantitative response y, the ENET separately fits two elastic-net linear regressions for two classes using training data, and then predicts y in the testing data based on its estimated Z. The tuning parameters of the CL and ENET methods are chosen by five-fold cross validation. Specifically, we split the training data into V sets with roughly equal size. In the ENET method, let $\hat{\boldsymbol{\beta}}^{(\nu)}_{\log}(\lambda)$ be the estimated coefficients of elastic-net logistic regression from the νth set of splitting data under given choice of two tuning parameters, $\nu \in \{1, \ldots, V\}$. Also let $\hat{\boldsymbol{Z}}^{(\nu)}(\lambda)$ be the predicted Z using $\hat{\boldsymbol{\beta}}^{(\nu)}_{\log}(\lambda)$ and training data excluding the νth set. The tuning parameters are chosen to minimize $\sum_{\nu=1}^{V} \|\hat{\boldsymbol{Z}}^{(\nu)}(\lambda) - \boldsymbol{Z}^{(-\nu)}\|_2^2$, where $\boldsymbol{Z}^{(-\nu)}$ represents the vector of values of Z in the training data without the νth set. Similarly, let $\hat{\boldsymbol{\beta}}^{(\nu)}_{\mathrm{lm}}(\lambda)$ be the estimated coefficients of elastic-net linear regression from the νth set of splitting data with two tuning parameters. Denote by $\hat{\boldsymbol{y}}^{(\nu)}(\lambda)$ the predicted y using $\hat{\boldsymbol{\beta}}^{(\nu)}_{\mathrm{lm}}(\lambda)$ and training data excluding the νth set. The tuning parameters are chosen to minimize $\sqrt{\sum_{\nu=1}^{V} \|\hat{\boldsymbol{y}}^{(\nu)}(\lambda) - \boldsymbol{y}^{(-\nu)}\|_2^2 / V}$, where $\boldsymbol{y}^{(-\nu)}$ represents the vector of values of y in the training data without the νth set. The cross-validation procedure of CL method for modeling the response Z is provided in detail in Cai and Liu [7]. The GLDA is implemented in the R software. The `glmnet`$(\cdot)$ function in R is used for the ENET method. The CL method is implemented by using `linprog`$(\cdot)$ function in the Matlab software. All numerical studies are carried out on an Intel Xeon Gold 6248 2.50 GHz processor.

Regarding the inverse covariance matrix $\boldsymbol{C}$, we consider the following five structures in the simulations, which are commonly used in the literature [47]:

Model 1. $\boldsymbol{C}_1 = \boldsymbol{I}$. $c_{ij} = 1$ if $i = j$ and 0 otherwise;

Model 2. $\boldsymbol{C}_2 = \mathrm{AR}(0.6)$. The conditional covariance between any two random variables is fixed to be $0.6^{|i-j|}$, $1 \le i, j \le p$;

Model 3. $\boldsymbol{C}_3$ is generated by randomly permuting rows and corresponding columns of the matrix $\boldsymbol{C}_2$;

Model 4. $\boldsymbol{C}_4 = \begin{pmatrix} \mathrm{CS}(0.6) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}$, where CS(0.6) represents a 5×5 compound symmetry matrix with diagonal entries 1 and others 0.6. $\boldsymbol{0}$ indicates a matrix with all entries 0;

Model 5. $\boldsymbol{C}_5 = \boldsymbol{\Theta} + \alpha \boldsymbol{I}$, where the diagonal entries of $\boldsymbol{\Theta}$ are zeros and $\boldsymbol{\Theta}_{ij} = \boldsymbol{\Theta}_{ji} = b * Unif(-1, 1)$ for $i \neq j$, where b is from the Bernoulli distribution with probability 0.15 equal 1. Each off-diagonal entry of $\boldsymbol{\Theta}$ is generated independently. The value of α is gradually increased to make sure that $\boldsymbol{C}_5$ is positive definite.

Model 1 is the simplest sparse matrix indicating that variables are independent of each other. Model 4 is a sparse matrix indicating that only the first 5 variables are correlated. This matrix includes more sparsity as the dimensionality increases. Models 2 and 3 are relatively dense matrices, and they also become more sparse when the dimensionality increases. All of these four matrices have sparse structures to some extent, while Model 5 is a general sparse matrix with no structure, which is similarly used in Bien and Tibshirani [4].

For the mean difference $\boldsymbol{\delta}_2$, we consider two different levels of sparsity. The $\boldsymbol{\mu}_1$ is the vector with all elements zeros, and the $\boldsymbol{\mu}_2$ is generated such that (S1): 25% of the elements in $\boldsymbol{\mu}_2$ are zeros; (S2): 75% of the elements in $\boldsymbol{\mu}_2$ are zeros. The positions of zeros in $\boldsymbol{\mu}_2$ are randomly distributed with its nonzero values independently generated from

Table 1: Averages and standard errors (in parenthesis) of misclassification errors (MEs) in percentage obtained from ENET, GLDA, CL and GAQQ methods for Models 1-5 based on 100 replications.

			Model 1	Model 2	Model 3	Model 4	Model 5
$p = 40$	S1	ENET	25.3(0.51)	25.4(0.44)	25.0(0.49)	24.8(0.46)	25.6(0.45)
		GLDA	9.28(0.42)	10.2(0.75)	9.40(0.66)	5.28(0.33)	9.13(0.33)
		CL	1.68(0.21)	10.9(1.04)	8.75(1.00)	4.58(0.51)	1.20(0.22)
		GAQQ	2.92(0.26)	9.93(0.41)	17.7(0.50)	2.33(0.24)	2.67(0.22)
	S2	ENET	25.3(0.45)	26.1(0.47)	24.3(0.42)	26.2(0.42)	25.1(0.41)
		GLDA	20.0(0.66)	9.78(0.41)	13.1(0.49)	22.8(0.63)	18.8(0.58)
		CL	6.92(0.39)	4.90(0.33)	8.63(0.50)	8.65(0.38)	8.48(0.40)
		GAQQ	5.32(0.32)	4.52(0.28)	7.10(0.28)	6.28(0.30)	5.88(0.29)
$p = 80$	S1	ENET	25.0(0.51)	24.4(0.44)	24.8(0.44)	23.9(0.48)	25.8(0.54)
		GLDA	7.88(0.43)	11.2(0.49)	12.6(0.52)	8.03(0.38)	11.3(0.49)
		CL	6.37(1.38)	10.4(0.66)	11.2(0.63)	5.63(1.48)	9.38(1.57)
		GAQQ	0.10(0.04)	2.97(0.22)	2.22(0.19)	0.07(0.03)	3.53(0.22)
	S2	ENET	24.8(0.42)	25.3(0.39)	25.2(0.43)	24.3(0.43)	23.9(0.52)
		GLDA	19.5(0.55)	26.7(0.68)	24.8(0.76)	16.1(0.63)	20.9(0.61)
		CL	4.67(0.37)	20.5(0.77)	18.2(1.01)	2.65(0.30)	14.5(0.92)
		GAQQ	1.08(0.13)	10.6(0.37)	5.42(0.30)	0.57(0.11)	3.33(0.22)
$p = 200$	S1	ENET	24.3(0.41)	24.9(0.52)	24.4(0.42)	25.2(0.43)	24.6(0.43)
		GLDA	2.25(0.20)	14.1(0.52)	13.6(0.42)	3.13(0.22)	7.23(0.36)
		CL	2.08(0.20)	2.88(0.18)	2.72(0.17)	2.12(0.21)	2.26(0.15)
		GAQQ	0.22(0.06)	0.47(0.09)	0.20(0.05)	0.23(0.07)	0.05(0.03)
	S2	ENET	25.3(0.40)	25.5(0.40)	24.6(0.47)	25.7(0.49)	25.3(0.51)
		GLDA	9.73(0.40)	20.5(0.55)	24.3(0.57)	9.08(0.40)	15.0(0.50)
		CL	1.46(0.14)	2.96(0.16)	2.29(0.11)	2.06(0.18)	2.38(0.16)
		GAQQ	0.01(0.00)	1.10(0.16)	1.55(0.16)	0.02(0.01)	0.17(0.05)

uniform distribution $Unif(0, 2)$. We consider $p \in \{40, 80, 200\}$, and generate $n_1 = 30$ observations from $N(\boldsymbol{\mu}_1, \boldsymbol{C}^{-1})$ as well as $n_2 = 30$ observations from $N(\boldsymbol{\mu}_2, \boldsymbol{C}^{-1})$ as the training set. The same procedure is employed to generate the testing data, which is used to evaluate the prediction performance of y and Z for different compared methods. We consider the root mean squared prediction error RMSPE $= \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}$ to measure the prediction accuracy for the quantitative response y, where $\hat{y}_i$ represents the predicted value. The prediction performance of the qualitative response Z is measured by the empirical misclassification error ME $= \sum_{i=1}^{n} I(z_i \neq \hat{z}_i)/n$, where $\hat{z}_i$ is the predicted value of z_i and $I(\cdot)$ is an indicator function. To make it easier for understanding the proposed methodology in the numerical studies, we summarize the implementation of GAQQ method in the following Algorithm 3 (based on Algorithms 1 and 2) for two-class settings of the qualitative response.

Algorithm 3

Step 1: Given (λ_1, λ_2), set $\hat{\boldsymbol{\delta}}_{2,0} = (\bar{\boldsymbol{w}}_1 - \bar{\boldsymbol{w}}_2)/2$.

Step 2: Given $\boldsymbol{\delta}_2 = \hat{\boldsymbol{\delta}}_{2,t}$, solve $\boldsymbol{C}$ in (7) by the Glasso at the tth iteration.

Step 3: Given $\boldsymbol{C} = \hat{\boldsymbol{C}}_t$, solve $\boldsymbol{\delta}_2$ in (9) by the Lasso at the tth iteration.

Step 4: Repeat Step 2 and 3 till both $\hat{\boldsymbol{C}}_t$ and $\hat{\boldsymbol{\delta}}_{2,t}$ converge.

Step 5: Choose the optimal values of $\boldsymbol{\delta}_2$ and $\boldsymbol{C}$ with respect to (λ_1, λ_2) via (10), denoted by $\hat{\boldsymbol{\delta}}_2$ and $\hat{\boldsymbol{C}}$ respectively.

Step 6: Calculate $\hat{\boldsymbol{\gamma}} = \bar{\boldsymbol{w}} + (n_2 - n_1)\hat{\boldsymbol{\delta}}_2/n$; compute $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\delta}}_2 + \hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\delta}}_2$.

Step 7: For a new observation $\boldsymbol{x}$, compute $\hat{y}_k = \hat{\mu}_{ky} - \hat{\boldsymbol{C}}_{Xy}^{\top}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{kX})/\hat{c}_y^2$ for $k \in \{1, 2\}$; obtain $\hat{p}_k$ by plugging $(\boldsymbol{x}^{\top}, \hat{y}_k)^{\top}$ into the density functions of $N(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{C}}^{-1})$.

Step 8: If $\hat{\pi}_1\hat{p}_1 > \hat{\pi}_2\hat{p}_2$, let $\hat{y} = \hat{y}_1$ and $\hat{Z} = 1$; otherwise let $\hat{y} = \hat{y}_2$ and $\hat{Z} = 2$.

Tables 1 and 2 report the averaged MEs in percentage and averaged RMPSE, as well as their corresponding standard errors in parenthesis for each approach over 100 replications. It can be seen from Table 1 that the proposed method generally outperforms other approaches with respect to MEs. Such an advantage becomes more significant as

Table 2: Averages and standard errors (in parenthesis) of root mean squared prediction errors (RMSPE) obtained from ENET, GLDA, CL and GAQQ methods for Models 1-5 based on 100 replications.

			Model 1	Model 2	Model 3	Model 4	Model 5
$p = 40$	S1	ENET	1.18(0.01)	1.62(0.02)	1.84(0.02)	1.91(0.01)	1.73(0.02)
		GLDA	1.82(0.03)	2.00(0.04)	2.03(0.04)	1.82(0.03)	1.82(0.03)
		CL	1.79(0.03)	1.97(0.06)	2.03(0.08)	1.65(0.03)	1.74(0.03)
		GAQQ	1.07(0.01)	1.21(0.01)	1.49(0.01)	1.02(0.01)	1.17(0.02)
	S2	ENET	1.59(0.01)	1.20(0.01)	1.42(0.02)	1.22(0.01)	1.12(0.01)
		GLDA	1.93(0.03)	1.96(0.03)	1.90(0.03)	1.85(0.03)	1.77(0.02)
		CL	1.82(0.03)	1.78(0.03)	1.70(0.03)	1.67(0.03)	1.58(0.03)
		GAQQ	1.07(0.01)	1.14(0.01)	1.37(0.01)	0.98(0.01)	1.09(0.01)
$p = 80$	S1	ENET	1.58(0.01)	1.77(0.02)	1.68(0.02)	1.37(0.01)	1.75(0.01)
		GLDA	2.01(0.03)	2.62(0.04)	2.38(0.04)	2.08(0.03)	1.92(0.03)
		CL	2.02(0.04)	2.63(0.07)	2.31(0.05)	1.88(0.03)	1.72(0.03)
		GAQQ	1.11(0.01)	1.26(0.01)	1.54(0.01)	1.11(0.01)	1.31(0.01)
	S2	ENET	1.08(0.01)	1.31(0.01)	1.44(0.02)	1.61(0.01)	1.11(0.01)
		GLDA	1.96(0.03)	2.56(0.05)	2.27(0.04)	2.04(0.03)	2.36(0.04)
		CL	1.76(0.03)	2.38(0.05)	2.10(0.04)	1.85(0.03)	2.20(0.04)
		GAQQ	1.02(0.01)	1.10(0.01)	1.39(0.01)	0.99(0.01)	1.11(0.01)
$p = 200$	S1	ENET	1.66(0.02)	1.24(0.01)	1.68(0.02)	1.07(0.01)	1.61(0.02)
		GLDA	1.24(0.01)	1.58(0.02)	1.62(0.02)	1.21(0.01)	1.40(0.02)
		CL	1.27(0.03)	1.60(0.02)	1.65(0.02)	1.28(0.02)	1.36(0.03)
		GAQQ	1.08(0.01)	1.27(0.01)	1.44(0.02)	1.06(0.01)	1.15(0.01)
	S2	ENET	1.03(0.01)	1.65(0.01)	1.62(0.01)	1.22(0.01)	1.17(0.01)
		GLDA	1.19(0.01)	1.67(0.02)	1.76(0.02)	1.22(0.01)	1.32(0.01)
		CL	1.19(0.01)	1.55(0.02)	1.74(0.02)	1.23(0.01)	1.33(0.01)
		GAQQ	1.01(0.01)	1.25(0.01)	1.43(0.01)	1.01(0.01)	1.15(0.01)

the underlying models are more sparse. Specifically, in the scenario of S1 = 25% and $p = 40$, the proposed GAQQ model does not perform as well as others, since the underlying models in this scenario are the least sparse, especially for the dense Models 2 and 3. In contrast, the proposed method produces relatively better comparison results in the scenario of S2 = 75% and $p = 40$, where the true mean difference is more sparse. Furthermore, this advantage of proposed GAQQ model is well evidenced in the scenario of $p = 80$, and even more notable when $p = 200$ with its substantially lower MEs than other methods.

From Table 2, we observe that the proposed method generally gives superior performance over other compared approaches for each scenario in predicting the quantitative response y. The possible explanations are in two folds. First, the proposed GAQQ method provides an accurate classification of the qualitative response Z. Second, the proposed GAQQ has a proper estimation of $\boldsymbol{C}$ by the regularization that is used in the prediction of quantitative response y according to (11), resulting in an improvement of the prediction accuracy. It is also seen that the CL and GLDA models are comparable in some cases, due to that both of them use the Moore-Penrose generalized inverse of the sample covariance of $\boldsymbol{X}$ for $\hat{\boldsymbol{\Sigma}}_X^{-1}$ in the prediction of quantitative response y in (11). But the CL method is generally better since it has more accurate classification results than the GLDA in Table 1.

4.2. Multi-class settings of the qualitative response

Now, we examine the performance of the proposed GAQQ method for multi-class settings of the qualitative response. We consider $p = 200$ and $K = 4$ classes of qualitative response Z with training sizes $n_1 = n_2 = n_3 = n_4 = 30$ for Models 1 - 5 of inverse covariance matrix $\boldsymbol{C}$. Let μ_{kj} represent the jth entry of the mean value $\boldsymbol{\mu}_k$. Generate $\mu_{kj} = 0.5 * k + u_{kj}$ for $j \in \{2k-1, 2k, 2k+1, \ldots, 2k+6\}$, otherwise $\mu_{kj} = 0$, where u_{kj} is from $Unif(-1, 1)$. The training data are generated from $N(\boldsymbol{\mu}_k, \boldsymbol{C}^{-1})$, and the testing data follow the same generation procedure. We summarize the proposed GAQQ model in terms of both estimation and prediction procedures in the following Algorithm 4 (based on Algorithms 1 and 2) for multi-class settings of the qualitative response.

Table 3: Averages and standard errors (in parenthesis) of MEs in percentage and RMSPE obtained from compared methods for multi-class settings of $p = 200$ based on 100 replications.

	Model 1	Model 2	Model 3	Model 4	Model 5
			ME		
GLDA	40.58 (0.46)	59.15 (0.56)	55.56 (0.51)	39.40 (0.48)	48.28 (0.55)
WT	17.26 (0.35)	43.73 (0.59)	43.01 (0.47)	17.90 (0.38)	33.17 (0.67)
CHWE	14.70 (0.32)	25.14 (0.40)	32.16 (0.50)	16.66 (0.44)	21.31 (0.44)
GAQQ	14.02 (0.32)	25.36 (0.53)	33.34 (0.50)	17.11 (0.42)	20.99 (0.49)
			RMSPE		
GLDA	1.64 (0.01)	2.09 (0.02)	2.02 (0.02)	1.66 (0.01)	1.71 (0.02)
WT	1.56 (0.01)	2.05 (0.02)	1.94 (0.02)	1.57 (0.01)	1.63 (0.02)
CHWE	1.56 (0.01)	2.01 (0.02)	1.92 (0.02)	1.55 (0.01)	1.61 (0.02)
GAQQ	0.99 (0.01)	1.11 (0.01)	1.27 (0.01)	1.01 (0.01)	1.39 (0.02)

Algorithm 4

Step 1: Given (λ_1, λ_2), set $\hat{\boldsymbol{\delta}}_{k,0} = (\bar{\boldsymbol{w}}_k - \bar{\boldsymbol{w}}_1)/K$.

Step 2: Given $\boldsymbol{\delta}_k = \hat{\boldsymbol{\delta}}_{k,t}$, solve $\boldsymbol{C}$ in (17) by the Glasso at the tth iteration.

Step 3: Given $\boldsymbol{C} = \hat{\boldsymbol{C}}_t$, $\boldsymbol{\delta}_g = \hat{\boldsymbol{\delta}}_{g,t}, g \neq k$, solve $\boldsymbol{\delta}_k$ in (19) by the Lasso at the tth iteration.

Step 4: Repeat Step 2 and 3 till $\hat{\boldsymbol{C}}_t$ and all the $\hat{\boldsymbol{\delta}}_{k,t}$ converge.

Step 5: Choose the optimal values of $\boldsymbol{\delta}_k$ and $\boldsymbol{C}$ with respect to (λ_1, λ_2) via (20), denoted by $\hat{\boldsymbol{\delta}}_k$ and $\hat{\boldsymbol{C}}$ respectively.

Step 6: Calculate $\hat{\boldsymbol{\gamma}} = \bar{\boldsymbol{w}} + \sum_{g=2}^{K} \hat{\boldsymbol{\delta}}_g - (K/n) \sum_{g=2}^{K} n_g \hat{\boldsymbol{\delta}}_g$, and $\hat{\boldsymbol{\mu}}_k = \hat{\boldsymbol{\gamma}} - \sum_{g=2}^{K} \hat{\boldsymbol{\delta}}_g + K\hat{\boldsymbol{\delta}}_k$.

Step 7: For a new observation $\boldsymbol{x}$, compute $\hat{y}_k = \hat{\mu}_{ky} - \hat{\boldsymbol{C}}_{Xy}^{\top}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{kX})/\hat{c}_y^2$ for $k \in \{1, \ldots, K\}$; obtain $\hat{p}_k$ by plugging $(\boldsymbol{x}^{\top}, \hat{y}_k)^{\top}$ into the density functions of $N(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{C}}^{-1})$.

Step 8: Let $t = \arg\max_k \hat{\pi}_k \hat{p}_k$, then $\hat{y} = \hat{y}_t$ and $\hat{Z} = t$.

We compare the proposed method with the GLDA, as well as the estimators proposed by Witten and Tibshirani [45] (WT) and Clemmensen et al. [9] (CHWE), where the latter two methods are designed for multi-class problems. We use the WT and CHWE models to first predict the class label Z for the testing data, and then the response y is estimated, by the multivariate normal property, as $\hat{\mu}_{ky} + \hat{\boldsymbol{\Sigma}}_{Xy}^{\top}\hat{\boldsymbol{\Sigma}}_X^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{kX})$ if their estimates $\hat{Z} = k$. The WT and CHWE methods are implemented by packages `PenalizedLDA` and `sda` from R software, respectively. The results of performance measures, ME and RMSPE, of the methods under comparison, are summarized in Table 3 based on 100 replications. One can see that the GAQQ method performs better than the GLDA as well as the WT method, and is comparable with the CHWE in terms of the MEs. Besides, the GAQQ method gives the best performance among the compared approaches with significantly lower values of RMSPE.

4.3. Case studies with real-world data

In this section, we apply the proposed GAQQ method to two real-data case studies. The first one is from the study of Heusler compounds in material science and the second one is from the study of molecular diagnostics of Ulcerative colitis and Crohn's disease. Although from different fields, both problems contain QQ responses with high-dimensional predictors, and the proposed GAQQ method appears to have much better performance in terms of prediction accuracy compared with other methods.

The case study on material sciences is regarding the Heusler compounds, which are a large family of intermetallics with more than 1000 known members. Many Heusler compounds have shown exotic properties, such as superconductivity and topological band structures, which have promising applications for quantum computing. Understanding the thermodynamic stability of Heusler compounds lays the foundation for exploiting the large chemical space to discover and design new functional Heusler materials [33, 36, 37]. To determine the thermodynamic stability of Heusler compounds, there are two key metrics: the mixing enthalpy (quantitative response) and the global stability based on hull energy (binary qualitative response). The comprehensive database of 180628 full Heusler structures was built by collecting the relevant structural and energetic data from the Materials Project [20], OQMD [41], and AFLOW [12]. These data were calculated using first-principles methods based on density functional theory, and it was extremely computationally expensive (taking hours) to generate one entry of the data. Therefore, a statistical model that can

accurately predict the thermodynamic stability for any elemental and compound features is a useful surrogate of the first-principle computation models.

Since there is an intrinsic relationship between two QQ responses, the proposed GAQQ method is suitable to improve the prediction accuracy by jointly fitting them together. To demonstrate the GAQQ method in the scenario when the number of predictors is large relative to the size of the data, we randomly choose 150 samples from each class of the binary response. We delete the predictor variables whose standard deviations are less than $1.0e^{-6}$, resulting in 157 predictors of elemental and compound features. To examine the prediction performance of the GAQQ method and other comparison methods, we randomly divide data into a training set with a size of 200 and a testing set with a size of 100. Table 4 reports the prediction performance results based on 50 random splits of the Heusler data. From the results, it is seen that the proposed GAQQ performs much better than other methods in comparison, with the smallest values for the misclassification error (ME) and the root mean squared prediction error (RMSPE).

Table 4: Averages and standard errors (in parenthesis) of MEs in percentage and RMSPE obtained from compared methods for Heusler and gene expression data based on 50 random splits.

		Heusler Data		
Methods	GLDA	ENET	CL	GAQQ
ME	27.27 (1.828)	11.87 (0.332)	16.20 (0.688)	10.49 (0.363)
RMSPE	1.797 (0.445)	0.317 (0.083)	1.046 (0.053)	0.142 (0.002)
		IBD Gene Data		
Methods	GLDA	WT	CHWE	GAQQ
ME	21.90 (0.800)	24.80 (0.583)	18.10 (0.555)	15.77 (0.584)
RMSPE	0.743 (0.014)	0.751 (0.014)	0.746 (0.014)	0.661 (0.011)

The second data for the case study considers the multi-class settings of the qualitative response. The IBD gene data [6] are gene expressions on Ulcerative colitis (UC) and Crohn's disease (CD), two of which are common inflammatory bowel diseases (IBD) producing intestinal inflammation and tissue damage. The IBD data set was collected at North American and European clinical sites from blood samples of 42 healthy individuals, 59 CD patients, and 26 UC patients with 22,283 genes. An exploratory analysis, similarly conducted as in Shao et al. [42], is performed as variable screening by one-way ANOVA with three levels (healthy individuals, CD patients, and UC patients). We choose the top 101 significant gene variables to form the data for methods comparison. To create a quantitative response, one gene variable is randomly chosen as the quantitative response from the 101 significant variables. The data set is then randomly partitioned into a training set with 67 samples and testing data with the rest 60 samples. Table 4 presents the comparison results by the GLDA, WT, CHWE, and proposed GAQQ methods based on 50 random splits of the data. We observe that the proposed GAQQ method performs substantially well with relatively lower values of ME and RMSPE, as well as their corresponding standard errors in the parenthesis. Such empirical results demonstrate that the proposed GAQQ method can achieve accurate predictions for both QQ responses in high-dimensional data.

5. Conclusions, limitations and future research

In this work, we propose a generative modeling approach to jointly model the data with QQ responses. By fully exploring the joint distribution of the QQ responses and predictor variables, the proposed method enables efficient parameter estimation, model prediction, and most importantly, lays a good foundation for investigating the asymptotic properties for QQ responses, which few works have studied so far. Note that there are some discussions in the literature about the comparable performance between the joint modeling and separate modeling under certain situations with $n > p$. When $n < p$, the proposed joint modeling approach can generally have better advantage than the separate modeling since there are relatively limited observations in the data.

The proposed GAQQ model in this work can be easily extended to accommodate the cases where the quantitative response $\boldsymbol{y} = (y_1, \ldots, y_q)^\top$ has multiple dimensions by assuming $(\boldsymbol{X}^\top, \boldsymbol{y}^\top)^\top$ is from different multivariate normal distributions, see assumption (1). The parameter estimation follows the same procedure, and the prediction on $\boldsymbol{y}$ is essentially via the multivariate normal properties, see (11). For the case of multiple qualitative responses, $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^\top$, the GAQQ model cannot be applied directly. The simplest solution is to replace $\boldsymbol{Z}$ with a single multi-class qualitative variable with $K = \prod_{j=1}^{m} K_j$ and K_j the number of classes for Z_j. Then we can apply the

GAQQ model discussed in Section 2.2. Another potential challenge is the situation that some of the $\boldsymbol{X}$ variables are quantitative whereas the others are qualitative. Then one cannot simply adopt multivariate normal distribution for $\boldsymbol{X}$. To accommodate this situation for the proposed method, one possibility is to consider the multivariate normal distribution for quantitative variables in $\boldsymbol{X}$ and proper nonparametric density distributions for the qualitative variables.

The research has several future directions. In this work, we assume that the observed data follow normal distribution for model parsimony in the large-dimensional setting. When the data are not normally behaved or contaminated with outliers, it will be important to consider robust statistical modeling [28]. One possible direction is to incorporate robust discriminant analysis [30, 38] into the proposed method, such as using robust estimation of mean and inverse covariance matrix. Another possible way is to use t distribution as the assumption for robust modeling. Additional research direction is to accommodate a more flexible structure on the joint distribution of QQ responses and predictors. For example, one can extend the LDA for the classification of the qualitative response to the quadratic discriminant analysis (QDA). However, its estimation for high-dimensional data would encounter difficulty due to a large number of parameters. Besides, additional research direction is to apply the generative modeling approach for the data with semi-continuous responses, or the ordinal and quantitative responses [48]. One may employ the ordinal regression for the ordinal response, and then derive its joint likelihood function with appropriate regularization.

Appendix

Below we provide the derivation from (18) to (19). The derivation from (8) to (9) is a special case with $K = 2$. For $\boldsymbol{\delta}_j$, $j \in \{2, \ldots, K\}$,

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{i\in G_k}(\boldsymbol{w}_i-\bar{\boldsymbol{w}}+\frac{K}{n}\sum_{g=2}^{K}n_g\boldsymbol{\delta}_g-K\boldsymbol{\delta}_k)^\top\boldsymbol{C}(\boldsymbol{w}_i-\bar{\boldsymbol{w}}+\frac{K}{n}\sum_{g=2}^{K}n_g\boldsymbol{\delta}_g-K\boldsymbol{\delta}_k)+\lambda_2|\boldsymbol{\delta}_j|_1\\
=&\sum_{k=1,k\neq j}^{K}\sum_{i\in G_k}\{\boldsymbol{C}^{1/2}(\boldsymbol{w}_i-\bar{\boldsymbol{w}}+\frac{K}{n}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g-K\boldsymbol{\delta}_k)+\boldsymbol{C}^{1/2}\frac{K}{n}n_j\boldsymbol{\delta}_j\}^\top\{\boldsymbol{C}^{1/2}(\boldsymbol{w}_i-\bar{\boldsymbol{w}}+\frac{K}{n}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g-K\boldsymbol{\delta}_k)+\boldsymbol{C}^{1/2}\frac{K}{n}n_j\boldsymbol{\delta}_j\}\\
&\quad+\sum_{i\in G_j}\{\boldsymbol{C}^{1/2}(\boldsymbol{w}_i-\bar{\boldsymbol{w}}+\frac{K}{n}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g)+\boldsymbol{C}^{1/2}(\frac{K}{n}n_j\boldsymbol{\delta}_j-K\boldsymbol{\delta}_j)\}^\top\{\boldsymbol{C}^{1/2}(\boldsymbol{w}_i-\bar{\boldsymbol{w}}+\frac{K}{n}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g)\\
&\quad+\boldsymbol{C}^{1/2}(\frac{K}{n}n_j\boldsymbol{\delta}_j-K\boldsymbol{\delta}_j)\}+\lambda_2|\boldsymbol{\delta}_j|_1\\
=&\sum_{k=1,k\neq j}^{K}\sum_{i\in G_k}\{\frac{2Kn_j}{n}(\boldsymbol{C}^{1/2}\boldsymbol{\delta}_j)^\top(\boldsymbol{C}^{1/2}\boldsymbol{w}_i-\boldsymbol{C}^{1/2}\bar{\boldsymbol{w}}+\frac{K}{n}\boldsymbol{C}^{1/2}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g-K\boldsymbol{C}^{1/2}\boldsymbol{\delta}_k)+\frac{K^2n_j^2}{n^2}\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{\delta}_j\}\\
&\quad+\sum_{i\in G_j}\{2K(\frac{n_j}{n}-1)(\boldsymbol{C}^{1/2}\boldsymbol{\delta}_j)^\top(\boldsymbol{C}^{1/2}\boldsymbol{w}_i-\boldsymbol{C}^{1/2}\bar{\boldsymbol{w}}+\frac{K}{n}\boldsymbol{C}^{1/2}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g)+K^2(\frac{n_j}{n}-1)^2\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{\delta}_j\}+\lambda_2|\boldsymbol{\delta}_j|_1+C\\
=&\frac{Kn_j}{n}\sum_{k=1,k\neq j}^{K}(2\boldsymbol{\delta}_j^\top\boldsymbol{C}\sum_{i\in G_k}\boldsymbol{w}_i-2n_k\boldsymbol{\delta}_j^\top\boldsymbol{C}\bar{\boldsymbol{w}}+\frac{2n_kK}{n}\boldsymbol{\delta}_j^\top\boldsymbol{C}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g-2n_kK\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{\delta}_k+\frac{Kn_jn_k}{n}\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{\delta}_j)\\
&\quad+Kn_j(\frac{n_j}{n}-1)\{2\boldsymbol{\delta}_j^\top\boldsymbol{C}(\frac{1}{n_j}\sum_{i\in G_j}\boldsymbol{w}_i)-2\boldsymbol{\delta}_j^\top\boldsymbol{C}\bar{\boldsymbol{w}}+\frac{2K}{n}\boldsymbol{\delta}_j^\top\boldsymbol{C}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g+K(\frac{n_j}{n}-1)\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{\delta}_j\}+\lambda_2|\boldsymbol{\delta}_j|_1+C\\
=&\frac{K^2n_j(n-n_j)}{n}\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{\delta}_j-\frac{2K}{n}\boldsymbol{\delta}_j^\top\boldsymbol{C}\{\sum_{k=1,k\neq j}^{K}(-n_j\sum_{i\in G_k}\boldsymbol{w}_i+n_jn_k\bar{\boldsymbol{w}}-\frac{Kn_jn_k}{n}\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g+Kn_jn_k\boldsymbol{\delta}_k)\\
&\quad-(n_j-n)\sum_{i\in G_j}\boldsymbol{w}_i+n_j(n_j-n)\bar{\boldsymbol{w}}-Kn_j(\frac{n_j}{n}-1)\sum_{g=2,g\neq j}^{K}n_g\boldsymbol{\delta}_g\}+\lambda_2|\boldsymbol{\delta}_j|_1+C\\
=&\frac{K^2n_j(n-n_j)}{n}\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{\delta}_j-\frac{2K}{n}\boldsymbol{\delta}_j^\top\boldsymbol{C}\boldsymbol{M}+\lambda_2|\boldsymbol{\delta}_j|_1+C,
\end{aligned}
\tag{32}
$$

where

$$M = -n_j \sum_{i \notin G_j} \boldsymbol{w}_i + n_j(n-n_j)\bar{\boldsymbol{w}} - \frac{Kn_j(n-n_j)}{n} \sum_{g=2,g\neq j}^{K} n_g \boldsymbol{\delta}_g + Kn_j \sum_{g=2,g\neq j}^{K} n_g \boldsymbol{\delta}_g - (n_j - n) \sum_{i \in G_j} \boldsymbol{w}_i + n_j(n_j - n)\bar{\boldsymbol{w}}$$
$$- Kn_j(\frac{n_j}{n} - 1) \sum_{g=2,g\neq j}^{K} n_g \boldsymbol{\delta}_g = (n-n_j) \sum_{i \in G_j} \boldsymbol{w}_i - n_j \sum_{i \notin G_j} \boldsymbol{w}_i + Kn_j \sum_{g=2,g\neq j}^{K} n_g \boldsymbol{\delta}_g.$$

Let $\tilde{y} = \boldsymbol{C}^{1/2} M / \{Kn_j(n-n_j)\} = \boldsymbol{C}^{1/2}\{(n-n_j) \sum_{i \in G_j} \boldsymbol{w}_i - n_j \sum_{i \notin G_j} \boldsymbol{w}_i + Kn_j \sum_{g=2,g\neq j}^{K} n_g \boldsymbol{\delta}_g\} / \{Kn_j(n-n_j)\}$. Hence, formula (32) is equal to

$$\frac{K^2 n_j(n-n_j)}{n} (\tilde{y} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_j)^\top (\tilde{y} - \boldsymbol{C}^{1/2}\boldsymbol{\delta}_j) + \lambda_2 |\boldsymbol{\delta}_j|_1 + C.$$

References

References

[1] T. Baghfalaki, M. Ganjali, A. Kabir, A. Pazouki, A bayesian shared parameter model for joint modeling of longitudinal continuous and binary outcomes, Journal of Applied Statistics in press (2020).

[2] N. M. Bello, J. P. Steibel, R. J. Tempelman, Hierarchical bayesian modeling of heterogeneous clusterand subject-level associations between continuous and binary outcomes in dairy production, Biometrical Journal 54 (2012) 230–248.

[3] P. J. Bickel, E. Levina, Covariance regularization by thresholding, Annals of Statistics 36 (2008) 2577–2604.

[4] J. Bien, R. J. Tibshirani, Sparse estimation of a covariance matrix, Biometrika 98 (2011) 807–820.

[5] P. Bühlmann, S. Van De Geer, Statistics for high-dimensional data, Springer, Verlag Berlin Heidelberg, 2011.

[6] M. E. Burczynski, R. L. Peterson, N. C. Twine, K. A. Zuberek, B. J. Brodeur, L. Casciotti, V. Maganti, P. S. Reddy, A. Strahs, F. Immermann, et al., Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells, The Journal of Molecular Diagnostics 8 (2006) 51–61.

[7] T. Cai, W. Liu, A direct estimation approach to sparse linear discriminant analysis, Journal of the American Statistical Association 106 (2012) 1566–1577.

[8] S. Chen, D. M. Witten, A. Shojaie, Selection and estimation for mixed graphical models, Biometrika 102 (2014) 47–64.

[9] L. Clemmensen, T. Hastie, D. Witten, B. Ersbøll, Sparse discriminant analysis, Technometrics 53 (2011) 406–413.

[10] D. Cox, F. O'Sullivan, Asymptotic analysis of penalized likelihood and related estimators, Annals of Statistics 18 (1990) 1676–1695.

[11] V. R. Craiu, A. Sabeti, In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes, Journal of Multivariate Analysis 110 (2012) 106–120.

[12] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, et al., Aflow: an automatic framework for high-throughput materials discovery, Computational Materials Science 58 (2012) 218–226.

[13] X. Deng, R. Jin, QQ models: Joint modeling for quantitative and qualitative quality responses in manufacturing systems, Technometrics 57 (2015) 320–331.

[14] D. B. Dunson, Dynamic latent trait models for multidimensional longitudinal data, Journal of the American Statistical Association 98 (2003) 555–563.

[15] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American statistical Association 96 (2001) 1348–1360.

[16] G. M. Fitzmaurice, N. M. Laird, Regression models for a bivariate discrete and continuous outcome with clustering, Journal of the American statistical Association 90 (1995) 845–852.

[17] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (2008) 432–441.

[18] R. V. Gueorguieva, A. Agresti, A correlated probit model for joint modeling of clustered binary and continuous responses, Journal of the American Statistical Association 96 (2001) 1102–1112.

[19] A. Guglielmi, F. Ieva, A. M. Paganoni, F. A. Quintana, A semiparametric bayesian joint model for multiple mixed-type outcomes: an application to acute myocardial infarction, Advances in Data Analysis and Classification 12 (2018) 399–423.

[20] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., Commentary: The materials project: A materials genome approach to accelerating materials innovation, Apl Materials 1 (2013) 011002.

[21] L. Kang, X. Kang, X. Deng, R. Jin, A bayesian hierarchical model for quantitative and qualitative responses, Journal of Quality Technology 50 (2018) 290–308.

[22] X. Kang, X. Chen, R. Jin, H. Wu, X. Deng, Multivariate regression of mixed responses for evaluation of visualization designs, IISE Transactions 53 (2021) 313–325.

[23] X. Kang, S. Ranganathan, L. Kang, J. Gohlke, X. Deng, Bayesian auxiliary variable model for birth records data with qualitative and quantitative responses, Journal of Statistical Computation and Simulation 91 (2021) 3283–3303.

[24] M.-H. Kao, H. Khogeer, Optimal designs for mixed continuous and binary responses with quantitative and qualitative factors, Journal of Multivariate Analysis 182 (2021) 104712.

[25] N. Klein, T. Kneib, G. Marra, R. Radice, S. Rokicki, M. E. McGovern, Mixed binary-continuous copula regression models with application to adverse birth outcomes, Statistics in Medicine 38 (2019) 413–436.
[26] E. Kürüm, R. Li, S. Shiffman, W. Yao, Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data, Statistica Sinica 26 (2016) 979–1000.
[27] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, Annals of Statistics 37 (2009) 4254–4278.
[28] K. L. Lange, R. J. Little, J. M. Taylor, Robust statistical modeling using the t distribution, Journal of the American Statistical Association 84 (1989) 881–896.
[29] W. Lee, Y. Pawitan, Direct calculation of the variance of maximum penalized likelihood estimates via em algorithm, The American Statistician 68 (2014) 93–97.
[30] Q. Li, L. Li, Integrative linear discriminant analysis with guaranteed error rate improvement, Biometrika 105 (2018) 917–930.
[31] L. Lin, D. Bandyopadhyay, S. R. Lipsitz, D. Sinha, Association models for clustered data with binary and continuous responses, Biometrics 66 (2010) 287–293.
[32] Y. Liu, Z. Ren, et al., Minimax estimation of large precision matrices with bandable cholesky factor, Annals of Statistics 48 (2020) 2428–2454.
[33] Z. Liu, L. Yang, S.-C. Wu, C. Shekhar, J. Jiang, H. Yang, Y. Zhang, S.-K. Mo, Z. Hussain, B. Yan, et al., Observation of unusual topological surface states in half-heusler compounds lnptbi (ln=lu,y), Nature Communications 7 (2016) 1–7.
[34] C. Luo, J. Liang, G. Li, F. Wang, C. Zhang, D. K. Dey, K. Chen, Leveraging mixed and incomplete outcomes via reduced-rank modeling, Journal of Multivariate Analysis 167 (2018) 378–394.
[35] J. Lv, Y. Fan, A unified approach to model selection and sparse recovery using regularized least squares, Annals of Statistics 37 (2009) 3498–3528.
[36] K. Manna, Y. Sun, L. Muechler, J. Kübler, C. Felser, Heusler, weyl and berry, Nature Reviews Materials 3 (2018) 244–256.
[37] Y. Nakajima, R. Hu, K. Kirshenbaum, A. Hughes, P. Syers, X. Wang, K. Wang, R. Wang, S. R. Saha, D. Pratt, et al., Topological rpdbi half-heusler semimetals: A new family of noncentrosymmetric magnetic superconductors, Science Advances 1 (2015) e1500242.
[38] A. M. Pires, J. A. Branco, Projection-pursuit approach to robust linear discriminant analysis, Journal of Multivariate Analysis 101 (2010) 2464–2485.
[39] G. Raskutti, B. Yu, M. J. Wainwright, P. Ravikumar, Model selection in gaussian graphical models: high-dimensional consistency of l_1-regularized mle, Advances in Neural Information Processing Systems 21 (2008) 1329–1336.
[40] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu, et al., Sparse permutation invariant covariance estimation, Electronic Journal of Statistics 2 (2008) 494–515.
[41] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd), Jom 65 (2013) 1501–1509.
[42] J. Shao, Y. Wang, X. Deng, S. Wang, Sparse linear discriminant analysis by thresholding for high dimensional data, Annals of Statistics 39 (2011) 1241–1265.
[43] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58 (1996) 267–288.
[44] H. Wang, R. Li, C. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika 94 (2007) 553–568.
[45] D. M. Witten, R. Tibshirani, Penalized classification using fisher's linear discriminant, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (2011) 753–772.
[46] P. Xu, J. Zhu, L. Zhu, L. Yi, Covariance-enhanced discriminant analysis, Biometrika 102 (2015) 33–45.
[47] M. Yuan, Y. Lin, Model selection and estimation in the gaussian graphical model, Biometrika 94 (2007) 19–35.
[48] X. Zhang, W. J. Boscardin, T. R. Belin, X. Wan, Y. He, K. Zhang, A bayesian method for analyzing combinations of continuous, ordinal, and nominal categorical data with missing values, Journal of Multivariate Analysis 135 (2015) 43–58.
[49] P. Zhao, B. Yu, On model selection consistency of lasso, Journal of Machine Learning Research 7 (2006) 2541–2563.
[50] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2005) 301–320.